



Language as Interface to Medical Image Representation Learning

Maxime Kayser
Harris Manchester College



University of Oxford
A thesis presented for the degree of
Doctor of Philosophy

Hilary 2025

Acknowledgements

I am grateful to my primary supervisor, Thomas Lukasewiecz, and co-supervisors Oana-Maria Camburu, Bartłomiej Papież and Seth Flaxman, who have supported me throughout my DPhil journey. Thomas created an environment in which I was able to pursue my research with a high degree of autonomy and consistently supported my ideas, which shaped me as an independent researcher. Bartek has been my supervisor-in-residence in Oxford and my link to the clinical world. I am especially thankful to Oana, who has played an instrumental role in shaping my research, both during her time as my official co-supervisor and after her departure from Oxford. Oana has been a constant source of intellectual insight and personal support, helping me navigate challenges and offering unwavering encouragement through moments of doubt.

Cornelius Emde, partner in crime and my de facto research lab, I can't imagine what this would have been like without you. From the early days together in the CDT to starting our DPhils at the height of Covid, marked by uncertainty and adaptation, I am immensely grateful to have faced it together.

To Aleksandr Petrov at Memorial Sloan Kettering, thank you for being both a mentor and advocate, for opening doors I didn't even know existed and for clearing obstacles before I even reached them.

I am also indebted to those who showed me this path was possible: Jens Rittscher, Shadi Albarqouni for being a wonderful BSc Thesis advisor, and Xavier Bresson for putting so much effort and care into his teaching, which has made this world accessible and inspired me ever since.

I want to thank Elsevier for funding my DPhil and Elia-Lima Walton for supporting me. My heartfelt appreciation goes to Ewoud Pons[†], whose kindness and generosity brightened the early days of my DPhil - wherever you are now, I hope you've found peace.

To the Health Data Science CDT directorate - Jim Davies, Nina Hallowell[†], Thomas Nichols, and Jens Rittscher - thank you for creating not just a program but a community. Nina, you touched our lives and your absence is felt by all of us. To my CDT cohort—Andres, Henrique, Claudia, Cornelius, Franck, Hang, Jack, Jacob, Kangning, Odhran, Xinchu, and Yifan—and to the wonderful Leema, thank you for making Oxford feel like home from the start. To my adop-

tive office family—Abeekh, Martin, Mohsen, Nicholas, Sonja, Thomasz, and Zhongbin—thank you for the warm space our office was every day.

I am also grateful to my collaborators—Bayar, Harini Veeraraghavan, Wendy, and Maksim—and to Max Bain, research advisor and surf companion. Heartfelt thanks to Susanne Gaube for being my lifeline out of the human subjects study rabbit hole - your sympathy saved my sanity. I owe immense gratitude to the clinicians who brought real-world perspective to my algorithms - Guy Parsons, Sarim Ather, my sister, and my father. My collaboration with Nathanael Swinburne has been a constant source of inspiration and motivation, fueled by his curiosity and openness.

Sports have given me a sense of belonging, an opportunity to grow, and ultimately kept me grounded throughout these years. Thank you to the wonderful people of Wadham College Boat Club, in particular Harriet, Morten, Luke, Rob M., Will, and Wissam, for making me fall in love with this sport and becoming friends beyond the boat. Yeah Wadham. To the football teams that I have been fortunate to be part of, Oxford University Futsal, Wadham, and Jesus/Magdalen (2025 MCR Cuppers champions forever!) - particularly Abdullah, Adam, Adrian, Marco, Ryan, and Shuailin - thanks for giving me something to look forward to every week.

I would not have been able to do this without the support and companionship of friends and family. Thank you to my family, Catherine, Claire, Isabelle, Jacques, and Martin, for always having my back. To the friends who made my transition into life at Oxford so safe and joyful, Kasia, Leander, and especially Emma C., whose friendship means the world, thank you. To my Oxford family, Emma, Lewis, Michelle, Rob F., you are like family to me. Thank you for making Oxford home. HMC friends Jan, Jan, Jonas, Josef, Lea, Leonie, Marie, Melody, Monty, and Penny, thank you for being a wonderful support network and a constant presence throughout this journey. Thank you to Annabel, Fé, and Tom, for your steady friendship and constant support. To Maxime, my academic twin in name, interests, and degrees - thanks for the solidarity of parallel journeys. To the Rewley Road crew - Allison, Franck, Lize, and Martin - who turned a house into a home. Last but not least, thank you to the four-legged friends for their wisdom and unconditional affection—Bella, Buna, Finn, and Lotti. And to Emma, I love you. You stood by me through the toughest part of this journey, and I could not have done it without you.

Abstract

Natural language inherently encodes human knowledge and reasoning. This thesis explores how language can be used as an interface between humans and artificial intelligence (AI) approaches in medical imaging. Language can function as a supervisory signal, or constraint, to learn meaningful representations that capture the human interpretation of medical images. At the same time, language can serve as a medium through which an AI system can communicate its decision-making to users.

First, I introduce the e-ViL benchmark, which systematically assesses vision-language models' capability to generate natural language explanations (NLEs). By introducing the largest dataset to date, consolidating existing datasets, and proposing a novel architecture, I demonstrate the ability of text-based explanations to constrain model reasoning and provide an explainability interface that aligns with human cognitive processes. This work also exposes the limitations of automated metrics for evaluating NLE quality and proposes a human evaluation framework to address this.

Building upon these insights, I extend the concept of NLEs to medical imaging, specifically chest X-ray analysis. I create novel datasets by extracting explanations directly from radiology reports. Models trained on these datasets not only produce explanations that mirror radiologists' reasoning but, by emphasizing scale and chain-of-thought prompting, also exhibit improved diagnostic accuracy.

Third, to better understand the role of language as a medium to communicate AI decisions to users, I conduct a large-scale user study with 85 clinicians interacting with AI under different AI explainability (XAI) conditions. The study reveals a critical dichotomy: Clinicians strongly prefer language-based explanations but tend to overrely on them, leading to a higher number of diagnostic errors. These findings underscore the complexities and caution needed in integrating such language-based AI systems into clinical settings.

Lastly, I generalize the use of language as supervision by leveraging radiology reports to train brain MRI models from scratch. This approach addresses the scarcity of labelled medical data in this domain and leads to improved performance on a suite of downstream tasks, showcasing the broad applicability and significant potential of language-guided learning in medical imaging.

This thesis advances our understanding of integrating language with medical AI models. It demonstrates substantial benefits in performance and interpretability while also highlighting critical issues to address for safe and effective clinical deployment of these approaches.

List of Publications

The thesis is largely based on the following papers:

1. [Kayser et al. \(2021\)](#): **Kayser, M.**, Camburu, O. M., Salewski, L., Emde, C., Do, V., Akata, Z., & Lukasiewicz, T. (2021). e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
2. [Kayser et al. \(2022\)](#): **Kayser, M.**, Emde, C., Camburu, O. M., Parsons, G., Papiez, B., & Lukasiewicz, T. (2022). Explaining Chest X-Ray Pathologies in Natural Language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.
3. [Kayser et al. \(2023\)](#): **Kayser, M.**, Camburu, Lukasiewicz, T. (2023). Can Large Language Models Reason about Chest X-Rays? In *Extended Abstract presented at Machine Learning for Health (ML4H) symposium, New Orleans, United States*.
4. [Kayser et al. \(2024a\)](#): **Kayser, M.**, Menzat, B., Emde, C., Bercean, B., Novak, A., Morgado, A., ... & Camburu, O. M. (2024). Fool Me Once? Contrasting Textual and Visual Explanations in a Clinical Decision-Support Setting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **EMNLP Outstanding Paper Award.**
5. [Kayser et al. \(2024b\)](#): **Kayser, M.**, Rangnekar, A., Petrov, A., Veeraraghavan, H., & Swinburne, N. C. (2025). A Copilot for Radiology: Vision-Language Models Achieve Accurate Brain MRI Interpretation. In *Abstract presented at RSNA 2024, Chicago, United States*.
6. **Kayser, M.**, Gridnev, M., Wang, W., Bain, M., Rangnekar, A., Chatterjee, A., Petrov, A., Veeraraghavan, H., & Swinburne, N. C. (2025). brat: Aligned Multi-View Embeddings for Brain MRI Analysis. *Under submission at ICCV 2025*.

During my DPhil, I also co-authored the publications shown below. These are not incorporated in this thesis.

- Nuthalapati, S. V., Chandradevan, R., Giunchiglia, E., Li, B., **Kayser, M.**, Lukasiewicz,

T., & Yang, C. (2021). Lightweight visual question answering using scene graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

- Emde, C., Paren, A., Arvind, P., **Kayser, M.**, Rainforth, T., Lukasiewicz, T., ... & Bibi, A. (2025). Shh, don't say that! Domain Certification in LLMs. In *International Conference on Representation Learning (ICLR)*.

In addition, I supervised the following Master's thesis during my DPhil:

- *Multimodal Modeling of Chest X-Rays*. Thesis written by Daniel Blasko for the degree of Master of Science/Engineering, INSA Lyon and Technische Universität Wien, 2024.

Table of Contents

1	Introduction	1
1.1	How Language Can Improve AI Models	1
1.2	Research Questions and Contributions	4
2	Background	7
2.1	Vocabulary and Notation	7
2.2	Modeling Vision and Language	8
2.2.1	Vision-Language Model Applications	9
2.2.2	Vision-Language Pre-Training Objective Functions	11
2.2.3	Image-Text Interaction	13
2.2.4	Pre-Training Datasets	14
2.2.5	Vision-Language Models in the Age of Scale	15
2.3	AI Interpretability	15
2.3.1	Types of Explanations	16
2.3.1.1	Post-hoc Explainers	16
2.3.1.2	Self-explaining Models	18
2.3.1.3	Natural Language Explanations	19
2.3.2	Properties of Explanations	20
2.3.3	Evaluating Explanations with Humans	21
2.4	Medical Image Analysis	23
2.4.1	Chest X-rays	24
2.4.2	Brain MRIs	24
2.4.3	Vision and Language	25
3	Multi-modal AI That Can Explain Itself in Natural Language	28
3.1	Related Work	30
3.2	e-SNLI-VE-2.0	33
3.2.1	Improving SNLI-VE	33
3.2.2	Adding Explanations to SNLI-VE	34
3.3	Approaches for the VL-NLE Task	38

3.3.1	VL-NLE Architectures	38
3.3.2	Training	41
3.4	The e-ViL Benchmark	44
3.4.1	Datasets	44
3.4.2	Evaluation	46
3.4.3	Results	49
3.4.4	Statistical Analysis and Alternative Metrics	54
3.5	Conclusion	59
3.6	Impact and What Happened Since	60
4	Natural Language Explanations for Chest X-ray Analysis	61
4.1	Related Work	63
4.2	MIMIC-NLE-v1	65
4.3	Benchmark v1	70
4.3.1	Models	70
4.3.2	Experimental Setup	70
4.3.3	Results	73
4.4	MIMIC-NLE-v2	76
4.5	Scale and Reasoning	86
4.5.1	NLE Generation Benchmark	86
4.5.2	Chain-of-Thought Reasoning for Chest X-rays	89
4.5.3	CoT and Generalization	91
4.6	Conclusion	94
4.7	Impact and What Happened Since	95
5	Evaluating Language as an Interface Between Humans and AI in a Safety-Critical Setting	96
5.1	Related Work	98
5.2	Study Design	100
5.2.1	Task Design	100
5.2.2	Survey & Qualitative Interview questions	106
5.2.3	Pilot Studies	106
5.2.4	Statistical model	107

5.3	Participant Selection and Outreach	112
5.3.1	Target Candidate & Sample Size	112
5.3.2	Recruitment Drive	117
5.3.3	Participant Demographics	118
5.4	Gathering Data for Study	121
5.4.1	AI Model Pipeline	121
5.4.2	Data Annotation & Selection	122
5.5	Study Outcome	128
5.5.1	Survey Results	128
5.5.2	Results for the Main Research Questions	131
5.5.3	Exploratory Analysis	135
5.5.4	Model Fit	140
5.6	Conclusion	142
6	Language as Signal for Brain MRI Representation Learning	143
6.1	Related Work	146
6.2	MSKBrain Dataset	148
6.2.1	Raw Dataset	148
6.2.2	Text processing	149
6.3	brat: A Vision-Language Model for Brain MRIs	154
6.3.1	Pairwise-View Alignment	154
6.3.2	Quality-Diversity via DPPs	156
6.3.3	brat as a Foundation Model	160
6.4	Experiments and Results	162
6.4.1	VLP Performance: Image-Text Retrieval	162
6.4.2	Downstream Tasks	164
6.5	Conclusion	169
7	Conclusions and Implications	170

List of Figures

3.1	Motivating example for multi-caption false neutral detection	34
3.2	Motivating example for keyword-based filtering	35
3.3	Motivating example for filtering based on premise-hypothesis similarity	36
3.4	High-level architectures of the models that are included in the e-ViL benchmark.	38
3.5	Representative examples from each e-ViL dataset.	45
3.6	A snapshot of the instructions that were provided to the annotators.	47
3.7	A snapshot of the interface through which annotators evaluated the explanations.	48
3.8	Pairs of annotated examples for each dataset	50
3.9	e-ViL scores with confidence intervals	55
3.10	Ordinal representation of e-ViL scores	56
3.11	Comparative e-ViL scores	57
4.1	MIMIC-NLE-v1 extraction process	65
4.2	MIMIC-NLE-v1 evidence-outcome relationship graph	68
4.3	Co-occurrence matrix for evidence-outcome pairs in MIMIC-NLE-v1.	69
4.4	Example representation of a single datapoint in MIMIC-NLE-v1.	69
4.5	The v1 model pipeline to provide an NLE for a prediction	70
4.6	MIMIC-NLE-v1 results examples	74
4.7	Example Representation of a single datapoint in MIMIC-NLE-v2.	76
4.8	MIMIC-NLE-v2 extraction process	77
4.9	RadGraph example	78
4.10	MIMIC-NLE-v2 evidence-outcome relationship graph	82
4.11	Co-occurrence matrix for evidence-outcome pairs in MIMIC-NLE-v2.	83
4.12	Distribution of labels in the MIMIC-NLE-v2 training set.	83
4.13	Distribution of labels in the MIMIC-NLE-v2 test set, CheXpert500, and RSNA	84
4.14	MIMIC-NLE-v2 examples	85
4.15	Example instruction provided to the vision-enabled LLM.	86
5.1	The flow of the user study that every participant goes through.	100
5.2	Snapshots of the different interfaces of the CDSS.	103
5.3	The instruction PDF that people have access to throughout the study	105

5.4	Screening survey test cases	112
5.5	The average scores obtained on the pre-screening survey based on radiology training level.	113
5.6	Power analysis results	116
5.7	The recruitment flyer used to reach out to participants to fill out the screening form.	117
5.8	Email sent to invited participants.	117
5.9	Self-assessed knowledge and frequency of use for different AI technologies . .	118
5.10	Participant demographics	119
5.11	Radiology and AI experience self-assessment	119
5.12	Effect of participant attributes on performance	120
5.13	The UI used by our radiologists to annotate the study examples.	124
5.14	Distribution of raw explanation correctness scores C_χ	126
5.15	Distribution of explanation correctness scores C_χ included in study	127
5.16	Four quadrants with examples	128
5.17	Five attributes of explainability methods, ranked on a 7-point Likert scale by the participants.	129
5.18	AI performance estimates and expectations	130
5.19	User study main results	132
5.20	Human accuracy given C_{AI} and C_χ , predicted with the model (5.1).	133
5.21	User study results across explanation correctness	134
5.22	Multiple testing adjusted results	135
5.23	Perceived usefulness ρ_U results	137
5.24	Perceived usefulness ρ_U lines	138
5.25	Positive certainty results	139
5.26	Decision speed results	140
5.27	Decision speed over the course of the study	141
6.1	PVA motivation figure	144
6.2	MSKBrain patient demographical insights	148
6.3	Raw report example	149
6.4	Report re-writing prompt	150
6.5	GPT-4 processing pipeline	151

6.6	Example #1 of processed report	151
6.7	Report information extraction prompt	152
6.8	Example #2 of processed report	153
6.9	Example #3 of processed report	153
6.10	brat architecture	154
6.11	Query token collapse	155
6.12	Multi-view embedding attention maps	155
6.13	DPP motivation figure 1	157
6.14	DPP motivation figure 2	158
6.15	brat for downstream tasks	160
6.16	MSKBrain retrieval examples	163
6.17	BIMCV-R quality issue examples	164
6.18	MSKBrain example generated reports	166
6.19	ADNI classification results	167
6.20	BraTS segmentation results	167

List of Tables

3.1	e-SNLI-VE-2.0 summary statistics.	33
3.2	Effects of e-SNLI-VE-2.0 filters	36
3.3	Human evaluation of VL-NLE datasets	37
3.4	e-ViL re-implementation results	42
3.5	Benchmark models hyperparameters	43
3.6	Summary of the model differences.	43
3.7	e-ViL benchmark scores	49
3.8	Annotated shortcomings of explanations	51
3.9	Automatic NLG metrics for all model-dataset pairs.	52
3.10	Correlation between human evaluation and automatic NLG metrics on NLEs.	53
3.11	Effect of explanations on task performance	54
3.12	Class-wise results on e-SNLI-VE-2.0.	54
4.1	MIMIC-NLE-v1 explanation keywords	66
4.2	MIMIC-NLE-v1 outcome labels	67
4.3	MIMIC-NLE-v1 rules to establish valid evidence-outcome relationships	68
4.4	MIMIC-NLE-v1 filters overview	69
4.5	MIMIC-NLE-v1 benchmark results	73
4.6	The UMLS classes grouped into labels used in MIMIC-NLE-v2.	81
4.7	MIMIC-NLE-v2 LLM performance	88
4.8	MIMIC-NLE-v2 indication ablations	88
4.9	MIMIC-NLE-v2 dataset version ablations	89
4.10	Summary of model responses under No CoT and With CoT scenarios.	90
4.11	MIMIC-NLE-v2 CoT results	91
4.12	CoT generalization results	93
5.1	Statements to evaluate participants' attitudes towards AI.	106
5.2	Statements capturing participants' subjective feedback on explanation types.	107
5.3	Power analysis assumptions	114
5.4	Statements capturing participants' self-assessed AI expertise.	118
5.5	Preference ranking of models.	129

5.6	Generalized Linear Mixed Effects Model Results	141
6.1	MSKBrain summary statistics	151
6.2	Implementation details of our pre-training.	161
6.3	Implementation details of our Alzheimer classification downstream task.	161
6.4	MSKBrain Retrieval Results	163
6.5	BIMCV-R retrieval results	164
6.6	MSKBrain report generation results	167
6.7	Detailed ADNI results	167
6.8	BraTS2021 results	168
6.9	BraTS2023-METS results	168

List of Abbreviations

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
NLE	Natural Language Explanation
VL	Vision-Language
VLP	Vision-Language Pre-Training
VLM	Vision-Language Model
LLM	Large Language Model
IAA	Inter-Annotator Agreement
SSL	Self-Supervised Learning
PVA	Pairwise View Alignment
QD	Quality-Diversity
DPP	Determinantal Point Processes

1 | Introduction

1.1 How Language Can Improve AI Models

Language is a fundamental instrument through which human beings construct and convey meaning. From everyday interactions to complex discourse, language shapes our understanding of the world, organizes our thoughts, and frames our perceptions of reality. This thesis, titled “Language as Interface to Medical Image Representation Learning,” explores how language can be harnessed as supervision, constraint, and explainability approach to act as a medium between humans and AI in medical imaging.

Language is not only a communication tool but also structures human thought and perception. [Wittgenstein \(1922\)](#) asserted that “the limits of my language mean the limits of my world”. Sapir-Whorf’s linguistic relativity hypothesis even claims that if something cannot be expressed in language, it cannot exist within our framework of understanding ([Whorf, 1956](#)), underlining that language affects how we perceive and categorize reality. The aspects of perception and cognition that image representation learning seeks to capture are thus already encoded in language structures. The ability of language to capture not only concrete visual features but also abstract conceptual information—such as relationships, uncertainty, and causality—makes it uniquely suited for guiding vision-based learning systems.

In clinical practice, language is the primary medium for describing findings, communicating diagnoses, and documenting patient histories. Clinical knowledge is largely textual, with medical records and diagnostic reports forming the backbone of patient care ([Weed, 1968](#)). The precision and nuance of language are critical in translating complex physiological and pathological states into communicable information that guides decision-making. By providing standardized vocabularies such as SNOMED CT and ICD classifications, medical terminology in itself structures clinical reasoning ([Stearns et al., 2001](#)). Given this centrality in clinical workflows, language is ideally suited for both learning representations of medical data and communicating these to clinicians.

Representation learning has rapidly progressed over the past decade, with deep learning emerging as the dominant paradigm in computer vision and natural language processing ([Krizhevsky et al., 2012](#); [Vaswani et al., 2017](#)). These advances have been driven by increasingly large-

scale neural networks, improved optimization techniques, and the availability of vast amounts of annotated data (Sun et al., 2017). Despite the impressive performance of AI models across diverse domains, several fundamental challenges persist, which limit their broader deployment and impact.

First, the complexity of deep learning models often results in a lack of interpretability, making it difficult to understand their decision-making processes and potential biases (Lipton, 2017; Rudin et al., 2022). This raises concerns in high-stakes applications, such as healthcare, where explainability is critical for trust, regulatory compliance, and clinical validation (Samek et al., 2017). Real-life failure cases such as a skin lesion detector that relied on detecting a ruler in the scan (Narla et al., 2018), a mortality risk predictor that considered asthma to reduce the risk for pneumonia patients (Caruana, 1997), models that rely on the text in radiographic scans to make a decision (Zech et al., 2018) or a study showing that a large majority of 145 COVID-19 prediction models were not sufficiently robust for clinical utilisation (Wynants et al., 2020), highlight the need for increased model transparency. Interpretability is widely considered a way to increase algorithms' readiness to be deployed (Piano, 2020; Brundage et al., 2020). One way to improve interpretability is by adopting model constraints, following Rudin et al. (2022): "An interpretable model is constrained, following a domain-specific set of constraints that make its reasoning process understandable". This thesis argues that natural language can provide such constraints through the use of natural language explanations (NLEs).

NLEs offer a powerful mechanism for enforcing human-like reasoning constraints in AI models. Human cognition relies heavily on language to structure thought, decompose complex problems, and communicate reasoning processes (Lakoff and Johnson, 1980). The ability to articulate explanations in natural language mirrors how humans reason and provides an intuitive way to encode that decision-making in AI (Doshi-Velez and Kim, 2017). Furthermore, linguistic supervision can facilitate model introspection and debugging, which can identify inconsistencies or biases in reasoning pathways (Camburu et al., 2018). Recent advances in reasoning-based models, such as chain-of-thought prompting (Wei et al., 2022), also highlight the effectiveness of language as a way of breaking down complex decision-making into more feasible steps, thereby improving model performance. These insights motivate Chapters 3 and 4 of the thesis, which look at NLEs for vision-language and chest X-ray analysis tasks.

Second, AI systems must be designed for usability and alignment with human needs. In med-

ical imaging, for instance, where, for the time being, AI tools should complement rather than replace expert judgment, model outputs should be clinically meaningful and intuitive to interpret (Tschandl et al., 2020). Poorly designed AI interfaces that lack meaningful explanations for their decisions may reduce adoption rates among clinicians and diminish potential benefits. The desired characteristics of interpretability methods differ depending on the application, target audience, and required robustness and faithfulness. Saliency maps are a common explanation type applied to medical imaging algorithms, which highlight the regions in an image that affected the model output the most. These have been criticized for conveying little information, being overly reassuring (confirmation bias) (Bornstein, 2016; Adebayo et al., 2018), and lacking robustness (Arun et al., 2021). Saliency maps also lack explanation capacity, as they only show where a model looks at, but not, e.g., if “an airspace opacity was important in the decision” or “if the shapes of the heart border or left pulmonary artery were the deciding factor” (Ghassemi et al., 2021). While initial work has advocated for the use of user-friendly, textual explanations in medical applications (Reyes et al., 2020), this thesis is the first to study the use of NLEs in clinical practice.

For AI to be effectively integrated into human decision-making processes, it must communicate in a way that aligns with human cognitive frameworks. The ability of AI to provide explanations in natural language, rather than relying solely on visual markers such as saliency maps, offers a more intuitive and accessible means for users to understand and interact with AI-generated outputs (Ehsan and Riedl, 2020; Liao et al., 2022). Despite these advantages, there is limited empirical evidence on the practical utility of language-based explanations compared to traditional visual methods. In Chapter 5, I conduct a large-scale user study to evaluate the effectiveness of NLEs versus visual explanation methods in a clinical decision-support setting. By assessing how different explanation modalities impact human trust, reliance, and diagnostic performance, this study provides critical insights into the role of language in AI interpretability and its implications for real-world deployment.

Lastly, AI development in medical imaging traditionally relies heavily on large datasets and high-quality annotations to embed human expertise within the models. This introduces challenges as medical data is scarce and annotating costs are high (Mazurowski et al., 2019). Various strategies have been proposed to mitigate this issue, including self-supervised learning (He et al., 2020), weak supervision (Ratner et al., 2017), and transfer learning from natural image

domains (Raghu et al., 2019). However, these approaches struggle to fully capture the domain-specific knowledge required for medical imaging.

Language, on the other hand, offers a naturally abundant and structured source of supervision. Free-text radiology reports, which are routinely collected in clinical practice, contain dense human knowledge encoded in natural language. These reports provide an ideal signal to learn rich image representations that reflect healthcare professionals’ medical expertise and reasoning without requiring additional annotation efforts (Zhang et al., 2022b; Boecking et al., 2022). Thus, leveraging the language in medical reports not only enhances the semantic depth of learned representations but also reduces the dependence on manual annotation. Chapter 6 explores how brain MRI models can be trained using language as a primary supervisory signal, demonstrating the potential of radiology reports to enhance image representation learning.

1.2 Research Questions and Contributions

In Chapter 3, the main research question is to *benchmark the capability of existing vision-language models to provide natural language explanations on complex tasks*. To achieve this, I first created the largest existing vision-language (VL) dataset with NLEs, called e-SNLI-VE-2.0. I then pool three robust and validated VL-NLE datasets, e-SNLI-VE-2.0, VCR (Zellers et al., 2019), and VQA-X (Park et al., 2018), to provide e-ViL, the first comprehensive comparison of existing approaches. I also propose a new model, e-UG, which significantly outperforms all existing models. To address the shortcomings of NLG metrics for evaluating NLEs, I provide an extensive evaluation of the models using crowd workers and additionally use these annotations to study the efficacy of different NLG metrics. This chapter is based on the publication below:

Kayser et al. (2021): Kayser, M., Camburu, O. M., Salewski, L., Emde, C., Do, V., Akata, Z., & Lukasiewicz, T. (2021). e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

The next research question, in Chapter 4, is to *introduce NLEs to chest X-ray analysis and verify whether they improve model performance*. As no NLE datasets exist for medical imaging, I created MIMIC-NLE-v1 and later MIMIC-NLE-v2. The observation that radiologists com-

monly provide evidence for the findings listed in a radiology report allowed me to extract NLEs directly from these notes. I then benchmark several models on MIMIC-NLE-v1, and evaluate them with clinicians. This highlighted several shortcomings of the dataset, such as a lack of patient context and hallucinated references to prior scans that were not available. This motivated an improved version of the dataset, MIMIC-NLE-v2, with significant quality improvements. Given the current ubiquity of large-scale VLMs, I also provide a new benchmark with scaled-up VLMs and find that chain-of-thought (CoT) reasoning improves chest X-ray analysis. This chapter is based on the publications below:

[Kayser et al. \(2022\)](#): **Kayser, M.**, Emde, C., Camburu, O. M., Parsons, G., Papiez, B., & Lukasiewicz, T. (2022). Explaining Chest X-Ray Pathologies in Natural Language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[Kayser et al. \(2023\)](#): **Kayser, M.**, Camburu, Lukasiewicz, T. (2023). Can Large Language Models Reason about Chest X-Rays? In *Extended Abstract presented at Machine Learning for Health (ML4H) symposium, New Orleans, United States*.

To understand whether AI models providing such NLEs can be helpful in clinical practice, in Chapter 5, I study *how real-world users engage with NLEs compared to visual explanation methods*. I set up a large-scale user study with 85 clinicians, where they collaborated with an AI under four different scenarios: providing no explanation at all, providing a saliency map, providing an NLE, or providing both the NLE and the saliency map. Every participant analyzes 80 different chest X-rays spread evenly across the tasks. By using both qualitative (e.g. “which explanations did you like the most?”) and quantitative metrics (which explanations increased human performance the most), I find that people heavily favour text-based explanations but tend to overrely on the AI when only NLEs are provided. With today’s ubiquity of powerful AI models that can explain their reasoning in natural language, this has important implications for the dangers of these models in safety-critical settings. This chapter is based on the publication below:

[Kayser et al. \(2024a\)](#): **Kayser, M.**, Menzat, B., Emde, C., Bercean, B., Novak, A., Morgado, A., ... & Camburu, O. M. (2024). Fool Me Once? Contrasting Textual and Visual Explanations in a Clinical Decision-Support Setting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*

(EMNLP). **EMNLP Outstanding Paper Award.**

Finally, in Chapter 6, I investigate the *use of radiology reports to train more capable AI models in a domain where data is especially scarce: brain MRIs*. I introduce the largest existing dataset of brain MRIs and paired radiology reports. *brat*, a newly proposed pre-training framework, leverages multi-view embeddings to capture the versatility and information richness of brain MRI reports. I show how this outperforms alternative pre-training methods and leads to strong vision backbones that can improve performance on several tasks, including tumour segmentation, Alzheimer’s classification, and radiology report generation.

[Kayser et al. \(2024b\)](#): **Kayser, M.**, M., Rangnekar, A., Petrov, A., Veeraraghavan, H., & Swinburne, N. C. (2025). A Copilot for Radiology: Vision-Language Models Achieve Accurate Brain MRI Interpretation. In *Abstract presented at RSNA 2024, Chicago, United States*.

Kayser, M., Gridnev, M., Wang, W., Bain, M., Rangnekar, A., Chatterjee, A., Petrov, A., Veeraraghavan, H., & Swinburne, N. C. (2025). *brat*: Aligned Multi-View Embeddings for Brain MRI Analysis. *Under submission at ICCV 2025*.

2 | Background

2.1 Vocabulary and Notation

Vocabulary This thesis addresses certain topics where terminology is sometimes used inconsistently across the literature. To ensure clarity, I define the key terms used in this thesis.

Representation learning refers to learning vector embeddings of any kind of data, e.g., images or text. In this thesis, I focus exclusively on deep learning-based approaches, where neural networks are used to extract these representations. I broadly use the term *AI (model)* to encompass all contemporary artificial intelligence methods.

“*Interpretability*” and “*explainability*” are used interchangeably in this thesis, collectively referred to as *explainable AI (XAI)*. We use it to refer to all methods designed to enhance the transparency of AI decision-making processes. The term *(X)AI explanations* denotes any form of explanation generated by these models to communicate their reasoning to humans. *Natural Language Explanations (NLEs)* are a specific subset of explanations that articulate model decisions in human-readable text, usually generated by the AI model itself. Alternative terms include *textual* or *text-based explanations*. In contrast, *saliency maps* or *heatmaps* refer to gradient-based visualization methods that highlight the image regions most influential to a model’s prediction.

Vision-language (VL) refers to tasks, datasets, and models that integrate both visual and textual modalities. *Vision-language models (VLMs)* are AI models designed to learn joint representations of both modalities. For large language models (*LLMs*), I adopt the convention of referring to models with at least 1 billion (1B) parameters as LLMs. However, some literature restricts this classification to models exceeding 10B parameters.

Notation For chapters with overlapping methodology, I use consistent notation. $\mathcal{L}$ denotes loss functions used for training models. In the context of AI models that provide explanations, I distinguish between the *task model* M_T , which produces predictions for a primary task T , and the *explanation module* M_E , which generates explanations for these predictions. Evaluation metrics for these models are categorized as follows. S_T is the performance metric for the primary task. S_E is the evaluation metric for explanations. S_O is an overall metric that integrates

both task and explanation performance.

2.2 Modeling Vision and Language

Deep learning emerged as a transformative force in 2012 with the success of AlexNet in the ImageNet Challenge (Krizhevsky et al., 2012). Using convolutional neural networks (CNNs), the authors demonstrated that deep neural networks could outperform traditional computer vision methods when given sufficient computational resources. This success, enabled by Graphics Processing Units (GPU) and large-scale datasets, reignited interest in neural networks.

Following AlexNet, more advanced CNN architectures were developed. VGG networks (Simonyan and Zisserman, 2015) used deeper layers with small convolutional filters to enhance performance. ResNet (He et al., 2016) addressed the vanishing gradient problem through skip connections, which enabled training much deeper networks. These innovations led to significant progress across computer vision tasks and enabled advances in fields such as object detection (Redmon et al., 2016; Ren et al., 2016) or segmentation (Ronneberger et al., 2015). The success of deep learning in computer vision also encouraged its wider use in natural language processing (NLP). Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) were used to model language as a sequence and led to significant improvements in fields such as machine translation (Sutskever et al., 2014; Cho et al., 2014).

Early applications of deep learning for combined vision and language modeling include image captioning and visual question answering. Notable works such as “Show and Tell” (Xu et al., 2015) and the VQA dataset (Antol et al., 2015) demonstrated the potential of jointly modeling images and text.

The introduction of the Transformer architecture in “Attention Is All You Need” (Vaswani et al., 2017) revolutionized both NLP and computer vision. By enabling parallel processing and improved modeling of long sequences, the self-attention mechanism introduced in Transformers led to state-of-the-art NLP models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018). In computer vision, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) demonstrated that pure Transformer models could achieve competitive results, challenging the dominance of CNNs. By allowing to model images and text through a unified architecture,

Transformers have accelerated the development of joint vision-language architectures ([Tan and Bansal, 2019](#)).

In this brief overview of vision-language (VL) approaches, I will first look at the different applications of VL models. Then, I will look at VL models through the lens of their objective functions used in training. I also classify VL architectures by examining how they fuse image and text data and consider how images can be “tokenized”. Finally, I will provide context of VL models in the era of LLMs. This section focuses on joint embedding and image-to-text architectures rather than text-to-image (e.g., Dall-E ([Ramesh et al., 2021](#))) models.

2.2.1 Vision-Language Model Applications

Vision-language models (VLMs) are capable of integrating visual and textual information, which is helpful for a wide range of tasks, such as image captioning or visual question answering (VQA). I also discuss how VLMs can be used for unimodal tasks, such as zero-shot image classification.

Image Captioning consists of generating textual descriptions of visual content. Before the widespread use of deep learning in sequence-to-sequence models, image captions were generated using templates and image classification pipelines ([Li et al., 2011](#); [Kulkarni et al., 2013](#)) or using retrieval-based methods ([Kuznetsova et al., 2012, 2014](#)). Works on n-gram prediction from images ([Joulin et al., 2016](#); [Li et al., 2017](#)) can also be considered precursors to modern image captioning. End-to-end neural architectures led to a big leap in performance, and the first methods were inspired by the use of encoder-decoder architectures in machine translation ([Vinyals et al., 2015](#); [Xu et al., 2015](#)). Analog to translating one language to another, image captioning can be viewed as translating an image to language. Thus, the encoder RNN can simply be replaced by a vision encoder. Important early benchmarks such as MS COCO ([Lin et al., 2014](#)) have spurred this development. Recent advancements in image captioning come primarily from scaling model size and pre-training data ([Alayrac et al., 2022](#); [Chen et al., 2023b](#)). **Visual question answering** requires models to answer natural language questions about images. The VQA dataset ([Antol et al., 2015](#)) was a key benchmark that supported early developments. Architectures like the bottom-up, top-down approach ([Anderson et al., 2018](#)) and LXMERT ([Tan and Bansal, 2019](#)) introduced sophisticated attention mechanisms to handle cross-modal interactions. Again, recent advancements are primarily driven by scale, with

general-purpose VLMs topping the leaderboards (Alayrac et al., 2022; Chen et al., 2023b). Another important task is **image-text retrieval**, which involves retrieving the most relevant image given a text query, or vice versa. The datasets used for this are generally the same as for image captioning, i.e., paired images and text meant to present the same semantic content. Recent models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have achieved state-of-the-art results in image-text retrieval by training on massive datasets to learn joint embeddings of paired images and text via contrastive loss.

Beyond these core domains, there is a wide range of sub-tasks to which VLMs are being applied. **Visual grounding** requires identifying specific objects in images based on textual descriptions (see benchmarks like RefCOCO (Kazemzadeh et al., 2014) and models like GLIP (Li et al., 2022b)). **Visual reasoning** tasks, like those evaluated on CLEVR (Johnson et al., 2017), require models to perform compositional reasoning over visual content, with Neural Module Networks (Andreas et al., 2016) and PaLM-E (Driess et al., 2023) being exemplary approaches. **Visual commonsense reasoning (VCR)** extends this to require commonsense knowledge (see the VCR dataset (Zellers et al., 2019)). **Visual entailment**, inspired by natural language inference tasks, requires models to determine whether an image supports or contradicts a given textual hypothesis (see SNLI-VE (Xie et al., 2019)).

VLMs can also be used for traditional vision tasks, where their language interface enables them to be easily adapted to new tasks. This is particularly useful for zero-shot vision tasks, where the unseen labels can be communicated via text. Models like CLIP (Radford et al., 2021) demonstrated the effectiveness of zero-shot classification by matching image embeddings with text embeddings of class labels. Zero-shot object detection emerged with models like ZSD (Bansal et al., 2018) and progressed to more advanced frameworks like GLIP (Li et al., 2022b). For segmentation, models like Segment Anything (Kirillov et al., 2023) have shown impressive zero-shot capabilities (Huang et al., 2024).

Over the last years, the pre-training paradigm that has proven successful in general NLP (i.e., pre-training on large text corpora, followed by fine-tuning on downstream tasks (Devlin et al., 2019)), has also proved successful in VLMs. For the remainder of this section, I will provide an overview of vision-language pre-training methods.

2.2.2 Vision-Language Pre-Training Objective Functions

Most modern VL approaches rely on models pre-trained on large datasets of paired images and text. The primary goal of these pre-training methods is to learn joint embeddings that represent images and text within the same semantic space, enabling models to learn a mapping between both modalities. Different pre-training objectives have been proposed to achieve this, with contrastive learning and image captioning being the main ones.

Image Captioning Generative image captioning is both a common application of VLMs and a common pre-training objective (Desai and Johnson, 2021; Saryildiz et al., 2020). By training models to generate descriptive captions for images, they learn to map visual features to textual representations. Most methods that rely purely on image captioning as a pre-training task focus on generative downstream tasks such as open-ended VQA and captioning and use the same complete encoder-decoder architecture that was pre-trained (Cho et al., 2021; Hu et al., 2022b). In other common VL downstream tasks, such as zero-shot classification, it has been shown that image captioning pre-training does not scale as well with the amount of data compared to other pre-training methods (Radford et al., 2021). Recent research, however, demonstrates that certain variants of image captioning losses, particularly parallel decoding, can enhance image captioning as a robust pre-training strategy for diverse downstream applications (Tschannen et al., 2023). The image captioning (IC) loss can be formalized as follows:

$$\mathcal{L}_{\text{IC}} = - \sum_{t=1}^T \log P(y_t | y_{1:t-1}, I)$$

where I represents the image input, y_t is the predicted token at time step t , and T denotes the total length of the caption. This loss function corresponds to maximizing the log-likelihood of the correct caption sequence given the image.

Image-Text Contrastive Learning Contrastive learning is another popular pre-training strategy for vision-language models, popularized in particular by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). These approaches enable scalable pre-training on millions of image-text pairs collected from the web. Contrastive learning effectively aligns visual and textual modalities by bringing matching image-text pairs closer in the embedding space and pushing

non-matching pairs apart. The image-text contrastive (ITC) loss can be expressed as:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(x_i, y_j)/\tau)} + \log \frac{\exp(\text{sim}(y_i, x_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(y_i, x_j)/\tau)} \right]$$

where x_i and y_i represent image and text embeddings, respectively, $\text{sim}(a, b) = \frac{a \cdot b}{\|a\| \|b\|}$ denotes cosine similarity, and τ is a temperature parameter. The loss is symmetric, measuring both image-to-text and text-to-image alignment.

Besides enabling zero-shot transfer to various image classification and retrieval tasks, contrastive loss also produces high-quality vision backbones that equal the best label-supervised models (Dong et al., 2022). Follow-up works that extend contrastive loss focus on scale (Cherti et al., 2023) or learning fine-grained alignments (Yao et al., 2022). Contrastive loss is also commonly combined with other objectives, such as image captioning or image-text matching (Yu et al., 2022; Kuo et al., 2023; Li et al., 2022a).

Other Pre-Training Objectives While image captioning and image-text contrastive loss are the primary pre-training objectives used in vision-language models (VLMs), a range of other pre-training methods have been explored to improve multimodal understanding. Masked language modeling (MLM), adapted from BERT, involves randomly masking a subset of textual tokens and training the model to predict these tokens based on the unmasked text and associated visual inputs (Tan and Bansal, 2019). Masked vision modeling is analogous to MLM, but instead, a portion of visual features is masked and needs to be reconstructed based on the remaining visual information and text (Chen et al., 2020b). Variants include masked feature regression, using L2 loss to regress masked features to original values (Chen et al., 2020b), or masked feature classification, where the semantic class of masked features needs to be predicted (Li et al., 2020). Vision-language matching requires classifying pairs of images and text as matching or not (Lu et al., 2019). Some strategies also incorporate downstream objectives like VQA (Wang et al., 2021; Huang et al., 2021c). Other pre-training tasks include grounding referring expressions, image-conditioned denoising autoencoding, or object detection (Zhou et al., 2020; Gu et al., 2022; Cherti et al., 2023).

2.2.3 Image-Text Interaction

Processing Images and Text Early VLMs were often based on encoder-decoder architectures that consisted of a CNN vision model and RNN language model, where images and text were processed in their “native” way. Text data is typically tokenized into subword units using algorithms like Byte Pair Encoding (BPE) (Sennrich et al., 2016) or WordPiece (Schuster and Nakajima, 2012). Each such token is then mapped to an index, which is embedded using an embedding layer. Images, on the other hand, are fed directly as $\text{channel} \times \text{width} \times \text{height}$ pixel tensors.

However, in modern, transformer-based vision-language models, both modalities are usually processed in the same architecture. The primary way to achieve this equivalence is by “tokenizing” images, similar to text. Initial methods relied on object extractors like Faster R-CNN (Ren et al., 2016), arguing that tokens (words) in text equate to objects in images (Chen et al., 2020b). Each bounding box, including labels and coordinates, is then embedded as a token. Later, it was found that grid features, i.e., feature maps from CNNs or patch embeddings from ViTs, would work just as well, simplifying the architecture (Hendricks et al., 2021). The most common approach is, hence, to decompose images into subregions, which are then tokenized and processed equivalently to (sub)words in text. Other work has looked into learnable query tokens that learn to extract relevant features from images and can then be fed into a multimodal transformer architecture (Jaegle et al., 2021; Alayrac et al., 2022; Li et al., 2022a).

Different architectural designs have been proposed to model the fusion of text and image data, aiming to learn an equivalence between the two. One of the key distinctions between existing methods is the stage at which both modalities interact.

Early Interaction Early interaction approaches aim to integrate visual and textual information early in the model architecture, enabling joint learning from the outset. These methods can be divided into single-stream and double-stream models, differing primarily in how the modalities are encoded and fused.

In **single-stream models**, a unified encoder processes both image and text inputs simultaneously, with shared model weights across both modalities. Sometimes, special tokens or token-type embeddings are used to distinguish between image and text tokens within a single input sequence. Representative examples include VisualBERT (Li et al., 2019b), VL-BERT (Su et al.,

2019), and UNITER (Chen et al., 2020b).

By contrast, **double-stream early interaction models** initially process images and text through separate, modality-specific encoders, followed by a fusion module. Early examples of this paradigm include many of the pre-transformer VLMs or ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019), where each stream independently encodes its respective modality before cross-attention mechanisms bring the streams together. This structure provides more flexibility in specialized feature extraction but still encourages early multimodal integration.

Late Interaction Late interaction methods rely on modality-specific encoders throughout most of the learning process and fuse visual and textual embeddings only at the final stage. This is usually achieved through contrastive learning. Unlike in double-stream early interaction, these late-fusion approaches do not allow intermediate layers of the encoders to cross-attend.

Cross-Modal Attention For early-interaction models, the manner in which the modalities attend to each other can vary. In **merged attention**, all query tokens (from both image and text) attend to keys drawn from every token in both modalities (Chen et al., 2020b). This fully merged attention can maximize cross-modal interactions at each layer. For **coattention**, queries from one modality (e.g., image features) attend only to keys from the other modality (e.g., language embeddings) (Lu et al., 2019). Attention can also be **assymetric**, with one modality’s queries attending to another modality’s keys, while the reverse does not necessarily happen. For example, in some architectures, only the text stream attends to the image stream, but the image stream does not attend back to text.

2.2.4 Pre-Training Datasets

The success of VLMs relies on the quality and scale of the datasets used for pre-training. Large-scale datasets containing paired visual and textual information allow models to learn rich, cross-modal representations (e.g., SBU (Ordonez et al., 2011), MS COCO (Lin et al., 2014), and Conceptual Captions (CC3M, CC12M) (Sharma et al., 2018; Changpinyo et al., 2021), LAION-400M (Schuhmann et al., 2021)). Specialized datasets like Visual Genome (with region-level descriptions and QA pairs) (Krishna et al., 2017), Matterport3D (RGB-D indoor scenes) (Chang et al., 2017), and Fashion-Gen (fashion attributes) (Rostamzadeh et al.,

2018) address domain-specific needs.

2.2.5 Vision-Language Models in the Age of Scale

With the rise of large language models (LLMs), the scale of VLMs has also increased. Many VLMs make use of LLMs for their world knowledge and fluency and propose different ways to integrate visual knowledge. Recent models like Frozen (Tsimpoukelli et al., 2021) and BLIP-2 (Li et al., 2023b) underscore the growing feasibility of “plug-and-play” multimodality, where large language models remain frozen while a vision encoder (or a compact mapping network) is trained to produce representations compatible with a text-only decoder. Similar strategies are adopted by systems like MiniGPT-4 (Zhu et al., 2024) and LLaVA (Liu et al., 2023a), where visual tokens are embedded into the semantic space of the language models, and the LLMs are fine-tuned to incorporate that knowledge. In most existing work, there is a clear discrepancy in representation capacity between visual and text encoders, i.e., the language component is typically much larger than the visual component (Tsimpoukelli et al., 2021; Li et al., 2023b; Zhu et al., 2024; Liu et al., 2023a).

The trend toward large-scale VLMs has motivated research into parameter-efficient tuning strategies that reduce the training time required to adapt these models. Prompt tuning, which gained popularity in language-only contexts (Brown et al., 2020), now extends to the multimodal field with approaches like Context Optimization (CoOp) (Zhou et al., 2022), where learnable text prompts help the system adapt to new image distributions and classification tasks, and visual prompting (Jia et al., 2022), which achieves similar efficiency gains by adjusting the visual input space. A similar rationale underlies LoRA (Hu et al., 2022a), which breaks down parameter updates into low-rank matrices to train more efficiently.

2.3 AI Interpretability

The high performance of deep learning methods often comes at the expense of scrutability, i.e., the complex feature spaces learned by these models make it challenging to understand how they process information and make decisions. AI interpretability methods aim to make these processes more transparent and shed light on the internal representations of these models. Interpretability has been identified as one of the key priorities for machine learning in medical applications (Langlotz et al., 2019). This challenge is especially pronounced in the larger neural

architectures required to process images or text.

There is no universally agreed-upon definition for AI interpretability and explainability, with both terms often used interchangeably (Doshi-Velez and Kim, 2017; Lipton, 2018; Jacovi and Goldberg, 2020). In this thesis, I will use them interchangeably and refer to the overall field as “explainable AI (XAI)”. A commonly accepted, general definition is by Doshi-Velez and Kim (2017), referring to XAI as “the ability to explain or to present [a model] in understandable terms to a human.” In this section, I will give a brief overview of the desired properties of AI explanations, the types of AI explanations relevant to this thesis, and how explanations can be evaluated from a human-computer interaction standpoint. The use cases illustrated will primarily focus on medical applications.

2.3.1 Types of Explanations

Various taxonomies have been proposed to classify interpretability methods based on methodology or desired outcomes. One widely adopted classification distinguishes between post-hoc explainers, i.e., a-posteriori attempts to explain a fully trained black-box model and intrinsically interpretable (or “self-explaining”) models, i.e., models designed to be inherently explainable (Lipton, 2018).

Another classification in the interpretability literature, particularly in NLP, differentiates between extractive and abstractive explanations. These terms are borrowed from text summarization: extractive explanations focus on identifying subsets of the input (such as tokens in text or regions in images) that influenced the model’s decision, whereas abstractive explanations involve more flexible descriptions, often in natural language, that articulate the reasoning behind a prediction.

2.3.1.1 Post-hoc Explainers

Post-hoc explainers can be classified into feature attribution and latent space disentanglement methods.

Feature Attribution Methods Feature attribution methods are generally based on either changing the inputs to the model and observing how it affects the outputs of the model, thereby trying to approximate how the model reasons about its inputs; or by differentiating the predic-

tion with respect to the input. These methods can be applied across a wide variety of black-box models and are among the most widely used XAI techniques in both academic research and practical applications.

One of the earliest feature attribution methods is LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016), which explains a model's prediction by perturbing the input and observing the corresponding changes in the output. LIME provides a linear approximation of the model's behavior around a specific instance by fitting a simple, interpretable model (such as a linear regression) to the locally perturbed data. Another foundational method is based on Shapley values (Lundberg and Lee, 2017), which provide an approach to distribute the contribution of each feature to the final prediction by considering all possible permutations of feature inclusion.

Gradient-based importance measures are another core category of feature attribution methods. These methods explain predictions by calculating the gradient of the output with respect to the input features, effectively measuring how a small change in the input would affect the model's output. Simonyan et al. (2014) introduced class-specific saliency maps, which visualize which pixels in an image are most influential for a specific prediction. Improvements for such saliency maps were introduced by Guided Backpropagation (Springenberg et al., 2015), Class Activation Mapping (CAM) (Zhou et al., 2016) and Grad-CAM (Gradient-weighted CAM) (Selvaraju et al., 2017). Saliency maps are a popular XAI tool in medical imaging (Irvin et al., 2019; Papanastasopoulos et al., 2020). Despite their widespread use, they have several shortcomings, such as eliciting confirmation bias (Adebayo et al., 2018; Bornstein, 2016), lacking robustness (Arun et al., 2021), and conveying limited information by just showing *where* the model looked, and not *why* it looked there.

Another type of feature attribution XAI method are counterfactual models, where the explanation aims to show how the input features need to change for the output to change (Wachter et al., 2017). For example, if a tumor scan is predicted as malign, a counterfactual explanation is the most similar image leading the model to label it as benign (Verma et al., 2020). Counterfactual explanations are usually generated using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) or by modifying the latent space of autoencoders. Schutte et al. (2021) used a GAN network to synthetically generate counterfactual images in knee osteoarthritis severity prediction and tumor detection on histology images of metastatic lymph nodes.

Beyond Feature Attribution Another type of post-hoc explainers are latent space disentanglement methods, which focus on mapping internal weights and representations to specific responsibilities and/or human-understandable concepts. Concept-based explainers attempt to interpret a model’s latent space by linking learned representations to predefined human concepts. A prominent example is TCAV (Testing with Concept Activation Vectors), which maps the latent space to a set of user-defined concepts (Kim et al., 2018a). Variants of TCAV, such as continuous concept detection (Graziani et al., 2018) and automated concept extraction (Ghorbani et al., 2019), have been applied to histopathology and segmentation tasks, respectively.

Alternatively, feature reduction techniques such as PCA or t-SNE can also be considered ways to disentangle the feature space by clustering similar images together (Dinsdale et al., 2021; Yang et al., 2022).

Mechanistic interpretability has recently emerged as a distinct direction in post-hoc explainability, focusing on reverse-engineering neural networks to understand their internal mechanisms. Unlike concept-based methods that map representations to human-understandable categories, mechanistic interpretability aims to identify specific neurons, attention heads, or “circuits” responsible for particular functions. The term was first coined by Olah et al. (2020) when discovering circuits—sub-networks within a model that replicate the full model’s performance on specific tasks. Other research has shown that certain neurons in generative language models are responsible for copying behaviors (Elhage et al., 2021), while attention heads have been linked to specific syntactic functions (Olsson et al., 2022).

2.3.1.2 Self-explaining Models

Self-explaining models provide insights into their decision-making process by design, i.e., they are meant to be inherently interpretable. These models can be divided into two broad categories: classical machine learning models that are inherently transparent and neural architectures that were modified to enhance transparency. Classical examples include decision trees and linear regression models, where interpretability is an intrinsic part of the model. In contrast, neural architectures usually require explicit constraints to achieve such levels of interpretability. These constraints can take various forms, including bottleneck representations, case-based reasoning mechanisms, or rule-based constraints.

Bottleneck architectures enforce interpretability by limiting the amount of information that

can pass through a specified layer in the network. [Koh et al. \(2020\)](#) introduced a concept bottleneck layer, where the model first predicts the presence of relevant human-understandable concepts relevant to osteoarthritis grading, and then bases its final prediction solely on these concepts. [LaLonde et al. \(2020\)](#) also developed a self-explaining concept-based model. They address the task of lung nodule malignancy prediction via an architecture that encodes human-understandable visual attributes (e.g. sphericity, margin, texture) within capsule vectors. [Shen et al. \(2019\)](#) developed a model for lung nodule malignancy classification that uses different submodules that each focus on a given set of features and concepts.

Prototype-based models (also called case-based explanations) predict by comparing input samples to a set of learned prototypes (e.g. examples from the trainset) ([Li et al., 2018b](#)). ProtoPNet ([Chen et al., 2019](#)) introduced a prototype layer that learns class-discriminative prototypes, balancing a cluster cost and separation cost to enhance interpretability. This architecture, and enhanced versions of it, have been used for Alzheimer’s disease classification ([Mohammad-jafari et al., 2021](#)), mammography classification ([Barnett et al., 2021](#)), and chest X-rays ([Kim et al., 2021](#)).

Other self-explaining models incorporate rule-based mechanisms that enforce interpretability. [Pisov et al. \(2019\)](#) incorporate a hand-crafted segmentation method based on structural knowledge about the brain with a deep learning approach to estimate brain midline shift more accurately and more interpretably. More precisely, they use structural knowledge to detect the midline and then predict the midline shift from that estimated curve. [Zhu and Ogino \(2019\)](#) use clinical guidelines about the features of lung nodules and incorporate them in their deep learning pipeline. Using KernelSHAP ([Lundberg and Lee, 2017](#)), they can then refer to clinical features that guided the decision.

2.3.1.3 Natural Language Explanations

Natural language explanations (NLEs), the explanation type that this thesis focuses on, are free-text explanations provided by a model to explain its decision-making. They are motivated by their user-friendliness in explaining model decisions to humans and their ability to convey complex reasoning. NLEs can generally be classified into either a *predict-then-explain* approach (first answering the task and conditioning the explanation generation on the answer and learned representation of the input) or the *explain-then-predict* framework (first generating

an explanation and extracting the answer from that, now also commonly referred to as Chain-of-Thought reasoning (Wei et al., 2022)) (Camburu et al., 2018). Besides early adoptions of textual explanations for traditional AI (Core et al., 2006; Lane et al., 2005), NLEs for deep learning models have first been adopted in vision (Hendricks et al., 2016). They are now most commonly applied in NLP problems (Camburu et al., 2018; Majumder et al., 2022; Narang et al., 2020; Kumar and Talukdar, 2020). *Explain-then-predict* NLE architectures generally consist of a task model M_T and an explanation generator M_E . M_T solves the task at hand, and M_E generates an explanation conditioned on the representation that M_T learned to solve the task. For *explain-then-predict*, M_T solves the task based on the output of M_E (in CoT, there is generally only a single model that generates reasoning steps and then provides an answer).

NLEs lie at the intersection of post-hoc explainers and self-explaining models. For the *predict-then-explain* framework, the model’s prediction does not directly rely on the generated explanation. However, the training process imposes a constraint on the model by ensuring that the learned input representations must also yield correct natural language explanations.

2.3.2 Properties of Explanations

Key properties to assess explanations include plausibility, faithfulness, and correctness. Each of these properties serves a distinct role in determining the utility and reliability of explanations.

Explanation Plausability Plausibility refers to the extent to which an explanation appears convincing or reasonable to a human observer. The concept is important for creating human-centred AI systems that foster trust and comprehension among users. As highlighted by Jacovi and Goldberg (2020), plausibility is sometimes conflated with terms such as readability or persuasiveness (Herman, 2017; Lage et al., 2019). Reliance on plausibility alone can pose significant risks. As noted by Wiegrefe and Pinter (2019), overly plausible explanations can mask model biases or errors, leading to inappropriate reliance on the system. This holds true in particular for natural language explanations (NLEs) (Atanasova et al., 2023; Agarwal et al., 2024).

Explanation Faithfulness Faithfulness refers to the degree to which an explanation accurately reflects the underlying reasoning process of the AI model. Jacovi and Goldberg (2020) define an explanation as faithful if it “accurately represents the reasoning process behind the

model’s prediction.” This definition highlights the critical distinction between explanations that merely sound convincing and those that truly reflect the model’s internal decision-making pathways.

Achieving faithfulness is elusive, and approaches across a wide range of explanation types have been shown to fail at achieving this, including post-hoc (Adebayo et al., 2018), interpretable-by-design (Jacovi and Goldberg, 2020), and CoT reasoning (Lanham et al., 2023) explanations.

Explanation Correctness I consider explanation correctness as analogous to notions of factual consistency in text generation and summarization tasks (Honovich et al., 2022; Zhang et al., 2020), i.e., whether the explanation’s content is factually accurate. Morrison et al. (2024) propose a similar definition, equating explanation correctness with factual accuracy. However, they conflate correctness with faithfulness, whereas I argue that these properties are distinct. An explanation can be correct without being faithful and vice versa. For example, consider a model that predicts pneumonia from an X-ray and generates the explanation: “Right lower lung opacity suggests pneumonia.” This explanation could faithfully reflect the model’s decision-making process but be factually incorrect if the opacity is actually on the left side. Conversely, the explanation could be factually correct, while the model internally identifies the opacity on the left side, rendering the explanation unfaithful. While explanation correctness may suggest faithfulness on the surface, they remain distinct properties.

2.3.3 Evaluating Explanations with Humans

Evaluating AI explanations is less straightforward than evaluating, e.g., prediction performance. The lack of unique ground-truth, the wide range of interpretability goals, and the human-computer interaction aspect make this more difficult. For these reasons, a growing body of work is evaluating XAI methods through the lens of human subject studies, generally following one of the four predominant evaluation approaches described below.

User Preference Some studies directly measure human participants’ preferences for XAI explanations. For instance, Adebayo et al. (2020) simulated a quality assurance context, requesting participants to assess the deployment readiness of AI algorithms, which came with different kinds of explanations. Kaur et al. (2020) ask participants how well they understand different types of explanations, how well they align with their intuition, and whether they think the algo-

rithms are ready to be deployed. However, [Hase et al. \(2020\)](#) demonstrated that user preference does not correlate with how well users can predict model behaviour, a proxy for how transparent the model is. Additionally, there are concerns that humans are prone to confirmation bias, i.e., focusing on evidence that confirms preexisting expectations in a model explanation ([Wang et al., 2019](#)). There is also evidence that XAI methods can unreasonably increase the confidence in a model's prediction ([Kunkel et al., 2019](#); [Schaffer et al., 2019](#); [Ghassemi et al., 2018](#); [Eiband et al., 2019](#)).

Bias Discovery Other studies measured the usefulness of XAI explanations by how well they reveal model biases to users, typically done by explicitly programming biases into the models. For instance, [Sixt et al. \(2021\)](#) modify a dataset to control which features, or biases, the model learns and then verifies whether users can find out what features are used by the model.

Model Predictability Arguably, the closest proxy for *full* model transparency is to measure how well humans can predict a model's predictions on unseen data. If users can correctly predict the model on all unseen data, it means the model is entirely transparent to them. The process typically involves exposing users to a sample of model predictions and their associated explanations and then they are assessed on their ability to predict the model's decisions on unseen samples. While some works opt for this method on simplified problems ([Alqaraawi et al., 2020](#); [Colin et al., 2022](#); [Yang et al., 2019](#); [Shen and Huang, 2020](#)), its applicability to radiology is limited, as predictions are highly nuanced and explanations are complex and label-specific.

Human-AI Team Performance Another approach to evaluate the usefulness of XAI explanations is to measure how much they improve human performance in the AI-human collaborative setting. The goal of XAI in this setting is to guide the user to appropriate evidence when the model is correct, or shed light on faulty AI reasoning when it is wrong. [Chu et al. \(2020\)](#) measured the impact of XAI methods in helping users predict age given images of human faces. [Kim et al. \(2022\)](#) analyzed performance changes in a bird classification task under the guidance of various XAI techniques. In clinical applications, where practitioners see a need for explanations to justify “their decision-making in the context of a model's prediction” ([Tonekaboni et al., 2019](#)), this evaluation method is particularly well suited and hence also used later in this thesis.

2.4 Medical Image Analysis

Medical imaging is a cornerstone of modern healthcare and is crucial in disease diagnosis, treatment planning, and patient monitoring. The primary imaging modalities include X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, and Positron Emission Tomography, each offering different advantages for visualizing anatomical structures and pathological conditions. Applying deep learning to medical imaging has resulted in significant advancements in the field, transforming both the technology and expectations surrounding AI in healthcare ([Zhou et al., 2021](#)).

Before the ubiquity of deep learning, automated medical image analysis relied on handcrafted features and classical machine learning (e.g. support vector machines and decision trees). While these traditional approaches had some success, they were limited by the need for expert-engineered features and often struggled with the complexity and variability of medical images.

The advent of deep learning in the early 2010s marked a transformative shift: Convolutional neural networks (CNNs) were able to learn features directly from image data, given sufficient training examples. Early, impactful applications of deep learning in medical imaging include CNNs for detecting diabetic retinopathy in retinal fundus images ([Gulshan et al., 2016](#)) and CheXNet, which was able to detect pneumonia from chest X-rays with performance comparable to radiologists ([Rajpurkar et al., 2017](#)). In oncology, models such as U-Net revolutionized image segmentation, particularly for tumor detection in MRI and CT scans ([Ronneberger et al., 2015](#)). [McKinney et al. \(2020\)](#) developed a deep learning model that achieved radiologist-level accuracy in detecting breast cancer in mammograms. However, the claims of human-level performance have been refuted in most cases ([Nagendran et al., 2020](#)). These results, while inspiring, also exemplified the need for careful evaluation and sparked discussion about benchmarking rigor and the importance of external validation. Given how protected and siloed healthcare data is, the size of public medical datasets is orders of magnitude smaller than the data used to train natural image models, meaning that the performance generally does not match their counterparts from natural images ([Varoquaux and Cheplygina, 2022](#)).

Despite numerous advances, deploying deep learning models in real-world clinical settings remains challenging. Significant barriers include issues such as data privacy, model generalizability ([Ktena et al., 2024](#)), and regulatory approval ([Mennella et al., 2024](#)). Moreover, the

black-box nature of many deep learning models has raised concerns regarding their reliability and trustworthiness in high-stakes medical decision-making ([Rasheed et al., 2022](#)). Recent efforts to develop more interpretable and transparent models, as well as frameworks for evaluating model fairness and robustness, are critical steps toward the responsible integration of AI into healthcare ([Esteva et al., 2019](#)). Below I will provide a background on chest X-rays and brain MRIs (the two modalities considered in this thesis), as well as the state of research in vision-language modeling for medical images.

2.4.1 Chest X-rays

Chest X-ray (CXR) imaging is one of the most widely used diagnostic tools for assessing thoracic conditions, primarily due to its cost-effectiveness, rapid acquisition, and ability to provide critical insights into a wide range of pulmonary and cardiac pathologies ([Speets et al., 2006](#)). The relatively simple nature of the imaging process has led to its ubiquitous application across healthcare settings, from high-resource hospitals to low-resource clinics ([DeStigter et al., 2021](#)). However, interpreting CXR images is challenging, often requiring highly skilled radiologists due to the complexities involved in detecting subtle pathologies amidst overlapping anatomical structures ([Hopstaken et al., 2004](#)).

The widespread use of chest X-rays has resulted in the availability of numerous large datasets, such as MIMIC-CXR ([Johnson et al., 2019](#)), CheXpert ([Irvin et al., 2019](#)), PadChest ([Bustos et al., 2020](#)), and ChestX-ray14 ([Wang et al., 2017](#)). Due to the wide availability of large-scale CXR datasets, they have become the most commonly used medical imaging modality for machine learning approaches ([Wang et al., 2017](#); [Johnson et al., 2019](#)). In clinical practice, it is common to document clinical findings of a medical image in a radiology report. These reports serve as a form of "free" annotation and are often included with these datasets (e.g., MIMIC-CXR, PadChest). This has enabled an extensive body of research on vision-language models for chest X-rays, with the predominant goal being automatic report generation.

2.4.2 Brain MRIs

Magnetic Resonance Imaging (MRI) is a brain imaging modality that uses radio waves and magnetic fields to induce changes in the spin directions of protons in water to provide detailed structural and functional insights vital for clinical diagnosis. Clinically, MRI is extensively

used to evaluate neurological conditions such as brain tumours, stroke, multiple sclerosis, and neurodegenerative disorders due to its superior soft-tissue contrast compared to other imaging techniques like CT scans (Hendee and Morgan, 1984). Over the decades, technological advancements have further enhanced MRI’s capabilities, enabling the visualization of subtle anatomical details, functional activity, and white matter connectivity (Le Bihan et al., 1986; Ogawa et al., 1990; Moseley et al., 1990).

Various MRI modalities have been developed, offering unique insights into brain structure and function. T1-weighted imaging provides excellent anatomical detail and is often used for volumetric analysis, while T2-weighted imaging is particularly sensitive to tissue water content and pathological changes (Hendee and Morgan, 1984). Fluid-attenuated inversion recovery (FLAIR) sequences suppress cerebrospinal fluid signals, enhancing the visibility of periventricular lesions (Hajnal et al., 1992). Diffusion-weighted imaging and diffusion tensor imaging allow for assessing water molecule diffusion, crucial for early detection of ischemic stroke and white matter tract analysis (Moseley et al., 1990; Basser et al., 1994).

Before the widespread use of deep learning, attempts to automatically analyze brain MRIs relied on extracting hand-crafted features from the images, for example, using watershed transforms and anatomical priors to segment gray and white matter (Grau et al., 2004). Nowadays, deep learning techniques are widespread and ushered significant improvements in various downstream tasks, such as classification (Parisot et al., 2018), metric learning (Ktena et al., 2018), segmentation (Havaei et al., 2017) or image registration (Balakrishnan et al., 2018).

2.4.3 Vision and Language

The intersection of computer vision and natural language processing, vision-language modeling, has opened a new frontier in medical imaging AI. By combining visual data (e.g., radiographs) with textual data (e.g., radiology reports), these models aim to achieve a higher level of understanding and provide more accessible outputs. Two prominent application areas have emerged: automated report generation and multimodal representation learning. Underlying all of these is the goal of creating joint representations of images and language, which enables AI models to connect images with the rich medical knowledge embedded in text. VL datasets occur naturally in the medical imaging domain, as radiologists routinely write reports to describe findings in medical scans. Large-scale public datasets are predominantly available for

chest X-rays (Irvin et al., 2019; Johnson et al., 2019; Bustos et al., 2020), and as such most VL approaches for medical imaging focus on this domain (Wang et al., 2018; Bannur et al., 2024). Recently there have been efforts to publish datasets in more advanced imaging modalities, such as lung CTs (Chen et al., 2024a), or extract datasets from medical literature (Zhang et al., 2023). As a consequence, an increasing number of VLMs are now being trained across multiple imaging modalities (Tu et al., 2024; Yang et al., 2024).

Report Generation When analyzing medical images, radiologists usually summarize the main findings, suggest a differential diagnosis, or suggest follow-up procedures in a radiology report (Pons et al., 2016). A long-standing goal is for models to generate these reports automatically. Initial attempts relied on traditional encoder-decoder architectures: a CNN would encode the image, and an RNN would decode a sequence of words, often augmented by attention mechanisms to focus on image regions while generating each phrase (Zhang et al., 2017; Gale et al., 2018; Wang et al., 2018). Later approaches adopted transformer-based architectures to achieve this (Chen et al., 2020c; Hou et al., 2021). As even free-text radiology reports often follow a template-like structure, some methods generate the reports in a semi-structured way (Gale et al., 2018; Li et al., 2018a; Tanida et al., 2023).

Pioneering works include MDNet (Zhang et al., 2017), which combined a ResNet with an LSTM to generate reports for bladder pathology images. TieNet (Wang et al., 2018) is another early attention-based CNN with RNN architecture that generated reports for chest X-rays. Advances over these baseline models came via the use of reinforcement learning to optimize for clinical metrics (Li et al., 2018a), augmenting the generation process with retrieved knowledge (Li et al., 2019a), or introducing transformer-based architectures (Chen et al., 2020c; Hou et al., 2021). Beyond chest X-rays, some works have looked at reports for retinal images (Huang et al., 2021a) or spinal MRI images (Han et al., 2021).

Since 2023, significant progress in radiology report generation has been driven by large multimodal models that improve both accuracy and interactivity. The Maira models, starting with Maira-1 (Hyland et al., 2023), have focused on generating more consistent radiology reports by leveraging improved vision encoders connected to LLMs. Models such as LLaVA-Med (Li et al., 2023a) were trained based on instruction following principles, meaning they can be used for zero-shot question answering. By fine-tuning on a sizable corpus of chest X-rays and reports, Flamingo-CXR (Tanno et al., 2025) achieves striking results. The authors claim that

in over 60% of cases, at least one radiologist preferred the AI-generated report to a human-written one. Med-Gemini (Saab et al., 2024) is another large-scale effort toward general-purpose biomedical AI, although the models are proprietary. In addition to general report generation, models like RGRG (Tanida et al., 2023) or Maira-2 (Bannur et al., 2024) have introduced region-guided approaches that allow users to query specific areas of medical images via bounding boxes or textual prompts. These developments mark a shift from static report generation toward more interactive, localized, and clinically adaptable systems.

Image Representation Learning By providing detailed information about the image, radiology reports can be used to learn richer representations of the image, which can ultimately lead to better performance on computer vision tasks. Early works such as TieNet (Wang et al., 2018) showed that joint training on images and reports had improved performance for classifying 14 different thoracic diseases on chest X-rays. Chauhan et al. (2020) showed that leveraging radiology reports can improve edema severity classification performance. The model trained jointly on images and reports (using an image-text matching loss) but only considers images during inference. However, the most effective and most commonly used approach these days has been contrastive learning, where a network is trained to maximize the agreement between true image-report pairs and minimize it between random pairs (Zhang et al., 2022b; Huang et al., 2021b; Wang et al., 2022). ConVIRT was a pioneering approach and demonstrated that with this form of self-supervision (no manual labels for diseases), the learned image encoder could be fine-tuned to perform classification or retrieval with far less labelled data than an ImageNet-pretrained model (Zhang et al., 2022b).

3 | Multi-modal AI That Can Explain Itself in Natural Language

Vision-language tasks, such as visual question answering, often require more complex reasoning than simple image classification tasks. NLEs are well-suited for this, as they are easy to understand for lay users and can explain fine-grained reasoning beyond just highlighting the important regions in an image. As such, this first chapter introduces vision-language models trained to explain their decision-making in natural language, which I call VL-NLE models.

While there have been increasing efforts to introduce such VL-NLE models, there is a lack of comparison between existing methods. Part of the reason is that VL-NLE is an emerging field, and only a few datasets exist. Moreover, existing datasets tend to be relatively small and unchallenging (e.g., VQA-X (Park et al., 2018)) or noisy (e.g., VQA-E (Li et al., 2018c)). Another limitation of the VL-NLE field is that there is no unified evaluation framework, i.e., no consensus on how to evaluate NLEs. NLEs are difficult to evaluate, as correct explanations can differ both in syntactic form and in semantic meaning. For example, “Because she has a big smile on her face” and “Because her team just scored a goal” can both be correct explanations for the answer “Yes” to the question “Does the girl look happy?” but existing automatic natural language generation (NLG) metrics are poor at capturing this. As such, the gold standard for assessing NLEs is human evaluation. Past works have all used different approaches for human evaluations, and therefore, no objective comparison between these methods exists.

In this chapter, I propose five main contributions to address the lack of comparison between existing work. (1) I propose e-ViL, the first benchmark for VL-NLE tasks. e-ViL spans three datasets of human-written NLEs and provides a unified evaluation framework designed to be reusable for future works. (2) Using e-ViL, I compare four VL-NLE models. (3) I introduce e-SNLI-VE-2.0, a dataset of over 430k instances, the currently largest dataset for VL-NLE. (4) I introduce a novel model called e-UG, which surpasses the state of the art by a large (and significant) margin across all three datasets. (5) I provide the currently largest study on the correlation between automatic NLG metrics and human evaluation of NLEs.

I will first give an overview of existing work (up until the time this work was completed) and situate this work within the literature (Section 3.1). In Section 3.2, I introduce e-SNLI-VE-2.0,

the e-UG model is introduced in Section 3.3, and lastly, the e-ViL benchmark is outlined in Section 3.4. This chapter is wholly based on [Kayser et al. \(2021\)](#).

3.1 Related Work

Explainability in Computer Vision Common approaches to explainability of deep learning methods in computer vision are saliency maps (Simonyan et al., 2014), attention maps (Xiao et al., 2015), and activation vectors (Kim et al., 2018a). Saliency and attention maps are meant to indicate where a model *looks* at. This may tell us what regions of the image are most important in the decision-making process of a model. Activation vectors are a way to make sense of the inner representation of a model, e.g., by mapping it to human-known concepts. However, these approaches often cover only a fraction of the reasoning of a model. On the contrary, NLEs can convey higher-order reasoning and describe complex concepts. For example, in the VCR examples of Figure 3.5, highlighted image regions or weights for different concepts would not be sufficient to explain the answer. Additionally, it has been shown that numerical or visual explanatory methods, in some cases, can be confusing even for data scientists (Kaur et al., 2020) and can pose problems even for explaining trivial models (Camburu et al., 2020).

NLE Datasets Before the ubiquity of LLMs, NLE generation relied on supervised training. Hence, they relied on training datasets with annotated explanations. There have been numerous efforts to build such datasets, with early ones including e-SNLI (Camburu et al., 2018) or CUB (Hendricks et al., 2016). For the CUB dataset, crowd workers provided explanations for why a particular bird species was identified in an image. These explanations aimed to capture distinguishing visual features that informed the classification decision (e.g., plumage colour or beak shape). e-SNLI (Camburu et al., 2018) provides explanations for natural language inference (NLI). Crowd workers were asked to justify why a given premise entails, contradicts, or is neutral to a hypothesis. Similarly, COS-E (Rajani et al., 2019) introduced common-sense QA explanations, where the explanations serve as rationales for the selected answers. While the dataset highlights the potential of text-based explanations, concerns have been raised about their quality, with many explanations being template-like or overly generic. The ECQA dataset (Aggarwal et al., 2021) addresses these issues by providing more diverse and robust explanations, emphasizing the importance of collecting multiple explanations per instance to capture the subjectivity inherent in natural language reasoning.

A handful of datasets have also been introduced for NLEs in vision-language tasks. The VQA-X dataset (Park et al., 2018) was introduced first and provides NLEs for a small subset of

questions from VQA v2 (Goyal et al., 2017). It consists of 33K QA pairs (28K images). However, many of the NLEs in VQA-X are trivial and could be guessed without looking at the image. For example, the NLE “because she is riding a wave on a surfboard” for the answer “surfing” to the question “What is the woman in the image doing?” can easily be guessed from the answer without looking at the image. VQA-E (Li et al., 2018c) is another dataset that builds on top of VQA v2. However, its explanations are gathered in an automatic way and were found to be of low-quality by Marasović et al. (2020), where the model-generated explanations obtain a human evaluation accuracy¹ that is only 3% short of the VQA-E ground-truth explanations (66.5%), suggesting that the dataset is essentially solved. The VCR dataset (Zellers et al., 2019) provides NLEs for VQA instances that require substantial commonsense knowledge. The questions are challenging, and therefore, both the answers and NLEs are given in the form of multiple-choice options. A survey on free-text rationales datasets by Wiegreffe and Marasovic (2022) shows that the vast majority of datasets are created via crowd-sourcing. For a more comprehensive overview, I point the reader to that survey. The proposed dataset, e-SNLI-VE-2.0, extends the range of VL-NLE datasets and addresses prior limitations around quality and scale. It contains over 430K instances for which the explanations rely on the image content (see examples in Figure 3.5). I will describe the dataset in more detail in Section 3.2. VCR and VQA-X are included in the e-ViL benchmark.

NLE Approaches First adoptions of NLEs have been in image classification (Hendricks et al., 2016) and were further extended to self-driving cars (Kim et al., 2018b), VQA (Park et al., 2018), and natural language processing (Camburu et al., 2018; Majumder et al., 2022; Narang et al., 2020; Kumar and Talukdar, 2020).

One of the first architectures able to provide NLEs in vision tasks was PJ-X (Park et al., 2018), which learns a multimodal, attention-weighted representation of the input image and text (e.g., a question). An LSTM-based decoder (Hochreiter and Schmidhuber, 1997) then generates an explanation from the multimodal representation, attention map, and answer. In Li et al. (2018c), the answer and explanation are generated simultaneously. The model FME by Wu and Mooney (2019) uses the UpDown architecture (Anderson et al., 2018) as a backbone for obtaining the multimodal representation of the input, which, together with the answer, is fed to an LSTM explanation generator. Marasović et al. (2020) use either an object detector, a

¹Percentage of explanations that, given the image and question, support the predicted answer.

situation recognition model, or VisualComet (Park et al., 2020) as backbone and then generate the explanation using GPT-2, a pre-trained language model (Radford et al., 2019). However, they do not predict the answers for the task itself but condition the explanations on the ground-truth answer. Their visual representation of the image is not trained but fixed. The proposed e-UG model leverages advances in multimodal transformers (Chen et al., 2020b) and language models (Radford et al., 2019) to provide both improved task and NLE performance. More details are provided in Section 3.3. The most important works in VL-NLE (Park et al., 2018; Wu and Mooney, 2019; Marasović et al., 2020) are included in this benchmark.

NLE Evaluation Evaluating natural language generation (NLG) is challenging and a much-studied field (Gatt and Krahmer, 2018). Evaluating NLEs is even more difficult, as sentences may not only differ in their syntactic form but also in their semantic meaning, e.g., there can be several different reasons why a hypothesis contradicts a premise. For this reason, current automatic NLG metrics, such as the BLEU score (Papineni et al., 2002), do not perform well in evaluating NLEs (Camburu et al., 2018). Several works have instead used human evaluation to assess their generated explanations (Park et al., 2018; Wu and Mooney, 2019; Marasović et al., 2020; Dua et al., 2021). However, they all used different evaluation rules, which does not make it possible to compare existing VL-NLE models. The main differences lie in the datasets used, the questions asked to annotators, whether the assessment is absolute or based on a ranking, and the formula used to calculate the final score. I select the best practices from existing evaluation schemes and develop a unified and reusable human evaluation framework for VL-NLE approaches.

3.2 e-SNLI-VE-2.0

This section introduces e-SNLI-VE-2.0², a large-scale dataset for visual-textual entailment with NLEs. It was built by merging the explanations from e-SNLI (Camburu et al., 2018) and the image-sentence pairs from SNLI-VE (Xie et al., 2019). Several filters and manual relabeling steps were used to address the challenges that arose from merging these datasets, which are described below. The validation and test sets were relabelled by human annotators. The dataset is publicly available³.

	Train	Validation	Test
# Image-Hypothesis pairs (# Images)	401,717 (29,783)	14,339 (1,000)	14,740 (1,000)
Label distribution (C/N/E, %)	36.0 / 31.3 / 32.6	39.4 / 24.0 / 36.6	38.8 / 25.8 / 35.4
Mean hypothesis length (median)	7.4 (7)	7.3 (7)	7.4 (7)
Mean explanation length (median)	12.4 (11)	13.3 (12)	13.3 (12)

Table 3.1: e-SNLI-VE-2.0 summary statistics. C, N, and E stand for Contradiction, Neutral, and Entailment, respectively.

3.2.1 Improving SNLI-VE

In SNLI-VE (Xie et al., 2019), an image and a textual hypothesis are given, and the task is to classify the relation between the image-premise and the textual hypothesis. The possible labels are *entailment* (if the hypothesis is true, given the image), *contradiction* (the hypothesis is false, given the image), or *neutral* (if there is not enough evidence to conclude whether the hypothesis is true or false). SNLI-VE builds off the SNLI (Bowman et al., 2015) dataset by replacing textual premises with Flickr30k images (Young et al., 2014). This is possible because the textual premises in SNLI are caption sentences of those images. However, as an image typically contains more information than a single caption describing it, this replacement led to labelling errors. Especially for the neutral class, it may be the case that, while a caption does not have enough evidence to suggest entailment or contradiction, the corresponding image does (see Figure 3.1). On a manually evaluated subset of 535 samples, I found a 38.6% error rate among the neutral labels. This subset is used to evaluate the effectiveness of our filters in the remainder of this chapter. As error rates for entailment and contradiction are reported to be under 1% (Xie et al., 2019), we focus only on correcting the neutral instances.

²2.0 refers to the improvement over a preliminary non-published version of the dataset (Do et al., 2020)

³<https://github.com/maximek3/e-ViL>

Do et al. (2020) relabelled the neutral examples in the validation and test sets using Amazon Mechanical Turk (MTurk). To ensure high-quality annotations, they used a series of quality control measures, such as in-browser checks, inserting trusted examples, and collecting three annotations per instance. In total, 39% of the neutral labels were changed to entailment or contradiction. The label distribution shifted from uniform to Ent/Neut/Cont of 39%/20%/41% and 39%/21%/40% for the validation and test sets, respectively.

For the training set, which is much larger, I propose an automatic way to remove false neutrals. I discovered that the five captions that come with each image often provide clues as to whether the label is indeed neutral or not. To leverage this, for every image-hypothesis pair i , I ran a natural language inference model m_{nli} on each caption-hypothesis pair $p_{i,c}$, where c is one of the captions. If the original label of image-hypothesis pair i is neutral, but $\sum_c m_{\text{nli}}(p_{i,c})$ indicates with high confidence that the label is not neutral, the label is deemed incorrect, and the instance is removed from the dataset. An example is shown in Figure 3.1. For

m_{nli} , I used Roberta-large (Liu et al., 2019c) trained on the MNLI dataset (Williams et al., 2018). Instances were removed if $\sum_c m_{\text{nli}}(p_{i,c})$ exceeded 2.0 for entailment and contradiction classes. On the 535-sample subset, this filter decreased the error of neutral labels from 39% to 24%. When validated against the relabeling on the validation set, the error decreased from 39% to 30%.



Hypothesis: A man and woman inside a church.
Textual premise: A man and woman getting married.
Original label: Neutral
Caption #2: A man and woman that is holding flowers smile *in the sunlight*.
Caption #4: A happy couple enjoying their *open air wedding*.

Figure 3.1: The original label of the textual premise-hypothesis pair in SNLI is neutral. However, by considering alternative captions describing the same image (#2 and #4), we can deduct that the neutral label for the multi-modal premise-hypothesis pair is false.

3.2.2 Adding Explanations to SNLI-VE

To create e-SNLI-VE-2.0, we sourced explanations from e-SNLI (Camburu et al., 2018), which extends SNLI with human-written NLEs. However, the explanations in e-SNLI are tailored to the textual premise-hypothesis pairs and are, therefore, not always well-suited for the image-hypothesis pair. After simply merging both datasets, I found that initially, 36%, 22%, and 42%

of explanations were low (incorrect), medium (correct, but there is an obvious better choice), and high quality (correct and relevant), respectively. I implemented several steps to detect and remove explanations of low and medium quality. The filters were designed to ensure an optimal trade-off between precision and recall (for flagging inadequate explanations) and with the constraint that the final dataset remains reasonably balanced.

Re-annotation Do et al. (2020) replaced the explanations for the neutral pairs in the validation and test sets with new ones simultaneously with collecting new labels via MTurk. In order to submit the annotation of an image-sentence pair, workers must choose a label, highlight words in the hypothesis, and use at least half of the highlighted words in the explanation. On a random sample of 100 explanations, they found that 83% of the newly annotated explanations are of high quality. The re-annotated explanations are used in this work.

Keyword Filter I used keyword filtering to detect explanations that refer to a linguistic feature of the textual premise, as these will not apply to the image. The keywords, which I manually defined, are “synonym”, “mention”, “rephrasing”, “sentence”, “way to say”, and “another word for”. The keyword filter removed 4.6% of all instances, and our 535-sample subset suggests that *all* filtered explanations were indeed of low quality (see example in Figure 3.2).

Similarity Filter The share of low-quality explanations was highest for entailment examples. I found that a frequent scenario for low-quality explanations was when the textual premise and hypothesis are almost identical, as then the explanation often just repeats both statements. To overcome this, I removed all examples where the ROUGE-1 score (a measure for sentence similarity (Lin, 2004)) between the textual premise and hypothesis was above 0.57, a threshold that was determined empirically on our annotated subset. This reduced the share of low-quality explanations for entailment by 4.2% (see example in Figure 3.3).

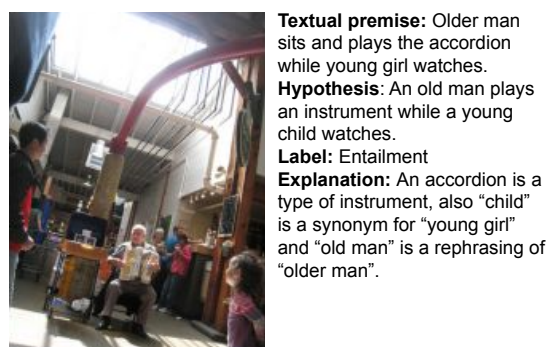


Figure 3.2: The use of the words “synonym” and “rephrasing” makes it clear that the explanation is overly focused on the linguistic features of the textual premise.

Uncertainty Filter Lastly, I found that image-hypothesis pairs with high uncertainty are correlated with low-quality explanations for contradictions. Uncertainty is defined as the standard deviation of the scores from $m_{\text{nli}}(p_{i,c})$ for the five image captions. m_{nli} is the same Roberta-large model that was described above. This filter reduced the share of low-quality explanations for contradiction examples by 5.1%.



Textual Premise: A mother stands in a kitchen holding a small baby
Hypothesis: A mother is holding a small baby.
Label: Entailment
Explanation: A mother standing in the kitchen holding a small baby is the same as a mother holding a small baby.

Figure 3.3: The textual premise and hypothesis are almost identical sentences, which led to a low-quality explanation.

Evaluation of e-SNLI-VE-2.0 A quantitative analysis of the effects of the filters on the dataset is provided in Table 3.2. The accuracies are obtained by validating against our hand-annotated subset of 535 examples. Explanations were annotated *bad*, *okay*, or *great*. If the label is wrong, the explanation is automatically deemed *bad*, as it will try to explain a wrong answer. Note that I obtained higher error rates for the Entailment and Contradiction classes (9.7% and 8.6%) than the SNLI-VE paper’s authors claimed (less than 1%). This could be explained by the inherent ambiguity in the task.

Note that in e-SNLI, the authors have found that the human-annotated explanations have an error rate of 9.6% (19.6% on entailment, 7.3% on neutral, 9.4% on contradiction), which serves as an upper bound of what could be achieved in terms of dataset cleaning.

	Dataset Size			% of Wrong Labels				% of Bad Explanations			
	Train Set	Val Set	Test Set	All	E	N	C	All	E	N	C
Raw	529,505	17,554	17,899	19.3	9.7	38.6	8.6	35.7	35.2	45.1	26.3
FN Removal	481,479	17,554	17,899	13.0	9.7	23.5	8.6	31.3	35.2	32.6	26.3
KW Filter	459,353	16,862	17,188	13.4	10.1	23.7	8.8	28.0	28.3	32.1	24.6
Uncertainty Filter	429,774	15,402	15,829	12.5	10.1	23.7	4.5	26.7	28.3	32.1	19.5
Similarity Filter	401,717	14,339	14,740	12.8	10.5	23.7	4.5	25.2	24.1	32.1	19.5

Table 3.2: Each row describes the state of the dataset upon application of the given filter. The share of wrong labels and bad explanations is only representative of the training split. The first row describes the state of the dataset in its raw form, i.e., before any of the automatic filtering steps. The second row describes the state of the datasets upon application of the false neutral (FN) removal filter, etc.

I also collected ground-truth evaluations from human annotators for e-SNLI-VE-2.0 and the other two benchmark datasets, VQA-X and VCR. For each dataset, 300 examples are evaluated by 12 annotators each, resulting in 3,600 evaluations. For each explanation, participants responded to the question “Given the image and the question/hypothesis, does the explanation justify the answer?” with *no*, *weak no*, *weak yes* or *yes*. The results in Table 3.3 show that e-SNLI-VE-2.0 comes close to the manually annotated datasets VCR and VQA-X (82.8% explanations with *yes* or *weak yes* vs. 87.9% and 91.4%). Besides the use of effective but imperfect automatic filters, another explanation for the higher share of lower-rated NLEs is the difficulty of the task (out of 100 human re-annotated explanations for neutral examples, we found that 17% had flaws, identical to the share of (weak) no in Table 3.3) and ambiguity inherent in the task (when three of us chose the labels for a set of 100 image-hypothesis pairs, we only had full agreement on 54% of examples).

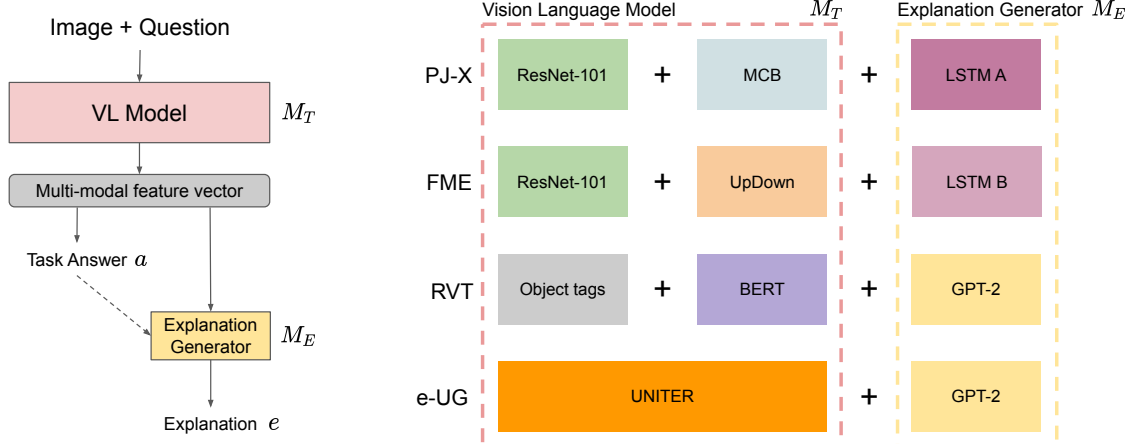
	No	Weak No	Weak Yes	Yes
e-SNLI-VE-2.0	10.3%	6.9%	27.7%	55.1%
VQA-X	4.1%	4.5%	25.1%	66.3%
VCR	6.9%	5.2%	36.6%	51.3%

Table 3.3: Human evaluation of the ground-truth explanations for the three datasets used in e-ViL. The question asked was: “Given the image and the question/hypothesis, does the explanation justify the answer?”

3.3 Approaches for the VL-NLE Task

This section contains an overview of VL-NLE architectures, details on the re-implementations done for our benchmark, and e-UG, a new model that leverages recent developments to set a new state-of-the-art performance.

3.3.1 VL-NLE Architectures



(a) High-level structure of VL models. (b) The components of the models that are evaluated.

Figure 3.4: High-level architectures of the models that are included in the e-ViL benchmark.

Task Formulation I denote a module that solves a VL task as M_T , which takes as input visual information V and textual information L . Its objective is to complete a task T where the outcome is a , i.e., $M_T(V, L) = a$. An example of a VL task is VQA, where V is an image, L is a question, and T is the task of providing the answer a to that question. This is extended by an additional task E , which requires an NLE e justifying how V and L lead to a , solved by the module $M_E(V, L) = e$. The final model M then consists of M_T and M_E . Thus, $M = (M_T, M_E)$ and $M(V, L) = a, e$.

Existing VL-NLE models follow a common high-level structure (Figure 3.4). First, a VL model learns a joint representation of the image and language inputs and predicts the answer. The models in this chapter then condition their explanation on different combinations of the question, image, their joint representation, and the answer. As M_T models, I consider MCB (Fukui et al., 2016), UpDown (Anderson et al., 2018), BERT (Devlin et al., 2019) with object tags, and UNITER (Chen et al., 2020b). The decoding of explanations is done via LSTM architectures or GPT-2.

PJ-X The PJ-X model (Park et al., 2018) provides multimodal explanations for VQA tasks and was originally evaluated on VQA-X. Its M_T module consists of a simplified MCB network that was pre-trained on VQA v2. This simplified version achieves performance similar to that of the original MCB network. An element-wise multiplication of visual and language features yields a multimodal embedding, which, combined with an attention mechanism, yields the final answer. To generate an NLE, the combined visual and language inputs, as well as the embedded answer predicted by the model, are fed to a separate visual attention operation over the visual features. The weights of the explanation attention are the visual part of the multimodal explanations, as they visually *point* to the relevant parts of the image. Similarly to the VQA part, the output of the attention operation is again combined with the embedded answer and the question embedding and is then used as the conditioning for a two-layer LSTM, which decodes the explanation auto-regressively.

PJ-X was re-implemented in PyTorch following the authors’ implementation in Caffe⁴ closely. To address numerical optimization problems, the L2 normalization in the decoder was replaced with layer normalization, as the original normalization zeroed gradients for earlier model parts. Additionally, I added gradient clipping of 0.1 to prevent too large gradients. To adapt PJ-X for multiple-choice question-answering (QA) in VCR, I followed the approach in the original VCR paper (Zellers et al., 2019). All other optimization settings were kept unchanged as in the original code.

FME The model introduced by Wu and Mooney (2019), referred to as FME (Faithful Multimodal Explanations), emphasizes producing faithful explanations by ensuring that the explanation uses the same visual features used to produce the answer. As their code is not publicly available, I re-implemented their base model according to the instructions in the paper. The base model was chosen as the modifications of the model variants were not feasible to re-implement from the descriptions in the paper, and, according to the authors, the added losses only sped up convergence. The re-implementation of FME is based on a frozen modified Up-Down (Anderson et al., 2018) VQA v2 pre-trained VL-model. A question-guided attention over the input image is then combined with a question embedding to produce a joint multimodal representation, which is used to predict the VQA answer. An LSTM conditioned on the answer embedding, visual features, and question embedding generates the explanation. Simi-

⁴<https://github.com/Seth-Park/MultimodalExplanations>

larly to PJ-X, I also train FME with gradient clipping of 0.1 and adopt the same approach for multiple-choice QA.

RVT The Rationale-VT Transformer (RVT) model (Marasović et al., 2020) uses a range of vision algorithms to extract information from an image and then feeds this information, the ground-truth answer, and the question to a pre-trained GPT-2 language model (Radford et al., 2019), which yields an explanation. As they omit the question-answering part, I extend their model by an answer prediction module to allow for a fair comparison and to get a sense of the overall performance. I use their overall most effective visual input⁵, which are the tags of the objects detected in the image. As task model M_T , I use BERT (Devlin et al., 2019), which takes as input the object tags and the question and predicts the answer.

e-UG Marasović et al. (2020) generate convincing explanations, but out of various M_T modules tested, including complex visual reasoning models, it obtains the best explanation accuracy when using object labels as the sole image information. However, in many cases, object labels do not contain sufficient information to explain the answer to the question. This suggests that GPT-2 is able to learn explanations mostly from the questions and answers alone. To condition more strongly on the image and, at the same time, use the general knowledge and fluency of GPT-2, I propose a novel model that combines GPT-2 with UNITER (Chen et al., 2020b), a powerful transformer-based VL model. Similar to BERT, UNITER leverages self-attention mechanisms to learn contextualized embeddings of image-text pairs. It uses tasks such as Masked Language Modelling, Masked Region Modelling, Image-Text Matching, and Word-Region Alignment to pre-train on large image captioning datasets. The outputs of UNITER are contextualized embeddings of the word tokens and image regions in the image-text pair. Words are embedded by tokenizing them into WordPieces and adding their position embedding. Images are embedded by extracting visual features of regions with Faster R-CNN (Ren et al., 2016) and encoding their location features. UNITER achieves SOTA on many downstream tasks when fine-tuned on them. For e-UG, I leverage these contextualized embeddings to condition GPT-2 on an efficient representation of the image and question. The embeddings of the image regions and question words are prepended to the textual question and predicted answer and then fed to GPT-2. GPT-2 is a decoder-only architecture that is pre-trained on conventional language modelling and, therefore, well-suited for language generation. As in Marasović et al.

⁵It obtained the highest *visual plausibility* score averaged across all datasets.

(2020), greedy decoding is used during inference.

3.3.2 Training

All models are trained separately on each dataset, meaning there are a total of 12 model-dataset combinations. To ensure comparability, image features for PJ-X and FME are obtained from the same ResNet-101 (He et al., 2016) pre-trained on ImageNet, which yields a 2048d feature representation for an image with 7×7 spatial dimension. To account for the small size of VQA-X, the VQA M_T models were pre-trained on VQA v2 for VQA-X and trained from scratch for the other two datasets. The object tags in RVT are obtained from a Faster R-CNN that was trained on ImageNet and COCO. For GPT-2, the pre-trained weights of the original GPT-2 with 124M parameters are used (Radford et al., 2019).

Joint or Separate Training All the VL-NLE models M in this work consist of M_T and M_E modules, which can either be trained jointly or separately. This allows to investigate the influence of the explanations on a model’s performance on the original task, i.e., whether the performance of M_T changes when trained jointly with M_E . Some works showed that NLEs helped improve their model’s performance (Rajani et al., 2019) while others show that, on the contrary, training both jointly has a (slightly) negative effect on the performance on the original task (Camburu et al., 2018). For the RVT model, training jointly would make no difference, as the explanation generation is not conditioned on a learnable representation in M_T (but instead on the fixed object tags for each image). For all other models, training jointly can be advantageous because the explanation loss backpropagates through the task model M_T , but this also comes at the risk of adverse effects on the optimization (Caruana, 1997). The authors of the PJ-X model mentioned that they tried both training approaches, but they did not specify which one worked best. Wu and Mooney (2019) only trained separately. It should be noted that PJ-X and FME were both solely run on VQA-X, where a much larger dataset (VQA v2) exists for task T . They pre-train M_T separately on this dataset, and it could be argued that, when training jointly, M_T runs the risk of becoming worse by overfitting on the smaller dataset VQA-X. For e-SNLI-VE-2.0 and VCR, no such pre-training dataset exists. For all models in this work, I experimented with training the M_T and M_E modules jointly and separately.

Reproducing Previous Results The code for RVT was publicly available, and only a classifier had to be added to it. The code of PJ-X is also publicly available, albeit in an outdated version of the Caffe framework, and therefore it was translated into Pytorch. For FME, no code is available, and thus, I re-implemented their model (as much as possible) according to the instructions given in the paper (Wu and Mooney, 2019). In Table 3.4, I show that the NLG metrics of our re-implementations come very close to those reported in the original papers.

Model		BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
PJ-X	<i>Original</i>	19.8	18.6	44.0	73.4	15.4
	<i>Ours</i>	20.1	18.3	43.0	71.8	15.3
FME	<i>Original</i>	23.5	19.0	46.2	81.2	17.2
	<i>Ours</i>	20.8	19.2	44.8	77.9	16.7

Table 3.4: A comparison (under the same settings) of automatic NLG metrics on VQA-X between our re-implementations (*Ours*) of PJ-X and FME and the results reported in the papers (*Original*).

Hyperparameters Choosing hyperparameters via human evaluation is prohibitively expensive. Instead, best hyperparameters were selected using a set of automatic NLG metrics (harmonic mean of the BERTScore F1 (Zhang et al., 2019) and NGRAMScore, defined as the harmonic mean of the n -gram NLG metrics ROUGE-L (Lin and Och, 2004), SPICE (Anderson et al., 2016), CIDEr (Vedantam et al., 2015), and METEOR (Banerjee and Lavie, 2005)). For PJ-X and FME, I choose the same hyperparameters as the authors across all datasets. For PJ-X, I also experimented with larger learning rates as I experienced convergence issues. For RVT and e-UG, I conducted grid search on three batch sizes, three learning rates, and three ways to combine the loss. I compared dynamic weight loss (with two loss temperatures $T = 2$ and $T = 0.5$) (Liu et al., 2019b) to simply adding both losses. However, this did not affect our results enough to warrant increased complexity. The best configuration was selected on VQA-X and then used to train on e-SNLI-VE-2.0 and VCR. For BERT on VCR, I had to use a higher batch size (128) to ensure convergence. The final hyperparameters for all four models are reported in Table 3.5. Architectural differences are summarized in Table 3.6.

Teacher forcing of answers For all the approaches, the explanation generator M_E is conditioned on the task’s answer. In the literature, there are three approaches to conceptualize this. In the first one (1), the predicted answer $a = M_T(V, L)$ is fed to the explanation generator M_E . M_E then predicts e , based off the final representation of M_T and the answer a . An issue with

	PJ-X	FME	RVT	e-UG
Batch Size	128	128	32* / 64	64
Learning Rate (LR)	7×10^{-4}	5×10^{-4}	5×10^{-5}	2×10^{-5}
Training Type	JOINT*	JOINT*	SEPARATE	JOINT
Loss Combination	$\mathcal{L}_T + \mathcal{L}_E$	$\mathcal{L}_T + \mathcal{L}_E$	N.A.	$\mathcal{L}_T + \mathcal{L}_E$
Optimizer	Adam	Adam	AdamW	AdamW for BERT
LR Scheduler	-	Step decay	Linear w/ warmup	Linear w/ warmup
Tokenization	Word	Word	WordPiece	WordPiece
Decoding	Greedy	Greedy	Greedy	Greedy

Table 3.5: Hyperparameters used for the different models across all datasets. $\mathcal{L}_T$ and $\mathcal{L}_E$ are the task loss and explanation loss, respectively. For RVT, the task batch size for VCR is 128, as 32 did not lead to convergence. *For PJ-X and FME, M_T and M_E were trained separately on VQA-X.

Model M	Vision Backbone	M_T	M_E	M_E Input
PJ-X	ResNet-101	MCB	LSTM (a)	image features, question, answer
FaiMu	ResNet-101	UpDown	LSTM (b)	image features, question, answer
RVT	Faster R-CNN	BERT	GPT-2	object tags, question, answer
e-UG	Faster R-CNN	UNITER	GPT-2	contextualized embeddings of image-question pair, question, answer

Table 3.6: Summary of the model differences.

that approach is that if a is predicted incorrectly, the model is trained to generate an explanation that is incoherent with the answer it is conditioned on. In the second option (2), explanations are only generated for examples where a was correct. That way, explanations are always conditioned on the right label, but we waste the training signal from all the explanations where the label is incorrect, which can be substantial. Finally, in the third option (3), we always condition M_E on the ground-truth answer during training. This way, we make use of all the training data and the explanations are always conditioned on the right answer. This is conceptually similar to teacher forcing in language modelling. For a fair comparison, I opt for method (3) and use it across all tasks. PJ-X used approach (1), and FM used approach (2). I have confirmed that the models' performances were not significantly impacted by changing this training regime.

3.4 The e-ViL Benchmark

In this section, I introduce the e-ViL (explainable **V**ision **L**anguage) benchmark, including its suite of datasets and automated and human evaluation frameworks used.

3.4.1 Datasets

Our benchmark uses the following three datasets, which vary in size and domain. Examples for each dataset are shown in Figure 3.5.

e-SNLI-VE-2.0 Our proposed e-SNLI-VE-2.0 dataset consists of over 430K image-hypothesis pairs, where the task T is to classify the pairs as entailment, neutral, or contradiction. More details on the dataset can be found in Section 3.2.

VQA-X VQA-X (Park et al., 2018) contains human written explanations for a subset of questions from the VQA v2 dataset (Goyal et al., 2017). The image-question pairs are split into train, dev, and test with 29.5K, 1.5K, and 2K instances, respectively. VQA-X only includes questions that require at least the intellect of an average 9-year-old. This ensures that overly easy questions, such as “What colour is the car?”, are excluded. The task T is a multi-label classification task of 3,129 different classes. One question can have multiple possible answers, as multiple people have annotated each example.

VCR Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) is a VL dataset that asks multiple-choice (single-answer) questions about image frames from movies. In addition to four answer options, it also provides four NLE options, of which one is correct. For the proposed VL-NLE task, I reformulate it as an explanation generation task. As the test set for VCR is not publicly available, I split the original train set into a train and dev set and used the original validation set as a test set. The splits are of size 191.6K, 21.3K, and 26.5K, respectively.

The representative examples in Figure 3.5 illustrate the key differences between the datasets. VCR has many questions that require substantial commonsense reasoning and general knowledge. For example, to explain the answer to the question in Figure 3.5e, one needs to know that person 1 is wearing a T-shirt of the classic rock band Guns N’ Roses. For VQA-X (Figures 3.5c and 3.5d), two examples are shown where a generic explanation, not necessarily grounded



Hypothesis: The man rode his bike. **Relation:** Contradiction

(a) e-SNLI-VE-2.0.



Hypothesis: A male has a hat on. **Relation:** Entailment

(b) e-SNLI-VE-2.0.



Question: What sport is this person doing? **Answer:** surfing

(c) VQA-X.



Question: What room is this? **Answer:** bedroom

(d) VQA-X.



Question: What kind of music does 1 like?
Answer: He likes classic rock.

(e) VCR.



Question: What is 2 doing?
Answer: 2 is communicating with someone outside of the room to give them instruction.

(f) VCR.

Figure 3.5: Representative examples from each e-ViL dataset.

in the image, will often suffice (this is a general limitation of this dataset). The explanations “Because there is a person on a surfboard” and “Because there is a bed in the room” will, in most cases, be correct with respect to the question and answer, regardless of the image. The examples for e-SNLI-VE-2.0 in Figures 3.5a and 3.5b both require the explanations to describe image-specific characteristics in order to be meaningful. In Figure 3.5a, a valid explanation would have to pick out a concrete element from the image to explain why it is a contradiction.

3.4.2 Evaluation

Evaluation Metrics In order to assess models on the VL-NLE task, I define separate evaluation scores S_T , S_E , and S_O for M_T , M_E , and M , respectively. S_T is the metric defined by the original VL task T , e.g., label accuracy for e-SNLI-VE-2.0 and VCR and VQA accuracy for VQA-X. S_E is defined as the average explanation score of the examples for which the answer a was predicted correctly. In line with previous work (Park et al., 2018; Wu and Mooney, 2019; Marasović et al., 2020), I assume the simplified scenario that an explanation is always false when it justifies an incorrect answer. The explanation score can be any custom human or automatic metric. Due to the limitations of current automated NLG metrics for evaluating NLEs, I developed a human evaluation framework for computing S_E , outlined later.

S_O is meant to summarize the performance of a model on both tasks T and E , to give an overall performance assessment of a VL-NLE model M . I define $S_O = S_T S_E$, which equates to the average of the scores of all explanations, but where the score of an explanation is set to 0 if its associated answer was predicted incorrectly. This can also be viewed as the explanation score S_E multiplied by a coefficient for task performance (accuracy, in most cases). This measure is introduced to avoid giving an advantage to models that purely optimize for generating a few good explanations but neglect the task itself.

Human Evaluation Framework As mentioned earlier, existing automated NLG metrics have strong limitations when evaluating NLEs. I therefore propose a human evaluation framework to compute S_E .

Human annotations were collected on the crowd-sourcing platform MTurk, where I asked the annotators to proceed in two steps. First, they have to solve the task T , i.e., provide the answer a to the question. This ensures the annotators reflect on the question first and enables us to do in-browser quality checks (since the answers are known). I disregard their annotation if they answered the VL task T incorrectly.

For each explanation, they are asked a simple evaluation question: “Given the image and the question/hypothesis, does the explanation justify the answer?”. This simple question is meant to capture one of the main goals of NLEs: to be a form of model explanation that lay users can easily understand and agree with. I follow Marasović et al. (2020) in giving four response choices: *yes*, *weak yes*, *weak no*, and *no*. Marasović et al. (2020) later merges *weak yes* with

yes and *weak no* with *no*, giving them only a binary evaluation for every explanation. I argue that keeping the four response options will allow for a more fine-grained evaluation. *yes*, *weak yes*, *weak no*, and *no* are mapped to the numeric scores of 1, $\frac{2}{3}$, $\frac{1}{3}$, and 0, respectively.

Annotators are also asked to select the explanations’ main shortcomings (if any). Three main limitations of explanations are considered. First, they can *insufficiently justify the answer*. For example, the sentence “because it’s cloudy” does not sufficiently justify the answer “the sea is not calm”. Second, an explanation can *incorrectly describe the image*, e.g., if a model learned generic explanations that are not anchored in the image. “There is a person riding a surfboard on a wave” is generally a good explanation for the answer “surfing” when asked “What activity is this?”, but the image may actually display a dog surfing. Lastly, the sentences can be *nonsensical*, such as “a man cannot be a man”. An example of the instruction and annotation interface that the MTurk workers were given is shown in Figure 3.6 and Figure 3.7.

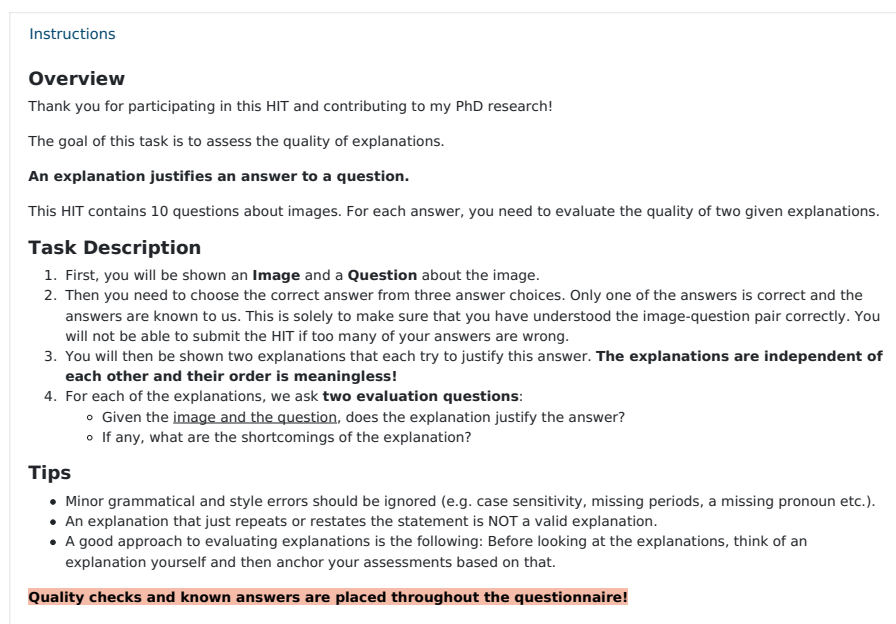


Figure 3.6: A snapshot of the instructions that were provided to the annotators.

For each model-dataset pair, I selected a random sample of 300 data points where the model answered the question correctly. Explanations, where the model answered incorrectly, are not evaluated, as a justification of a wrong answer is assumed to be incorrect. Every sample contains only unique images. For VCR, I ensure a balanced distribution across all movies from the dataset. Note that it is not possible to evaluate all models on exactly the same instances, as they do not all answer the same questions correctly. Taking a subset of examples where *all* models answered correctly is disadvantageous for two reasons. First, this makes the benchmark less

reusable, as future methods might not answer the same questions correctly. Second, this would bias the dataset towards the questions that the weakest model answered correctly.

Several additional measures are put in place to ensure robustness and re-usability of the framework. In order to account for annotator subjectivity, every generated NLE is evaluated by three different annotators. The final score per explanation is given by the average of all evaluations. In addition, each worker evaluates one model at a time to avoid potential anchoring effects between models (e.g., an annotator could evaluate one model more favourably because poor explanations from a different model have reduced their standard)⁶. This ensures that the e-ViL benchmark can be re-used without re-running it on all previous models. To implicitly induce a uniform anchoring effect, the annotators evaluate both the ground-truth explanation

(which is invariant to the model) and the explanation generated by a model for every image-question pair. They do not know which is which and are not asked to compare them. This implicitly ensures that all evaluations have the same anchor (the ground-truth), and it allows us to compute S_E in different ways, as outlined in Section 3.4.4. This also allows the detection of workers who did not understand the task correctly. Overall, over 200 individual annotators were employed for the benchmark, and all of them had to have a 98% prior acceptance rate on MTurk. The results are supported with statistical tests outlined in Section 3.4.4. The cost to

Image:



Question: What is the person doing?

What is the correct answer to the question?

☐ main

☐ ivory

☐ snowboarding

Explanation #1: He leans his body forward to glide down the mountain.

a) Given the above image and question, does this explanation justify the answer to the question?

☐ Yes

☐ Weak Yes

☐ Weak No

☐ No

b) What are the shortcomings of the explanation?

☐ Incorrect description of the image

☐ Insufficient justification

☐ Confusing sentence

☐ None

Figure 3.7: A snapshot of the interface through which annotators evaluated the explanations.

⁶I cannot enforce that the same annotator completes tasks for more than one model

evaluate *one* model on *one* dataset was 108-117\$. For re-usability, the questionnaires used in our benchmark were made public⁷.

3.4.3 Results

In this section, I provide the human evaluation results, a study on the correlation between different automatic NLG metrics and human evaluation scores, and a look at the effect that training with explanations has on the performance on task T .

	Overall	VQA-X			e-SNLI-VE-2.0			VCR		
	S_E	S_O	S_T	S_E	S_O	S_T	S_E	S_O	S_T	S_E
PJ-X	59.2	49.9	76.4	65.4	41.2	69.2	59.6	20.6	39.0	52.7
FME	60.1	47.7	75.5	63.2	43.1	73.7	58.5	28.6	48.9	58.5
RVT	62.8	46.0	68.6	67.1	42.8	72.0	59.4	36.4	59.0	61.8
e-UG	68.5	57.6	80.5	71.5	54.8	79.5	68.9	45.5	69.8	65.1
GT	79.3	–	–	84.5	–	–	76.2	–	–	77.3

Table 3.7: e-ViL benchmark scores. GT denotes the ground-truth explanations, which were also evaluated. Reported scores are multiplied by 100 for readability. The best results are in bold.

Human Evaluation The explanation scores S_E obtained from the e-ViL human evaluation framework are displayed in Table 3.7. e-UG outperforms existing methods on all datasets, with an average S_E score 5.7 points higher than the second-best model, RVT. Despite leveraging little image information, RVT achieves higher scores than PJ-X and FME on average, reflecting the ability of GPT-2 to learn to generate convincing explanations without much anchoring on the image. There is still a significant gap between S_E scores of generated explanations and ground-truth (GT) explanations. For VQA-X, S_E scores are higher overall, suggesting that the dataset is easier. In terms of the overall score S_O , the gap between e-UG and the rest increases further as UNITER achieves higher performance on VL tasks than the M_T modules of the other models. Figure 3.8 shows the explanations generated by the models (and the ground-truth) for two images of each dataset, including the e-ViL human evaluation scores. In some of the images, such as in Figures 3.8b and 3.8f, we can see that e-UG provides better, more image-grounded explanations.

As a second question, annotators were asked to select the shortcomings (if any) for each NLE.

⁷<https://github.com/maximek3/e-ViL>



Hypothesis: A dog is playing with a cat.
Relation: Contradiction
GT Explanation: A man running and a dog playing with a cat are two very distinct activities.
PJ-X: a dog is not a cat
FME: a dog is not a cat
RVT: A cat is not a dog.
e-UG: A dog is not a football player.

Human Evaluation:

0.00
0.17
0.00
0.56

(a) e-SNLI-VE-2.0.



Hypothesis: The lady is the owner of the store.
Relation: Neutral
GT Explanation: We cannot tell from this picture if the lady is the owner of the store.
PJ-X: a woman looking at a microscope does not imply that she is looking for the store
FME: a woman can be a man or a woman
RVT: Just because a lady is holding a book does not mean she is the owner of the store.
e-UG: Just because a lady is working at a store does not mean she is the owner.

Human Evaluation:

0.56
0.17
0.67
1

(b) e-SNLI-VE-2.0.



Question: What sport are they playing?
Answer: tennis
GT Explanation: they are holding tennis rackets inside the court.
PJ-X: the man is holding a tennis racket
FME: they are holding tennis rackets
RVT: the man is holding a racket
e-UG: they are holding tennis rackets

Human Evaluation:

1
0.89
0.89
1

(c) VQA-X.



Question: What game is the guy playing?
Answer: frisbee
GT Explanation: he is about to catch one in his hand.
PJ-X: he is holding a frisbee
FME: the man is holding a frisbee
RVT: he is holding a disc
e-UG: he is throwing a frisbee.

Human Evaluation:

0.22
0.22
0.33
0.67

(d) VQA-X.



Question: What is 1 doing ?
Answer: 1 is confronting 3.
GT Explanation: 3 looks uncomfortable and 1 looks as if he is not happy.
PJ-X: 1 is looking at 3 with a smile on his face
FME: 1 is looking at 3 with a frown
RVT: 1 is looking at 3 with a look of anger on his face.
e-UG: 1 is facing 3 and is speaking to him.

Human Evaluation:

0.00
0.78
0.5
0.89

(e) VCR.



Question: What is in wineglass ?
Answer: There 's champagne in wineglass.
GT Explanation: There 's a bottle of champagne near it.
PJ-X: there is a plate of food in front of her and she is holding a fork
FME: wineglass is a wine glass and wineglass is a wine glass
RVT: wineglass is a wineglass and wineglass is a wineglass.
e-UG: The glass is full of champagne and there is a champagne bottle in front of it.

Human Evaluation:

0.33
0.44
0.67
1

(f) VCR.

Figure 3.8: Pair of examples from the test set of each dataset, with ground-truth (GT) explanation, the generated explanations of each model, and human evaluation scores S_E .

Results for this are given in Table 3.8. The most frequent shortcoming is an insufficient justification of the answer. Least frequent, with around 10% prevalence, explanations can be nonsensical (e.g., “a woman is a woman”). All models struggle similarly with producing explanations that sufficiently justify the answer. e-UG and PJ-X are better at producing coherent sentences. e-UG is significantly superior in terms of the explanations that accurately describe the image content. This empirically confirms the effectiveness of the enhanced conditioning on the image. On a dataset level, we see that it is easiest for all models to provide explanations that make grammatical sense and justify the answer on VQA-X, reinforcing the notion that the explanations of VQA-X are easier and less elaborate.

Model	Untrue to Image	Lack of Justification	Non-sensical Sentence
PJ-X	25.0%	26.4%	8.9%
RVT	20.4%	24.2%	12.0%
FME	21.8%	23.1%	13.7%
e-UG	15.9%	25.0%	7.4%
Dataset			
e-SNLI-VE-2.0	21.3%	28.7%	12.8%
VCR	21.0%	31.2%	11.7%
VQA-X	20.0%	15.4%	7.4%

Table 3.8: Main shortcomings of the generated explanations, by models and by datasets. Human evaluators could choose multiple shortcomings per explanation. The percentages represent the share of samples, where a sample is given by an NLE-evaluator pair, for each dataset or model where the given shortcomings were selected by an evaluator. The best model results are in bold.

Correlation of NLG Metrics with Human Evaluation Automatic NLG metrics are reported in Table 3.9. They are computed for all the explanations from the test sets where the predicted answer was correct. A quick observation is that these automatic metrics do not always reflect the human evaluation results. For example, on the VCR dataset, FME, and not e-UG, obtains the highest S_E score when using automatic NLG metrics. Some tendencies are reflected nonetheless, such as the fact that e-UG is the best model overall, and that e-UG consistently outperforms RVT (albeit by a small margin).

In order to verify the intuition that the object labels used by RVT provide very little information about the image, I trained a GPT-2 language model that is only conditioned on the question and answer, ignoring the image (called *GPT-2 only* in Table 3.9). Without having any image input, this model closely shadows the performance of RVT on most metrics. RVT is still slightly

better in most cases, indicating that the object labels do provide some minor improvement. This suggests that RVT is not able to use visual information effectively and learns the explanations mostly based on spurious correlations and not based on the image.

<i>VQA-X</i>	e-ViL Scores (auto)			<i>n</i> -gram Scores								Emb. Score
	S_O	S_T	S_E	B1	B2	B3	B4	R-L	MET.	CIDEr	SPICE	BERTScore
PJ-X	32.1	76.4	42.1	57.4	42.4	30.9	22.7	46.0	19.7	82.7	17.1	84.6
FME	33.0	75.5	43.7	59.1	43.4	31.7	23.1	47.1	20.4	87.0	18.4	85.2
RVT	26.8	68.6	39.1	51.9	37.0	25.6	17.4	42.1	19.2	52.5	15.8	85.7
GPT-2 only	N.A.	N.A.	37.8	51.0	36.4	25.3	17.3	41.9	18.6	49.9	14.9	85.3
e-UG	36.5	80.5	45.4	57.3	42.7	31.4	23.2	45.7	22.1	74.1	20.1	87.0
<i>VCR</i>												
PJ-X	7.2	39.0	18.4	21.8	11.0	5.9	3.4	20.5	16.4	19.0	4.5	78.4
FME	17.0	48.9	34.8	23.0	12.5	7.2	4.4	22.7	17.3	27.7	24.2	79.4
RVT	15.5	59.0	26.3	18.0	10.2	6.0	3.8	21.9	11.2	30.1	11.7	78.9
GPT-2 only	N.A.	N.A.	26.3	18.0	10.2	6.0	3.8	22.0	11.2	30.6	11.6	78.9
e-UG	19.3	69.8	27.6	20.7	11.6	6.9	4.3	22.5	11.8	32.7	12.6	79.0
<i>e-SNLI-VE-2.0</i>												
PJ-X	26.5	69.2	38.4	29.4	18.0	11.3	7.3	28.6	14.7	72.5	24.3	79.1
FME	29.9	73.7	40.6	30.6	19.2	12.4	8.2	29.9	15.6	83.6	26.8	79.7
RVT	31.7	72.0	44.0	29.9	19.8	13.6	9.6	27.3	18.8	81.7	32.5	81.1
GPT-2 only	N.A.	N.A.	43.6	29.8	19.7	13.5	9.5	27.0	18.7	80.4	32.1	81.1
e-UG	36.0	79.5	45.3	30.1	19.9	13.7	9.6	27.8	19.6	85.9	34.5	81.7

Table 3.9: Automatic NLG metrics for all model-dataset pairs. The S_E based on automatic NLG metrics is the harmonic mean that was used to select the best model during validation. B1 to B4 stand for BLEU-1 to BLEU-4, R-L for ROUGE-L, and MET for METEOR.

To better understand to what extent automatic NLG metrics are able to mirror human judgment of explanations, I computed the Spearman correlation of different NLG metrics with the human evaluation scores. The human evaluation score is averaged and normalized (across all annotators) for each explanation. I have human evaluation scores for a total of 3,566⁸ generated explanations, which makes it the currently largest study on the correlation of NLG metrics with human evaluation in NLEs. The results in Table 3.10 show that BERTScore and METEOR exhibit significantly higher correlations with human annotators across all datasets, reaching a maximal value of 0.293 (which is still a relatively low correlation). The reliability of automatic metrics also differs by dataset. They are highest on VQA-X and lowest on VCR. This could be explained by the fact that explanations in VCR are generally semantically more complex or more speculative (and, therefore, there are more different ways to explain the same thing) than those in VQA-X. It is noteworthy that some *n*-gram metrics, such as BLEU, ROUGE, or

⁸There are four models, three datasets of 300 examples, and, therefore, 3,600 explanations. However, 34 had to be discarded as all three annotators answered the task question incorrectly.

CIDEr, have no statistically significant correlation with human judgment on VCR.

Metric	All datasets	VQA-X	e-SNLI-VE-2.0	VCR
BLEU-1	0.222	0.396	0.123	<i>0.032</i>
BLEU-2	0.236	0.412	0.142	<i>0.034</i>
BLEU-3	0.224	0.383	0.139	<i>0.039</i>
BLEU-4	0.216	0.373	0.139	<i>0.038</i>
METEOR	0.288	0.438	0.186	0.113
ROUGE-L	0.238	0.399	0.131	<i>0.050</i>
CIDEr	0.245	0.404	0.133	<i>0.093</i>
SPICE	0.235	0.407	0.162	0.116
BERTScore	0.293	0.431	0.189	0.138
BLEURT (Sellam et al., 2020)	0.248	0.338	0.208	0.128

Table 3.10: Correlation between human evaluation and automatic NLG metrics on NLEs. All values, except those in *italic*, have p-values < 0.001 .

Explanations as Learning Instructions Training a model jointly on the tasks T and E can be viewed as a form of multi-task learning (Caruana, 1997). The explanations e augment the data points of task T by explaining why an answer a was given. The module M_T (which solves task T) may benefit from this additional signal. Indeed, the model is forced to learn a representation of the image and question from which both the answer and explanation can be extracted, which could improve the model’s representation capabilities. Chain-of-Thought reasoning (Wei et al., 2022) has later confirmed the potential benefits of this. To verify this hypothesis, I compared the task scores of modules M_T that trained only on task T and those that, together with M_E , were jointly trained on tasks T and E . I do this for e-UG on all three datasets and for FME and PJ-X on VCR and e-SNLI-VE-2.0 (because a larger pre-training dataset exists for VQA-X). The results in Table 3.11 show that, without any adaptations, the task performance for joint training is equal or better in all but one model-dataset combination. These results suggest that explanations may have the potential to act as “learning instructions” and thereby improve the classification capabilities of a model.

Detailed Results for e-SNLI-VE-2.0 More detailed results for the e-SNLI-VE-2.0 dataset, i.e., task accuracy and explanation scores for each of the three different answer classes, are given in Table 3.12. For all models, we observe significantly lower accuracies and explanation scores for the neutral class. There are two potential explanations for this. First, the neutral class can be harder to identify than the other classes. In image-hypothesis pairs, entailment and con-

Model	M_T model	VQA-X		SNLI-VE		VCR	
		M_T only	Joint	M_T only	Joint	M_T only	Joint
PJ-X	MCB	N.A.	N.A.	<u>69.7</u>	69.2	38.5	<u>39.0</u>
FME	UpDown	N.A.	N.A.	71.4	<u>73.7</u>	35.7	<u>48.9</u>
e-UG	UNITER	80.0	<u>80.5</u>	79.4	79.5	69.3	<u>69.8</u>

Table 3.11: Comparison of task scores S_T (e.g., accuracies) when the models are trained only on task T vs. when trained jointly on tasks T and E . Scores are underlined if their difference is greater than 0.5. Results on VQA-X are not provided for PJ-X and FME as their vision backbones were pre-trained on VQA v2, a superset of VQA-X.

tradition examples can sometimes be reduced to more straightforward yes/no classifications of image descriptions. For the neutral class, there always needs to be some reasoning involved to decide whether the image does (not) contain enough evidence to neither indicate entailment nor contradiction. A second reason is that, despite our best efforts to clean the dataset, the neutral class is still more noisy and less represented in the training data.

	Entailment			Neutral			Contradiction		
	Acc.	MET.	BERTS.	Acc.	MET.	BERTS.	Acc.	MET.	BERTS.
PJ-X	74.4	14.0	79.2	61.5	12.4	77.4	72.8	15.9	79.3
FME	77.3	15.1	79.8	67.3	13.5	77.9	77.2	16.3	79.8
RVT	74.6	17.9	81.3	63.3	19.0	80.7	79.4	19.4	81.4
e-UG	80.3	19.6	81.6	71.7	18.5	80.9	87.5	20.9	82.6

Table 3.12: Class-wise results on e-SNLI-VE-2.0 for the different models. NLG metrics are only shown for METEOR and BERTScore, as those correlate most with human judgement.

3.4.4 Statistical Analysis and Alternative Metrics

To ensure the robustness of the results, a set of in-browser checks was put in place to prevent the annotators from submitting low-quality evaluations. Checks include making sure they answered the task question correctly and that they cannot simultaneously give an explanation of the lowest score and say that it has no shortcomings or give it the highest score and select shortcomings.

I collected 19,194 MTurk evaluations in total, half of which are for ground-truth explanations and the other half for model-generated explanations. There are evaluations for 264 to 299 unique question-image pairs for every model-dataset combination, meaning explanations are

missing for only 3.3% of questions. There are 82.1 evaluations per annotator on average ($SD = 170.1$), ranging from 16 to 1,244, with a median of 34. After pooling annotations of the same explanation, 6,494 annotations remain.

In Figure 3.9, I add standard errors to the numerical S_E scores given in Table 3.7. This figure confirms that e-UG uniformly outperforms the other models.

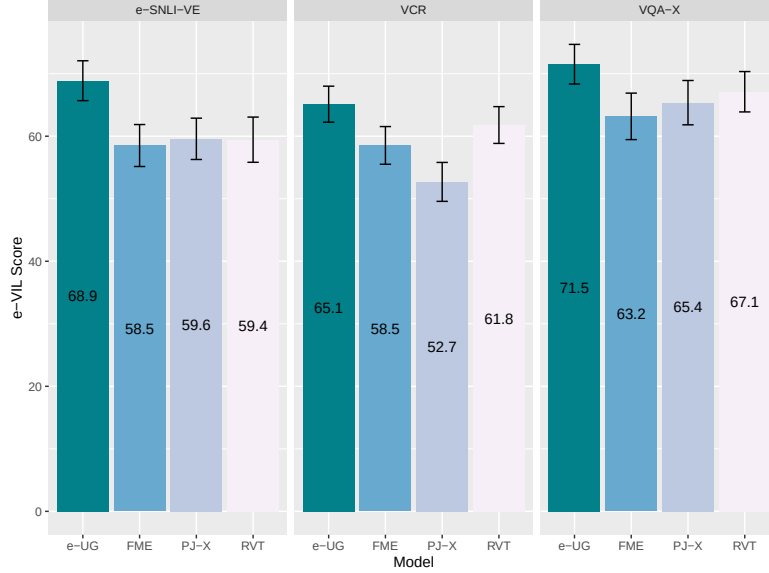


Figure 3.9: Human evaluation framework: e-ViL scores S_E . This plot shows the main e-ViL scores (based on numerical average) for the different model-dataset pairs. Error bars show $\pm 2SD/\sqrt{n}$ for each group.

To further establish the robustness of the e-ViL benchmark, I conducted a statistical analysis of our S_E scores by using a Linear Mixed Model (LMM) that predicts S_E from the model-dataset pairs, with the model as a fixed effect and dataset as a random effect. The Likelihood-Ratio Test (LRT) of the fixed effect indicates that the model significantly contributes to explaining S_E scores, $\chi^2(3) = 37.462, p < 0.001$. To further investigate differences between models, I conducted post-hoc pairwise comparisons, which indicate that e-UG significantly outperforms the remaining models, with $p < 0.001$. RVT significantly outperforms PJ-X ($p = 0.007$). To control for multiple comparisons, the significance level was adjusted for a family-wise type I error rate of $\alpha = 0.05$ using Bonferroni-Holm adjustments.

Alternative S_E Scores The questionnaire used for evaluating NLEs allows for multiple ways to compute the e-ViL score S_E . The key differences between the scoring methods are in how to aggregate the up to three evaluations we have for each explanation. In the main paper, S_E is computed by mapping the four evaluation choices to numerical values, then taking the

average for every explanation in the sample and then the sample average to get our S_E score. Below, I propose two alternative ways to compute S_E . While they lead to different values, the performance differences between our models remain relatively consistent.

In **median pooling**, the score for each explanation is obtained by taking the median of its up to three ordinal evaluations (as opposed to taking a numerical average). I always interpolate with rounding off, meaning that the median of $(Yes, Weak\ Yes) \mapsto Weak\ Yes$ and $(Yes, No) \mapsto Weak\ No$. This allows us to plot the distribution of *No*, *Weak No*, *Weak Yes*, and *Yes* for every model-dataset pair, as displayed in Figure 3.10.

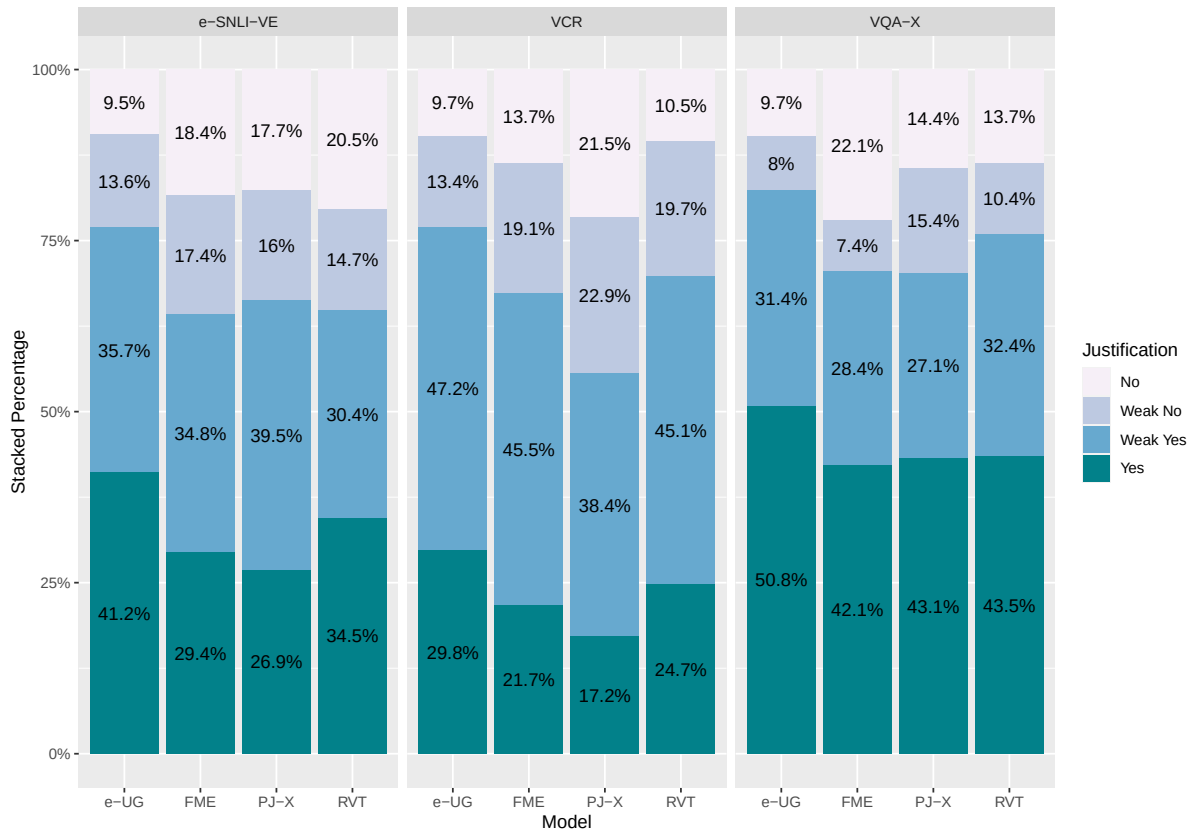


Figure 3.10: Human evaluation framework: Ordinal representation of the evaluations. Median responses for each question-image pair given by participants to the evaluation question “Given the image and the question/hypothesis, does the explanation justify the answer?”.

We observe that e-UG performs better across all datasets, with RVT second in place for the VCR and VQA-X datasets. The differences between the PJ-X, FME and RVT are relatively small. I analyze our results using a Cumulative Link Mixed Model (CLMM) with a logit link and flexible thresholding. The model predicts annotator responses, with the dataset as a random effect and the VL-NLE model as a fixed effect. The LRT indicates that the model significantly influences ratings ($\chi^2(2) = 42.4, p < 0.001$), as shown by a comparison between the full

model and a nested model that includes only the dataset as a predictor. The model predictor is dummy-coded with e-UG as the reference class, allowing us to interpret the model’s coefficients as pairwise contrasts between e-UG and all other models. All coefficients yield p -values below 0.001, indicating that e-UG significantly outperforms all other models.

For the **comparative** S_E score, the generated NLEs are compared directly to the corresponding ground-truth. For every image-question pair, the annotator has to evaluate both the ground-truth and the generated explanation without knowing which is which. This enables us to see, for every generated explanation, if it was deemed equal, better, or worse than the ground-truth. This mimics the approach in [Park et al. \(2018\)](#) and [Wu and Mooney \(2019\)](#), where annotators were explicitly asked if the generated explanation was worse, equally good, or better than the ground-truth. An advantage of this method is that it accounts for the criticalness of each annotator. The disadvantage is that we do not get *absolute* measurements of the quality of the explanations. The generated explanations are scored as 1 if they are rated equal or better than the ground-truth and 0 otherwise. The pooled score is obtained via median pooling with rounding off.

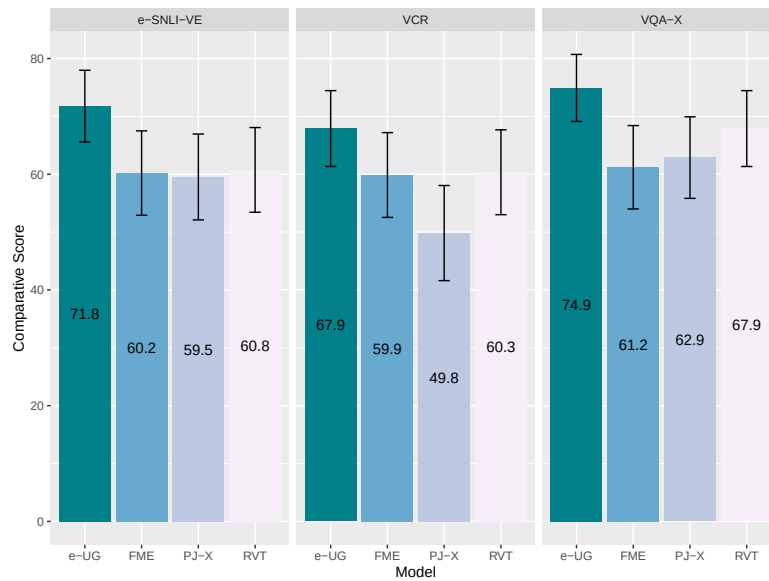


Figure 3.11: Human evaluation framework: Comparative scores. This figure displays the comparative scores (with respect to the ground-truth) of the explanations for the different model-dataset pairs. Error bars show $\pm 2SD/\sqrt{n}$ for each group.

Figure 3.11 displays the comparative score. We can observe that e-UG scores are strongest across all datasets, matching human-written explanations in between 67.9% and 74.9% of cases. The other three models are performing similarly to each other, except on the VCR

dataset, where PJ-X performs worse than the other models. For our statistical analysis, I fitted a Generalized Linear Mixed Model (GLMM) on the full unpooled annotation set, predicting whether an explanation was rated positively (compared to the ground-truth). The model includes the dataset and annotator as random effects and the VL-NLE model as a fixed effect, utilizing a logit link. The model parameter significantly predicts the evaluations ($\chi^2(3) = 67.366, p < 0.001$). Post-hoc tests (Tukey contrasts with Bonferroni-Holm adjusted significance) indicate that e-UG outperforms all other models ($p < 0.001$) and that RVT outperforms PJ-X ($p = 0.011$). All other pairwise comparisons are not statistically significant. Extending the model to include ground-truth explanations as a model category confirms that all model-generated explanations are rated significantly lower than ground-truth explanations.

3.5 Conclusion

This chapter addressed the lack of comparison between existing VL-NLE methods by introducing e-ViL, a unified and reusable benchmark for evaluating vision-language models. The framework was used to assess four frontier VL-NLE approaches via human evaluation. The human evaluations were then used to study the correlation between NLG metrics and human assessment of NLEs, highlighting the limitations of these automated metrics. To accommodate the need for annotated datasets for VL-NLE approaches, I introduced e-SNLI-VE-2.0, the largest existing VL dataset with human-written explanations. Additionally, a new state-of-the-art VL-NLE model was introduced, leveraging advances in vision-language pre-training (UNITER) and autoregressive language modelling pre-training (GPT-2).

The findings highlight the potential of NLEs to provide easy-to-understand and informative explanations for visual reasoning tasks and indicate that training with NLEs may lead to improved task performance. In medical imaging, model transparency (due to safety concerns) and visual reasoning (diagnoses based on radiographs often involve multiple reasoning steps connecting evidence to outcomes) are vital. In the next chapter, I will investigate how NLEs can be introduced to medical image analysis tasks, chest X-rays, in particular, and how they can lead to increased transparency and performance.

3.6 Impact and What Happened Since

At the time of writing this thesis, the e-ViL paper ([Kayser et al., 2021](#)) has been cited 108 times⁹, facilitating a wide body of research on the VL-NLE task. Notable works re-use its human evaluation framework ([Marasović et al., 2022](#); [Sammani and Deligiannis, 2023](#); [Chakrabarty et al., 2022](#); [Majumder et al., 2022](#)), cite it as evidence of the inadequacies of NLG metrics to be used for NLEs ([Wiegrefe et al., 2022, 2021](#); [Marasović et al., 2022](#); [Majumder et al., 2022](#)), train and evaluate their models on e-SNLI-VE-2.0 ([Chen et al., 2023a](#); [Hao et al., 2022](#); [Sammani and Deligiannis, 2023](#); [Majumder et al., 2022](#)), provide upgrades and extensions to e-UG ([Sammani and Deligiannis, 2023](#); [Majumder et al., 2022](#)), and use the e-ViL benchmark ([Sammani and Deligiannis, 2023](#); [Majumder et al., 2022](#)).

It is worth acknowledging that, with the advent of large language models (LLMs), certain aspects of the dataset extraction pipelines used in e-SNLI-VE-2.0 can be improved or automated. Modern LLMs are capable of producing high-quality data annotations with minimal human supervision, even surpassing the quality of human-annotated datasets in some domains ([Gilardi et al., 2023](#)). This can drastically reduce the time and cost associated with manual annotation, particularly in domains where human expertise is scarce or expensive. For instance, the “CoT Collection” ([Kim et al., 2023](#)) dataset generated 1.84M rationales on top of existing NLP datasets using large language models with in-context learning.

⁹Google Scholar, 19.03.2025

4 | Natural Language Explanations for Chest X-ray Analysis

A major obstacle to the adoption and regulatory approval of deep learning models in medical imaging is their lack of explainability (Langlotz et al., 2019). Explainability is especially important due to concerns of poor model robustness (Zech et al., 2018; Ktena et al., 2024), the risk of algorithmic bias (algorithms are prone to amplifying inequalities that exist in the world) (Obermeyer et al., 2019; Hajian et al., 2016) and the high-stakes nature of clinical decision-making (Vayena et al., 2018). Establishing trust in AI systems is, therefore, a critical requirement for their deployment (Leslie, 2019), and interpretability is widely seen as a key requirement to build that trust (Piano, 2020; Brundage et al., 2020). For computer vision models, transparency is often enhanced using saliency maps (Saporta et al., 2022), or other types of visual explanations (Puyol-Antón et al., 2020; Schutte et al., 2021; Kim et al., 2021). However, there is evidence that clinicians, the end-users of such models, prefer explanations in the form of text, i.e., NLEs (Tonekaboni et al., 2019).

Chapter 3 demonstrated the benefits of NLEs in terms of their user-friendliness, ability to convey complex reasoning, and ability to teach reasoning to the model. In medical imaging, different to many natural image classification tasks, classifying an image (i.e., suggesting findings or a diagnosis) often entails intermediate reasoning steps. For example, to suggest the diagnosis “pneumonia”, a radiologist might highlight certain areas in the lung, describe them, say why they would suggest pneumonia, and propose differential diagnoses. These parallels between medical image analysis and visual reasoning tasks suggest that NLEs may translate well to this domain. In this chapter, I first show how clinicians provide NLEs in routine practice and how that can be leveraged to train models to do the same.

In order to study the effectiveness of NLEs for medical image analysis, I introduce MIMIC-NLE-v1 and later MIMIC-NLE-v2, the first large-scale medical imaging dataset with NLEs. They contain over 38,000 resp. 52,000 NLEs, which explain the presence of various thoracic pathologies and chest X-ray findings. The NLEs were extracted from ground-truth radiology reports, meaning they reflect the actual reasoning of the radiologists assessing the scans. Our extraction process introduces little noise, on par or better than for NLE datasets in natural images presented in Chapter 3. MIMIC-NLE-v2 brings several advancements over its predecessor

MIMIC-NLE-v1. The main improvements are centred around extracting more explanations, including for negative findings, incorporating patient context, adding localizations of the evidence, and removing explanations that require hallucinating information, especially references to unavailable patient history. The datasets are publicly available¹.

I then propose a general approach to solving the task and establish an evaluation framework, comparing three strong baselines. The evaluation by a clinician validates the feasibility of the task but also highlights shortcomings that motivated the second version of the dataset. As recently introduced large language models (LLMs) have showcased remarkable reasoning capabilities (Ouyang et al., 2022; Touvron et al., 2023), e.g. with techniques like Chain-of-Thought (CoT) reasoning (Wei et al., 2022), where a model is prompted to think “step-by-step” about a problem, I also study a range of VLMs and try to understand how LLMs can develop reasoning skills over medical images and how this can enable a more transparent and robust approach to medical image analysis.

I will first put this work in the context of the literature (Section 4.1). Section 4.2 introduces MIMIC-NLE-v1, and Section 4.3 proposes a generalized architecture for the task and compares a suite of state-of-the-art models. These two sections are based on Kayser et al. (2022). After introducing the updated dataset in Section 4.4, Section 4.5 examines how recent developments around scale and reasoning manifest on this benchmark. The latter two sections are partially based on Kayser et al. (2023).

¹<https://github.com/maximek3/MIMIC-NLE>

4.1 Related Work

Saliency maps, which attribute importance weights to regions in an image, are the most commonly used form of explainability in medical imaging. They have a range of shortcomings, such as lacking consistency and reproducibility (Arun et al., 2021), conveying limited information and being prone to confirmation bias (Bornstein, 2016; Adebayo et al., 2018), as well as only telling us *how much* highlighted regions affect the model’s output, and not *why* (Ghassemi et al., 2021). Additionally, it has been shown that numerical or visual explanatory methods, in some cases, can be confusing even for data scientists (Kaur et al., 2020) and can pose problems even for explaining trivial models (Camburu et al., 2020). NLEs, on the other hand, would be able to fully capture how the evidence in a scan relates to the diagnosis. Furthermore, saliency maps are post-hoc explainers, whereas NLEs constrain the model during training to mimic how radiologists explain diagnoses (Miller, 2019; Gale et al., 2018).

Previous attempts to augment medical image classification with textual information rely on template-generated sentences (Gale et al., 2018; Lee et al., 2019) or sentences from other images based on image similarity (Kougia et al., 2021). Furthermore, their focus lies mostly on adding descriptive information about a pathology (e.g., its location) instead of explaining the diagnoses. Gale et al. (2018) design a model that accompanies a hip fracture classification with a template-like explanation that indicates the location, severity, and type of fracture. The explanations were manually annotated by a radiologist (as the reports were deemed too inconsistent in their structure). They show that doctors prefer natural language explanations over saliency maps. Lee et al. (2019) design a CNN and RNN attention-based architecture to train a model that is able to classify a mass in breast mammograms as benign or malignant and provide textual explanations. The textual explanations were trained on sentences that the authors generated using the BI-RADS mass lexicons. They evaluate the sentences via automatic NLG metrics and provide neither a score for the classification performance nor a clinical validation of the explanations. Kougia et al. (2021) propose the RTE system to rank and classify reports and propose natural language text to accompany the label. Their approach consists of retrieving the most similar report from the training set, using cosine similarity of the feature vector of the images, and ensuring the report contains the same labels as were detected in the image.

That report is then used as a diagnostic sentence. Their texts have lower NLG scores than early image captioning models.

In contrast, this work leverages the free-text nature of chest X-ray radiology reports, which provide rich and diverse descriptions and justifications for the diagnoses in the X-rays. As such, these reports do not only provide the severity and location of pathologies but also the degree of certainty of diagnoses, differential diagnoses, and justifications of how visual evidence in a scan relates to them.

Recently, Chain-of-Thought (CoT) reasoning has become a dominant paradigm within the field of NLEs, with the primary goal being to improve models' reasoning capabilities through structured, intermediate reasoning steps. CoT involves generating a rationale in natural language before providing a final prediction (Wei et al., 2022), which is identical to the “predict-then-explain” NLE framework proposed in Camburu et al. (2018). This can be achieved even in zero-shot settings through simple prompts such as *"Let's think step by step"* (Wei et al., 2022; Khot et al., 2023). Subsequent work has proposed various improvements to CoT prompting, such as using tree search to explore multiple reasoning paths (Yao et al., 2024). While very large language models (over 100B parameters) benefit significantly from these prompting strategies, it has been shown that smaller language models require additional training to achieve similar gains. For example, Zelikman et al. (2022) demonstrated that smaller models finetuned on structured reasoning datasets perform substantially better on various reasoning benchmarks. At the same time, other research found that gains from CoT are small outside of math-focused domains (Lanham et al., 2023).

CoT reasoning has only seen very few applications to clinical problems, in particular to medical images. Existing approaches focus on general medical text or knowledge-based problems, where large-scale models (100B+ parameters) are effective, and CoT can be elicited without any specific training (Miao et al., 2024). Tu et al. (2024), who use VLMs with up to 562B parameters, employ a one-shot CoT approach and show limited insights that this may be helpful for very large VLMs. This chapter introduces the first attempt to teach clinician reasoning on medical images to smaller-scale VLMs via supervised CoT finetuning.

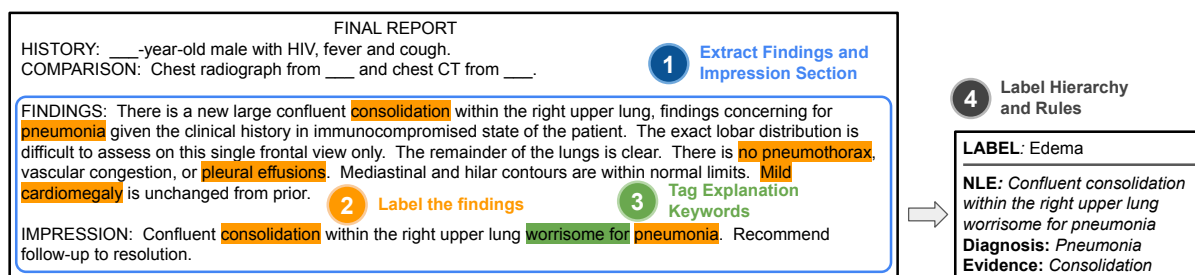


Figure 4.1: The steps required to extract NLEs from raw radiology reports. I first extract the Findings and Impression sections, which contain the descriptive part of the report. Next, class labels and sentences that contain explanation keywords are tagged. Based on this information, the rules described in Table 4.3 extract valid NLEs, as well as their diagnosis and evidence labels.

4.2 MIMIC-NLE-v1

Gathering NLEs is expensive, especially when radiologic expertise is required. To address this, I show that it is possible to automatically distil NLEs from radiology reports, as radiologists typically explain their findings in these reports. As a starting point, I use the radiology reports from MIMIC-CXR (Johnson et al., 2019), a publicly available chest X-ray dataset with 227,827 radiology reports. By applying various filters, labellers, and label hierarchies, I extract 38,003 *image-NLE* pairs, or 44,935 *image-diagnosis-NLE* triplets (as some NLEs explain multiple diagnoses). The extraction process is summarized in Fig. 4.1. The remainder of this section provides additional details on the dataset creation.

By exploring the reports and through conversations with clinicians, I observed that a small selection of phrases, such as “compatible with” or “worrisome for” (see the full list in Table 4.1), are accurate identifiers of sentences where a potential pathology is explained by observations made on the scan (i.e., an *explanatory* sentence). I use regular expressions to ensure that all inflexions of the keywords are covered. The selection of keywords is based on in-house knowledge, empirical evidence, and a survey of radiologists conducted in Mityul et al. (2018). Table 4.1 provides insight into the frequency of all the explanations keywords considered.

As clinical reports use relatively structured language, this small set of keywords covers the major part of all explanations. However, upon inspecting the extracted sentences, I noticed that in some cases, the explanation keywords relate to imaging techniques (e.g., “Assessment is somewhat limited *due* to patient positioning and exclusion of the left costophrenic angle.”) and follow-up suggestions (“Continuous monitoring is suggested”). I, therefore, remove sentences

Table 4.1: The occurrence of different explanation keywords in the unfiltered sentences from MIMIC-CXR. The list is displayed in three columns.

Keyword	Count	Keyword	Count	Keyword	Count
suggest	19,787	compatible with	6,025	worrisome for	1,685
reflect	17,878	likely represent	5,239	indicate	1,568
due	17,771	relate	5,241	account	1,193
consistent with	14,654	may represent	5,141	suspicious for	684
concerning for	10,243	potentially	3,160		

referring to follow-up procedures (keywords “recommend”, “perform”, “follow”) and patient positioning (“CT”, “technique”, “position”, “exam”, “assess”, “view”, “imag”), and lastly also the keyword “finding”, as it is often used to refer to a finding given in an earlier sentence. After removing 79,068 sentences containing anonymized data, a further 297,463 sentences were removed based on the keywords.

Next, the CheXbert labeler (Smit et al., 2020), which can extract fourteen different chest X-ray labels from clinical text, is used to identify the findings mentioned in each sentence. CheXbert extracts the following 14 labels from free-text: “Enlarged Cardiomedastinum”, “Cardiomegaly”, “Lung Opacity”, “Lung Lesion”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”, “No Finding”. Table 4.2 shows how often different diagnosis labels appear in the final dataset, both in combination with others or isolated. The fourteen labels correspond to those extracted for the CheXpert (Irvin et al., 2019) dataset and were chosen for their frequency and clinical relevance in chest X-rays. For each label, CheXbert predicts whether it is “Positive”, “Uncertain”, “Negative”, or not mentioned. I only extract labels for which the uncertainty label is “Positive” or “Uncertain”. For 17,982 explanations, CheXbert finds labels in an NLE that were not annotated on the image. Manual verification showed that this was mostly due to the poor performance of CheXpert (which is used to get image labels in MIMIC-CXR) compared to CheXbert (an improved version of CheXpert).

For each sentence in the MIMIC-CXR radiology reports, we now know whether it is an explanatory sentence, which labels it refers to, and if the labels have a negative, positive, or uncertain (i.e., they are *maybe* present) mention. As the goal of NLEs is to explain an outcome with evidence, it needs to be established which of the labels mentioned in an NLE are the outcome and which are the evidence. To determine the evidence relationships present in

Table 4.2: The distribution of diagnosis label sets annotated per explanation. The sets are ordered by occurrence. # of sentences corresponds to the number of unique image–sentence pairs. The list is displayed in two halves.

Diagnosis Labels	# of sentences	Diagnosis Labels	# of sentences
Atelectasis	10,616	Atel., Edema, Pneumonia	104
Pneumonia	9,032	Atelectasis, Edema	103
Edema	5,098	Pl. Eff., Pneumothorax	65
Atelectasis, Pneumonia	4,773	Pleural Effusion, Pneumonia	55
Pleural Effusion	4,585	Edema, Pleural Effusion	31
Consolidation	916	Atelectasis, Pleural Other	17
Pleural Other	846	Atel., Edema, Pl. Eff.	9
Edema, Pneumonia	623	Edema, Pl. Eff., Pneumonia	8
Atelectasis, Pleural Effusion	397	Atel., Edema, Pl. Eff., Pneum.	6
Pneumothorax	311	Edema, Pleural Other	6
Pleural Effusion, Pleural Other	195	Pleural Other, Pneumonia	4
Atel., Pl. Eff., Pneumonia	190	Other	13

NLEs, I propose a limited set of clinically validated, known relationships. For example, if an explanatory sentence in a radiology report indicates the presence of both *Atelectasis* and *Consolidation*, it is not apparent whether *Consolidation* is evidence for *Atelectasis* (e.g., “Right upper lobe new consolidation is compatible with atelectasis with possibly superimposed aspiration.”) or whether they are both explained by a different finding (e.g., “A persistent left retrocardiac density is again seen reflecting left lower lobe atelectasis or consolidation.”). However, for other label combinations, such as *Consolidation* and *Pneumonia*, the evidence relationship is usually straightforward, i.e., *Consolidation* is the evidence for *Pneumonia*. The proposed evidence graph in Figure 4.2 depicts the label relationships with a known and high-confidence evidence-outcome relationship (i.e., their co-occurrence in a sentence lets us deduct an evidence relationship with high probability). The graph was constructed with an external radiologist and by empirically validating the co-occurrences of these labels in MIMIC-CXR. I also explored models that annotated cause-effect relationships in sentences but have found them to be insufficient without additional annotation of the data (it only provided annotations for 10-20% of the explanations, with many being incorrect.) (Li et al., 2021b).

Based on the evidence graph and manual inspection of every label combination that appears at least 25 times in an explanatory sentence (121 combinations in total), I established 12 mutually exclusive rules that mark a sentence as a valid NLE. The rules take into account the label combination, their uncertainty label (i.e., *Uncertain* or *Positive*), and the presence of explanation keywords. The rules are defined in Table 4.3. For each sentence in MIMIC-CXR, these rules

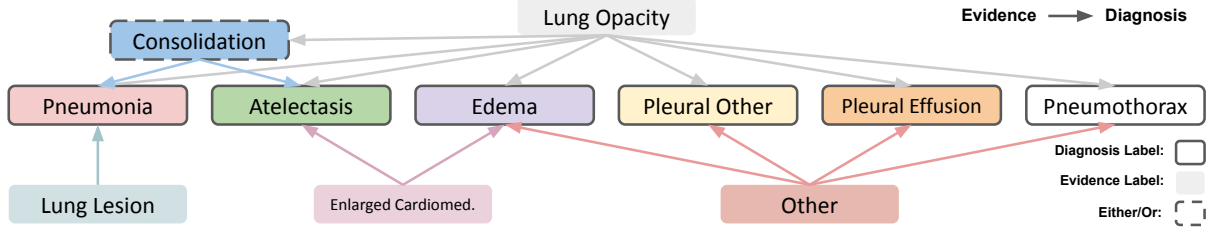


Figure 4.2: The evidence graph that visualizes the possible evidence-outcome relationships in the dataset.

Table 4.3: Rules to establish valid evidence-outcome relationships. The “kw req.” column specifies whether an explanation keyword is additionally required to validate a relationship. For labels not denoted by U or P , both positive and uncertain labels are allowed. Here, A^U and B^U denote the sets A and B with all labels uncertain, and $\mathcal{P}_{\geq 2}(A^U)$ is the power set of A^U containing subsets of size at least two.

Evidence	Outcome(s)	kw req.
<i>Other / misc.</i>	$d \in A = \{\text{Pleural Eff.}, \text{Edema}, \text{Pleural Other}, \text{Pneumoth.}\}$	Yes
<i>Other / misc.</i>	$s \in \mathcal{P}_{\geq 2}(A^U)$	Yes
Lung Opacity	$d \in B = A \cup \{\text{Pneumonia}, \text{Atelectasis}\}$	No
Lung Opacity	$s \in \mathcal{P}_{\geq 2}(B^U)$	No
Lung Opacity	Consolidation	No
Consolidation	Pneumonia	No
{Lung Op., Cons.}	Pneumonia	No
Lung Lesion	Pneumonia	Yes
Lung Opacity	$\{\text{Atelectasis}^P, \text{Pneumonia}^U\}$	No
Consolidation	$\{\text{Atelectasis}^U, \text{Pneumonia}^U\}$	No
Enlarged Card.	Edema	Yes
Enlarged Card.	Atelectasis	Yes

are used to determine whether the label combination referred to in the sentence makes it a valid NLE and to determine which of the labels are part of the evidence and the diagnosis. Some label combinations, such as *Consolidation* and *Pneumonia*, are considered a valid NLE even if they do not contain an explanation keyword (as the keywords are not exhaustive). However, for label combinations with less frequent evidence relationships, such as *Enlarged Cardiomediastinum* and *Edema*, the presence of an explanation keyword is required. It is worth noting that only explanations for positive and uncertain cases are included at this stage. An overview showing the frequency of the evidence-outcome relationships is provided in Figure 4.3.

Based on these rules, I obtained 38,003 NLEs. An overview of the different steps in the pipeline is provided in Table 4.4. Figure 4.4 shows an exemplary datapoint. On a subset of 100 NLEs, I found an accuracy of 92% (compared to between 49% to 91% for natural images NLE datasets, as shown in Chapter 3). An NLE was deemed correct if it correctly used the extracted evidence

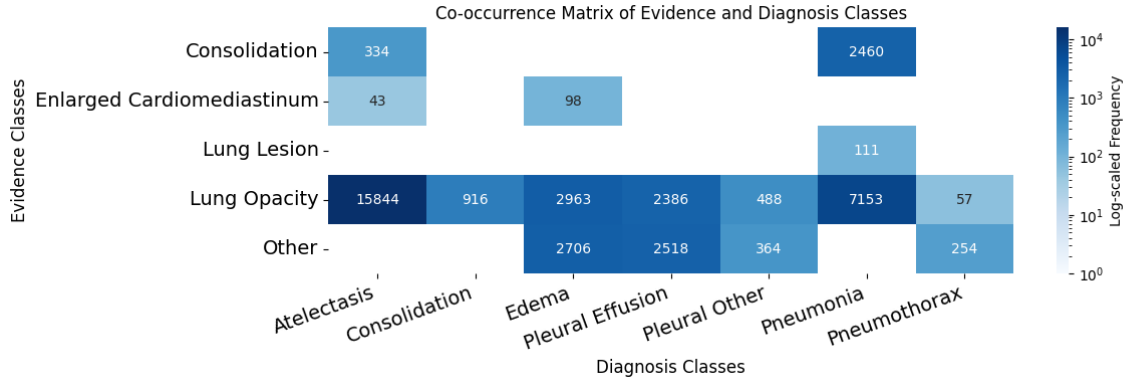


Figure 4.3: Co-occurrence matrix for evidence-outcome pairs in MIMIC-NLE-v1.

Table 4.4: Filters applied to extract image–outcome–NLE triplets from MIMIC-CXR. The number of sentences corresponds to unique image–sentence pairs. “Non-descriptive aspects” are aspects that cannot be derived from the image itself. Duplicates refer to sentences from the same report that mention the same labels.

Filter	# of sentences
Extract Findings and Impression sections	1,383,533
Remove sentences with anonymized data	1,304,465
Remove sentences referring to non-descriptive aspects:	1,007,002
Recommendations: “recommend”, “perform”, “follow”	
Technical: “CT”, “technique”, “position”, “exam”, “assess”, “view”, “imag”	
Filter by rules from Table 4.3	43,612
Remove duplicates	39,094
Remove studies without AP or PA images	38,003

labels to explain the extracted outcome labels. The main reasons for incorrect NLEs were failures of the CheXbert labeller. Thus, with improved labellers, the accuracy could be further improved. Using the train, dev, and test splits in MIMIC-CXR (Johnson et al., 2019), we get split sizes of 37,016, 273, and 714 for MIMIC-NLE-v1, respectively. Note that, as MIMIC-NLE-v1 is the first dataset of its kind, I focused on precision rather than recall, prioritizing verifiable evidence-outcome relationships over large dataset size.

```
{
  "sentence_ID": "s51038639#2", // unique NLE identifier
  "patient_ID": "p11662490", // unique patient ID
  "report_ID": "s51038639", // unique radiology report ID
  "nle": "Subtle lower lobe opacities may reflect atelectasis.",
  "outcome_label": "Lung Opacity", // outcome label explained
  "evidence_label": "Atelectasis", // evidence label for diagnosis
  "img_labels": [[0, 1, 0], [1, 0, 0], ..., [1, 0, 0]] // image-wide labels
}
```

Figure 4.4: Example representation of a single datapoint in MIMIC-NLE-v1.

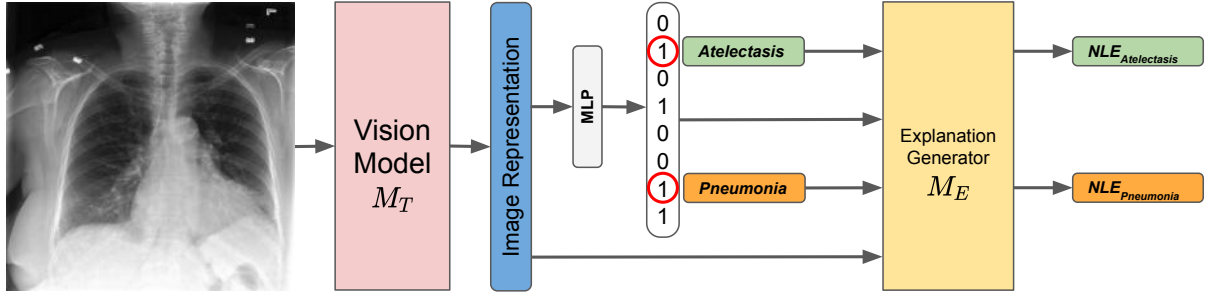


Figure 4.5: The model pipeline to provide an NLE for a diagnosis prediction.

4.3 Benchmark v1

4.3.1 Models

The high-level approach for generating NLEs illustrated in Figure 4.5 is similar to that described in Section 3.3. The task model M_T is given by an image classification model. In our case, $M_T(x) = Y$, where x is a radiographic scan, and Y is the prediction vector $Y \in R^{n_{\text{unc}} \times n_{\text{path}}}$, with $n_{\text{unc}} = 3$ being the number of certainty levels, and $n_{\text{path}} = 10$ being the number of pathologies. This follows the *U-MultiClass* approach from Irvin et al. (2019), i.e., for each pathology, the image is classified as *negative*, *uncertain*, or *positive*.

To the best of my knowledge, this was the first application of NLEs to multi-label classification. I address this problem by generating an NLE for every label that was predicted as *uncertain* or *positive* and is considered an outcome label in our evidence graph (see Figure 4.2). Given a set of outcome pathologies P , an NLE is generated for every pathology $p_j \in P$. For every pathology p_j , I condition the NLE generation on the M_T prediction, i.e., we have the explanation generator $M_E(x_{\text{REP}}, Y, p_j) = e_j$, where x_{REP} is the learned representation of the image x , and e_j is the NLE that explains the classification of p_j . By backpropagating the loss of M_E through M_T , M_T is constrained to learn representations that embed the correct reasoning for each diagnosis. During training, I adopt diagnosis teacher forcing, i.e., I condition M_E on the ground-truth (GT) pathology p_j and prediction vector Y .

4.3.2 Experimental Setup

This section evaluates three baselines on the MIMIC-NLE-v1 dataset using both automatic and expert-based evaluations.

Baselines First, I adapt two state-of-the-art chest X-ray captioning architectures to follow the NLE generation approach outlined in Figure 4.5. I re-implemented TieNet (Wang et al., 2018), following their paper and a publicly available re-implementation² and RATCHET (Hou et al., 2021), a recent transformer-based captioning method. DenseNet-121 (Huang et al., 2017) is used as the vision model M_T for each approach. DenseNet-121 is a CNN widely used for chest X-ray classification (Irvin et al., 2019). For TieNet, contrary to the original model, I do not predict the labels from the learned text representations but condition the NLE generation on the labels. TieNet’s decoder, an LSTM, is conditioned on the pooled features of the image x_{REP} via an attention layer on the prediction vector Y by adding its embedding to x_{REP} , and on the diagnosis by initializing the hidden state of the decoder with an encoding of the predicted diagnosis p_j in textual form. RATCHET considers 49×1024 feature maps as x_{REP} , and the embedding of Y is modelled as an additional, 50-th, feature map. I condition on p_j in the same way as in TieNet by initializing the hidden state of the decoder.

I also propose an additional baseline referred to as DPT (DenseNet-121 + GPT-2), which is inspired by e-UG, the at-the-time state-of-the-art NLE approach for natural image reasoning. DPT leverages a DenseNet-121 as M_T and a vision-enabled GPT-2 language model (Radford et al., 2019) as M_E . An attention layer enforces M_E to focus on the same regions in the scan as M_T . The NLE generation is modelled as a sequence-to-sequence translation from the 49 feature maps outputted by Densenet-121, a token for Y , and the diagnosis p_j in text form to an NLE. For a fair comparison, I use the same GPT-2 vocabulary for all models and initialize all word embeddings with the pre-trained GPT-2 weights. All models use DenseNet-121 from TorchXRyVision (Cohen et al., 2022), which was pre-trained on CheXpert (Irvin et al., 2019). All text generation was done via greedy decoding.

Model Training All models were implemented in PyTorch 1.10.1 and trained on Nvidia TITAN Xp and RTX TITAN GPUs. I used a maximum sequence length of 38 during training, which corresponds to the 98th percentile of the training set. I used the Adam optimizer with weight decay for the transformer models and without weight decay for TieNet (as it failed otherwise). I also used a linear scheduler with warmup for the learning rate. For each model, I experimented with different learning rates and batch sizes as hyperparameters. Using the dev set, I selected the following hyperparameters: for DPT, a learning rate of 5×10^{-4} and batch

²<https://github.com/farrell236/RATCHET/tree/tienet>

size of 16; for RATCHET, a batch size 16 and learning rate 5×10^{-5} ; for TieNet, a batch size of 32 and learning rate of 1×10^{-3} . They were selected based on the product of the task score S_T , CLEV score, and the average of the BERTScore (Zhang et al., 2019) and METEOR (Banerjee and Lavie, 2005) score, all outlined below.

Evaluation Metrics All models were evaluated based on their ability to solve the task, i.e., image multi-label classification, and their ability to explain how they solved the task, i.e., providing NLEs. The task score S_T is given by the weighted AUC score, where I consider *uncertain* and *positive* as one class (following Irvin et al. (2019)). For the NLE scores, again, only NLEs that explain correctly predicted labels are considered. Chapter 3 has shown that automated natural language generation (NLG) metrics are not well suited for NLEs, as the same answers can be explained in different syntactic forms and even via different semantic meanings. I, therefore, propose the CLEV (CLinical EVIDence) score, which verifies whether an NLE refers to the correct clinical evidence. It uses the CheXbert labeller to extract the evidence labels referred to in the GT and generated NLEs to verify if they are the same. For example, if the GT NLE mentions *Lung Opacity* and *Consolidation* as evidence for *Pneumonia*, the generated NLE is expected to contain the same findings. The CLEV score is the accuracy over all the generated NLEs, i.e., what share contains exactly the same evidence labels as the corresponding GT. The NLG metrics BLEU, Rouge, CIDEr, SPICE, BERTScore, and METEOR, as in Chapter 3, are provided too. For BERTScore, the weights were initialized with a BERT pre-trained on clinical text.³

Clinical Evaluation Given the limitations of evaluating NLEs with automatic NLG metrics, I also provide an assessment of the NLEs by a clinician (an internal medicine emergency doctor). They were presented with 50 X-ray scans, a diagnosis to be explained, and four different NLEs: the ground-truth and one for each of our three models. The NLEs are shuffled for every image, and the clinician does not know their origin (whether ground-truth or from either of the models). They were then asked to judge how well each NLE explains the diagnosis given the image on a Likert scale of 1 (*very bad*) to 5 (*very good*) (exact instruction: “How well do the explanations explain the diagnosis given the image? Likert scale of 1 to 5 corresponds to [Very Bad, Bad, Neutral, Good, Very Good]”).

³https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

4.3.3 Results

Table 4.5 contains the AUC task score, the evaluation conducted by a clinician, and the automatic NLG metrics. The AUC is also provided for DenseNet-121 only, i.e., without an M_E module, which shows that providing NLEs can improve task performance, e.g., for RATCHET. The AUC score obtained by DenseNet-121 is slightly lower than on similar datasets. Besides different dataset distributions and different classes being considered, a possible explanation is that instances for which there are NLEs may be less obvious and easy to detect.

Table 4.5: The S_T score, clinical evaluation, and NLG scores for our baselines on the MIMIC-NLE-v1 test set. $\geq$ GT reflects the share of generated NLEs that received a rating on par or better than the GT. Clin.Sc. reflects the average rating of 1 (lowest) to 5 (highest) that was given to the NLEs by a clinician. R-L refers to Rouge-L, and B_n to the n -gram BLEU scores. Best results are in bold. As only NLEs for correctly predicted diagnoses were evaluated, our NLG metrics cover 534, 560, and 490 explanations for RATCHET, TieNet, and DPT, respectively.

	AUC	$\geq$ GT	Clin.Sc.	CLEV	BERTS.	MET.	B1	B4	R-L	CIDEr	SPICE
<i>Ground-truth</i>	-	-	3.20	-	-	-	-	-	-	-	-
DenseNet-121	65.2	-	-	-	-	-	-	-	-	-	-
RATCHET	66.4	48%	2.90	74.7	77.6	14.1	22.5	4.7	22.2	37.9	20.0
TieNet	64.6	40%	2.60	78.0	78.0	12.4	17.3	3.5	19.4	33.9	17.2
DPT	62.5	48%	2.66	74.9	77.3	11.3	17.5	2.4	15.4	17.4	13.7

The results show that the CLEV score neither correlates to expert evaluation nor NLG metrics. One explanation could be that the evidence labels in MIMIC-NLE-v1 are highly imbalanced, i.e., predominantly *Lung Opacity*. Therefore, as the CLEV score does not take into account much of the diversity that is inherent in our NLEs, such as the location of findings and their size and appearance, a model that generates generic NLEs that make reference to *Lung Opacity* will yield a good CLEV score. The CLEV score is highest for TieNet, which performs less well on most other metrics. Upon inspecting the NLEs generated by TieNet, it often generates very generic NLEs referring to opacities as evidence, thereby inflating the CLEV score. Since the release of this benchmark, more advanced metrics such as RadCliQ (Yu et al., 2023), RadFact (Shen et al., 2024), or GREEN (Ostmeier et al., 2024) have been released, which may be a better option to evaluate the clinical accuracy in an automated way. Overall, the NLG metrics are on par with the NLG metrics for report generation (Hou et al., 2021). Hence, the difficulty of generating longer texts (reports) seems to be offset by the degree of difficulty of information-dense but shorter NLEs. The results also confirm that the NLG metrics are generally poor at reflecting expert judgment. More precisely, BERTScore, which showed the highest correlation

	LABELS: Atelectasis (Positive)		
	Natural Language Explanations for <i>Atelectasis</i>:		Clinical Evaluation:
	Ground-Truth: Opacification at the right base again is consistent with collapse of the right middle and lower lobes.		5
	RATCHET: There is a new opacity at the right lung base which may represent atelectasis.		4
	DPT: Bibasilar opacities likely represent atelectasis.		1
	TieNet: Retrocardiac opacity likely reflects atelectasis.		1
	LABELS: Edema (Positive)		
	Natural Language Explanations for <i>Edema</i>:		Clinical Evaluation:
	Ground-Truth: Indistinct appearance of the pulmonary vasculature is compatible with pulmonary edema.		2
	RATCHET: Findings suggesting mild pulmonary edema.		1
	DPT: Pulmonary edema and extensive bibasilar opacification appear slightly worse.		3
	TieNet: Diffuse bilateral pulmonary opacities, likely edema.		5
	LABELS: Atelectasis (Uncertain) and Pneumonia (Uncertain)		
	Natural Language Explanations for <i>Pneumonia</i>:		Clinical Evaluation:
	Ground-Truth: Interval appearance of patchy opacity at the left base could represent early pneumonia, although aspiration or patchy atelectasis would also be in the differential.		3
	RATCHET: Patchy opacities in the lung bases may reflect atelectasis, but infection is not excluded in the correct clinical setting.		2
	DPT: Peribronchial deformities in the right apex and elevation of the right hemidiaphragm are likely due to scarring. No acute osseous disease or infection.		1
	TieNet: Patchy bibasilar airspace opacities, likely atelectasis.		2

Figure 4.6: The GT NLEs and three model-generated NLEs explaining three different diagnosis labels on three different scans. The clinical evaluation is given on a Likert scale, where 5 is the highest, and 1 is the lowest score.

with human judgment for natural images in Chapter 3, has poor correlation for medical NLEs. One reason could be that automatic NLG metrics generally put equal emphasis on most words, while certain keywords, such as the location of a finding, can make an NLE clinically wrong, but only contribute little to the NLG score. While the GPT-2-based architecture of DPT proved very effective for natural images, its performance is less convincing on medical images. A reason could be that GPT-2 relies on embedded commonsense knowledge to generate high-quality NLEs, which is less helpful on highly specific medical text. Example NLE generations are provided in Figure 4.6.

The ground-truth NLEs received a relatively low average score of 3.2/5, only slightly surpassing the highest-scoring model (RATCHET, at 2.9). This low rating highlights significant inter-annotator disagreement between our annotator—an emergency internal medicine doctor—and the original authors of the reports. A qualitative review of all data examples with our annotator revealed several potential causes for this discrepancy. A key one was the lack of patient context: In clinical practice, doctors typically compare findings to previous scans and incorporate patient context from the indication section. This issue is exacerbated by the fact that

many NLEs in our dataset include explicit comparisons to prior scans (e.g., “*Improved right perihilar consolidation likely representing infection.*”). The presence of these “hallucinated” references to prior scans (which the model did not have access to) negatively influenced our annotator’s perception of the NLEs. Interestingly, despite being trained on these NLEs, the model-generated explanations referenced changes in pathology less frequently.

To better understand the inter-annotator disagreement, I conducted a second evaluation on the same data points with a different clinician (a practising radiologist). I used Intraclass Correlation Coefficients (ICCs), a metric to compute the reliability of individual rater scores, and inter-annotator agreement with Cohen’s kappa, a metric for inter-annotator agreement on categorical variables. The ICC1, which measures the reliability of single raters with absolute agreement, was 0.370 [95% CI: 0.31, 0.43] with a p-value of 9.84×10^{-27} . This indicates poor reliability, suggesting substantial variation in how individual annotators scored the explanations. In contrast, Cohen’s kappa for binary labelling was 0.464, which reflects “moderate agreement” between the annotators. For binary labelling, I retained only examples where both annotators classified the explanation as either 1 or 2 (“bad explanation”) or 4 or 5 (“good explanation”). Interestingly, the variance between annotators was lower for ground-truth examples compared to model-generated ones. Specifically, ground-truth examples achieved an ICC1 of 0.412 and a Cohen’s kappa of 0.542, while model-generated examples had ICC1 values ranging from 0.325 to 0.377 and Cohen’s kappa ranging from 0.348 to 0.506.

To address the concerns of lack of patient context, non-image-grounded information (such as comparison to prior scans), as well as the limitations of the CheXbert labeller, I created an updated version of the dataset: MIMIC-NLE-v2.

4.4 MIMIC-NLE-v2

Based on the findings from the previous benchmark and human evaluation, I propose a new and improved NLE dataset for chest X-rays, called MIMIC-NLE-v2, containing 54,378 *image-evidence-outcome-NLE* tuples. This dataset brings several advancements over its predecessor MIMIC-NLE-v1, including extracting more explanations, adding negative findings, incorporating patient context, grounding the evidence in the image via bounding boxes, and removing explanations that contain information that is not grounded in the image, such as references to unavailable patient history. A representative data sample is shown in Figure 4.7.

```
{
  "sentence_ID": "s51038639#2", // unique NLE identifier
  "patient_ID": "p11662490", // unique patient ID
  "report_ID": "s51038639", // unique radiology report ID
  "image_ID": "5a32886d-a4653f96-53ae3fbd-4903075b-320b865d", // unique image ID
  "gender": "F", // gender of patient ('F' or 'M')
  "age_decile": "80-90", // patient age range
  "indication": "chest pain and dyspnea. Rule out infiltrate.", // patient context
  "nle": "Subtle lower lobe opacities may reflect atelectasis.",
  "outcome_label": "atelectasis", // outcome label explained
  "evidence_label": "opacity", // evidence label for diagnosis
  "certainty": "Uncertain", // diagnosis certainty (certain or uncertain)
  "img_labels": [[1, 0], [1, 0], ..., [0, 1]], // image-wide labels
  "all_bboxes": [ // bounding boxes for all anatomical locations with:
    {
      "name": "right hilar structures", // anatomical location
      "coords": {"x1": 60, "x2": 91, "y1": 99, "y2": 134},
      "bbox_labels": [[0, 1], [0, 1], ..., [1, 0]], // bbox findings
      "bbox_descriptions": [ // bbox-grounded sentences
        "Right hilar consolidative opacities are new.",
        "Pulmonary vascular markings are prominent."
      ]
    },
    ..., { ... }
  ],
  "evidence_bboxes": [ // evidence bounding boxes:
    {
      "name": "left lower lung zone", // anatomical location
      "coords": {"x1": 120, "x2": 176, "y1": 87, "y2": 123}
    },
    ..., { ... }
  ],
  "outcome_bboxes": [...] // outcome bounding boxes
}
```

Figure 4.7: Example Representation of a single datapoint in MIMIC-NLE-v2.

The improved extraction process is summarized in Fig. 4.8. The filters consist mainly of removing sentences that make reference to previous scans or to the view, i.e., to information not available to the model. Again, only frontal, i.e., anteroposterior (AP) and posteroanterior (PA) scans are included. MIMIC-NLE-v2 builds on concurrent datasets RadGraph (Jain et al., 2021) and Chest Imagenome (Wu et al., 2021) to extract highly accurate, grounded explanations.

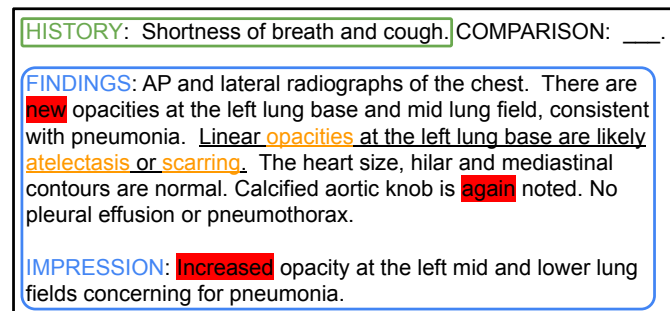


Figure 4.8: An illustration of the v2 NLE extraction process. First, the report is processed to extract the indication section (green). NLEs are extracted from the findings and impression sections (blue). All sentences for which an *suggestive_of* relation was found by the RadGraph labeller (underlined), which connects an evidence and an outcome entity (orange), are considered as raw NLEs. Next, all sentences that contain references to previous scans (red) are filtered out. UMLS matching, evidence-outcome relationship filtering, and duplicate removal are then applied to get our final 54,378 NLEs. Note that we will get two NLEs for the example above, one with outcome class *atelectasis* and one with outcome class *pleural/parenchymal scarring*.

Extracting Explanatory Sentences from RadGraph The RadGraph dataset (Jain et al., 2021) contains annotations of *suggestive_of* relations in MIMIC-CXR radiology reports. *suggestive_of* is defined as a “relation between two observation entities indicating that the presence of the second observation is inferred from that of the first observation”. Hence, a sentence containing this relation has a high chance of being an NLE, i.e., a sentence that explains the presence of one observation by the presence of another. They label MIMIC-CXR automatically by first training a model on a set of 500 radiology reports that a board-certified radiologist annotated. That model then labels 170,992 *suggestive_of* relations in MIMIC-CXR. For each *suggestive_of* relation, they also label the text spans of the connected entities via this relation. This supervised approach surpasses our keyword-based approach from the earlier version of the dataset. Another advantage is that evidence and outcome labels are explicitly labelled and do not have to be assumed, like in the earlier version.

I use these annotations as the foundation for extracting NLEs from MIMIC-CXR, thereby replacing the approach using explanation keywords and the CheXbert labeler (Smit et al., 2020), used in Section 4.2. The raw NLEs are extracted by taking the minimal set of sentences which contain both entities connected by *suggestive_of*, i.e. the *evidence* and *outcome* entities. An example is given in Figure 4.9. Typically, this is a single sentence, but sometimes the evidence for an observation can be in a previous sentence, which leads to multi-sentence NLEs. I restrict the dataset to NLEs composed of two sentences or fewer, as NLEs consisting of three or more sentences frequently result from inaccurate annotations in the RadGraph dataset. Each entity

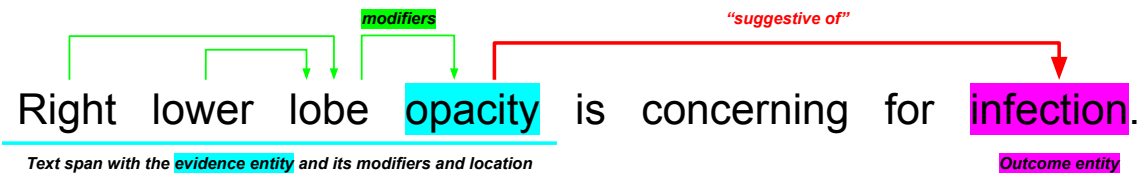


Figure 4.9: An example sentence from the RadGraph dataset. The basis of MIMIC-NLE-v2 is formed of sentences where two radiographic findings are connected via a *suggestive_of* relation.

also comes with a label whether it is mentioned as a positive (i.e., it is present), negative (i.e., it is not present), or uncertain (i.e., it is *maybe* present).

The RadGraph model’s performance in identifying the relation *suggestive_of* is insufficient to create a robust NLE dataset, with an F1 Score of 0.685 reported in their paper. To avoid having incorrect NLEs in our dataset, several steps are proposed to address this: mapping entities to UMLS terms (to ensure they are relevant clinical entities), consulting with clinicians to validate evidence-outcome relationships, applying filters to remove irrelevant sentences, and a clinician-led independent evaluation of the NLE dataset. These steps increase the accuracy (i.e., the share of correctly extracted NLEs) to 99% on a random set of 100 extracted NLEs.

The initial pool of raw NLEs corresponds to the 170,992 *suggestive_of* relations annotated in RadGraph. After removing all sentences that either contain anonymized data (i.e., where some information is replaced with an underscore “_”; 10,171 sentences), for which there is no *AP* or *PA* image available (i.e. frontal chest X-ray; 3,234 sentences) or where the evidence and outcome label are the same (727 sentences), we are left with 153,938 NLEs.

Keeping Only Image-Grounded NLEs Some NLEs contain references to previous studies that the model cannot access and thus is not expected to generate. Following the convention in most existing work (Hou et al., 2021), I had not initially removed these in our previous approach. However, it is clear that teaching a model to hallucinate how radiographic findings have evolved from prior, unseen scans is a significant limitation in the current field of radiology captioning. Specifically, in our previous dataset, it was one of the contributing factors to the low score that our ground-truth NLEs obtained in a human evaluation. In this updated version of the dataset, I remove sentences that contain references to previous scans.

This problem is starting to be recognized by others, too. A recent study (Ramesh et al., 2022) attempted to address the issue by training a model to eliminate keywords referring to previous

scans (e.g. “increased consolidations” or “interval change of opacities”). However, the approach showed limited performance and, in many instances, resulted in invalid sentence structure by removing certain words. Additionally, modifying adjectives, especially in the evidence, can be crucial in the evidence-outcome relationship. For instance, “increasing opacities” may provide stronger evidence for pneumonia in certain patient contexts than simply “opacities”.

For the above reasons, I do not change the sentences by removing the reference keyword; instead, I completely remove them. A clinician validated our list of temporal keywords. I ensured that all keywords were covered by iteratively adding new keywords to a filter until a random sample of 200 filtered sentences no longer contained any temporal keyword. In the end, I found that 72,144 NLEs made comparisons to prior scans and 7,153 referred to the specific X-ray view or type (some had both). In total, 77,724 make reference to neither. Our final temporal keywords were: “improv”, “worse”, “increas”, “decreas”, “change”, “unchange”, “interval”, “persist”, “again”, “new”, “remain”, “still”, “constant”, “recur”, “previous”, “chronic”, “progress”, “stable”, “recent”, “additional”, “cleared”, “develop”, “continue”, “history”, “prior”, “known”, “similar”, “resolv” and “CT”, with the latter being case sensitive. For view keywords, I just used “view” and “radiograph” (e.g. “lingular opacity seen on the lateral view”).

Providing Patient Context Another reason why the ground-truth NLEs in the initial version of this dataset obtained a low human evaluation score is that in clinical practice, it is rare that radiologists will look at chest X-rays without any patient context, such as their age, and the reason for why the scan was requested (i.e., what they should look for). Upon discussing the results of our human evaluation with the clinicians, one of the conclusions was that this *prior* of knowing a patient’s status and symptoms is crucial for interpreting their radiographic scan. In most work on automated chest X-ray analysis existing at the time, including MIMIC-NLE-v1, this information is, however, not included (Yuan et al.; Liu et al., 2019a; Hou et al., 2021).

To address this shortcoming, in this version of the dataset, I included the patient’s gender, age (given by a 10-year decile, e.g. “80-90”), and the indication section of the radiology report. The indication section typically contains a brief description of the patient’s symptoms as well as some specific requests for what the radiologists should look for in the scan (e.g. “___-year-old woman with chest pain, rule out pneumonia or pulmonary edema.”). This information was obtained by connecting to the Chest Imagenome dataset, as explained in the section below.

Additional patient context, such as lab results, medical prescriptions and other electronic health record (EHR) data, could be added by fusing the MIMIC-CXR dataset with the MIMIC-IV dataset (Johnson et al., 2023). However, upon discussing this with clinicians, they argued that only in very complex cases would they look at such patient data. Usually, they just consider the context provided in the report (i.e., the indication section). For this reason, I opted not to add EHR data as context to the dataset.

Matching RadGraph and Chest Imagenome The raw NLE annotations obtained from RadGraph consist of two entities linked by a *suggestive_of* relation, along with their modifiers and locations (refer to Figure 4.9). The entities are given as extracted text spans, not clinical classes. Mapping evidence and outcome spans to a set of classes serves two purposes: ensuring the tagged entities are valid radiographic findings and their relationships are clinically relevant. This filters out incorrect NLEs and allows for the task to be framed as generating NLEs for class predictions.

Wu et al. (2021) extracted scene graphs from MIMIC-CXR radiology reports and anchored them in the images. They mapped all the nodes (i.e., entities) to a Concept Unique Identifier (CUI) in the Unified Medical Language System (UMLS) (Bodenreider, 2004) and proposed several semantic categories for findings. Based on clinical relevance, MIMIC-NLE-v2 includes the following categories: technical assessment (e.g. *low lung volumes*), anatomical finding (e.g. *lung opacity*), tubes and lines (e.g. *chest tube*), devices (e.g. *prosthetic valve*), and diseases (e.g. *pneumonia*). Based on consultations with a radiologist and an internal medicine emergency doctor, UMLS classes are further grouped to reduce the number of classes in the dataset from 76 to 44, as shown in Table 4.6. For example, “density”, “lung opacity”, and “airspace opacity” are all mapped to “opacity”. These classes are closer to the level of granularity used in most chest X-ray analysis work (Irvin et al., 2019). Chest Imagenome also includes bounding boxes for these nodes. Mapping the RadGraph entities of the NLEs to Chest Imagenome classes, therefore, also yields anatomically grounded bounding boxes for all the evidence and outcome classes of the NLEs.

Since the Chest Imagenome vocabulary grouping to map report sentences to UMLS entities was not publicly accessible nor available on request, I developed a custom algorithm. The algorithm generates a three-tier hierarchical representation of each evidence or outcome span. The first tier includes only the entity, the second tier includes the entity’s “modifiers” (“relation between

MIMIC-NLE-v2 Label	UMLS classes
opacity	density, lung opacity, airspace opacity
atelectasis	atelectasis, linear/patchy atelectasis, lobar/segmental collapse
pulmonary edema	pulmonary edema/hazy opacity
vascular congestion	vascular congestion, vascular redistribution
lung lesion	lung lesion, mass/nodule (not otherwise specified), multiple masses/nodules, calcified nodule
fracture	rib fracture, clavicle fracture, spinal fracture
subdiaphragmatic air	subdiaphragmatic air, sub-diaphragmatic air
medical implants and devices	sternotomy wires, CABG grafts, aortic graft/repair, prosthetic valve, cardiac pacemaker and wires, chest tube, mediastinal drain, pigtail catheter, endotracheal tube, tracheostomy tube, PICC, IJ line, chest port, subclavian line, Swan-Ganz catheter, intra-aortic balloon pump, enteric tube

Table 4.6: The UMLS classes grouped into labels used in MIMIC-NLE-v2.

two [...] entities indicating that the first entity modifies the scope of, or quantifies the degree of, the second entity” (Jain et al., 2021)), and the third tier additionally includes the location of the entity (“located_at”, “relation between an observation entity and an anatomy entity indicating that the observation is related to the anatomy” (Jain et al., 2021)). This hierarchy ensures I have comprehensive information about the entities to accurately match them with Chest Imagenome entities. For example, when the entity is given solely as “structure”, its modifiers “nodular calcified structure” are required to map it to the class *calcified nodule*. When the entity is solely given as “thickening”, its location modifiers “parahilar bronchial walls thickening” are needed to map it to *bronchiectasis*.

RadGraph entity spans are mapped to UMLS classes by computing the cosine similarity of the embeddings (based on a medical transformer⁴) of the respective spans and class names. For each entity span in RadGraph I get $k = 10$ candidate classes based on the three-level hierarchical presentation and highest cosine similarity. The span is then mapped to the most similar candidate that has also been tagged in the same sentence by Chest Imagenome. 15,800 NLEs, which in most cases were erroneous (which may have led to the low 0.685 F1 score in RadGraph), were removed as no such match could be found.

I also removed duplicate NLEs with the same evidence and outcome class for the same chest X-ray. This occurs either when a sentence is repeated both in the findings and impression section of a report or when two different RadGraph findings are mapped to the same Chest Imagenome finding (for example “lung opacity” and “density” would be mapped to “opacity”).

⁴<https://huggingface.co/pritamdeka/PubMedBERT-mnli-snli-scinli-scitail-mednli-stsb>

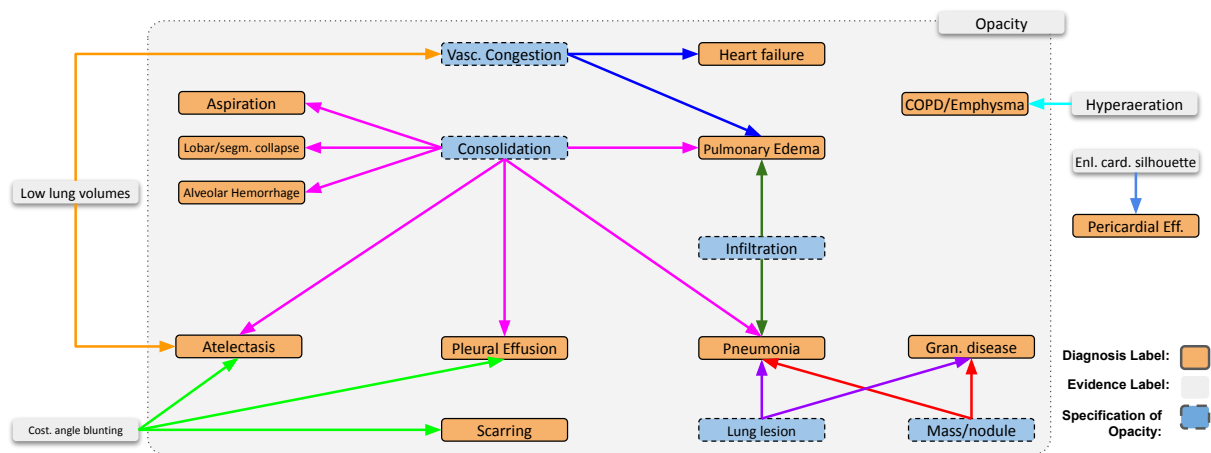


Figure 4.10: The hierarchy of permitted evidence-outcome relationships. The arrows point from evidence classes to outcome classes. The three colours represent the three levels of radiographic findings: image-level evidence (grey), specifications of opacity (blue), and outcome/-diagnosis classes (orange). All classes in the grey box have *opacity* as an evidence class.

At this stage, there are 58,356 NLEs with UMLS evidence and outcome classes and bounding box localizations of these classes.

Permitted Radiographic Hierarchies As a final step to further filter out incorrect NLEs, a filter based on an evidence-outcome relationship hierarchy, which formalizes the clinical validity and evidence hierarchy of the most common pathology relationships in MIMIC-CXR, is applied. Only relationships that occur at least 25 times and where both the evidence and outcome occur at least 100 times are considered. Examples are then reviewed for each label pair, and in consultation with a clinician, the label pair is deemed valid or not. The final relationship graph is shown in Figure 4.10.

After all extraction steps, there are 54,378 NLEs featuring 20 distinct evidence and outcome classes and 37 unique relationships between evidence and outcome classes. Figure 4.14 shows how often all evidence-outcome pairs appear. The image labels and anatomical boundary labels are restricted to the same set of 20 classes. The class labels are categorized as positive or negative, with positive referring to instances where the observation is either definitely present or maybe present. Figures 4.12 and 4.13 show the label distributions for the train and test set.

To validate our NLE extraction pipeline, a clinician evaluated a random sample of 100 NLEs. None of these NLEs were referencing information that is not grounded in the image, and only for a single NLE the outcome label was incorrect (“pulmonary edema” for the NLE “Hazy opacification at the left base is likely atelectasis, although an underlying pneumonia cannot be

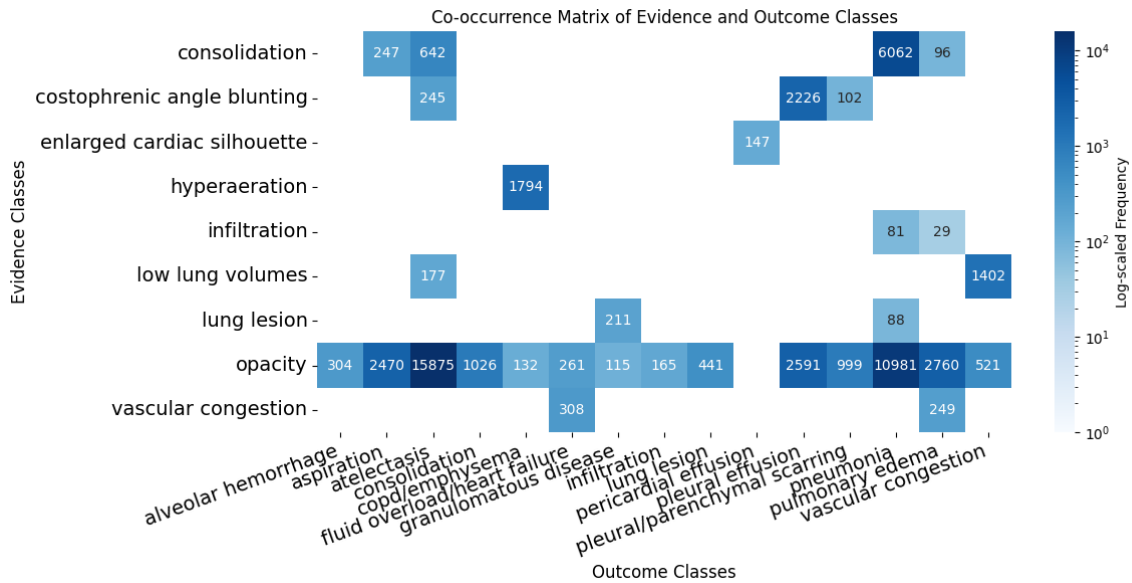


Figure 4.11: Co-occurrence matrix for evidence-outcome pairs in MIMIC-NLE-v2.

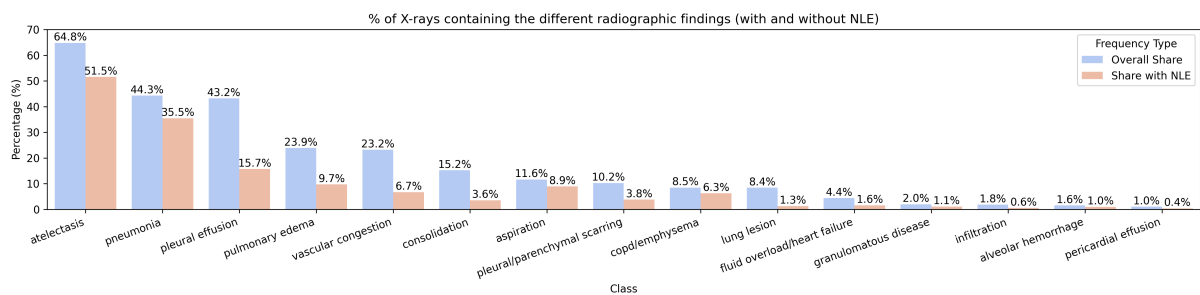


Figure 4.12: Distribution of labels in the MIMIC-NLE-v2 training set.

excluded.”). This suggests that the extraction process has almost perfect accuracy.

Dataset splitting I adhered to the official train/dev/test split provided by MIMIC-CXR, resulting in 53,093 NLEs in the training set, 386 NLEs in the development set, and 899 NLEs in the test set. Although the development and test set size is rather small, this ensures that models pre-trained on the MIMIC-CXR train set can be used for this benchmark without worrying about leakage.

The dataset also includes negative NLEs, which describe why the absence of evidence findings suggests the absence of an outcome class. These examples are not included in the main benchmark as these NLEs are generally very template-based, with over 95% of the cases involving the evidence-outcome classes of opacity/consolidation and pneumonia and lacking descriptive details as the findings are not present. Nonetheless, they are still included in the dataset for future research on negative NLEs. The benchmark version consists of 48,429 NLEs, split 47,277,

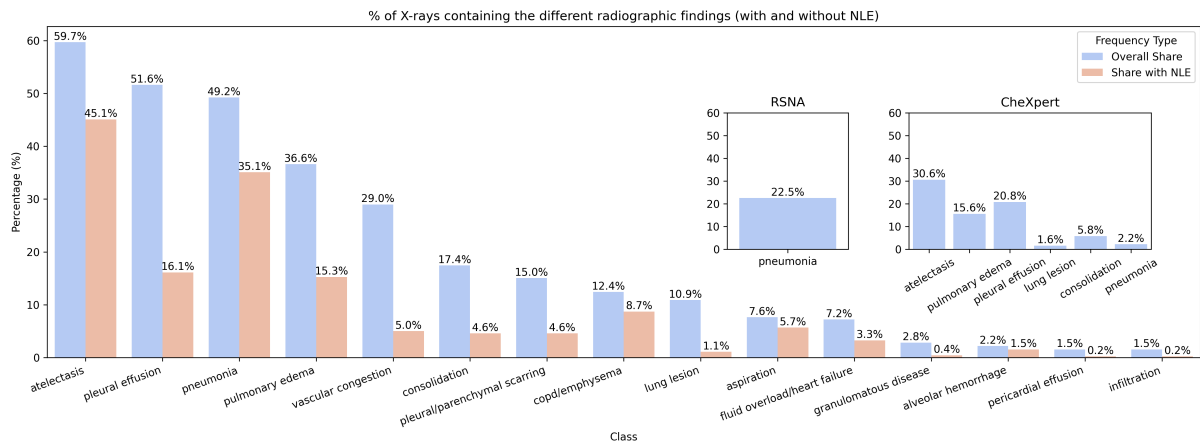


Figure 4.13: Distribution of labels in the MIMIC-NLE-v2 test set, as well as the external test sets CheXpert and RSNA Pneumonia. The share of pathologies is higher in MIMIC-NLE-v2, as I only include images with at least one positive NLE, which naturally increases the share of images with pathologies.

348, 804 across the three sets.

Grounding Evidence and Outcome Classes MIMIC-NLE-v2 includes bounding boxes that ground the evidence and outcome classes in each NLE (as shown in Figure 4.14). These bounding boxes come from Chest Imagenome, which uses an atlas-based approach to detect 29 different anatomical locations on chest X-rays from the MIMIC-CXR dataset. They then manually corrected these bounding boxes and trained a Faster R-CNN on the corrected set to annotate the entire dataset.

Using an NLP pipeline to extract radiographic findings and their locations from radiology reports, they mapped findings to specific anatomical locations. For example, when a report mentions “bibasilar opacities,” the finding “lung opacity” is mapped to both the *left lower lung zone* and *right lower lung zone* (since “bibasilar” refers to both lung bases). The precision of these bounding boxes is therefore limited to the granularity of the anatomical location. Additionally, since each bounding box corresponds to an anatomical location, findings that appear on both sides of the lung will have separate bounding boxes for *Left Lung* and *Right Lung*.

Chest Imagenome’s approach to assigning radiographic findings often results in hierarchical redundancies. For instance, “bibasilar opacities” might be assigned to both *Right Lower Lung Zone* and the broader *Right Lung*, since the lower lung is part of the whole lung. However, while *Right Lung* is technically correct, its bounding box contains much more area than just where the opacities are located. The more precise bounding boxes of *Right Lower Lung Zone*

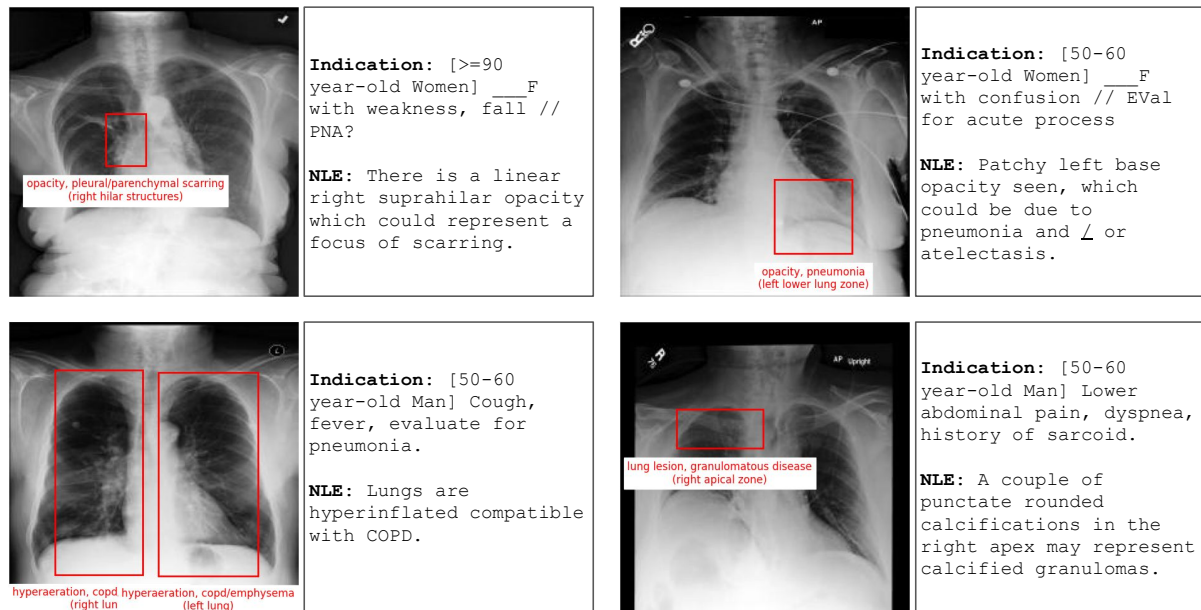


Figure 4.14: Four examples from the final dataset with grounded NLEs and patient context.

and *Left Lower Lung Zone* are sufficient for localizing the *opacity*.

Similarly, general statements like “Lungs are hyperinflated” will be assigned to *Right Lung* and *Left Lung*, but not to all their subregions. This hierarchical relationship means that for overlapping bounding boxes of the same finding, we can prioritize the more fine-grained anatomical location. To resolve overlapping bounding boxes, I identified nested relationships where one box is enclosed within or significantly overlaps another (with an overlap exceeding 5% of the subregion’s area). Using established anatomical rules, I distinguished between valid nesting relationships (such as “lower lung zone” within “lung”) and invalid ones (like “cardiac silhouette” within “lung”). In valid nested pairs, I removed the outer bounding boxes, keeping only the minimal boxes that provide the most precise localization for each finding.

4.5 Scale and Reasoning

In Section 4.3, I provided an initial benchmark of the capabilities of various medical VLMs to generate NLEs for their classifications on chest X-rays. Within the timeframe that the improved version of the dataset was created, two trends started taking shape in AI: scale (language modelling performance continually increases with model size) and eliciting reasoning capabilities from these large models (Chain-of-Thought). In this updated benchmark, which uses the improved dataset, I focus on the following two questions: How do larger VLMs perform on NLE generation, and can the success of CoT reasoning in LLMs be replicated for medical images?

Diagnosing medical images requires reasoning: Radiologists usually identify different findings on a scan and then integrate them to form an overall diagnosis in light of the patient’s condition. At the same time, LLMs have demonstrated remarkable language reasoning skills (Ouyang et al., 2022; Touvron et al., 2023). In this section, I first assess the capabilities of up-to-8-billion parameter, vision-enabled LLMs in reasoning about chest X-rays, i.e., generating NLEs. I then show how these reasoning capabilities can lead to improvements in medical image classification, i.e., medical image reasoning tasks.

4.5.1 NLE Generation Benchmark

To construct a large-scale VLM for chest X-ray analysis, I follow the framework proposed in MAIRA-2 (Bannur et al., 2024). The feature maps from a vision encoder are connected to a language model via a bridging module (a four-layer MLP adapter in this case). The instruction given to the language model includes the task, optionally patient context, and the image feature maps as tokens (see example in Figure 4.15).

Instruction

You are an expert radiologist. USER: <image> $\times 256$ Does the X-ray indicate any evidence of pneumonia? INDICATION: M with hypoxia, recent cough. ASSISTANT: Patchy opacities in the lung bases may reflect areas of atelectasis, but infection particularly in the left lung base cannot be completely excluded.

Figure 4.15: Example instruction provided to the vision-enabled LLM. <image> $\times N$ is a placeholder for the N image tokens. The text following “ASSISTANT” is the response to the instruction.

Image Processing I largely follow the image processing in MAIRA-2, except, for compute reasons, 224x224 sized images are used instead of 518x518, meaning the images are represented as 256 tokens instead of 1369. Research suggests that performance gains from using larger image resolutions are marginal and not consistent across chest pathologies (Haque et al., 2023). Images are min-max scaled to [0, 255] and z-score normalization with mean and variance derived from MIMIC-CXR are applied. The same processing is applied to the external datasets, presented in Section 4.5.3.

Vision Backbone The same vision backbone is used as in MAIRA-2: Rad-DINO-MAIRA-2, a 87M-parameter ViT-B model (Dosovitskiy et al., 2020). Rad-DINO-MAIRA-2 is trained on a total of 1,404,383 chest X-rays from 6 different data sources. DINOv2 (Oquab et al., 2023) is a state-of-the-art self-supervised learning method for vision transformers (ViTs), leveraging a teacher-student siamese network to learn robust global and local image representations *without* text captions.

Language Model Results are provided for a wide range of language model sizes, from the smallest GPT-2 (124M parameters) to frontier LLMs such as Llama-3.1-8B (8B parameters) (Grattafiori et al., 2024). This allows us to study the effect of model size on medical image reasoning capabilities. For all these LLMs, I take their pre-trained weights and fine-tune a vision-enabled version of the task. Including GPT-2 allows us to isolate the effect of increasing the scale of the language models from the effect of the new vision model (as I used a less advanced vision backbone in the benchmark presented in Section 4.3).

Results A straightforward approach is used to evaluate NLE generation capabilities: Each data point consists of an image-finding pair, and the models are trained to predict the NLE using the instruction shown in 4.15. For simplicity, I do not separate the task into T and E , requiring a separate M_T and M_E . Essentially, in this section, only results for M_E on the explanation generation task E are provided. All models in Table 4.7 used a learning rate of 0.0002, a batch size of 1024, and included the indication section in the prompt. The results show that, overall, NLG metrics are much higher than in the benchmark provided for the earlier version of the dataset (Table 4.5). This is the case even with GPT-2, where the only change to DPT-2 is using a different vision encoder, suggesting that the stronger vision encoder and/or dataset quality differences improve performance. Still, increasing the scale of the LLM leads to further

improvements. The optimal LLM size seems to be 3 billion parameters, with performance decreasing again for the 8B version, nearing similar NLG metrics as the significantly smaller GPT-2. The results also show that keeping the image encoder frozen is beneficial across all set-ups.

Rad-Dino	LLM	μ NLG	Meteor	Cider	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge
<i>trained</i>	GPT-2	0.298	0.157	0.555	0.262	0.161	0.107	0.073	0.293
<i>frozen</i>	GPT-2	0.314	0.166	0.659	0.258	0.164	0.113	0.083	0.298
<i>trained</i>	Llama 1B	0.319	0.161	0.704	0.247	0.159	0.112	0.082	0.312
<i>frozen</i>	Llama 1B	0.321	0.167	0.706	0.253	0.162	0.112	0.081	0.312
<i>trained</i>	Llama 3B	0.320	0.163	0.688	0.263	0.164	0.113	0.083	0.308
<i>frozen</i>	Llama 3B	0.329	0.165	0.755	0.254	0.166	0.118	0.087	0.318
<i>trained</i>	Llama 8B	0.276	0.148	0.482	0.237	0.141	0.091	0.062	0.274
<i>frozen</i>	Llama 8B	0.315	0.164	0.653	0.261	0.166	0.114	0.083	0.308

Table 4.7: Performance comparison of different LLMs across various NLG metrics. Best values in each column are highlighted in bold. μ NLG refers to the average score across metrics.

In order to understand whether the dataset upgrades play a role in this increased NLE generation performance, I conducted ablation studies. First, I investigate the role of the indication section. Table 4.8 shows that excluding the indication section leads to significant drops across all metrics, highlighting the importance of patient context in not just diagnosing patients, but also explaining findings. We also see that randomizing the indication at inference time, i.e., randomly associating a patient context from a different patient, further reduces performance. This additionally underlines the importance of patient context in chest X-ray reasoning.

LLM	Indication	μ NLG	Meteor	Cider	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge
Llama 1B	included	0.321	0.167	0.706	0.253	0.162	0.112	0.081	0.312
Llama 1B	none	0.288	0.150	0.578	0.214	0.134	0.093	0.069	0.290
Llama 1B	random	0.278	0.153	0.479	0.229	0.136	0.088	0.061	0.284

Table 4.8: Performance comparison of Llama-3.2-1B under different indications across various NLG metrics. Best values in each column are highlighted in bold.

Next, in order to isolate the effect of data quality from the effect of just adding the indication section, I propose to use subsets of both v1 and v2 datasets where only classes present in both versions are included (i.e., atelectasis, consolidation, pulmonary edema, lung lesion, pleural effusion, and pneumonia). These subsets can then be used to compare the performance of

models trained on either MIMIC-NLE-v1 or MIMIC-NLE-v2. MIMIC-NLE-v2 is used as an evaluation dataset as it is of higher quality and, therefore, a better benchmark for evaluating the quality of the NLE generation model. Whilst I also ran this experiment with the indication section included, these results may be misleading as they explicitly disadvantage the model trained on v1, which will only see the indications at test time. Table 4.9 shows that, even without the indication, the higher quality v2 dataset still leads to overall performance improvements. As expected, this becomes more significant once the indication section is included. Overall, the NLG metrics are lower than in Table 4.8, which may be due to the smaller training set and/or because the subsets of labels included here could be more difficult to reason about.

LLM	Train subset	μNLG	Meteor	Cider	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge
Llama 1B	trained on v1	0.254	0.133	0.416	0.201	0.116	0.070	0.045	0.276
Llama 1B	trained on v2	0.267	0.142	0.414	0.235	0.138	0.087	0.059	0.274
Llama 1B	trained on v1 (+ ind.)	0.228	0.115	0.346	0.162	0.088	0.052	0.032	0.251
Llama 1B	trained on v2 (+ ind.)	0.284	0.156	0.504	0.241	0.143	0.091	0.061	0.290

Table 4.9: Performance comparison of Llama-3.2-1B trained on subsets of either the v1 or v2 training dataset, and evaluated on a v2 subset. “+ ind.” means that the indication section is included during training (when available) and during evaluation. Best values in each column are highlighted in bold.

4.5.2 Chain-of-Thought Reasoning for Chest X-rays

Chain-of-Thought (CoT) reasoning has been shown to significantly improve the performance of LLMs on reasoning tasks (Wei et al., 2022). CoT consists of first reasoning about a task before providing an answer. The main difference to NLE approaches discussed earlier in this thesis is that instead of M_T first providing an answer to a task and then M_E being conditioned on this answer, now M_E first provides an explanation for the task and the answer is then given based off that explanation (referred to “explain-then-predict” versus “predict-then-explain” in Camburu et al. (2018)). Inherent CoT, i.e., the ability to improve the performance of a model by prompting it to think step-by-step, without specific training of that kind, generally is only observed in models with more than 100B parameters (Suzgun et al., 2023), a scale not common in biomedical applications. On the other hand, Supervised CoT, where the model is trained to generate such rationales using ground-truth annotations, can translate these benefits to smaller language models (Kim et al., 2023).

In this section, I use supervised CoT to study the effectiveness of using NLEs as supervision

for intermediate reasoning steps in improving chest X-ray analysis. The same instruction as shown in 4.15 is used. For each label, the model is tasked with answering the question:

“Is there evidence for [label] on this X-ray?”

The answer options are as follows:

- **No CoT (Baseline):** Here, a simple binary response is used, either *Yes* or *No*.
- **With CoT:** Here, the response depends on the following:
 - If evidence exists (positive case) and an NLE is available, the response is the NLE.
 - If evidence exists (positive case) but no NLE is available, the response is: *“There is indeed evidence.”*
 - If the label is not present in the report (negative case), the response is: *“There is no evidence.”*

Table 4.10 further illustrates our set-up. Including the yes/no baseline enables us to isolate the benefits of the CoT reasoning over simply using an LLM on top of a vision encoder for classification. I also added a “vision encoder only” baseline, where the vision encoder is fine-tuned on the classification task. For the VLM approaches, the vision encoder is frozen. To obtain the AUC scores shown in Table 4.11 and 4.12, I use the likelihood of predicting “There is no evidence.” as logit for predicting a class as negative. This ensures that no NLE information is used at test time. All models compared in this section were selected based on their lowest loss on the MIMIC-NLE-v2 development dataset.

Is Positive	NLE Available	Response (No CoT)	Response (With CoT)
Yes	Yes	Yes	[NLE]
Yes	No	Yes	Yes, there is indeed evidence.
No	N/A	No	There is no evidence.

Table 4.10: Summary of model responses under No CoT and With CoT scenarios.

In Table 4.11, we observe that the VLMs using a 1B LLM (Llama-3.2-1B) are outperforming the vision encoder-only approach, highlighting how, arguably, the size of the LLMs can help improve chest X-ray classification performance. For some of the included classes the share of NLEs is quite low. For example, for lung lesions, only 15.5% of the positive occurrences of the class in the dataset come with an NLE. For others, the share can be much higher (e.g. 80.1%

	Avg.	NLE ↓	NLE ↑	Pneum.	Pl. Eff.	Lesion	Conso.	Atel.	Vasc. C.	Ed.	Scarr.	Inf.	Gran. D.	HF	COPD	Asp.	Alv. H.	Per. Eff.
Share with NLE	55.9%	29.9%	71.5%	80.1%	36.4%	15.5%	23.3%	79.5%	28.9%	40.7%	37.2%	31.1%	55.1%	36.0%	73.8%	76.9%	63.5%	37.4%
Vis. Encoder Only	0.733	0.717	0.718	0.680	0.860	0.679	0.714	0.691	0.734	0.804	0.790	0.528	0.720	0.716	0.914	0.597	0.704	0.858
+ 1B LLM, yes/no	0.786	0.780	0.776	0.729	0.874	0.671	0.754	0.738	0.745	0.799	0.802	0.827	0.693	0.788	0.899	0.670	0.924	0.870
<i>No Indication</i>	0.732	0.720	0.710	0.669	0.870	0.643	0.714	0.702	0.745	0.817	0.770	0.601	0.707	0.720	0.906	0.537	0.738	0.843
+ 1B LLM, CoT	0.788	0.771	0.793	0.727	0.855	0.659	0.791	0.760	0.739	0.813	0.774	0.841	0.736	0.793	0.912	0.717	0.908	0.802
<i>No Indication</i>	0.746	0.745	0.718	0.678	0.861	0.670	0.745	0.722	0.760	0.805	0.795	0.821	0.684	0.762	0.921	0.568	0.736	0.659
1B cLLM ³ , CoT	0.779	0.781	0.759	0.711	0.854	0.694	0.775	0.736	0.745	0.802	0.778	0.846	0.715	0.815	0.917	0.664	0.815	0.811
GPT-2, CoT	0.608	0.597	0.598	0.628	0.731	0.525	0.642	0.631	0.639	0.628	0.731	0.503	0.595	0.579	0.813	0.566	0.355	0.557

Table 4.11: Classification performance (AUC scores) comparison of different models with and without CoT reasoning. The “Share with NLE” row specifies the share of positive cases where our dataset has a clinician-written NLE. The NLE ↓ and NLE ↑ columns group all the classes where either less or more than half of the examples come with an NLE. “cLLM” stands for an LLM specifically trained on clinical text. The following abbreviations were used for the class names: “Pneum.” is pneumonia. “Pl. Eff.” is pleural effusion. “Conso.” is consolidation. “Atel.” is atelectasis. “Vasc. C.” is vascular congestion. “Ed.” is pulmonary edema. “Scarr.” is pleural/parenchymal scarring. “Inf.” is infiltration. “Grad. D.” is granulomatous disease. “HF” is fluid overload/heart failure. “Asp.” is aspiration. “Alv. H.” is alveolar hemorrhage. “Per. Eff.” is pericardial effusion.

for pneumonia). Generally, we observe that the benefits of the CoT approach only manifest for the subset of classes with a *relatively* high share of NLEs (0.793 v. 0.776 AUC on the classes with a share of NLE exceeding 50%, see NLE ↑) versus those with less than 50% (0.771 v. 0.780 AUC, see NLE ↓). We also see that leaving out the indication section significantly harms classification performance. Further, the benefits of using a VLM over a vision encoder-only apply only when scaling up the VLMs. For the 124M GPT-2, performance is significantly lower than for the vision encoder only. Using a clinical LLM does not provide additional benefits, however, the model weights used were not from a peer-reviewed publication.

4.5.3 CoT and Generalization

Chest X-ray analysis presents two significant challenges. First, as demonstrated in my earlier experiments, the ground-truth can be very ambiguous, with large inter-annotator disagreement. This raises concerns regarding the reliability and utility of these benchmarks, especially when the ground-truth is automatically extracted from radiology reports. In theory, the gold standard ground-truth should be obtained through laboratory results, e.g., blood tests or sputum analysis confirming pneumonia. Unfortunately, no such datasets are publicly available. An alternative solution is to validate the ground-truth with multiple expert annotators, thereby reducing ambiguity. Second, clinical models must be validated on external datasets to ensure robust gen-

eralization and prevent reliance on hospital-specific spurious correlations (Ktena et al., 2024). To address these issues, I further evaluate the CoT framework on the two external datasets outlined below, where the ground-truth labels were established through multi-expert annotation rather than being extracted from single-author radiographic reports.

Chexpert500 The CheXpert test set consists of 500 studies from 500 patients, with each study annotated by five radiologists. The ground-truth was determined by a majority vote among five radiologists. I binarized the labels by grouping “present” and “uncertain-likely” as positive and “absent” and “uncertain-unlikely” as negative. For each case, only a single frontal image is included. Only classes included in MIMIC-NLE-v2 (the training set) are considered: “pneumonia”, “pleural effusion”, “lung lesion”, “consolidation”, and “atelectasis”.

RSNA Pneumonia The RSNA dataset (Shih et al., 2019), a subset of the NIH CXR8 dataset, consists of 30,000 frontal chest radiographs with annotations for pneumonia detection. Eighteen board-certified radiologists from 16 academic institutions, with a mean of 10.6 years of experience, performed the annotations. The radiologists underwent an initial training phase, annotating 50 radiographs to calibrate and standardize their approach. They used a bounding box system to localize pneumonia-like opacities and categorized findings as high, medium, or low probability, with only medium and high probabilities included in the final dataset. Although 4,527 test images were triple-annotated with majority voting by three radiologists, the labels for these images were, unfortunately, not publicly released. Instead, I run the evaluation on the entire released train dataset.

Results Table 4.12 shows that the benefits of CoT finetuning are much more pronounced on these external datasets. This suggests that the CoT reasoning improves the model’s capability to generalize, potentially because the constraint to provide NLEs reduces reliance on spurious correlations. Using CoT improves performance over the “yes/no” by 3.3 AUC points on Chexpert500, and 5.4 AUC points on RSNA. Interestingly, even though these external datasets do not have patient context, meaning that the model versions trained with indication only see this patient context at train time, including the indication at train time still leads to minor test time performance gains. Overall these results show that, when using large VLMs, supervised CoT fine-tuning using our MIMIC-NLE-v2 datasets leads to higher performing and more robust chest X-ray analysis models.

	CheXpert500						RSNA
	Avg.	Pneum.	Pl. Eff.	Lesion	Conso.	Atelact.	Pneum.
Vision Enc. Only (87M)	0.865	0.864	0.926	0.804	0.891	0.805	0.732
+ LLM (1B), yes/no	0.878	0.846	0.932	0.890	0.886	0.836	0.785
<i>No Indication</i>	<i>0.863</i>	<i>0.864</i>	<i>0.929</i>	<i>0.840</i>	<i>0.887</i>	<i>0.794</i>	<i>0.768</i>
+ LLM (1B), CoT	0.911	0.935	0.938	0.925	0.924	0.833	0.839
<i>No Indication</i>	<i>0.903</i>	<i>0.932</i>	<i>0.923</i>	<i>0.880</i>	<i>0.914</i>	0.866	<i>0.780</i>
+ cLLM (1B), CoT	0.888	0.896	0.934	0.858	0.925	0.828	0.846
+ GPT-2 (124M), CoT	0.805	0.808	0.885	0.854	0.712	0.763	0.696

Table 4.12: Classification performance (AUC scores) impact of CoT reasoning on external datasets. “cLLM” stands for an LLM that was specifically trained on clinical text. The following abbreviations were used for the class names: “Pneum.” is pneumonia. “Pl. Eff.” is pleural effusion. “Conso.” is consolidation. “Atelact.” is atelectasis.

4.6 Conclusion

This chapter established the use of NLEs in medical imaging and provided a dataset and general method to achieve this. The introduced MIMIC-NLE-v1 is the first NLE dataset for the medical domain. A comparison of at-the-time, leading vision-language models, including a clinical evaluation, reveals the importance of considering patient context, removing non-image-grounded information from the reports, and the general high ambiguity and inter-annotator disagreement in chest X-ray analysis. These findings informed the upgraded and improved version of the dataset: MIMIC-NLE-v2. This new dataset allows the training of significantly stronger NLE generation models. I also used it to study the effectiveness of scaled-up LLMs and Chain-of-Thought reasoning in chest X-ray analysis. The results on external, gold-standard datasets indicate that both scale and reasoning can improve model capabilities, suggesting that our dataset paves the way to introduce advanced reasoning capabilities in medical image analysis.

The findings highlight the potential of NLEs to both communicate the reasoning of AI models to humans and to enforce human reasoning onto the AI models. While many arguments advocate for using NLEs in medical applications, no studies exist that verify how people actually engage with them in safety-critical settings. In an age where powerful LLMs are ubiquitous, it is crucial to not only study the correctness or faithfulness of such explanations but also understand how they affect users. In the next chapter, I conduct a large-scale user study to investigate this.

4.7 Impact and What Happened Since

The MIMIC-NLE-v1 dataset was the first dataset providing NLEs for a medical task, and its publication at MICCAI 2022 ([Kayser et al., 2022](#)) has been cited 27 times to date⁶. The dataset has been used to enhance instruction-following datasets to train improved radiology foundation models ([Chen et al., 2024b](#); [Pellegrini et al., 2023](#)), to benchmark NLE approaches for medical imaging ([Hamza et al., 2024](#)), to motivate extending NLEs to other medical image modalities ([Thomas et al., 2024](#)), and to advocate for using NLEs in real life decision-support settings ([Dayanandan and Lall, 2024](#)).

The dataset extraction process can be viewed as equivalent to filtering techniques used in modern LLM training pipelines, where there has been an increasing emphasis on the importance of high-quality data selection ([Grattafiori et al., 2024](#)). Several works demonstrate that filtering out for semantically rich, reasoning-heavy text can significantly enhance a model’s ability to perform complex tasks ([Liu et al., 2024](#)). The datasets presented in this chapter thus provide a useful tool to apply these principles to medical image analysis models. As in Section 3.6, it is worth noting that many of the dataset processing steps proposed could be further improved with the LLMs that have been released since.

⁶Google Scholar, 19.03.25

5 | Evaluating Language as an Interface Between Humans and AI in a Safety-Critical Setting

AI models have progressed rapidly in recent years and are increasingly used across various domains, including medical applications ([Moor et al., 2023](#)). At the same time, a remaining significant barrier to the adoption and regulatory approval of deep learning models in medical imaging is the limited transparency of the reasoning processes underlying these models ([Hassija et al., 2024](#)). To address this, explainable AI (XAI) techniques, such as the NLEs introduced in Chapter 3 and 4, aim to make these models safer to use by making their inference process more transparent.

However, current explainability methods are seldom evaluated in the way they are intended to be used: by real-world end users. While there is an abundance of literature and regulatory frameworks that advocate the importance of interpretability in medical applications ([Frasca et al., 2024](#)), the practical utility of AI explainability methods that aim to address this remains poorly understood. There can often be several correct ways to explain a decision, and the criteria for judging their quality are diverse (e.g., plausibility, faithfulness, clarity ([Jacovi and Goldberg, 2020](#))). Since one of the primary benefits of explanations is their utility to end-users, evaluating them with human subjects is crucial. As explanations can lead to confirmation bias and user preference frequently does not align with desired quality requirements (e.g., a user might prefer an explanation type even if they are more likely to misinterpret the AI), explanation usefulness needs to be evaluated via proxy measures ([Ehsan and Riedl, 2020](#); [Liao et al., 2022](#); [Liao and Varshney, 2021](#); [Ehsan et al., 2021](#)). Only a few studies attempt this, with some suggesting that many XAI methods may not work as well as expected ([Adebayo et al., 2018](#); [Shen and Huang, 2020](#); [Hoffmann et al., 2021](#); [Margeloiu et al., 2021](#)).

This chapter addresses this by carrying out a large-scale human subject study to evaluate the usefulness of natural language explanations (NLEs), saliency maps, and a combination of both in the setting of imperfect AI and imperfect XAI. Saliency maps, which attribute importance weights to regions in an image, are the prevailing mode of interpretability in medical imaging ([Van der Velden et al., 2022](#)). NLEs, on the other hand, are becoming more widespread with recent advances in large language models ([Wei et al., 2022](#)) and have been advocated for deployment in clinical practice ([Reyes et al., 2020](#)). In addition, the communication interface of

generative AI is often language-based (Gabriel et al., 2024), which offers a human-like mode of interaction. Some research suggests that this linguistic interface “humanizes” these AI systems and thereby increases reliance on them (Breum et al., 2024). Whilst some work indicates a preference of humans for text-based explanations (Tonekaboni et al., 2019; Reyes et al., 2020), it is not well understood how they actually impact human-AI interaction. I also studied the combination of both explanation modalities to understand if they can complement each other. I consider both AI and XAI to be *imperfect*, to reflect real-world applications, where both the AI predictions and the AI explanations can contain errors. Specifically, I investigate how different types of explanations, taking into account both AI and XAI *correctness*, affect users in a clinical decision-support system (CDSS) environment. As the focus is on AI that enhances medical practitioners (Langlotz, 2019), rather than replacing them, the proxy for the usefulness of explanations is how much they improve human performance in human-AI collaborative chest X-ray analysis. In the final study, 85 doctors and medical students analysed 80 unique images each, distributed across four different CDSS set-ups: either of the three explanation types or the “no explanation” control condition. The study design is illustrated in Figure 5.1.

The study outcomes highlight the pitfalls of language-based explanations, which lead to over-reliance. Interestingly, however, saliency maps and NLEs complement each other, and their combination is the most useful explanation type. I also find that explanation correctness, and how it aligns with AI correctness, is an important factor in determining whether explanations are helpful or harmful to users. When they misalign, such as when the AI is correct but the explanation contains many inaccuracies, this negatively impacts human performance. Conversely, when they align, explanations improve our participants’ task performance.

This chapter will first discuss related work (Section 5.1), then present the study design (Section 5.2), discuss the participants (Section 5.3), the study data (Section 5.4), and finally the study outcomes (Section 5.5).

5.1 Related Work

The rapid advancements in AI have spurred discussions about its capability to automate processes and outperform humans in specific tasks. However, a parallel discourse is centred on how AI can enhance, rather than replace, humans, a domain referred to as human-AI collaboration. This is especially applicable to medical imaging, where concerns around safety and trust make autonomous deployment of AI models challenging. CDSSs, where AI models offer recommendations to humans for specific tasks, are a common form of human-AI collaboration in clinical practice.

Existing studies investigate these systems through the lens of human-AI team performance by looking at how the sequential order of human and AI decisions affect performance (Fogliato et al., 2022), what influence the assertiveness of AI suggestions has (Calisto et al., 2023), or which kind of users benefit the most from it (Gaube et al., 2023). A recent large-scale study conducted by Agarwal et al. (2023) shows that, in most cases, human performance is enhanced when using CDSSs. Tanno et al. (2025) look at collaborative report generation, i.e., a CDSS with language as an interface, and show that this leads to improved outcomes.

However, few works have looked at the usefulness of XAI in clinical applications. Du et al. (2022) consider a simple, 5-feature set-up to compare explanation-based and feature attribution methods in a CDSS setting. Rajpurkar et al. (2020) and Ahn et al. (2022) provide visual explanations when evaluating the usefulness of a CDSS, but they do not look at the effect that XAI explanations had. Gaube et al. (2023) find that visual explanations improve the diagnosis performance for non-task experts, but they do not compare it to other XAI methods. Tang et al. (2023) look at AI tools for lung nodule detection in chest X-rays and find that localisation maps do not improve performance.

Despite limited studies investigating the role of XAI in CDSSs through the lens of human-AI team performance, several studies suggest that XAI is an important need for clinicians. In a study by Tonekaboni et al. (2019), when clinicians were exposed to AI-based health risk evaluations in simulated settings, they expressed a strong interest in understanding the criteria the AI model relied upon to derive its conclusions. Similarly, in a laboratory experiment, Cai et al. (2019b) examined how clinicians interacted with an AI cancer diagnosis support system. Results showed clinicians sought not only specific outputs but also broader understanding of the

AI's capabilities, limitations, and design objectives. In a separate experiment, [Cai et al. \(2019a\)](#) studied pathologists' use of an AI system to find similar pathology scans. They observed that pathologists creatively used the provided tools, not just for their intended purpose but also to dissect the core functioning of the search mechanism.

In this study, I include both post-hoc explainers (saliency maps) and self-explainable models (NLEs), as well as a combination of both types. Despite most of today's frontier models explaining their decision-making in natural language, besides this work, only [Morrison et al. \(2024\)](#) have evaluated NLEs as an interpretability tool using human subject studies. This work differs by the task (safety-critical CDSS vs. bird classification), by looking at the combination of visual and textual explanations, and by extending explanation correctness to be continuous (rather than binary) and defined even for incorrect AI. This is also the first attempt to compare the effect of different explanation types and take into account their interaction with diagnosis and explanation correctness in a clinical, real-world set-up.

5.2 Study Design

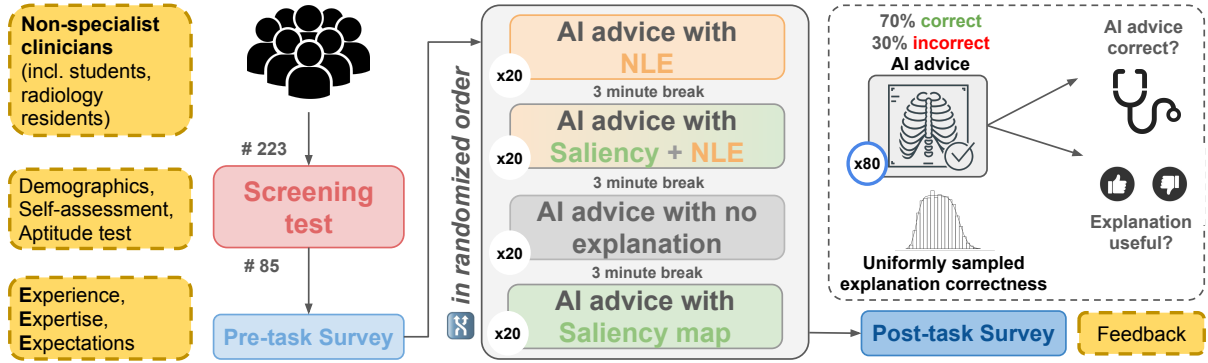


Figure 5.1: The flow of the user study that every participant goes through.

5.2.1 Task Design

The pre-registered and IRB-approved¹ user study was designed to evaluate the usefulness of NLEs, saliency maps, and their combination in a clinical decision-support context. It also controls for AI advice correctness $C_{AI} \in \{0, 1\}$ and explanation correctness $C_X \in [1, 7]$. The main research question is how different explanation types, in the context of advice correctness C_{AI} and explanation correctness C_X , affect human performance on the task of classifying chest X-rays, where explanation *usefulness* equates human performance. It involved 85 clinical participants and was developed through iterative pilot studies and consultations with expert clinicians. The *usefulness* of explanations is considered to be their ability to help a user discern whether a model’s prediction is correct or not. Figure 5.1 provides an outline of the study design.

Definition of explanation correctness C_X : C_X captures to what extent the information provided in an explanation is clinically, factually correct. An explanation can contain a substantial amount of incorrect information even when the AI prediction was correct, and vice versa, similar to definitions from [Honovich et al. \(2022\)](#) and [Zhang et al. \(2020\)](#). Note that this is different from other explanation criteria such as *faithfulness* (i.e., how accurately it reflects the true reasoning process of the model) and *plausibility* (i.e., how convincing the explanation is to humans) ([Jacovi and Goldberg, 2020](#)) (for more details, see Section 2.3.2). The ground-truth for both advice and explanation correctness is obtained from annotations by three expert radiologists. Further details on the annotation process are given in Section 5.4.2.

¹osf.io/nf52s; Approval Nr. CS_C1A_23_018_001

The following four conditions are studied: (i) X_{None} (participants receive the AI model’s advice without any explanation), (ii) X_{SM} (participants receive the model’s advice and a saliency map), (iii) X_{NLE} (participants receive the model’s advice and an NLE), (iv) X_{Comb} (participants receive the model’s advice, a saliency map, and an NLE). To strike a balance between having a reasonable representation of correct and incorrect model predictions and not making the model appear overly unreliable, I simulated an environment where the model has an accuracy of 70% under each condition. I conducted pilot studies using AI assistants with 60%, 70%, and 80% accuracy (see Figure 5.6), and found that 70% represented the best trade-off between realism and feasibility within our budget constraints. One may argue that 70% accuracy is lower than what would be expected of a deployed AI assistant, potentially influencing participants’ overall perception of the model. The main concern is that participants might entirely disregard the AI due to its perceived inadequacy. However, this risk is mitigated by the use of non-expert participants, who may be less attuned to the model’s accuracy. Moreover, even if their overall impression is somewhat negative, this is acceptable: the results show that, in certain conditions, participants still exhibit overreliance on the AI. This suggests that they continue to engage with the assistant despite its limitations. Per condition, participants are shown 20 cases, which consist of a chest X-ray, the patient context, the AI advice, e.g., “Pneumonia”, and a condition-specific explanation. They are then asked to express their agreement with the AI advice (“Not present”, “Maybe present”, or “Definitely present”). I also ask them whether they found the explanation useful in their decision-making (e.g. “How useful was the AI model’s explanation in helping you decide whether the AI was right or wrong in suggesting pneumonia.”). This is meant to encourage them to engage with the explanation, and it allows us to quantify the relationship between *perceived* and *actual* explanation usefulness. Image-explanation pairs are sampled in a way that ensures that the overall distribution of C_X scores is as uniform as possible (so that all C_X levels are well represented). The distributions are shown in Figure 5.14 and 5.15.

Robustness Measures Several robustness measures and design considerations are implemented to improve the study’s reliability.

In order to maximize the engagement with the AI, and thereby with the explanation types, participants were **not asked to make a diagnosis before they were shown the AI advice**. This has been shown to lead to higher reliance on the AI (Fogliato et al., 2022). In addition,

participants are given limited information about the model, as evidence suggests that people are more likely to find AI advice reasonable when no information about system performance is given to them (Kim and Song, 2020).

The designed CDSS provides a suggestion (the *AI advice*) for each image, consisting of a **single radiographic finding** predicted by the AI. Even though chest X-rays may often contain multiple findings, the design is simplified by focusing on one finding per image. It is clearly communicated to participants that this finding is neither the only nor necessarily the most important finding. There were three main reasons for this abstraction of the diagnosis process: First, the NLE pipeline is not optimized for multi-label classification, which could result in an out-of-scope scenario where different NLEs for different labels can contradict each other. Second, listing all findings for an image would introduce false negatives, i.e., it could become evident that a model missed a prediction. As I only consider NLEs for positive predictions for the time being, this would again introduce conditions that cannot be accounted for. Lastly, the clinicians involved in the study assured us there is a benefit in CDSSs that mention specific findings rather than providing a holistic interpretation of the scan (for example, to ensure no findings were missed). I also piloted a study design with multiple findings per image and concluded that the UI design would become too complicated and its design itself would become a significant factor affecting the performance.

Other concerns were **carry-over effects and reader fatigue**. Reader fatigue could set in because participants engage with 80 different chest X-rays and AI decisions and explanations, which takes them an hour on average. Carry-over effects occur when people's engagement with XAI explanations is affected by the AI model they have encountered previously in the study. For example, it could be that explanation type A reduces a user's trust in the system and, therefore, the user would be more likely to start ignoring the model suggestions, even if the type of explanation is changed. Ideally, participants would engage with each type of explanation in a separate session, with a washout period in between, in order to reduce these effects. However, as they complete the evaluation four times, once for each scenario, it is not feasible to have a washout period that is long enough to ensure they forget about the images between each task. To mitigate order effects and user fatigue, I randomize the order of the conditions for each participant. I also enforce three-minute breaks between each condition, where participants are given the option to follow a guided meditation. It is emphasized multiple times throughout the

study that the users are engaging with different AI models in each condition to avoid a person’s engagement with explanation types carrying over. Lastly, I introduce an incentive of doubling the compensation for participants who perform in the top 20%. This aims to ensure that users are dedicated and motivated throughout the 80 cases.

Designing a Clinical Decision Support System To conduct our online study, I built a decision-support system online application using Streamlit², showcased in Figure 5.2. I iteratively built the application over several months to make it as simple, intuitive, and user-friendly as possible. Participants first had to enter the email address with which they signed up for the study and were then automatically assigned a user ID in our backend. Every participant has to watch a short, 3-minute instruction video³ before they start the study. The video is based on a visual instruction PDF with explanations (see Figure 5.3), and both the video and the PDF are accessible to the participants at any point in the study. The platform also contained all the pre- and post-task survey questions. To further encourage dissociation between the models and explanation types, I use different randomized AI names and colours and specific transition pages (e.g., “Proceed to Model ABC”, see examples in Figure 5.2). Throughout the study, I measured time to enforce minimum breaks (3 minutes) and minimum time spent per example (10 seconds).

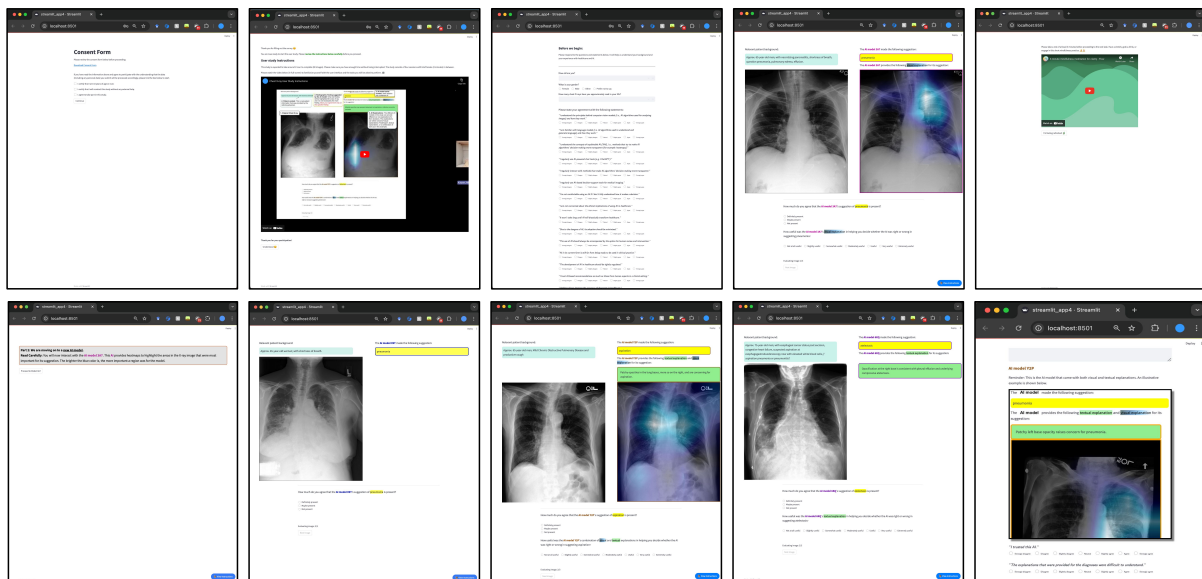


Figure 5.2: Snapshots of the different interfaces of the CDSS.

For each case, participants are shown an X-ray, patient context, an AI suggestion, and an AI explanation (except for the no-explanation scenario). The X-rays are shown as high-resolution

²<https://streamlit.io/>

³<https://youtu.be/siRH1uffSM8>

JPEGs (between Full HD and 4K resolution). I used JPEGs rather than DICOMs as they are easier to display and correspond to the same image modality that the models see, and our three radiologist annotators and an emergency medicine doctor assured us that the quality of our images was sufficient for this use case. [Trieu et al. \(2023\)](#) also confirms that “most radiologists rated DICOM and JPEG images clinically and statistically equivalent with respect to difficulty of interpretation” in a mammography setting. For patient context, I used the indication section that is provided with the radiology reports and added age (10-year range) and gender from the metadata. It was given to participants in the following format: “[Approx. X-year-old male/female] *indication section*”. As the indication sections in our datasets contain removed information (anonymization) and many abbreviations, all 80 indication sections were rewritten to remove these. For example “[Approx. 85-year-old female] ___F with ams transfer for oncology work-up // r/o pna” was re-written to “[Approx. 85-year-old woman] with altered mental status transfer for oncology work-up // rule out pneumonia”. Section 5.4 gives more information on the clinical cases included in the study.

For each X-ray, participants answer one or two questions, depending on whether the AI provides an explanation or not. The main question is always whether the participants agree with the AI, posed as “How much do you agree that the AI model *ABC*’s suggestion of *label* is present?”. The answer can be selected from “Definitely present”, “Maybe present”, or “Not present”. These three answer options were adopted as they are clinically relevant and common in clinical practice: Often, there are signs for a finding with inconclusive evidence, however, it still needs to be reported as follow-ups may be required. The participants are provided the following guidelines on how to interpret the three answer options:

- Not present: The finding cannot be seen and does not need to be highlighted in the radiology report.
- Maybe present: While the evidence is inconclusive and/or there is some ambiguity, it is worth mentioning in the radiology report that the finding may be present.
- Definitely present: The finding is clearly present and has to be noted in the radiology report.

The second question is whether they found the explanation useful in their decision-making, posed as, e.g., “How useful was the AI model *ABC*’s visual explanation in helping you decide

Relevant patient background:

Approx. 65-year-old woman; with shortness of breath

2. Patient context: This is real patient information that was provided by the referring physician.

4. Radiographic finding suggested by the AI model: Only one finding will be highlighted for every X-ray, and it is not necessarily the main finding. *Base your agreement only on this specific finding.*

The AI model Y2P made the following suggestion:

3. AI model serial number: each session has a different AI

atelectasis

The AI model Y2P provides the following **textual explanation** and **visual explanation** for its suggestion:

Bibasilar opacities may represent atelectasis, but aspiration or infection cannot be excluded.

5. AI Explanations: The different AI models in this study can provide different (or no) explanations for their decisions. The explanations can either consist of visual explanations ("heatmaps"), textual explanations, or a combination of both (as in this example).

1. Original Chest X-ray

6. Agreement rating: To what extent do you agree with the AI whether the finding is present in the X-ray? You have the following options:

- **Not present:** The finding cannot be seen and does not need to be highlighted in the radiology report.
- **Maybe present:** While the evidence is inconclusive and/or there is some ambiguity, it is worth mentioning in the radiology report that the finding may be present.
- **Definitely present:** The finding is clearly present and has to be noted in the radiology report.

Here we ask solely you agree with the finding suggested by the AI, not whether you agree with the explanation the AI provides!

How much do you agree that the AI model Y2P's suggestion of atelectasis is present?

☐ Definitely present
☐ Maybe present
☐ Not present

How useful was the AI model Y2P's combination of visual and textual explanations in helping you decide whether the AI was right or wrong in suggesting atelectasis?

☐ Not at all useful ☐ Slightly useful ☐ Somewhat useful ☐ Moderately useful ☐ Useful ☐ Very useful ☐ Extremely useful

Evaluating image: 1/3

[Next Image](#)

7. Explanation Usefulness: If the AI provided an explanation, you will rate how useful it was in deciding whether you agree with the AI. This is not necessarily the same as agreeing with explanation itself (for example: if an explanation helps you to see that the AI suggestion is incorrect).

Click [HERE](#) to rewatch the instruction video

Figure 5.3: The instruction PDF that people have access to throughout the study, which overlays instructions onto the actual study UI.

whether the AI was right or wrong in suggesting *label?*”. Following best practices, the answer options are provided on a 7-point Likert scale going from “Not at all useful” to “Extremely useful” (Finstad, 2010).

5.2.2 Survey & Qualitative Interview questions

Before and after going through the four CDSS conditions, participants must fill out a survey. Besides gathering basic demographic information, I ask them about their experience with the different AI explanations and measure how their attitude towards AI has been affected.

Questions about attitude towards AI: Below in Table 5.1 are the nine statements used to evaluate participants’ attitudes towards AI in terms of trust, ethical concern, and performance expectations. I used multiple statements for similar topics to enhance our measurements’ reliability and mitigate the influence of any single item’s bias. The response is given on a 7-point Likert scale.

Category	Statements
Trust	(1) I’m not comfortable using an AI if I don’t fully understand how it makes a decision. (2) The use of AI should always be accompanied by the option for human review and intervention. (3) I trust AI-based recommendations as much as those from human experts in a clinical setting.
Ethical Concerns	(4) I am not concerned about the ethical implications of using AI in healthcare. (5) Due to the dangers of AI, its adoption should be minimised. (6) The development of AI in healthcare should be tightly regulated.
Performance Expectations	(7) It won’t take long until AI will drastically transform healthcare. (8) AI in its current form is still far from being ready to be used in clinical practice. (9) I believe AI can improve the accuracy of diagnoses in healthcare.

Table 5.1: Statements to evaluate participants’ attitudes towards AI.

Explanation Type Feedback Questionnaire: To capture participants’ subjective feedback on explanation types after the study, they were asked the questions outlined in Table 5.2 for each XAI type (for “No XAI”, they were only asked (1) and (6)).

5.2.3 Pilot Studies

Several pilot studies were conducted to guide the design of our user interface and to get estimates for the effect sizes needed to compute the required sample size. The first two pilots were

Statements
(1) I trusted this AI.
(2) The explanations that were provided for the diagnoses were difficult to understand.
(3) It was transparent to me how the AI came to a diagnosis.
(4) I did not rely on the AI's explanations to decide whether I agree with the diagnosis or not.
(5) I have learned something from the AI's explanations and they helped me become more proficient in reading chest X-rays.
(6) How accurate do you think this AI was (in %)?

Table 5.2: Statements capturing participants' subjective feedback on explanation types.

conducted in video conference sessions with physicians from our personal networks. This allowed us to observe how they engage with the UI and helped us improve several aspects of the study: the way certain instructions are framed, how clear and detailed the instructions are, and the way different aspects of the UI are emphasized. Another insight I got was that agreement with the AI persistently increased over time, i.e., over the course of the study. The interviews conducted after the pilots led us to conclude that this was mainly due to fatigue and decreased motivation, causing increased overreliance on the AI over time. I introduced the monetary incentive explained earlier in this chapter to counteract this. In the actual study, this fatigue effect was no longer observed. Our discussions have also confirmed that participants did not believe they were interacting with the same AI model throughout the different stages. Once I iterated sufficiently on our study design and finalized it, two additional pilot studies were conducted for further verification and to improve our estimates of the effect sizes.

5.2.4 Statistical model

Our study outcomes are modelled using a Generalized Linear Mixed-Effects Model (GLMM) that predicts human accuracy for each case. I chose GLMMs because they offer a flexible and robust framework to handle non-normally distributed outcome variables and account for both fixed and random effects. The model below carefully accounts for our complex study design, including the triple interaction terms (explanation type, correctness, and advice correctness) and missing values by design (no explanation correctness for χ_{None}). I followed best practices presented in Koch et al. (2023). Explanation type is given as χ . The binomial GLMM is given below:

$$\begin{aligned}
l_{ij} = & \beta_0 \\
& + \beta_{AC} C_{AI} \\
& + \beta_{\chi} \chi \\
& + \beta_{\chi \times AC} (\chi \cdot C_{AI}) \\
& + \beta_{\chi \times EC} (\chi \cdot C_{\chi}) \\
& + \beta_{\chi \times EC \times AC} (\chi \cdot C_{\chi} \cdot C_{AI}) \\
& + u_{Participant} \\
& + u_{Image}
\end{aligned} \tag{5.1}$$

This model predicts the log-odds of the human accuracy l_{ij} for the i -th participant on the j -th image. As fixed effects, I considered advice correctness $C_{AI} \in \{0, 1\}$, explanation type $\chi \in \{\chi_{None}, \chi_{SM}, \chi_{NLE}, \chi_{Comb}\}$, explanation correctness $C_{\chi} \in [-3, 3]$ (mean-centred from 7-point Likert scale), and different interactions of these effects. As random effects, I included the participants $u_{Participant}$ (which can have different skill levels) and the images u_{Image} (which can have different difficulty levels). Note that here χ is one-hot encoded, meaning that β_{χ} actually represents multiple coefficients, one for each explanation type. The model fit is shown in Table 5.6.

A rationale for the different interaction terms is given below:

- $\chi \times C_{AI}$: I am assuming that different explanation types have a different impact on human accuracy when advice is correct or incorrect. For example, explanation types prone to confirmation bias will have a particular effect when the advice is incorrect.
- $\chi \times C_{\chi}$: Note that I do not include C_{χ} as a main effect. This is because C_{χ} between different explanation types is not directly comparable (e.g. NLEs contain more specific information and therefore can contain both more correct and more false information). Therefore, I consider C_{χ} as a type-specific metric and need to include the interaction term.
- $\chi \times C_{\chi} \times C_{AI}$: This interaction needs to be modelled as C_{χ} is strongly correlated to C_{AI} . This is because incorrect advice generally has explanations with a lot less correct information, and therefore C_{χ} is much lower when $C_{AI} = 0$.

This model is fitted to our study outcome to interpret the effects and test different hypotheses that align with our research questions. Due to data dependencies, I opted for a mixed model approach to test adjusted means rather than performing inferential statistics on observed means. Model parameters of this three-way full factorial model GLMM are hard to interpret because the model probability l_{ij} is the log-odds of human accuracy, because of the random effects, and because our various interaction terms that make it hard to isolate different factors. For this reason, I do not directly discuss effect sizes and significance values for individual model terms, although they are given in Table 5.6. For example, $\beta_{\chi_{\text{NLE}}}$ does not correspond to the overall effect of the explanation type NLE on human accuracy. However, it corresponds specifically to the difference in human accuracy compared to the baseline (no explanation) when the advice is incorrect ($C_{AI} = 0$), and the explanation correctness is average ($C_{\chi} = 0$). Instead, I used the model to predict human accuracies on our observations and test differences via contrasts. Most results in this chapter, such as in Figure 5.19, are based on hypothesis testing using this model.

I tested the model statistically and found that both random and fixed effects should be included. A likelihood ratio test (LRT) between the model in Equation (5.1) and a baseline model disregarding explanation correctness and interactions shows that the full model yields a significantly better fit $\chi^2_{12} = 28.21$, $p = .005$. The study design strongly suggests the inclusion of random effects u_{Image} and $u_{\text{Participant}}$. For the model (5.1), I find $\text{Var}(u_P) = 0.059$ and $\text{Var}(u_I) = 0.295$, showing the variance between participants and images. The LRT comparing our model with and without random effects is significant, $\chi^2_2 = 144.43$, $p < .0001$. Including each random effect individually also yields significant improvements.

Hypothesis Testing It can be challenging to interpret the parameters of a GLMM directly; they are connected to the modelled probability via the non-linear inverse-logit link, there are random effects, and, in our case, multiple interactions. As such, to answer our research questions, I do not attempt to interpret the parameters directly but instead rely on model predictions and the distribution of the dependent variable across different experimental conditions.

To evaluate the effect of explanation type on human accuracy, I used the *marginalEffects* package in R (Arel-Bundock et al., 2024), which provides a framework for extracting and interpreting contrasts from statistical models. Specifically, I employed a set of pairwise contrasts to assess the differences between different types of explanations in how they influence human decision accuracy.

For the research questions, I formulated contrasts to explore the effect of explanation type under different conditions. I posed them as hypothesis tests. For example, in order to verify whether, overall, NLEs have a different effect on human accuracy than saliency maps, we can verify the hypothesis $H : \beta_{\chi_{\text{NLE}}} - \beta_{\chi_{\text{SM}}} = 0$. We cannot use simpler approaches like paired t-tests for this because the residuals for t-tests must be independent and identically distributed. Since multiple participants evaluate the same image, the individual measurements are not independent. They are only independent given the image, making the image a random factor. The same applies to a participant who evaluates multiple images. This means that the independence of residuals assumption is violated.

I consider the following **main research questions**:

1. *RQ1*: Effect of Explanation Type on Human Accuracy.

The first set of contrasts was designed to evaluate the overall effect of explanation type on human accuracy. This analysis allowed me to determine whether the different explanation types (none, saliency maps (SM), natural language explanations (NLE), and a combination of both (Comb)) lead to statistically significant changes in human accuracy compared to each other. The following contrasts were tested:

$$H_1 : \beta_{\chi_i} - \beta_{\chi_j} = 0 \quad \forall \chi_i, \chi_j \in \chi_{\text{None}}, \chi_{\text{SM}}, \chi_{\text{NLE}}, \chi_{\text{Comb}}; i \neq j \quad (5.2)$$

2. *RQ2*: Effect of Explanation Type on Human Accuracy, Given Advice Correctness.

To evaluate whether the effect of explanation type on human accuracy varies depending on advice correctness (C_{AI}), I tested contrasts for both correct and incorrect advice conditions:

$$H_2 : \beta_{\chi_i|C_{AI}=k} - \beta_{\chi_j|C_{AI}=k} = 0 \quad i \neq j, k \in \{0, 1\} \quad (5.3)$$

These contrasts were calculated separately for conditions where advice was either correct ($C_{AI} = 1$) or incorrect ($C_{AI} = 0$), allowing me to determine if the accuracy of the provided advice influences explanation effectiveness.

3. *RQ3*: Effect of Explanation Type on Human Accuracy, Given Advice Correctness and Explanation Correctness Alignment.

Finally, I formulated contrasts to assess the impact of explanation type when considering both advice correctness and the level of explanation correctness. Specifically, explanation

correctness was categorized as high or low, where high corresponds to being in the top 50th percentile for that explanation type and low corresponds to being in the bottom 50th percentile. The generalized contrast for RQ3 can be represented as:

$$\begin{aligned}
 H_3 : \quad & \beta_{\chi_i | C_{AI}=k, C_\chi \in q} - \beta_{\chi_j | C_{AI}=k, C_\chi \in q} = 0 \\
 & \forall \chi_i, \chi_j \in \{\chi_{\text{None}}, \chi_{\text{SM}}, \chi_{\text{NLE}}, \chi_{\text{Comb}}\}, \\
 & i \neq j, \quad k \in \{0, 1\}, \quad q \in \{\text{high}, \text{low}\}
 \end{aligned} \tag{5.4}$$

These contrasts allowed me to investigate how explanation type affects human accuracy when considering both the correctness of advice and the level of explanation correctness.

If they are present, provide the locations of the radiographic findings in the chest X-ray below. *

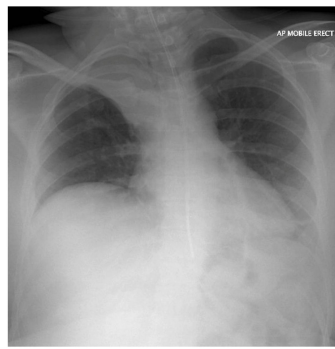
[FYI: The right side on the image is the left side of the patient]



	Not present	Right lower L.	Right middle...	Right upper L.	Left lower lo...	Left upper lo...
Pneumonia	☐	☒	☐	☐	☐	☐
Lung collapse	☐	☐	☐	☐	☐	☐
Pleural Effus...	☐	☐	☐	☐	☐	☐

If they are present, provide the locations of the radiographic findings in the chest X-ray below. *

[FYI: The right side on the image is the left side of the patient]



	Not present	Right lower L.	Right middle...	Right upper L.	Left lower lo...	Left upper lo...
Pneumonia	☐	☐	☐	☒	☐	☐
Lung collapse	☐	☐	☐	☒	☐	☐
Pleural Effus...	☐	☐	☐	☐	☐	☐

If they are present, provide the locations of the radiographic findings in the chest X-ray below. *

[FYI: The right side on the image is the left side of the patient]



	Not present	Right lower L.	Right middle...	Left lower lo...	Right upper L.	Left upper lo...
Pneumonia	☐	☒	☐	☐	☐	☐
Lung collapse	☐	☐	☐	☐	☐	☐
Pleural Effus...	☐	☒	☐	☐	☐	☐

Figure 5.4: The three test cases included in the screening survey⁴. These X-rays contain examples of pneumonia, pleural effusion, and lobe collapse (atelectasis), which are the most common classes in the dataset. Participants can select the anatomical locations for each class and each case.

5.3 Participant Selection and Outreach

5.3.1 Target Candidate & Sample Size

As I aim to study the effect of different explanation types in an imperfect (X)AI setting, recruited participants need to have a foundational competence in reading chest X-rays, i.e., be knowledgeable enough not to rely wholly on the AI system but still be likely to engage with the AI's predictions and explanations. [Kim et al. \(2022\)](#) argue that for these studies, it is important that participants rely on the model to a certain extent and not just on their prior knowledge. Indeed, [Gaube et al. \(2021\)](#) found that increasing expertise in radiology leads to an increased likelihood of dismissing AI suggestions. Furthermore, CDSSs are generally seen as most useful for people who have medical training but are not experts in the task at hand ([Bussone et al., 2015](#)). This is particularly relevant in scenarios where there is a scarcity of expert radiologists, and non-expert clinicians benefit from collaborating with AI systems ([Gaube et al., 2023](#)).

For the above reasons, the primary target group for this study are medical students and doctors who have undergone training in reading chest X-rays but who are not specialist radiologists. This includes radiologists in training. Participants' radiology proficiency is validated via a screening form, which contains a self-assessment form as well as a quiz on three chest X-rays that fulfil the medical student curriculum of the Royal College of Radiologists (UK) (exam-

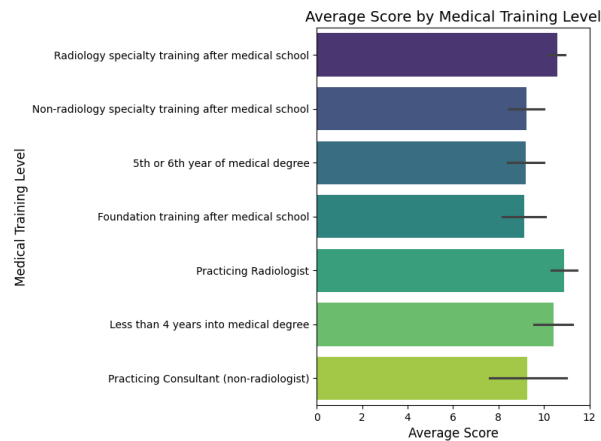


Figure 5.5: The average scores obtained on the pre-screening survey based on radiology training level.

ples are shown in Figure 5.4). The screening form also contains questions about their level of medical training (options are “I have no formal medical training”, “Less than 4 years into medical degree”, “5th or 6th year of medical degree”, “Foundation training after medical school”, “Non-radiology speciality training after medical school”, “Radiology speciality training after medical school”, “Practicing Consultant (non-radiologist)”, or “Practicing Radiologist”) and their geographic backgrounds (“In which country have you spent the most time studying and practising medicine?”).

I compute the score as follows:

- For each finding 1 point is given if the participant has either a true positive or a true negative.
- An additional point is given if they give the right location. As only a single finding is correct for each image, the maximum points per image is thus 4 points.
- If a user provides additional locations for a finding where they already got a point for the right location, I subtract -1 if they listed another location that is on the wrong side or -0.5 for another location that is at the wrong height.

Based on these results, applicants with a score below 8 (which corresponds to getting at least two of the three images perfectly right) were filtered out. Practicing radiologists are also excluded.

Estimating required sample size An important consideration when running a human subjects study is ensuring that the study has high statistical power. Statistical power is the probability that an effect is detected with a significant p -value when present in the underlying data. It is typically desired to have at least 80% power. This would mean that if there is indeed a different effect of explanation types on human performance, we would have an 80% likelihood of detecting this effect with statistical significance. Simulating statistical power is, therefore, a useful tool to estimate the desired user sample size before launching a study.

The key requirement to simulating power is having well-informed estimates of the true effects. After simulating the experiment N times (given the stochasticity of the process) we can measure how often the resulting data distribution leads to effects detectable with statistical significance. This means that, during the power analysis, the entire analysis framework (or a good approximation of it) already needs to be implemented. This helps ensure that the study and statistical models are thoroughly verified before the actual study is launched.

P_{ij} Condition			Condition	Other Evidence and Assumptions	Param.	Coefficients		
Weak	Pilot	Strong				Weak	Pilot	Strong
40%	13%	13%	$\chi = \chi_{\text{None}}, C_{AI} = 0$	38% in our initial pilots, 44% in a not-yet-published head CT study by Gaube et al. The 12.5% obtained in our two pilots seems overly strong.	β_0	-0.41	-1.95	-1.95
80%	83%	80%	$\chi = \chi_{\text{None}}, C_{AI} = 1$	78% in the first pilots; 83% in chest X-ray study; 79% in head CT study.	β_{AC}	1.79	3.53	3.33
48%	34%	38%	$\chi = \chi_{\text{SAL}}, C_{AI} = 0, C_{\chi} = 0$	86% in chest X-ray study for correct advice with correct visual explanations.	$\beta_{\chi_{\text{SAL}}}$	0.33	1.27	1.44
75%	75%	83%	$\chi = \chi_{\text{SAL}}, C_{AI} = 1, C_{\chi} = 0$		$\beta_{\chi_{\text{SAL}} \times AC}$	-0.61	-1.76	-1.24
48%	34%	38%	$\chi = \chi_{\text{NLE}}, C_{AI} = 0, C_{\chi} = 0$	Pilots showed the same effect as for saliency when AI advice is incorrect.	$\beta_{\chi_{\text{NLE}}}$	0.33	1.27	1.44
84%	90%	95%	$\chi = \chi_{\text{NLE}}, C_{AI} = 1, C_{\chi} = 0$		$\beta_{\chi_{\text{NLE}} \times AC}$	-0.05	-0.66	0.12
45%	28%	33%	$\chi = \chi_{\text{COMB}}, C_{AI} = 0, C_{\chi} = 0$	Pilots showed the same effect as for NLE when AI advice is correct.	$\beta_{\chi_{\text{COMB}}}$	0.21	1.01	1.22
84%	90%	95%	$\chi = \chi_{\text{COMB}}, C_{AI} = 1, C_{\chi} = 0$		$\beta_{\chi_{\text{COMB}} \times AC}$	0.07	-0.39	0.34
50%	55%	60%	$\chi = \chi_{\text{SAL}}, C_{AI} = 0, C_{\chi} = -3$		$\beta_{\chi_{\text{SAL}} \times EC}$	-0.03	-0.29	-0.31
53%	60%	65%	$\chi = \chi_{\text{NLE}}, C_{AI} = 0, C_{\chi} = -3$		$\beta_{\chi_{\text{NLE}} \times EC}$	-0.07	-0.36	-0.38
46%	30%	55%	$\chi = \chi_{\text{COMB}}, C_{AI} = 0, C_{\chi} = -3$		$\beta_{\chi_{\text{COMB}} \times EC}$	-0.01	-0.03	-0.31
87%	95%	95%	$\chi = \chi_{\text{NLE}}, C_{AI} = 1, C_{\chi} = +3$		$\beta_{AC \times EC}$	0.15	0.61	0.38

Table 5.3: Assumptions and estimates of the distribution parameters used for our power simulation. P_{ij} | Condition refers to the estimated human accuracy under the given condition. The middle column uses the actual results obtained in our pilot studies ($n=2$), whereas the *weak* and *strong* assume a generally weaker or stronger effect of our conditions of interest (explanation type, explanation correctness, etc.). Rationales for these estimates are provided based on existing work and earlier pilot studies. The order (from top to bottom) follows the order in which I solved equation 5.5 to extract the coefficient values given in the last three rows. When available, I relied on [Gaube et al. \(2023\)](#) or a not-yet-published study from the same author to get the estimates. $C_{\chi} = 0$ corresponds to average explanation correctness.

In order to estimate the true distribution of our data, we need to make certain assumptions about human accuracy in different scenarios and then solve for the parameters in equation 5.1.

For instance, let us consider the estimated human accuracy for incorrect AI advice with no explanation. For these conditions ($C_{AI} = 0, C_{\chi} = 0, \chi = \chi_{\text{None}}$), and assuming a participant with average skill and an image with average difficulty (so random effects are 0), equation 5.1 simplifies to $l_{ij} = \beta_0$. By estimating the probability P_{ij} (where $l_{ij} = \text{logit}(P_{ij})$) that our average participant for an image of average difficulty will be correct, we can thus solve for the intercept β_0 . Once we have β_0 , we can then come up with additional scenarios to estimate the human accuracy given some conditions to solve for the remaining parameters. Table 5.3 shows how I considered different scenarios in a specific order, which enabled us to solve for all the parameters in the simplified statistical model given in 5.5. By mean-centring C_{χ} , I ensured that there was a condition where $C_{\chi} = 0$ (average explanation correctness), which was required for solving all the parameters. For the random effects, I assume that human expertise and image difficulty level follow a normal distribution where the 95% confidence interval is between 0.2 and 0.9 human accuracy for images and between 0.4 and 0.9 human accuracy for participants. This results in estimated coefficients $\bar{u}_{\text{Participant}} = 0.128$ and $\bar{u}_{\text{Image}} = 0.179$.

I use a simplified model here without the three-way interaction so that we can solve for all the parameters. The model is given by:

$$\begin{aligned}
 l_{ij} = & \beta_0 \\
 & + \beta_{AC}C_{AI} \\
 & + \beta_{\chi}\chi \\
 & + \beta_{\chi \times AC}(\chi \cdot C_{AI}) \\
 & + \beta_{\chi \times EC}(\chi \cdot C_{\chi}) \\
 & + u_{\text{Participant}} \\
 & + u_{\text{Image}}
 \end{aligned} \tag{5.5}$$

Based on the *weak*, *pilot*, and *condition* estimated parameters, I ran the power analysis for which the results are shown in Figure 5.6. The power is given by number of participants, share of incorrect AI advice, and effect strength. I found that 80 participants should provide significant power. I ended up recruiting 85 participants, as I sent out extra invitations to account for dropouts.

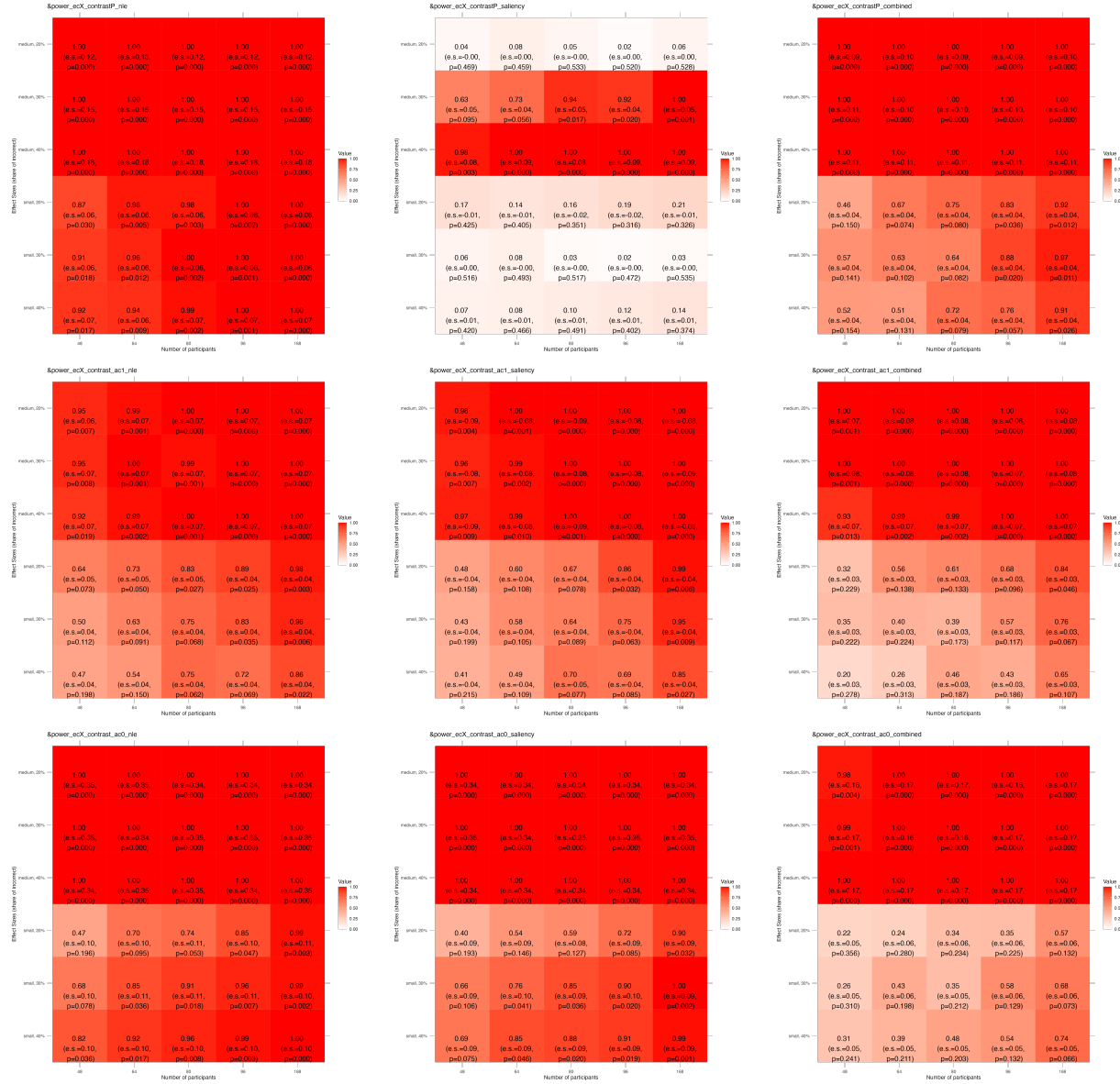


Figure 5.6: The heatmaps used to explore the power of different set-ups. In this figure, I consider number of participants (x-axis), assumed effect sizes (medium is pilot, small is weak, y-axis), and share of incorrect predictions of the model (y-axis). Inside the cells I print power as well as average effect size (e.s.) and p-value (p). I ran $N = 3000$ simulations. Row 1 shows the overall effect of the three explanation types, whereas rows 1 and 2 show correct and incorrect advice, respectively. Based on this analysis, I targeted a participant pool of 80 people to achieve overall satisfactory power.

5.3.2 Recruitment Drive

Participants were recruited via mailing lists and personal networks focusing mainly on the United Kingdom and Romania. Initial outreach consisted of emails and social media posts with a flyer (see Figure 5.7) containing a screening survey link. Participants were compensated for their time with a voucher worth an equivalent of £30 for the one-hour study, which was advertised. In total, 223 people filled out our form, including the three test cases. The participants range from medicine students to radiology residents. Once selected for the study, I followed up with participants and provided them with the study login details (see the email sent out in Figure 5.8).

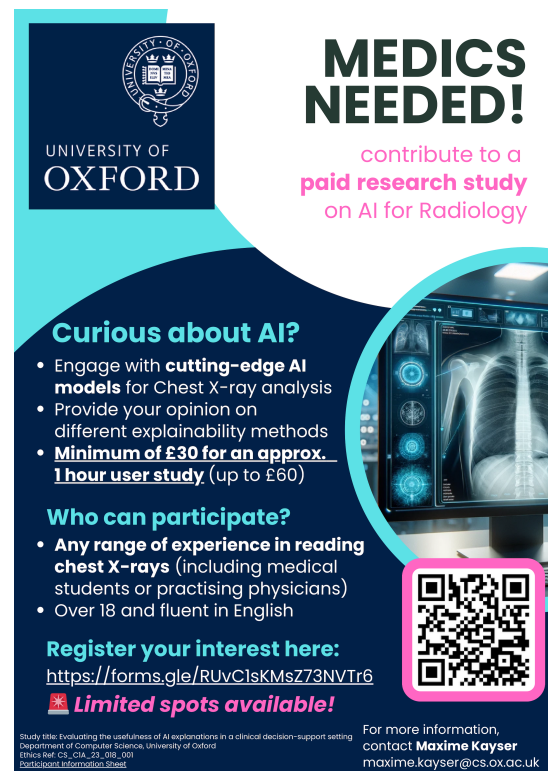


Figure 5.7: The recruitment flyer used to reach out to participants to fill out the screening form.

Dear Study Participant,

This is a friendly reminder to complete the **AI 4 Radiology** online task. We look forward to your contribution to the study.

Below are the login details to complete the online task, which is meant to take around 1 hour.

Task website: <https://f0ak.short.gy/ai4radiology>

Username: <the email address with which you received this mail>

Password: YlOupqi03I84IV1tVoNT

To compensate you for your time and effort, you will receive a voucher worth £30 (can go up to £60), a certificate of participation, and optionally an acknowledgement in a subsequent publication. For UK-based participants, we also issue CPD points.

If you want more information about the study, please refer to the Participant Information Sheet or contact me.

Thank you and happy X-ray reading!

Best wishes,

Maxime Kayser

PhD student at the University of Oxford

Figure 5.8: Email sent to invited participants.

5.3.3 Participant Demographics

The application form collected basic demographic information regarding geography, age, gender, and medical training (displayed in Figure 5.10) as well as self-assessed radiology expertise and experience (see left of Figure 5.11). During the pre-task survey, I also collected basic information on the participants' AI expertise and experience levels. Table 5.4 shows the statements used for participants to express their agreement on a 7-point Likert scale from “Strongly Disagree” to “Strongly Agree”. Results for this are shown on the right of Figure 5.11 and in Figure 5.9. Figure 5.12 shows that experience and attitudes towards AI did not significantly affect task performance of the participants.

Statements
(1) I understand the principles behind computer vision models (i.e., AI algorithms used for analyzing images) and how they work.
(2) I am familiar with language models (i.e. AI algorithms used to understand and generate language) and how they work.
(3) I understand the concepts of explainable AI (XAI), i.e., methods that try to make AI algorithms' decision-making more transparent (for example: heatmaps).
(4) I regularly use AI-powered chat tools (e.g. ChatGPT).
(5) I regularly interact with methods that make AI algorithms' decision-making more transparent.
(6) I regularly use AI-based decision-support tools for medical imaging.

Table 5.4: Statements capturing participants' self-assessed AI expertise.

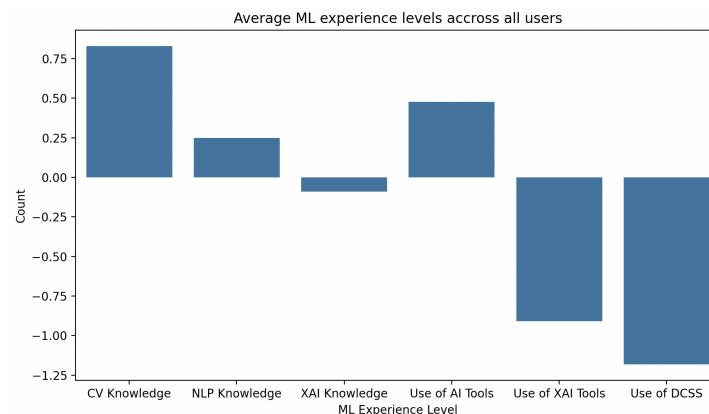


Figure 5.9: Self-assessed knowledge and frequency of use for different AI technologies (-3 to +3 Likert scale). “CV” stands for computer vision, “NLP” for natural language processing.

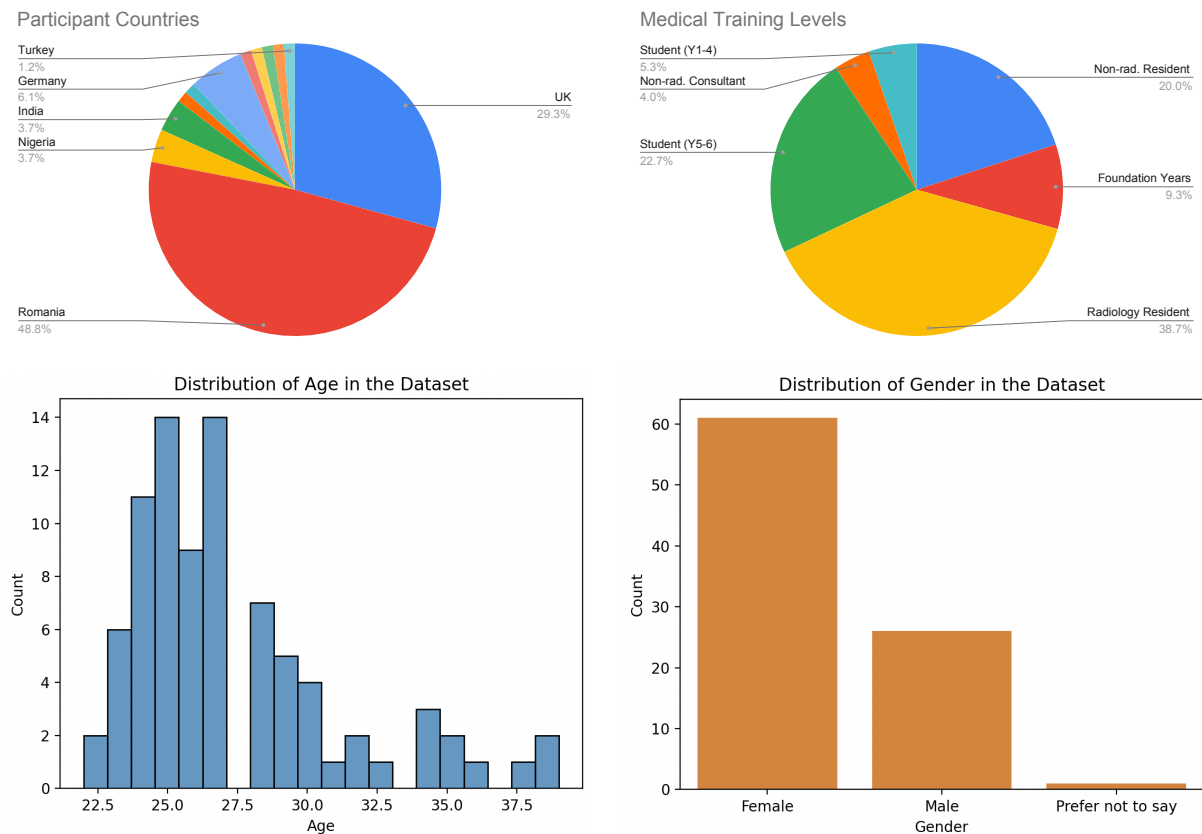


Figure 5.10: Participant demographics. *Top-left*: Countries where participants have spent the most time “studying or practising” medicine. *Top-right*: Medical Training Level of Participants. YN is the year of medical school. Foundation years refer to the general training right after medical school (two years in the UK). *Bottom-left*: Our participants are relatively young, which is unsurprising as non-specialist clinicians were targeted. *Bottom-right*: Distribution of participant genders.

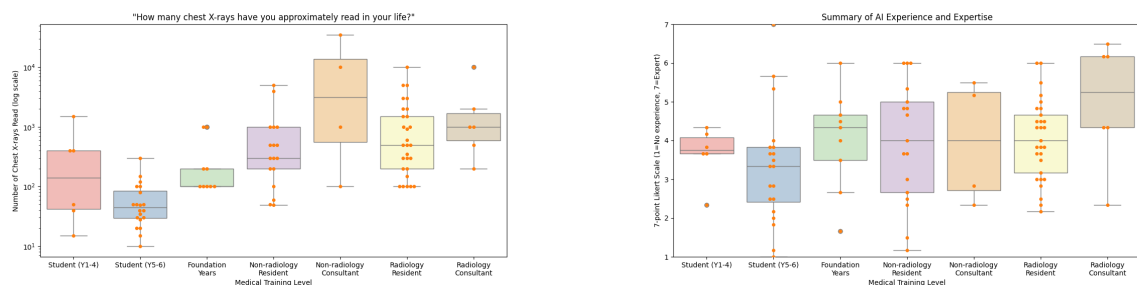


Figure 5.11: *Left*: Self-assessed estimates of how many X-rays the participants have read in their life. *Right*: Self-assessed levels of experience and expertise in AI (summarized across computer vision, NLP, explainable AI, and clinical decision-support systems) for different medical training levels. The plot shows box plots and all individual data points (orange). YN is the year of medical school. Foundation years refer to the general training right after medical school (two years in the UK). These figures include radiology consultants, although these were not included in the actual study.

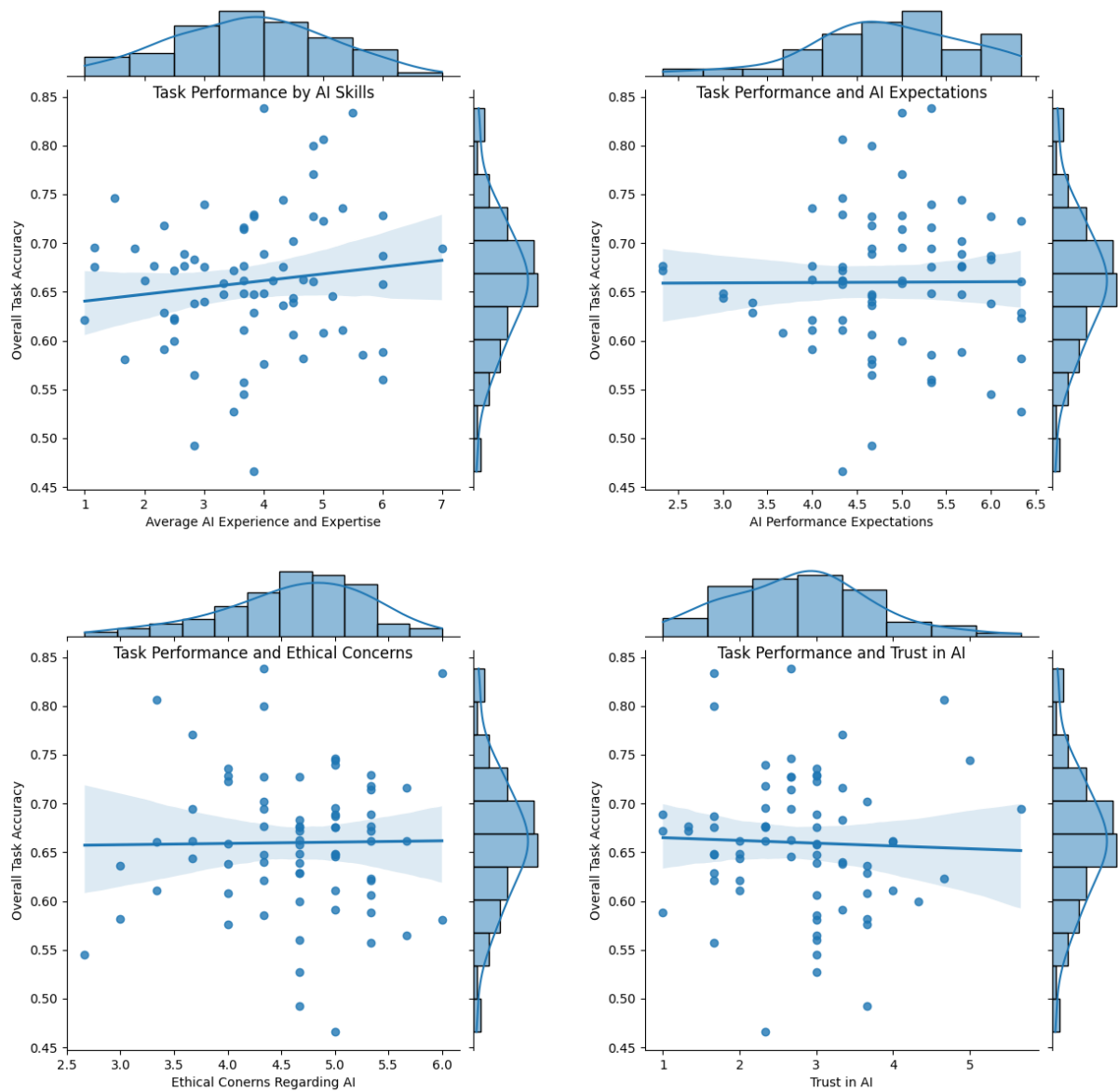


Figure 5.12: *Top-left*: A participant's AI experience and understanding compared to their diagnostic accuracy across all tasks. *Top-right*: A participant's expectation of AI compared to their diagnostic accuracy across all tasks. *Bottom-left*: Participant's level of ethical concerns regarding AI compared to their diagnostic accuracy across all tasks. *Bottom-right*: A participant's trust in AI compared to their diagnostic accuracy across all tasks.

5.4 Gathering Data for Study

5.4.1 AI Model Pipeline

In the eyes of the study participants, they are presented with four different AI models throughout the study. In reality, to ensure comparability, the backbone vision classifier is the same for all models. I trained a transformer-based vision-language model (VLM) based on the Ratchet architecture (Hou et al., 2021), as described in Section 4.3.2, on MIMIC-NLE-v2. The model obtained a weighted AUC of 0.75. Note that the main purpose was not to maximize model performance. Instead, I specifically focus on the case of imperfect AI, where a model, for various reasons, such as limited or biased data, does not perform optimally. This differs from existing work in human-AI collaboration, where they often consider AI models that outperform humans to investigate how they could improve human performance (Tschandl et al., 2020; Fogliato et al., 2022). Nonetheless, the model still performs favourably on existing benchmarks, ensuring that our model and the generated explanations are of a realistic standard (Irvin et al., 2019).

This VLM then yields the four models introduced in our four conditions. X_{None} consists only of the backbone Densenet-121 (Huang et al., 2017), X_{SM} consists of the backbone Densenet and saliency maps extracted from this backbone, X_{NLE} consists of the entire VLM, without saliency maps, and X_{Comb} adds saliency maps to the VLM.

The NLEs that the model generates are learned in a purely supervised way. They, therefore, capture the nuances around assertiveness and the certainty of findings that naturally occur in clinical practice. For this reason, I consider assertiveness an integral part of the NLEs, as opposed to a design factor that can be studied by itself (Calisto et al., 2023).

The models perform multi-label classification, assigning a single logit to each class. In order to get “positive” predictions from the model, I established class-specific thresholds by maximizing the Youden Index, optimizing the balance between sensitivity and specificity.

Saliency maps are obtained using Grad-Cam (Selvaraju et al., 2017) following Goldenblat and contributors (2021). Grad-CAM was chosen as it is widely used, and previous work has shown that, out of the commonly used saliency techniques, it is the most accurate one for medical imaging (Saporta et al., 2022). They find that for Densenet-121, out of Grad-CAM,

Grad-CAM++, integrated gradients, Eigen-CAM, DeepLIFT, layer-wise relevance propagation (LRP), and “occlusion”, Grad-CAM localized pathologies the best. I have also qualitatively verified this by comparing it to Grad-Cam++, HiResCam, AblationCAM, and XGradCAM on our images. The usefulness and adequacy of saliency maps have previously been called into question ([Adebayo et al., 2018](#)). To address this, I used the most widely established type of saliency maps and additionally obtained expert annotations of their *correctness* from three radiologists (as detailed in the following section). This approach allows us to disentangle the overall effect of saliency maps from the effect of their individual quality.

5.4.2 Data Annotation & Selection

Upon consultations with radiologists, I selected the following subset of labels based on their clinical significance and detectability in chest X-rays alone: “pneumonia”, “atelectasis”, “pulmonary edema”, “fluid overload/heart failure”, “aspiration”, and “alveolar hemorrhage”. Classes that were removed based on these criteria are: “copd/emphysema”, “granulomatous disease”, “pleural effusion”, and “pleural/parenchymal scarring”. I could have influenced what kind of examples I sourced for the study by sampling based on their logit values. Higher logits could indicate higher model confidence, which could suggest easier examples, higher quality saliency maps ([Saporta et al., 2022](#)), and potentially better NLEs; lower logits could suggest harder or more ambiguous examples. However, as these relationships are not sufficiently understood, I refrained from intervening in the study in such a manner and instead sampled randomly from all the examples where the logits are above the threshold.

Even though the chest X-rays are paired with human-written radiology reports, I followed existing work ([Gaube et al., 2023](#); [Ahn et al., 2022](#); [Seah et al., 2021](#)) and had three experienced radiologists annotate the correctness of the AI advice and explanations. This has several advantages. First, there is typically low inter-annotator agreement in chest X-ray evaluation (as discussed in Chapter 4), and, therefore, having multiple experts reduces ambiguity. Second, since the X-rays used are from ICU patients, routine findings may be omitted from reports; thus, it is unclear whether an unmentioned finding is absent or simply not reported. Third, it ensures that the annotations are obtained under the same conditions as our study: limited to a single view and without comparison to previous chest X-rays (having only the indication subsection as patient context). Finally, the annotation process is crucial to obtain quantitative measurements of the explanation correctness.

To ensure that I will ultimately obtain a set of 80 unique images (four tasks with 20 images each) with AI advice at 70% accuracy (56 correct, 24 incorrect), a larger subset of 160 examples was annotated. This initial subset was carefully chosen to reflect a class distribution similar to that of the overall dataset.

When evaluating the AI advice, annotators are presented with a chest X-ray and a single class predicted by the AI (e.g. “pneumonia”). They are then asked whether they think the class is “Not present” (the finding can not be seen, so it is not worth mentioning, or it can be mentioned negatively. For example: “No signs of pneumonia.”), “Maybe present” (while the evidence is inconclusive and/or there is some ambiguity, it is worth mentioning in the radiology report that the finding may be present. For example: “Bibasilar opacities may represent atelectasis or pneumonia.”), or “Definitely present” (the finding is clearly present and will be noted in the radiology report. For example: “There are clear signs for pneumonia.”), following a common convention in evaluating the presence of chest X-ray findings (Irvin et al., 2019). They are given the same instructions on how to interpret these three labels as the study participants. An illustration of the annotator interface can be found in Figure 5.13.

The annotators also evaluated the textual explanation and saliency map for each prediction. Given that explanations can vary significantly in information richness, I argue that a continuous scale is better suited than a binary correctness label (Rivera-Garrido et al., 2022), as opposed to Morrison et al. (2024). Suppose our annotators deem the AI advice (e.g. “pneumonia”) correct (“Definitely present” or “Maybe present”). In that case, they are asked “How correctly does the NLE (or heatmap) explain the AI advice pneumonia in this image?” with their response recorded on a 7-point Likert scale. I also asked them “If you consider the heatmap and the NLE as a joint explanation, how correctly do they explain the AI advice pneumonia in this image?” to obtain a correctness score for the combined explanation. In case they think the AI prediction is incorrect, I still want to get a measure of how much correct information an explanation contains and ask them the following: “How correctly does the (e.g.) heatmap highlight radiographic findings that would be relevant for the AI advice *pneumonia* in this image?”. Figure 5.14 shows the distribution of explanation correctness scores C_χ . As can be seen, saliency maps are generally ranked higher than NLEs.

The consensus annotations were obtained by selecting *advice correctness* C_{AI} via majority vote and *explanation correctness* C_χ via averaging. I mean-centred C_χ for each type of explanation

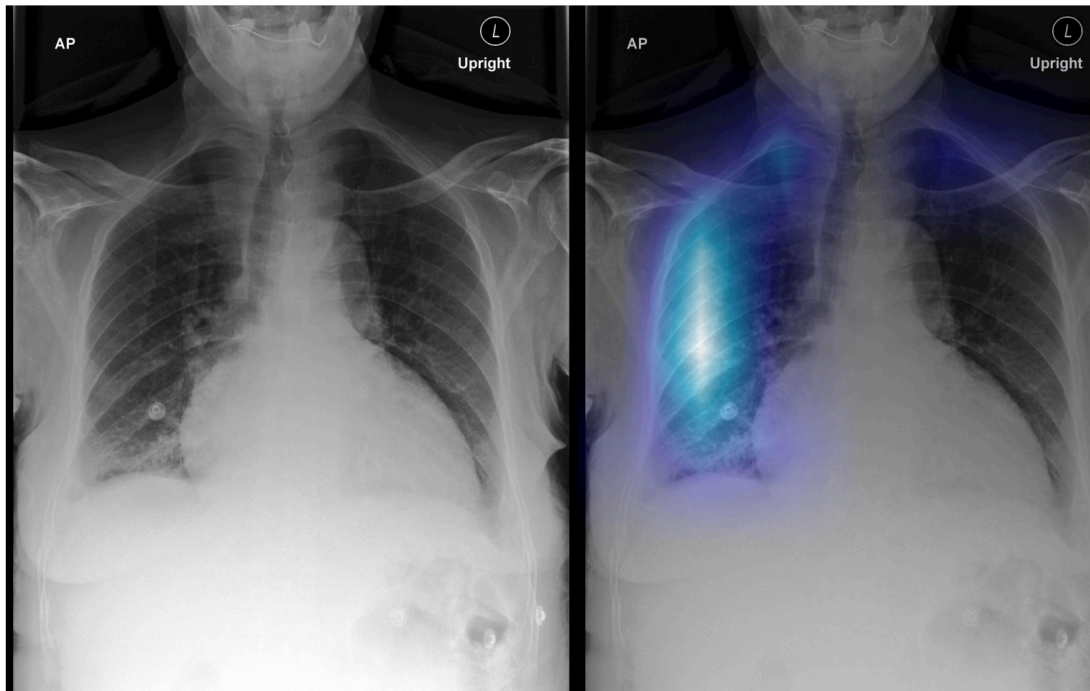
Patient context:

[age: 60-70, gender: M] Altered mental status, fall.

The AI model made the following suggestion:

pneumonia

The AI provides the following explanations for its suggestion:



Right basilar opacity may reflect atelectasis, but aspiration or infection cannot be excluded.

The diagnosis pneumonia is:

- ☐ Not present
☐ Maybe present
☐ Definitely present

How correctly does the NLE explain pneumonia in this image?

1

2

3

4

5

6

7

How correctly does the heatmap explain pneumonia in this image?

1

2

3

4

5

6

7

If you consider the heatmap and the NLE as a joint explanation, how correctly do they explain pneumonia in this image?

1

2

3

4

5

6

7

Next Image

Evaluating image: 6/161

Figure 5.13: The UI used by our radiologists to annotate the study examples.

to facilitate our statistical modelling. Mean-centering also ensures that we look at explanation correctness as a relative, type-specific metric where the absolute values cannot be compared across types. The information content and specificity between, e.g., saliency maps and NLEs are very different; therefore, an absolute comparison of these scores is difficult.

Our samples thus consist of real AI predictions, including correct ones and failures, accompanied by their actual explanations. This is in contrast to existing work such as [Kim et al. \(2022\)](#), where they obtain explanations for wrong predictions by making the model explain a decision it would not have made itself.

Selecting 80 representative cases for the study Now that there are 160 images with evaluated AI advice and explanations, the subset of 80 to be included in the study was selected based on several criteria. First, I removed highly ambiguous cases, which manifest as substantial disagreement among the three annotators. Sixteen examples that contain both an annotation of “Not present” and “Definitely present” were removed. As single-view chest X-rays do not always provide sufficient evidence for certain findings, this ensures that I remove cases that are too difficult to detect. Next, I wanted to ensure that the share of false positives for each class is as close to 70% as possible. This is to prevent any class bias, such as having *all* pneumonia cases labelled as incorrect or *all* atelectasis cases labelled as correct. Finally, within the existing constraints, I sampled examples such that the distribution of explanation correctness C_x scores is as uniform as possible. This ensures a representative sample across the spectrum of explanation quality scores. The final distributions, including mean-centering, are shown in Figure 5.15. As expected, C_x for positive predictions is much higher than for negative predictions.

For our selected sample, I obtained pairwise kappa scores of 0.451, 0.458, and 0.502 between the three annotators when grouping “Maybe present” and “Definitely present” as positive. These kappa scores correspond to “moderate” agreement. Note that if I leave out “Maybe present” votes, I get perfect kappa scores because of the ambiguity exclusion criteria.

After selecting the 80 images, I distribute them across tasks and participants. Every participant interacts with all 80 unique images through four tasks. However, as every explanation type was annotated for each image-advice pair, the same images were distributed to different tasks (explanation types) for different participants, further ensuring robustness. I distributed images across user-task pairs as randomly as possible whilst ensuring the following two conditions:

(i) each image must appear equally often in each task (this prevents the images from biasing interactions within different explanation types), and (ii) for each explanation type a participant reviews, the proportion of correct AI advice is maintained at 70% (this guards against bias that could arise if some explanation types were associated with a higher or lower likelihood of correct AI predictions).

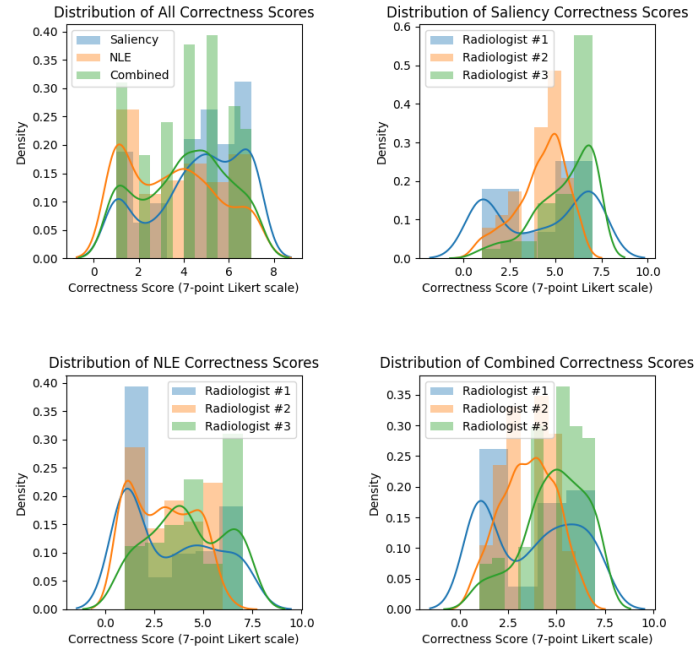


Figure 5.14: Distribution of C_χ scores assigned to different expl. types by our annotators.

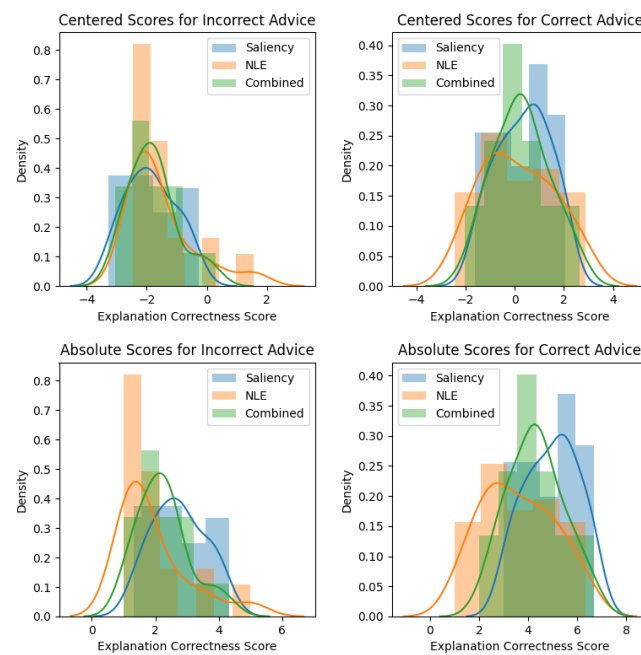


Figure 5.15: An illustration of the distribution of C_χ scores included in the study.

5.5 Study Outcome

In this section, I will discuss the results of the study. The result section is separated into three parts: the qualitative results from the survey (Section 5.5.1), the analysis for the main research questions (Section 5.5.2), and the additional analysis listed as explorative in our pre-registration (Section 5.5.3).

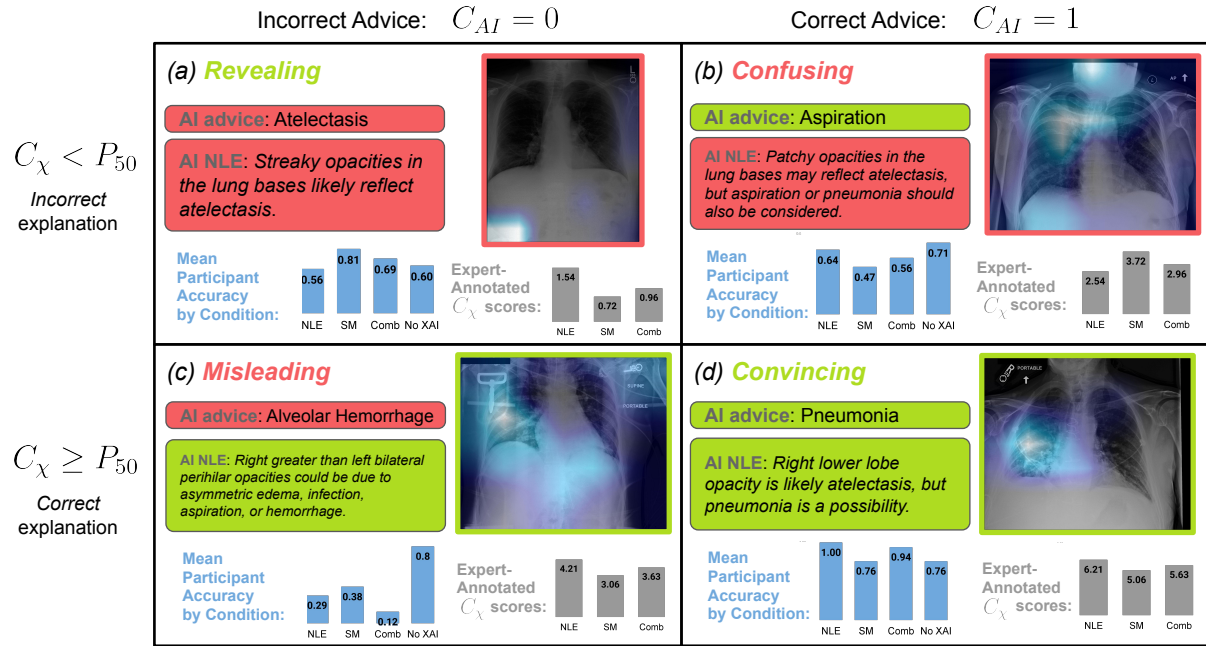


Figure 5.16: This Figure depicts a representative example for each of the four quadrants of advice and explanation correctness interaction. The provided mean participant accuracy and C_{χ} scores are specific to the shown example. P_{50} denotes the 50-th percentile. (a) *Revealing* ($C_{AI} = 0$, low C_{χ}): The AI incorrectly suggests atelectasis, but the poorly rated explanations help clinicians identify the error, leading to higher accuracy compared to relying on the AI prediction alone. (b) *Confusing* ($C_{AI} = 1$, low C_{χ}): The AI correctly identifies aspiration but provides low C_{χ} explanations, leading to lower diagnostic accuracy compared to the No XAI setting. (c) *Misleading* ($C_{AI} = 0$, high C_{χ}): The AI incorrectly suggests alveolar hemorrhage but provides highly rated explanations, misleading participants to agree with the incorrect AI when explanations are provided. (d) *Convincing* ($C_{AI} = 1$, high C_{χ}): The AI correctly identifies pneumonia and provides highly rated explanations, resulting in high diagnostic accuracy, especially for NLEs.

5.5.1 Survey Results

All study participants filled out a questionnaire both before and after the study. The survey has three aims: (1) to understand how participant attributes are associated with how they interact with the AI, (2) to see how their views on AI/XAI change after interacting with the different

models, and (3) to get some qualitative insights on how participants perceive the different explanation types. I gauge their experience, expertise, and expectations regarding radiology in general, AI, and XAI. I followed best practices when creating these questions by (a) giving statements that can be answered via a Likert scale (instead of questions) and (b) providing multiple similar statements to capture the information fully (e.g. “I believe AI can improve the accuracy of diagnoses in healthcare” and “It won’t take long until AI drastically transforms healthcare.”). All the questions are provided in Section 5.2.2. At the end of the study, I also gave the option to provide free-form text as feedback on the different explanation types and the study as a whole. Overall, there was praise for the study design (“Really good study, easy to use format and easy to answer questions”, “Very good study - and very well designed, thank you”, “very nice study”).

Table 5.5: Preference ranking of models.

	μ Rank	#1	#2	#3	#4
NLE	1.85	38.9%	38.9%	20.0%	2.21%
Comb.	2.05	40.0%	23.3%	27.8%	8.90%
No XAI	2.98	14.4%	21.1%	16.7%	47.8%
SM	3.11	6.72%	16.7%	35.6%	41.1%

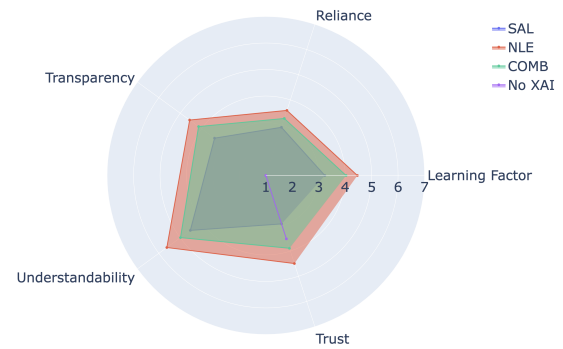


Figure 5.17: Five attributes of explainability methods, ranked on a 7-point Likert scale by the participants.

Table 5.5 shows that the survey revealed a strong preference for NLEs, with saliency maps liked the least. Participants, who did not know that all models have the same share of incorrect AI predictions, also perceived the model with saliency maps to be, on average, 17% less accurate than the model with NLEs (see Figure 5.18). In the feedback box, people expressed general criticism of saliency maps (“The visual explanations are quite irrelevant and not so accurate.”, “Often the heat map was focused outside of the lung area but still assumptions about lung-related diagnosis was made. This made it less trustworthy, but i liked having the ability to see the area of AI focus ”), and largely praise for the NLEs (“This one was the easiest to understand”, “Quite liked the comments - it was intuitive, and I could rapidly decide whether I agreed”, or “the descriptions we helpful in that they highlighted the uncertainties in some cases.”). For the combined explanations, some express confusion about the misalignment of text and visual explanations (e.g. “I did not understand why the heat map and the comment did

not always have the same focus. To me, the heat map should show the focus and the comment should give additional information, not one talking about the right and the other one of the left side.”, “there were discrepancies between the heatmap and the textual explanation”, “I think that from all of the models I saw in this project this one was a little bit hard to work with.”). Furthermore, participants expressed a desire for the text-based explanations to explicitly incorporate the provided patient background (e.g. “Please build model with consideration of patient’s medical history. physical exam and background” or “the AI clearly didnt seem to be going on the textual history, seeming to rely predominantly on its recognition of images/patterns- we all know context means so much in medicine”). Participants also evaluated each explanation type across five key characteristics of explanations, with language-based explanations scoring the highest on all 5, as shown in Figure 5.17 (the exact questions can be found in Section 5.2.2). Across the board, there is a strong preference for language-based explanations among users.

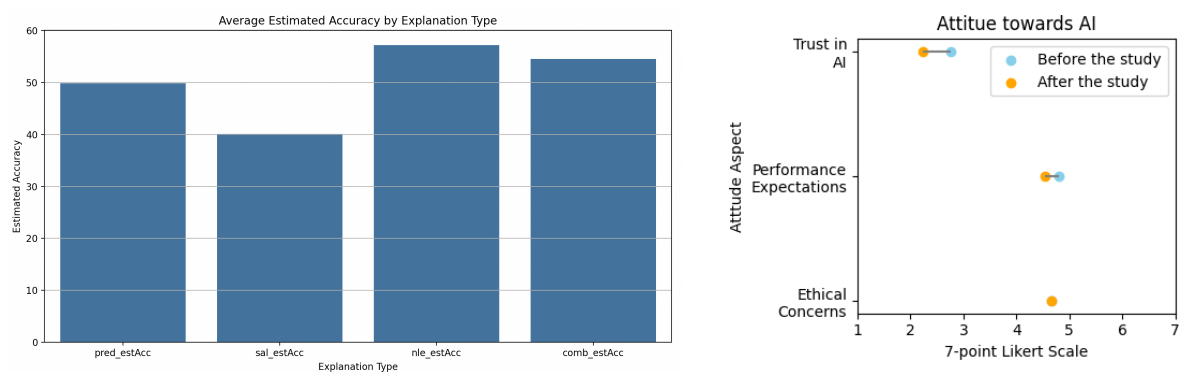


Figure 5.18: *Left*: Participants were asked to estimate the accuracies of all four models/explanation types. The actual accuracy is 70% for all four. *Right*: Performance expectations and trust in AI were reduced after completing our study, which is expected as I present an imperfect AI with imperfect explanations.

Overall, people underestimate the accuracy of the models. Given that I specifically investigate imperfect AI and XAI, it also makes sense that participants’ attitudes towards AI changed towards more caution, shown on the right of Figure 5.18. To capture more qualitative insights, I initially planned for this user study to use a “think-out-loud” approach, including in-person interviews while the participants complete the task. However, I opted against this as the conversation would be a kind of intervention that could introduce a bias into how the participants interact with the different explanation types. Instead, I selected a subset of participants with whom an interview was conducted right after the task. These results are processed by a collaborator and will be written up at a later date.

In the following section, rather than looking at subjective preferences and experiences, I will look at what I consider *actual* explanation usefulness, i.e., how much they improve human-AI team performance on the task of chest X-ray analysis.

5.5.2 Results for the Main Research Questions

This section contains the results for the main research questions established in Section 5.2.4. The effect of different explanation types under specified conditions can be studied by considering the contrasts in Equation 5.1. The analysis highlights three main findings: (1) AI and XAI correctness need to align, (2) NLEs, on their own, lead to overreliance on the AI, and (3) when aligned, a combination of NLEs and saliency maps is the most useful explanation type. As can be seen on the left of Figure 5.21, the differences between our four conditions (no XAI, saliency maps, NLEs, and combined) cancel each other out when looked at across all C_{AI} and C_X values. When the AI is correct, there are also no significant differences between explanation types across explanation correctness scores. Furthermore, the participants generally underperform the AI, i.e., their average performance would be better if they would always blindly agree with the AI. Most of the main results are drawn from Figure 5.19, where I consider the alignment between AI and XAI. Further elaborations on the main results are provided below.

C_X **needs to align with** C_{AI} . AI and XAI are considered aligned when XAI explanations' factual correctness aligns with the AI advice's factual correctness. I propose the following framework to capture how the correctness of AI advice and AI explanations can interact:

- Explanations are *insightful* when their correctness **aligns** with advice correctness: *Convincing* explanations are explanations with high C_X when the AI advice is correct; *Revealing* explanations are explanations with low C_X when the AI advice is incorrect.
- Explanations are *deceptive* when their correctness **misaligns** with advice correctness: *Misleading* explanations are explanations with high C_X when the AI advice is incorrect; *Confusing* explanations are explanations with low C_X when the AI advice is correct.

Figure 5.16 shows real examples from our user study that correspond to each of these four interaction types. Explanations are split into “incorrect” (C_X in lower 50% percentile) and “correct” (C_X in upper 50% percentile). Although best efforts have been made to ensure a uniform distribution of C_X scores in our study data, this split should rather be viewed as capturing

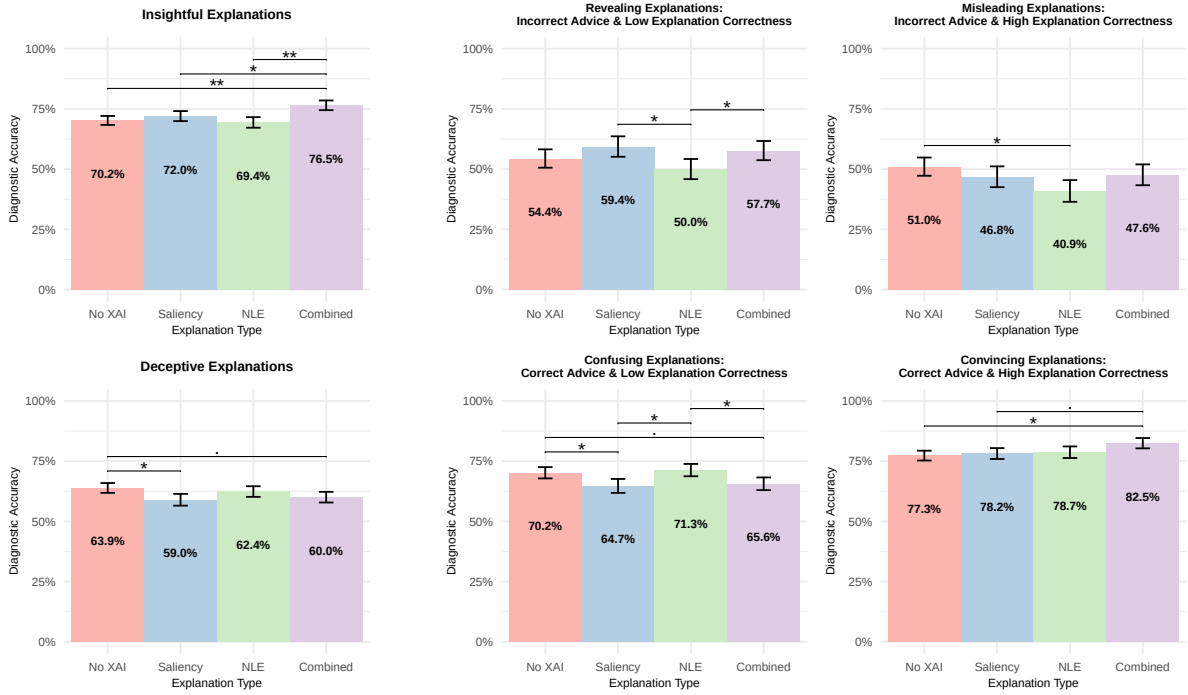


Figure 5.19: The bar charts represent model-based predictions of human accuracy under different conditions. For example, the model predicts a 76.5% “expected probability” of correct user decisions for “insightful explanations” with NLEs (top-left plot). p -values are derived from hypothesis testing, comparing human accuracy between explanation types for specific data subsets. The error bars represent standard errors. ., *, ** ($p < 0.1, 0.05, 0.01$)

relative explanation correctness, rather than *absolute* explanation correctness based on some C_χ threshold.

The results confirm that insightful explanations, i.e., where C_χ aligns with C_{AI} , are helpful in a decision-support setting. Figure 5.20 illustrates how higher C_χ scores harm human accuracy when the AI prediction is incorrect (*deceptive* explanations) and benefits human accuracy when the AI advice is correct (*insightful* explanations). These effects are, however, less strong for NLEs than for the visual methods.

In Figure 5.19, I look at human accuracy by explanation type for the four C_{AI} - C_χ alignment scenarios described earlier. To obtain human accuracy for “No XAI”, where no explanation correctness scores are given, I simply consider all the images where the average C_χ of saliency, NLE, and combined is in the upper or lower half.

As a general trend, human accuracy is negatively impacted when explanations are *deceptive*, and people would be better off seeing no explanation. For saliency maps, human accuracy goes down 4.9% ($p < .05$) when C_{AI} and C_χ do not align (i.e., the AI is correct, but the saliency map

is incorrect, or vice versa). For combined explanations, it goes down 3.9% ($p = .06$). On the contrary, for insightful explanations, human accuracy goes up 4.3% ($p < .005$) for combined explanations. Thus, overall, at least for visual explanation types, alignment between C_{AI} and C_χ is crucial in determining whether explanations are useful or not.

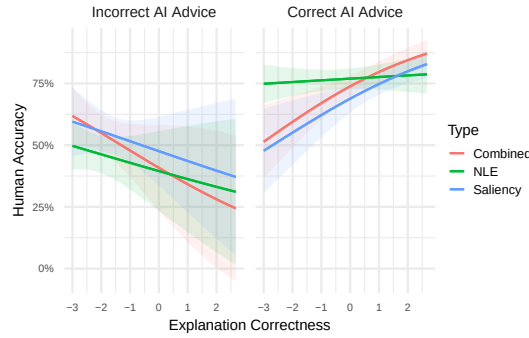


Figure 5.20: Human accuracy given C_{AI} and C_χ , predicted with the model (5.1).

Natural language explanations, on their own, lead to overreliance. Whilst I have shown that the users express a strong preference for text-based explanations, our study shows that, actually, for incorrect advice, there is a significant drop in human accuracy when NLEs are given compared to combined (-7.3% , $p < .05$) and saliency maps (-6.2% , $p < .05$) (Figure 5.21, middle). This suggests that NLEs make people more likely to agree with the AI when it is actually incorrect, i.e., people over-rely on the AI. It is alarming that text-based explanations made people overestimate the model’s capabilities and trust the model *too much*. Especially when C_χ is comparatively high, but the AI advice is incorrect, i.e., the explanation highlights findings that are somewhat clinically correct but are not evidence for the diagnosis, people are 10.1% ($p < 0.05$) more likely to agree with the AI when it is accompanied by an NLE than without explanation (top-right in Figure 5.19). This also means that for the scenario of correct advice and comparatively low C_χ explanations, NLEs lead to higher performance (6.6%, $p < .05$ vs. saliency maps and 5.7% $p < .05$ vs. combined), as people are more likely to agree with low C_χ NLEs than other explanation types with low correctness. Overall, people agree with the AI 67.3% of the time when it is accompanied by an NLE, compared to 63.8% on average for the other explanation types. This could suggest that the assertiveness (Calisto et al., 2023) and/or human-like nature (Breum et al., 2024) of language-based explanations may lead people to overly trust and rely on AI.

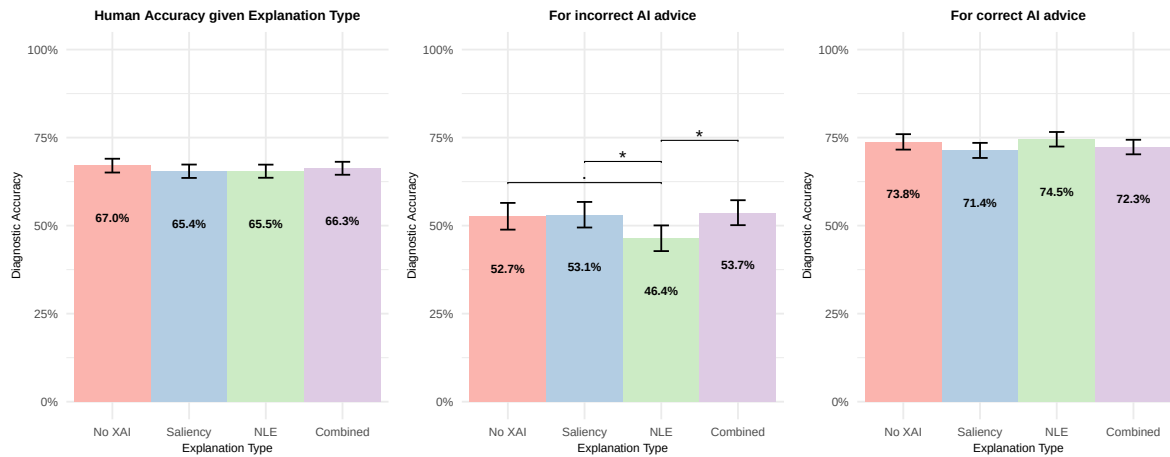


Figure 5.21: Human accuracy given explanation types overall (left), for incorrect advice (middle), and for correct advice (right).

When aligned, combine saliency maps and NLEs. Although abstractive, text-based explanations lead to significant overreliance on their own, combining them with saliency maps can make them more effective than saliency maps on their own and, importantly, than having no explanation. For insightful explanations, i.e., where correctness aligns with AI correctness, combining saliency maps and NLEs provides significant improvements compared to the other conditions: 6.3% ($p < .005$) over “No XAI”, 7.1% ($p < .005$) over NLEs alone, and 4.5% ($p < .05$) over saliency maps alone. This combination seems useful in particular when the AI is incorrect, where low C_χ visual explanations lead to significantly higher human performance, even when accompanied by NLEs. At the same time, when the AI is correct, high C_χ combined explanations have a boost over visual-only explanations. This suggests that participants can integrate the information from visual and textual cues to identify when an AI is wrong or right more accurately. Besides different modalities (text versus visual), the compared XAI methods also differ in either being “feature attribution” (saliency maps) or “abstractive” explanations (NLEs). The study results suggest that combining these different types of explanations may alleviate their respective shortcomings.

Multiple testing adjustments I ensured the robustness of our results by pre-registering our study, aligning with best practices to avoid p -hacking (Wicherts et al., 2016), having a rigorous model selection process, guided by AIC and BIC in addition to likelihood ratio tests, and by minimizing the number of subsequent tests. I also reported effect sizes and confidence intervals over p -values where possible, focusing on practical significance (Nakagawa, 2004). Nonethe-

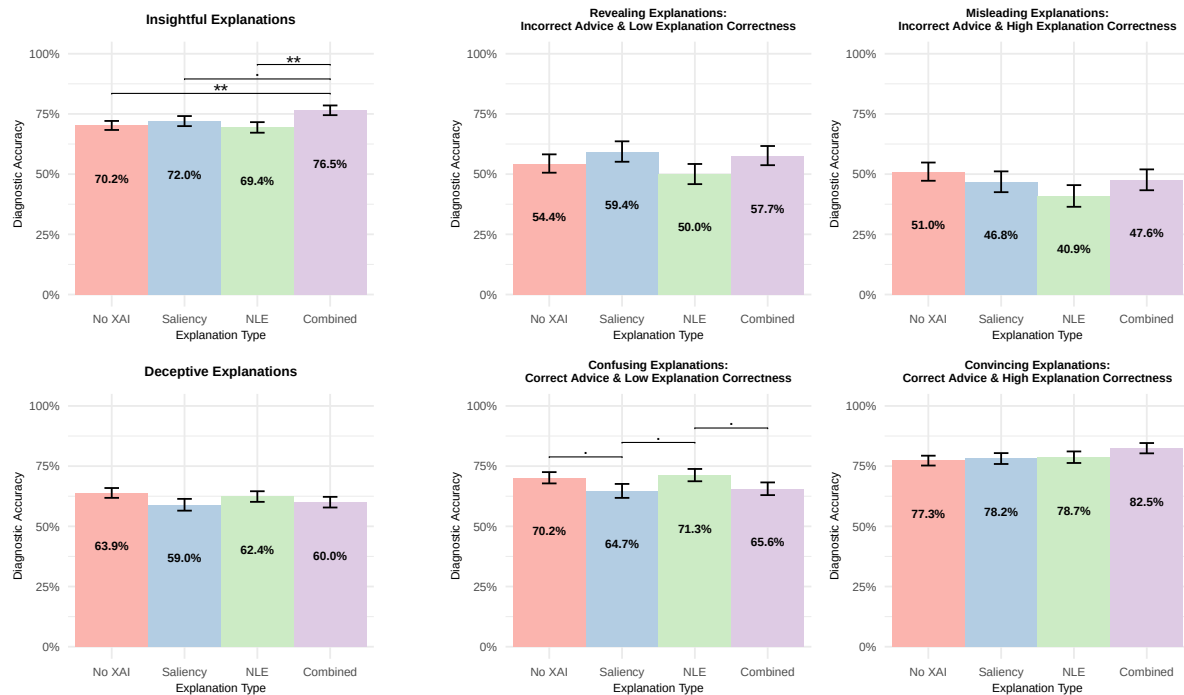


Figure 5.22: Multiple testing adjusted results. The bar charts and error bars represent model-based predictions of human accuracy under different conditions. p -values are derived from hypothesis testing, comparing human accuracy between explanation types for specific data subsets and using Benjamini-Hochberg’s corrections for multiple testing. The error bars represent standard errors. ., *, ** ($p < 0.1, 0.05, 0.01$)

less, I also include Benjamini-Hochberg’s corrections for multiple testing (Figure 5.22). While some effects are no longer significant, combined explanations still provide a significant boost when explanations are insightful.

5.5.3 Exploratory Analysis

In addition to the main research questions, I also measured “perceived usefulness” (how useful are the explanation types perceived by participants), “decision speed” (how much do the explanation types affect the time taken per case), and “positive certainty” (how much do the explanation types affect whether users say “maybe” or “definitely” present). These research questions were listed as exploratory in the pre-registration.

Perceived Usefulness Perceived usefulness is a subjective measure of how useful participants find an explanation (7-point Likert scale response to “How useful was the AI model’s explanation in helping you decide whether the AI was right or wrong in suggesting (e.g.) pneumonia”). This allows us to measure subjective preference on a per-instance level and juxtapose

it to “objective”, actual usefulness.

In order to understand the role of perceived usefulness, I consider a similar model to Equation (5.1), but instead, I predict perceived usefulness ρ_U and add effect A . A is given as $A = 1$ when the participant agrees with the AI advice and $A = 0$ otherwise (Equation (5.6)). This model was validated in a similar fashion to the main model, as explained in Section 5.2.4.

$$\begin{aligned}
 \rho_{ij} = & \beta_0 \\
 & + \beta_a C_{AI} + \beta_t \chi + \beta_p A \\
 & + \beta_{t \times a} (\chi \times C_{AI}) + \beta_{p \times a} (A \times C_{AI}) \\
 & + \beta_{t \times e} (\chi \times C_\chi) + \beta_{t \times p} (\chi \times A) \\
 & + \beta_{t \times e \times a} (\chi \times C_\chi \times C_{AI}) \\
 & + \beta_{t \times e \times p} (\chi \times C_\chi \times A) \\
 & + u_{Participant} + u_{Image}
 \end{aligned} \tag{5.6}$$

Across all scenarios NLEs obtain the highest ρ_U scores, as shown in Figure 5.23. This is in line with our post-survey findings, where participants expressed a strong preference for NLEs. We also note that *perceived usefulness is higher both when participants agree with the AI advice (versus disagree) and when they are correct (versus when they are wrong)*. This suggests that perceived and actual usefulness are somewhat aligned. I also found that the difference in ρ_U when disagreeing vs. agreeing with the AI is significantly larger for NLEs than for saliency maps ($p < .001$) and for combined explanations than for saliency maps ($p < .001$).

In order to find out if there are significant differences between ρ_U across explanation types when participants are correct or wrong, i.e., whether perceived usefulness is associated with actual usefulness significantly more or less, between different explanation types, I fit a model where I replace C_{AI} with human accuracy in Equation (5.6). I find no significant differences between explanation types in this regard, *suggesting that perceived usefulness is equally associated with actual usefulness for all explanation types*.

In cases where participants correctly *disagree* with the AI (top-left plot of Figure 5.24), ρ_U of saliency maps increases with decreasing explanation correctness, suggesting that *participants are able to perceive that incoherent saliency maps can be helpful in detecting false predictions*. This is not the case for NLEs or combined explanations. The bottom-left plot aligns with our

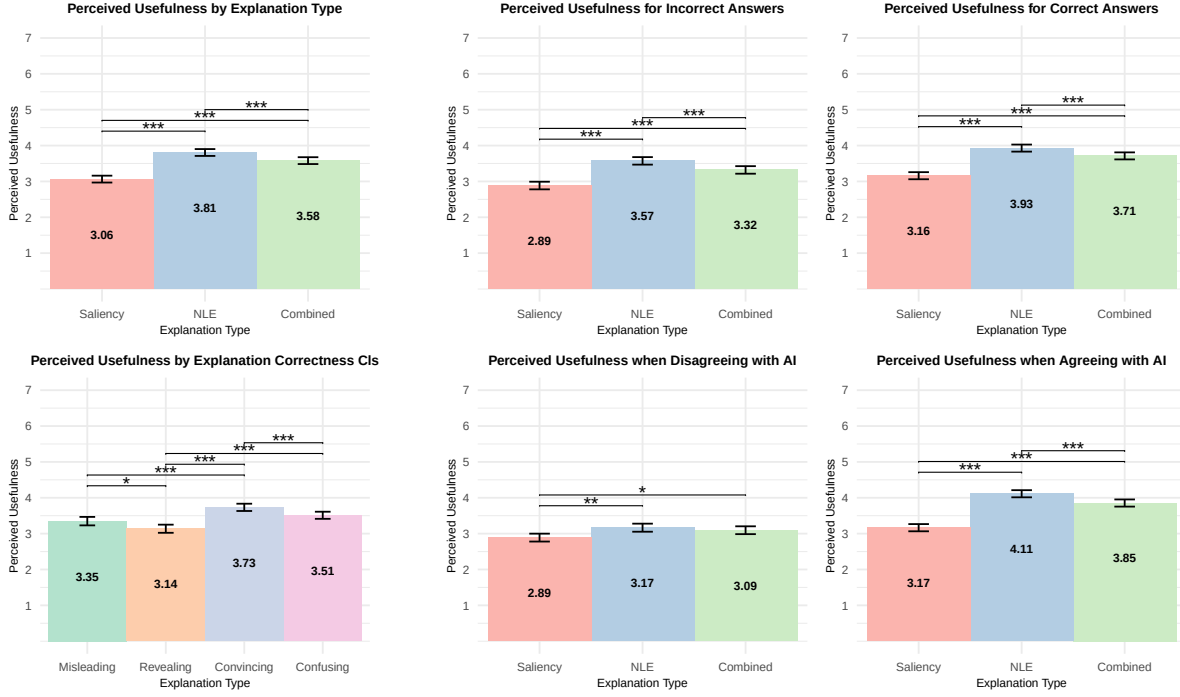


Figure 5.23: Perceived usefulness ρ_U . The upper left shows overall ρ_U with respect to explanation types. The lower left shows ρ_U with respect to the explanation correctness quadrant, averaged across all types. The remaining four plots show ρ_U for when participants are wrong or right or when they agree or disagree with the AI advice. The error bars represent standard errors. ., *, ** ($p < 0.1, 0.05, 0.01$).

intuition: when participants agree with the AI, even when it is wrong, they are more likely to rate factually correct explanations as useful. For the case of agreeing with correct AI advice, I observe that ρ_U is by far the most highly correlated with explanation correctness for combined explanations.

Positive certainty I define positive certainty as the share of times participants say a finding is “Definitely” instead of “Maybe present” (for negative, we only have “Not present”, so we cannot measure the degree of certainty). Again, I consider the same model as in (5.6), but instead, I predict positive certainty.

$$\begin{aligned}
 \gamma_{ij} = & \beta_0 \\
 & + \beta_a C_{AI} + \beta_t \chi \\
 & + \beta_{t \times a} (\chi \cdot C_{AI}) + \beta_{t \times e} (\chi \cdot C_\chi) \\
 & + \beta_{t \times e \times a} (\chi \cdot C_\chi \cdot C_{AI}) \\
 & + u_{Participant} + u_{Image}
 \end{aligned} \tag{5.7}$$

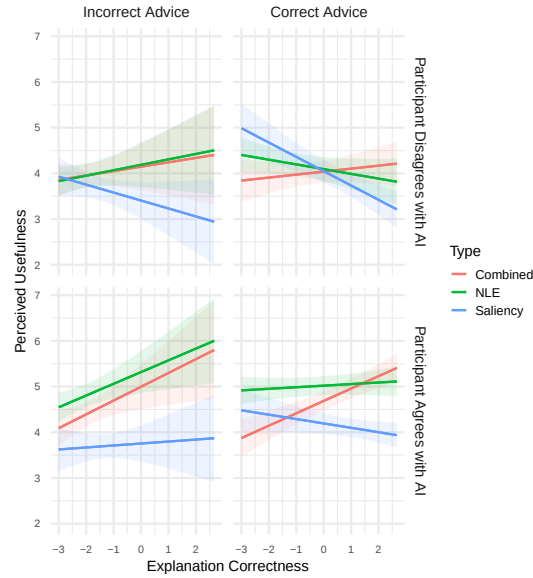


Figure 5.24: Perceived usefulness ρ_U by AI advice correctness C_{AI} , user agreement A , and explanation correctness C_X .

The model from Equation 5.7 has a poor fit, suggesting that there is no clear relationship between positive certainty and explanation types and explanation correctness. Figure 5.25 confirms that *explanation types do not significantly affect positive certainty*. However, when subdividing into explanation correctness quadrants, I find that, unsurprisingly, *convincing explanations (correct AI advice and correct explanation) lead to the highest positive certainty*, significantly higher than all other quadrants ($p < .01$).

Decision Speed Decision speed is the time taken to provide an answer for a single chest X-ray. I removed cases where the time taken is above 2 minutes, as this likely suggests participants were interrupted (this removes 5.1% of cases). Again, I consider the same model as in (5.6), but I predict decision speed. I also found that adding participant agreement A leads to a better fit.

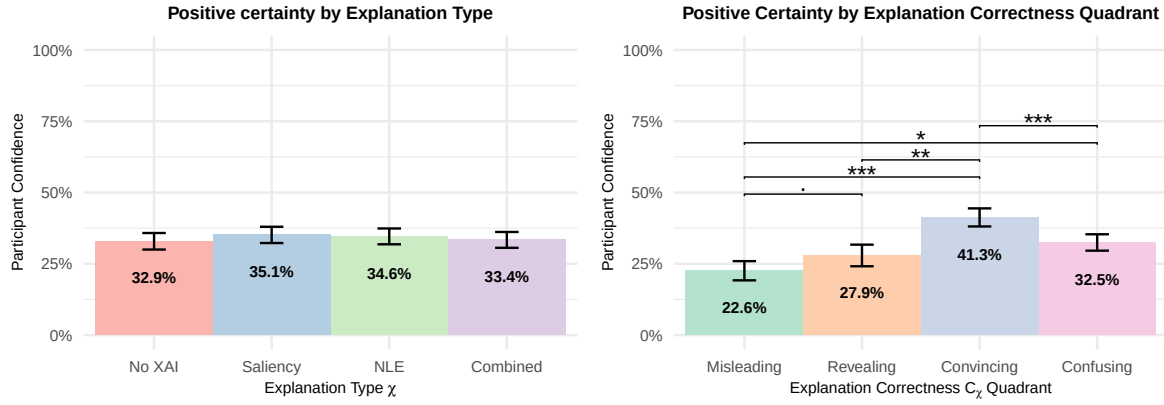


Figure 5.25: Positive certainty. The left plot shows overall positive certainty with respect to explanation types χ , and the right plot shows positive certainty with respect to explanation correctness C_χ quadrant. The error bars represent standard errors. ., *, ** ($p < 0.1, 0.05, 0.01$).

$$\begin{aligned}
\delta_{ij} = & \beta_0 \\
& + \beta_a C_{AI} + \beta_t \chi + \beta_p A \\
& + \beta_{t \times a} (\chi \times C_{AI}) + \beta_{p \times a} (A \times C_{AI}) \\
& + \beta_{t \times e} (\chi \times C_\chi) + \beta_{t \times p} (\chi \times A) \\
& + \beta_{t \times e \times a} (\chi \times C_\chi \times C_{AI}) \\
& + \beta_{t \times e \times p} (\chi \times C_\chi \times A) \\
& + u_{Participant} + u_{Image}
\end{aligned} \tag{5.8}$$

The top-left plot of Figure 5.26 shows that *decision speed increases with increasing complexity of the explanation type*. There is a significant increase in time spent ($p < .001$) between each increasing complexity step, except between NLEs and combined explanations. *Time taken ranges from 36.0 seconds (no explanation) to 43.1 seconds (combined explanations)*. Whilst this confirms some of the comments I received, that NLEs make people engage more with the AI, i.e., increase time for making a decision, it is interesting that combining visual and textual explanations does not lead to an additional overhead in time taken. It is promising that this combination, whilst being a much more useful explanation type and seemingly more complex, does not increase the time clinicians spend on a case. The duration is not significantly affected by whether participants are right or wrong or agree or disagree with the AI. Interestingly, *explanation correctness quadrants do not show a significant effect on decision speed*. I also find that *explanation correctness has no significant effect on decision speed*, suggesting that participants

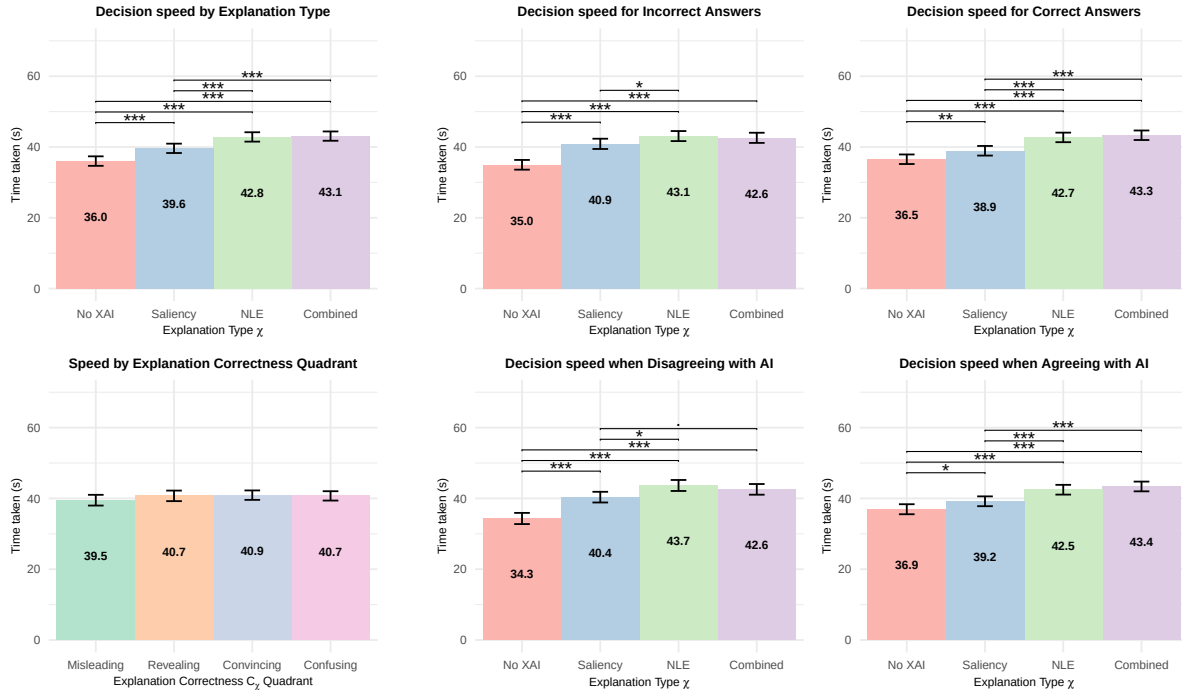


Figure 5.26: Decision speed. The top-left plot shows overall decision speed with respect to explanation types χ , and the bottom-left plot shows decision speed with respect to explanation correctness C_χ quadrant, averaged across all types. The remaining four plots show decision speed for when participants are wrong or right or when they agree or disagree with the AI advice. The error bars represent standard errors. ., *, ** ($p < 0.1, 0.05, 0.01$).

do not spend more time on cases where the explanation is correct or incorrect. Figure 5.27, as expected, shows that decision speed decreases over the course of the study, with spikes at the start of each new task.

5.5.4 Model Fit

Albeit I mainly use contrasts to interpret our statistical model, I provide the model fit in Table 5.6. As expected, the strongest overall effect is for advice correctness (β_{AC}). The three-way interaction between visual explanations (COMB and SAL) and advice and explanation correctness is also significant, whereas it is not for NLEs. This confirms the findings that explanation correctness is not as noticeable to participants for NLEs.

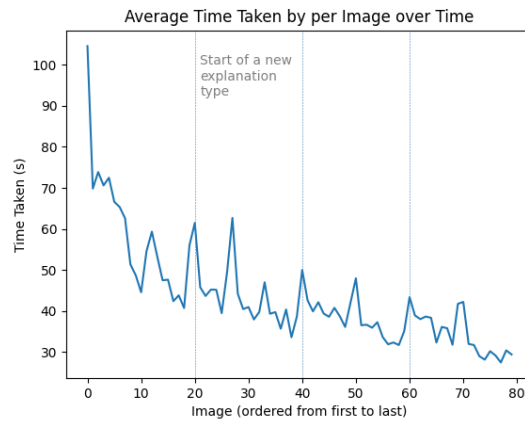


Figure 5.27: This plot shows the average decision speed (time taken per image) and how it changed over time. The overall trend is that participants become faster over time. We can also see spikes at the start of each new task when they are introduced to a new explanation type.

Table 5.6: Generalized Linear Mixed Effects Model Results

Fixed Effects				
Predictor	Estimate	SE	<i>z</i>	<i>p</i>
β_0 (Intercept)	0.090	0.162	0.559	0.576
$\beta_{\chi_{\text{COMB}}}$	-0.480	0.364	-1.321	0.187
$\beta_{\chi_{\text{NLE}}}$	-0.535	0.341	-1.570	0.117
$\beta_{\chi_{\text{SAL}}}$	-0.206	0.279	-0.738	0.461
β_{AC}	1.053	0.198	5.313	<0.001***
β_{EC}	-0.161	0.163	-0.986	0.324
$\beta_{\chi_{\text{COMB}} \times AC}$	0.355	0.380	0.934	0.350
$\beta_{\chi_{\text{NLE}} \times AC}$	0.579	0.361	1.602	0.109
$\beta_{\chi_{\text{SAL}} \times AC}$	-0.163	0.308	-0.531	0.595
$\beta_{\chi_{\text{COMB}} \times EC}$	-0.124	0.203	-0.609	0.543
$\beta_{\chi_{\text{NLE}} \times EC}$	0.023	0.194	0.117	0.907
$\beta_{\chi_{\text{COMB}} \times AC \times EC}$	0.612	0.197	3.104	0.002**
$\beta_{\chi_{\text{NLE}} \times AC \times EC}$	0.176	0.159	1.110	0.267
$\beta_{\chi_{\text{SAL}} \times AC \times EC}$	0.455	0.186	2.449	0.014*
Random Effects				
Group	Variance	SD		
Participant ID	0.060	0.245		
Image ID	0.333	0.578		

AIC = 6499.1, BIC = 6604.9

*** $p < .001$, ** $p < .01$, * $p < .05$

5.6 Conclusion

In this chapter, I conducted a large-scale user study simulating a real clinical decision support setup, including in-domain and clinical experts, and juxtaposed textual (NLEs) and visual (saliency maps) explanations. This is the first attempt in the literature to provide a quantitative assessment of how users interact with text-based AI explanations in a safety-critical setting. The study highlights the importance of alignment between AI advice and XAI for explanations to be useful. Another key finding was that NLEs, that is, text-based explanations, are strongly preferred by end users but equally lead to overreliance on AI. This is a very significant finding with respect to modern generative AI, where the main interface between AI models and users is natural language. The humanization of AI, due to using natural language instead of more numerical feature attribution methods to communicate, may lead to more serious issues around automation bias (Gabriel et al., 2024). Given how common it is for current generative AI tools to explain their decision-making in natural language (e.g., OpenAI’s o1 and o3, Deepseek R1), understanding the effect this has on users in safety-critical applications is crucial. Our study also shows that this overreliance can be alleviated by combining NLEs with saliency maps. Saliency maps, as opposed to abstractive text-based explanations, are feature attribution-based explainability methods. They are arguably more faithful, at the cost of scrutability, than text-based explanations. Combining these types of explanations is a promising avenue to address the risks of overreliance and humanization of NLEs while still providing user-friendly explanations. In addition, I have also shown that this apparent added complexity does not increase the time taken per case compared to text-only explanations. An alternative path to address overreliance caused by generative text-based explanations could be to work on the faithfulness of the explanations. If explanations more faithfully explain decision-making, they could potentially make it easier for users to realize when a model is wrong. I hope the findings in this chapter will enable future research on optimizing such explanations to lead to safer use of AI.

In terms of limitations, while I attempted to replicate clinical practice as realistically as possible, the study cannot fully mirror the conditions under which clinicians work. It is also not a longitudinal study, meaning I do not explore how interaction with the models changes over time. Recruitment biases such as self-selection can also affect the generalizability of the participant pool. While our cohort of participants is fairly diverse, it is likely not representative of the global population as a whole.

6 | Language as Signal for Brain MRI Representation Learning

The three previous chapters focused on NLEs and showed that such textual explanations, representing human knowledge and reasoning, can make AI models learn better representations of images. Radiology reports, rich in such human knowledge and source of the NLEs, can thus be used to learn image representations from scratch. This is especially helpful in domains where data, especially annotated data, is scarce, such as for magnetic resonance imaging (MRI).

MRIs are standard-of-care imaging performed for the diagnosis and management of neurodegenerative diseases as well as cancers occurring in the brain. Despite their clinical importance, developing effective AI models for MRI analysis presents significant challenges due to the complex 3D nature of these scans and the limited availability of annotated datasets.

Clinical reports for medical images, especially detailed cross-sectional (3D) scans such as CT and MRI, often contain rich, diverse information, with multiple sections addressing different aspects of the scan. As illustrated in Figure 6.1, these reports provide detailed descriptions that correspond to specific regions of the brain MRI.

Modern vision-language models (VLM) and vision-language pre-training (VLP) have shown the capability to generate descriptive free-text for medical images (Yang et al., 2024) and produce strong vision backbones (Huang et al., 2021b). However, the scarcity of large-scale medical image-text datasets and the limited generalization of conventional vision-language methods designed for 2D images have hindered the development of effective VL approaches for complex 3D imaging modalities. Furthermore, existing methods that aim to learn joint representations of medical images and paired text often overlook the complexity and length of clinical reports. Instead, it is common to adopt architectures similar to those used in natural image captioning, where descriptions usually consist of a single sentence and are represented via a single embedding.

To address this, I introduce the largest existing multimodal brain MRI dataset, containing approximately 80,000 3D MRIs and paired reports, and propose a new VLP framework called **brat** (**brain report alignment transformer**). **brat** is based on the premise that similar to documents, clinical reports can consist of extensive text describing a wide range of findings. Multi-view

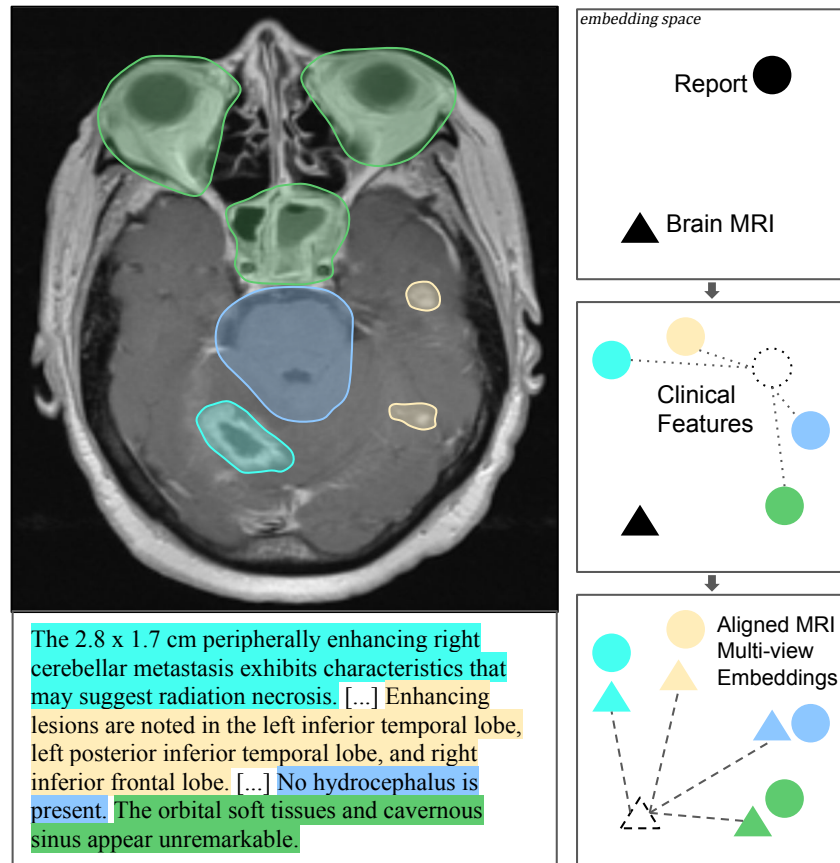


Figure 6.1: (Left) Brain MRI reports contain rich and diverse information that relates to different aspects of the image. Colours are used to associate report sentences with corresponding regions on the scan. The report was cut off ([...]) to only contain findings that can be seen on this 2D slice. (Right) Drawing parallels to multi-vector retrieval (Zhang et al., 2022a), I align multi-view embeddings of the MRI with clinical features (sentences) given in the reports.

embedding approaches have been used in document retrieval to represent the diversity of information in a document, which can then be mapped to sub-units of the user query (Khattab and Zaharia, 2020). In brat, learnable multi-view embeddings represent the 3D brain MRI, whereas the corresponding reports are subdivided into sentence embeddings or sub-units. By aligning these sub-units with brain MRIs via contrastive learning, I implicitly encourage the multi-view image embeddings to represent the clinical features described by the sentences (see Figure 6.1). I ensure that distinct image views are aligned with sub-unit sentence embeddings through the proposed Pairwise View Alignment (PVA) matching algorithm. Diversity of the multi-view embeddings is further enhanced via quality diversity (QD) loss based on Determinantal Point Processes (DPPs) (Kulesza and Taskar, 2012). The results show that brat leads to improved image-text and text-image retrieval and that vision backbones trained via brat lead to significant performance improvements on downstream tasks, such as report generation, tumour

segmentation, or Alzheimer’s classification.

This chapter will first discuss related work in Section 6.1. In Section 6.2, I introduce MSKBrain, the proposed dataset. After that, the `brat` framework is presented in Section 6.3, and results are discussed in Section 6.4.

6.1 Related Work

Brain MRI Datasets The effectiveness of deep learning methods in brain MRI analysis relies on the availability of large and high-quality datasets. Recent work in the field has been enabled partly through private datasets (Kamnitsas et al., 2017) or by pooling small public datasets (Munk et al., 2024). With ADNI (Petersen et al., 2010), an Alzheimer’s dataset, being the largest with 8,019 imaging sessions, brain MRI datasets are much smaller in size than datasets in natural images (Deng et al., 2009) or other medical imaging modalities such as chest X-rays (Johnson et al., 2019).

Brain MRI Pre-Training The scarcity of large-scale brain MRI datasets (and lack of image-text datasets) has led researchers to pool smaller public datasets (Munk et al., 2024) and focus on SSL. SSL tasks involve reversing various image augmentations (Zhang et al., 2016; Noroozi and Favaro, 2016), masked image modeling (Bao et al., 2022; He et al., 2022; Tang et al., 2022; Jiang et al., 2022), and contrastive losses (Chen et al., 2020a). Other efforts aim to learn models that can be generalized across MRI modalities (Konwer et al., 2023; Xu et al., 2024). Downstream applications often focus on segmentation, with notable datasets being the BraTS series (Baid et al., 2021; Moawad et al., 2023), on which I also evaluate our models. This chapter introduces the largest pre-training dataset for brain MRIs to date and provides the first VLP framework for brain MRI representation learning.

Multi-vector Document Retrieval Document retrieval, which involves retrieving documents based on user queries, has seen significant improvements through the use of multi-vector retrieval methods, which outperform traditional single-vector approaches. In single-vector retrieval, documents are represented as a single embedding, limiting the ability to capture the diversity of information in documents (Karpukhin et al., 2020). In contrast, multi-vector embeddings offer more versatile query-document interactions, which better represent the variety of information in documents. One of the important works is ColBERT (Khattab and Zaharia, 2020; Santhanam et al., 2022), which computes query-document similarity by selecting the most similar document token for each query token and aggregating the similarities across a document. However, these ideas have not yet been applied to image-text datasets, even though text like radiology reports is often of document length. Zhang et al. (2022a) emphasized that documents typically contain multiple semantic units, each potentially relevant to different queries,

and proposed using multi-view embeddings to represent these diverse aspects. Drawing inspiration from their method, I treat brain MRI scans as “documents” and report sentences as corresponding “queries”, to enhance representation and retrieval.

Quality-Diversity of Learned Features Increasing diversity to avoid informational collapse is common in SSL, where information maximization methods are used to achieve this (Zbontar et al., 2021; Bardes et al., 2022). In terms of diversity for representing different aspects of an input, Zhang et al. (2022a) in document retrieval has a similar goal and use a loss that penalizes high pairwise similarity between embedding vectors. Determinantal point processes (DPPs), on the other hand, enable the measurement of the overall spread of diversity and allow the integration of quality notions seamlessly (Kulesza and Taskar, 2012). Whilst they have been used for increasing quality and diversity in recommender systems, summarisation, or dataset/batch sampling (Li et al., 2021a; Saran et al., 2023), they have not yet been applied to multi-view representation learning. I show that they are well-suited for this and lead to better performance than simple pairwise similarity reduction.

6.2 MSKBrain Dataset

6.2.1 Raw Dataset

I collected a dataset of brain MRI scans and their corresponding clinical reports from Memorial Sloan Kettering Cancer Center, covering the period from 2012 to 2017. These scans were primarily obtained to monitor brain metastases and tumours in cancer patients, resulting in a dataset rich in positive findings (89.7% of scans show abnormalities, an average of 134 words or 8 sentences per report) and representative of a diverse patient population. I also collected the clinical reports corresponding to these images, as well as extensive demographic data, primary diagnosis, ongoing chemotherapy and radiotherapy treatments, and survival information, which will be utilized in future studies.

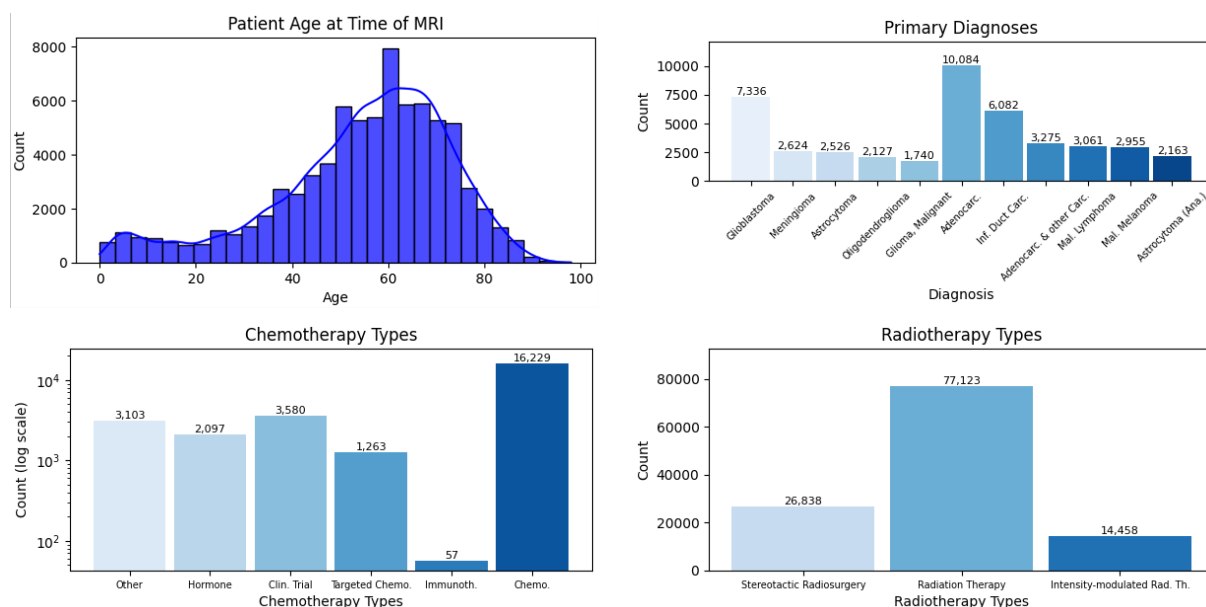


Figure 6.2: Patient demographics. Representation of all ages. Primary diagnoses refer to the primary cancer diagnosis for the patients for whom the scan was ordered. Chemotherapy and radiotherapy types show a count of all the types of chemo/radio sessions assigned to the patients in this dataset.

MRI sessions typically consist of multiple MRI modality scans (e.g. FLAIR, T2, etc.); however, in this first iteration, I focus on T1-post contrast MRIs, the most informative type of scan. To extract T1 post-contrast scans, I generated a long, clinician-validated list of keywords typically used to refer to these scans. The list contained over 50 expressions such as “Axial T1 post SENSE” or “Ax T1 POST”. I removed around 3,000 sessions that did not include T1 post-contrast imaging. DICOM medical images were converted and saved as 3D NIFTI files. I first

sorted and loaded the DICOM slices into a 3D array, and then, for each image stack, I preserved relevant metadata, such as pixel spacing and slice thickness, in the affine matrix. To standardize the intensity values, I thresholded the images at the 99th percentile and rescaled them to a range of 0-800, converting the final values to 16-bit integers. I saved the processed 3D volumes as compressed NIFTI files (.nii.gz). I then connect the MRI sessions to patient data stored in a REDCap database. For each imaging session, I attached the patient's demographic information and added calculated fields like age and time to death. I then mapped treatments and diagnoses to imaging sessions by finding the closest diagnosis date and checking which medications and radiation therapies were active at the time of imaging. I provide an overview of key patient data in Figure 6.2. The final dataset includes 77,228 brain MRI image-report pairs from 24,262 unique patients. To develop our model, I performed a patient-wise split of the data into 75,142 examples for training, 945 for development, and 1,141 for the test set.

6.2.2 Text processing

Again status post left-sided craniotomy with stable postoperative changes and with slight increase in the heterogeneously enhancing mass lesion centred in the left temporal lobe which now measures 7.5 x 4.8 cm on image 13 series 14 from 6.7 x 4.7 cm, though the enhancement within it is more irregular and less intense than before. The mass is not completely imaged on the perfusion sequence but there is hyperperfusion inferiorly within the nodular enhancing component which is incompletely demonstrated. The surrounding hyperintense T2/FLAIR infiltrating nonenhancing signal abnormality is stable consistent with nonenhancing tumor/edema. No new discontinuous suspiciously enhancing brain lesions. There is slightly increased dilatation of the ventricles with slightly increased hyperintense T2/FLAIR signal in the periventricular white matter particularly about the frontal horns and atrium, suggesting transependymal flow of CSF from a communicating hydrocephalus. Stable mild midline shift to the right without significant downward herniation. No acute intracranial hemorrhage, infarct, or new extra-axial collections.

Figure 6.3: An example report showing references to prior scans in blue and descriptions of findings not visible on T1 post-contrast scans in yellow.

As this chapter focuses on learning image representations from brain MRIs, I had to ensure that all report content was visually grounded in the corresponding images. For example, keyword filtering revealed that 94% of reports make references to prior scans. Figure 6.3 contains an example report showing the two main types of information not detectable from the image: references to prior scans and description of findings not visible on T1 post-contrast images. Inspired by recent work demonstrating GPT-4 performs well on radiology report processing (Liu

You are a highly experienced radiologist. Re-write the given brain MRI report and only modify the following:

- (a) Leave out any details not visible on T1-weighted post-contrast images. Note that T2/FLAIR hyperintensities can often be seen on T1 Images. Observations related to e.g. perfusion, plasma volume or K trans cannot be seen and should be excluded.
- (b) Leave out any terms that suggest temporal change or progression (e.g. dates, “new”, “increased”, “previous”, “now”, “compared to”, “since last”, “more”, “less”, etc.)
- (c) Remove any PHI.

Figure 6.4: The final prompt that was used to re-write the reports and remove PHI and information not visible from the T1 post-contrast images.

[et al., 2023b](#)), I developed an on-premise GPT-4-based report processing pipeline. This pipeline enabled us to achieve three objectives without using expensive human annotators: the data was anonymized by removing protected health information (PHI), the reports were re-written to remove the aforementioned references, and I extracted structured information. Through iterative prompt engineering and radiologist feedback, I arrived at a 2-stage approach that extracted information from the reports with a 96% accuracy on a gold standard set of 50 manually annotated reports. This accuracy was calculated by counting the share of exact matches for all structured data points. By evaluating the second step of the two-step approach, we also implicitly ensure that the intermediate re-written report captures all the relevant information from the original report. Annotating all the reports costs approximately \$1,600, which is significantly lower than the cost of expert annotation. Figure 6.4 shows the report re-writing prompt and Figure 6.7 shows the information extraction prompt. Example of this two-step processing are shown in Figures 6.6, 6.8, and 6.9.

As there were around 80,000 long radiology reports, I used Python’s asyncio framework to process multiple reports in parallel with GPT-4. Each report underwent two API calls: one for re-writing and one for extracting structured information. I managed the API rate limits by processing reports in batches and including sleep time between batches. I also added a logit bias to avoid certain temporal medical terms (e.g., “increase”, “new”) in the answer. The temperature was set to 0.0 and top_p to 1.0 for deterministic outputs. With parallelization, processing 80,000 reports took around 48 hours. GPT-4 performed significantly better than GPT-3.5 during preliminary comparisons. GPT-4o or later had not yet been released at this time.

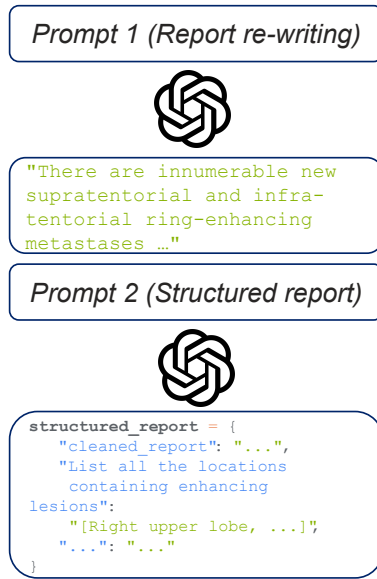


Figure 6.5: The two-step GPT-4 based report processing pipeline. Prompt 1 and 2 are in Figure 6.4 and 6.7, respectively.

Table 6.1: Brain MRI dataset characteristics. For reference, the Conceptual Captions (Sharma et al., 2018) dataset has 10 tokens (less than 10 words) per image.

Characteristic	Value
Word Count (Q1, Median, Q3)	115, 134, 156
Sentence Count (Q1, Median, Q3)	7, 9, 11
Age (Q1, Median, Q3)	45, 58, 68
Any Abnormality (%)	87.9
Prior Surgery (%)	38.1
Enhancing Lesions (%)	47.6
Midline Shift (%)	5.4
White Matter Changes (%)	43.6
Pituitary Gland Abnormality (%)	2.3
Hydrocephalus (%)	2.6
Biggest Mass Length (%)	<1cm (17.7), 1-2cm (14.5) >2cm (18.8)
Enhancing Lesion Count (%)	1 (27.0), 2-6 (23.0) 7-15 (1.6), >15 (2.3)
# of Unique Surgeries	32,428
# of Enhancing Lesion Locations	95,815

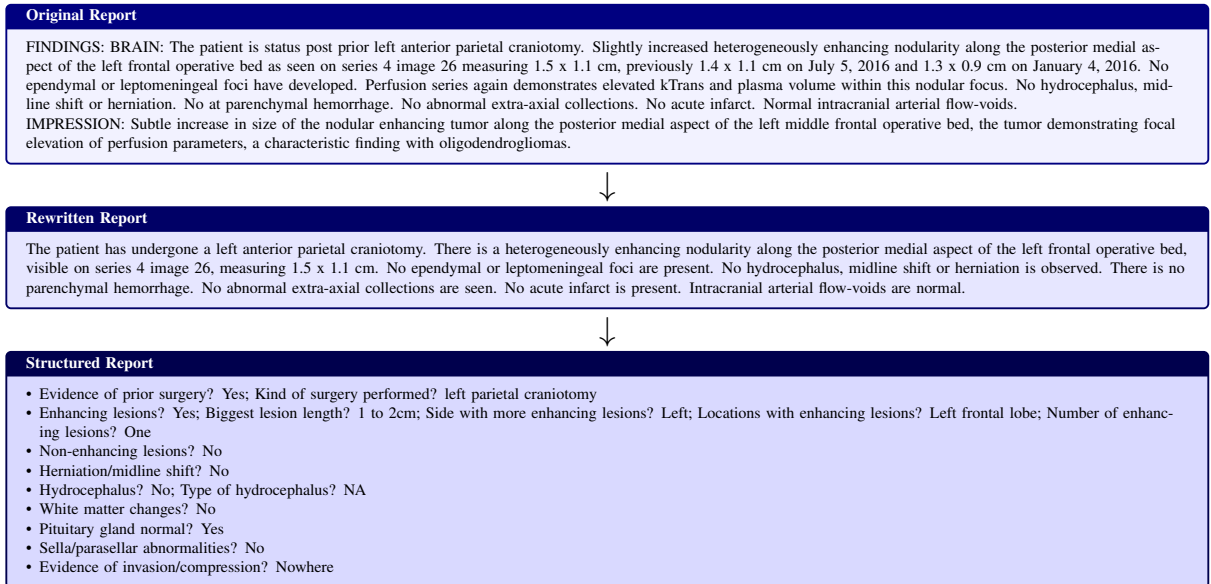


Figure 6.6: An example of our automated report processing. The re-writing correctly rephrases the sentences, making references to changes in findings and removes references to, e.g. kTrans findings. The structured information extraction correctly answers all our instructions.

You are a highly experienced radiologist. Accurately answer the questions below based on the given brain MRI report. Your output must be in json format.

- (a) For each question, choose the appropriate answer (wording must match exactly). If answers are mutually exclusive, choose one. If multiple answers can apply, list all that are true, separated by semicolons (";").
- (b) If the MRI report does not contain information to answer a specific question, use the default answer indicating a normal status.
- (c) Note the following assumptions: meningiomas are considered enhancing lesions; burr holes and ventriculostomy and Ommaya catheters are considered prior surgery; punctate lesions are less than 1cm.

Questions (Answer options):

Is there evidence of prior surgery? (Yes / No)

What kind of surgery was performed? (NA / left frontal craniotomy; right frontal craniotomy; left parietal craniotomy; right parietal craniotomy; left temporal or pterional craniotomy; right temporal or pterional craniotomy ; left occipital craniotomy; right occipital craniotomy)

Are there any enhancing lesions? (Yes / No)

What is the length of the biggest mass lesion? (NA / Less than 1cm / 1 to 2cm / More than 2cm)

Which side of the brain has more enhancing lesions? (NA / Left / Right)

List all the locations that contain enhancing lesions. (NA / Left frontal lobe; Right frontal lobe; Left parietal lobe; Right parietal lobe; Left temporal lobe; Right temporal lobe; Left occipital lobe; Right occipital lobe; Left thalamus or basal ganglia; Right thalamus or basal ganglia; Cerebellum; brainstem; cervical spinal cord)

How many enhancing lesions are there? (NA / One / Between 2 and 6 / Between 7 and 15 / More than 15)

Is there a herniation or midline shift? (Yes / No)

Are there any signs of white matter disease (e.g., leukoaraiosis or leukoencephalopathy)? (Yes / No)

Is the pituitary gland normal in appearance? (Yes / No)

Are there abnormalities in the sella or parasellar regions? (Yes / No)

Where is there evidence of invasion into or compression of adjacent structures? (Nowhere / ventricles; brainstem; cranial nerves)

Brain MRI report: \${Insert processed report}\$

Figure 6.7: The final prompt that was used to extract structured information from the reports.

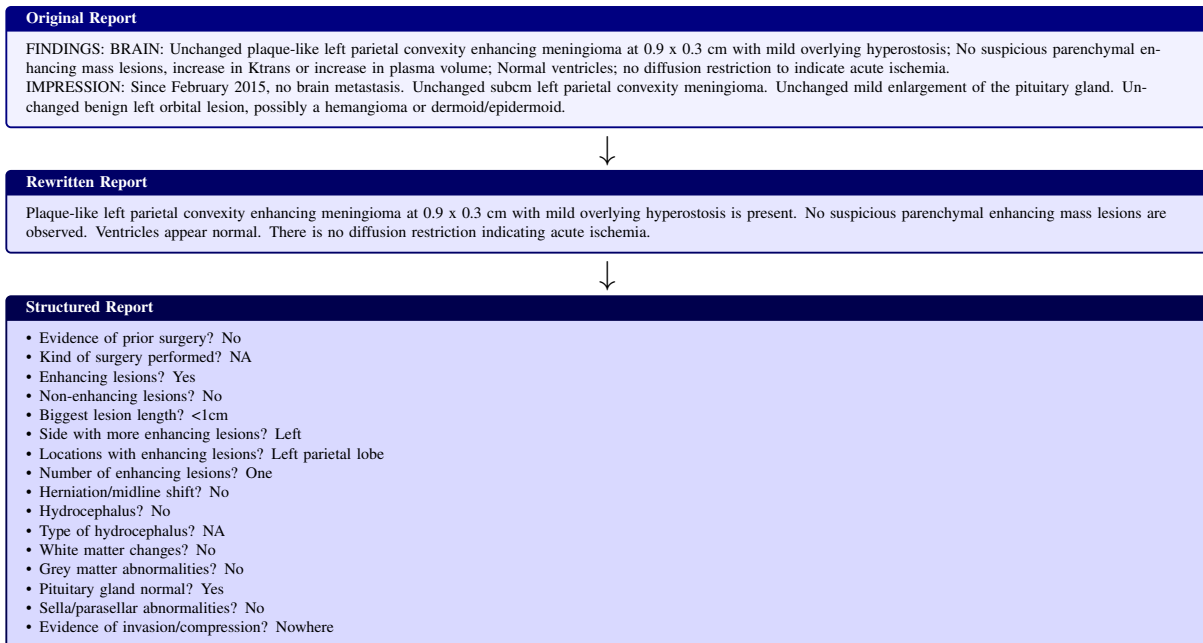


Figure 6.8: An example of our automated report processing. Here, the intermediate re-written report omits the enlargement of the pituitary gland and left orbital lesion.

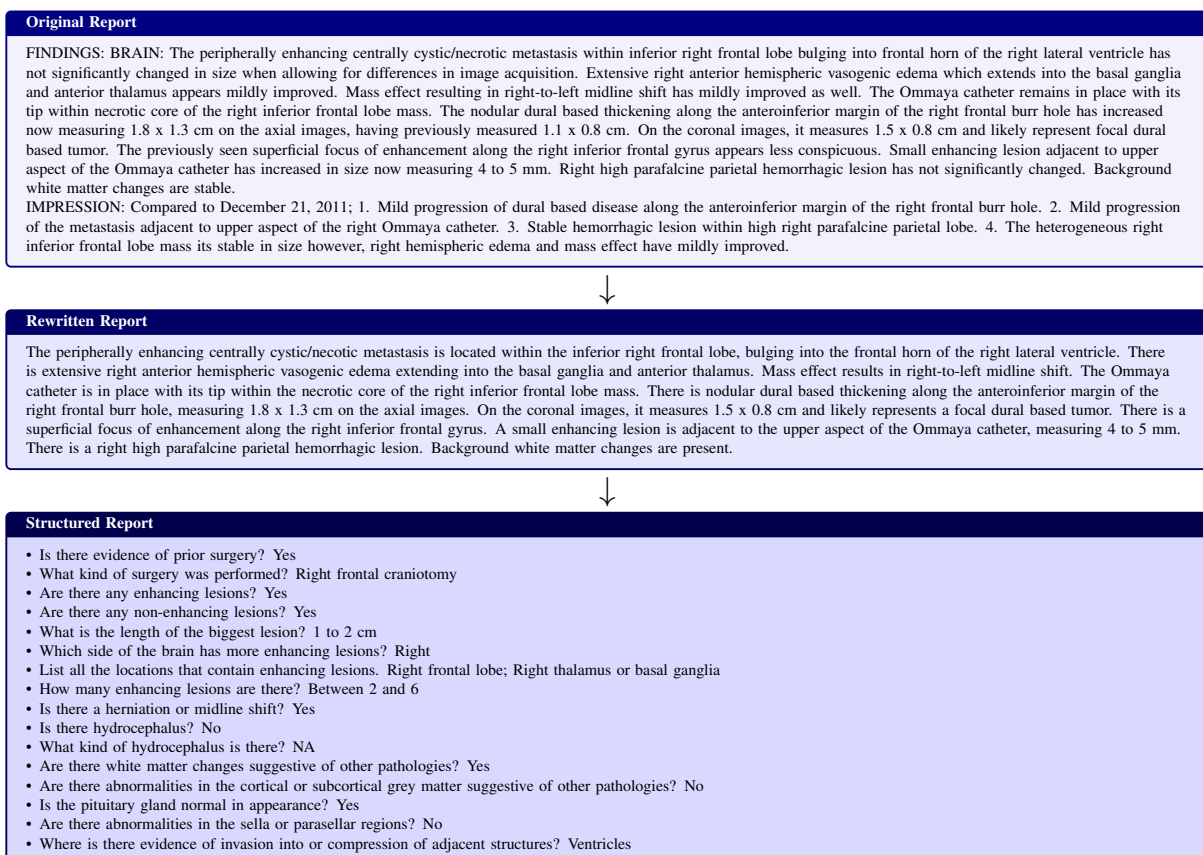


Figure 6.9: An example of a long and complex report that was processed successfully.

6.3 brat: A Vision-Language Model for Brain MRIs

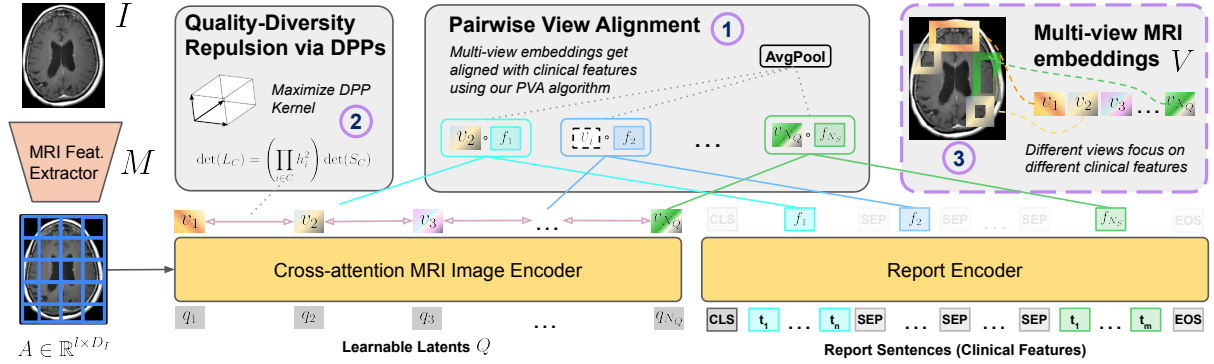


Figure 6.10: Our brat architecture. Pairwise View Alignment (PVA) and quality-diversity via Determinantal Point Processes (DPPs) lead to aligned multi-view embeddings of the MRI.

brat is a vision-language contrastive pre-training framework that represents images via aligned multi-view embeddings. To obtain the multi-view embeddings, I adopt a similar base architecture as Q-Former (Li et al., 2023b). brat uses learnable latents that extract multi-view embeddings by cross-attending to the MRI features (see Fig. 6.10). A 3D vision model M is used to extract features from an MRI image I , resulting in a set of feature maps $M(I) = A \in \mathbb{R}^{l \times D_I}$ with l feature maps of dimension D_I . I mainly use Densenet-121 (Huang et al., 2017) as M , which outperformed ViT and Resnet-50 in our preliminary experiments. The set of learnable latent tokens $Q = [q_1, \dots, q_{N_Q}]$ where $q_l \in \mathbb{R}^{D_Q}$, interact with the image encoder features A to extract a set of image-informed multi-view embeddings $E_I(Q, I) = V$, where $V = [v_1, \dots, v_{N_Q}]$ with $v_i \in \mathbb{R}^{D_V}$. The text encoder E_R takes a brain MRI radiology report and returns sentence embeddings that capture the clinical features described in them: $E_R(R) = F$, where $F = [f_1, \dots, f_{N_S}]$ with $f_i \in \mathbb{R}^{D_F}$ representing the i -th sentence. I obtain sentence embeddings f_i by averaging all token embeddings of the sentence. As $D_F = D_V$, D_F is used in the rest of the paper for clarity. The remainder of this section will discuss our alignment algorithm and quality-diversity protocol.

6.3.1 Pairwise-View Alignment

Existing approaches that use latent variables to learn to extract image features, such as Q-Former, often exhibit embedding collapse, where the learned latents converge into a single representation (Chauhan et al., 2024), as shown in Figure 6.11. However, given the complexity of brain MRIs, which frequently contain a diverse range of findings analogous to how doc-

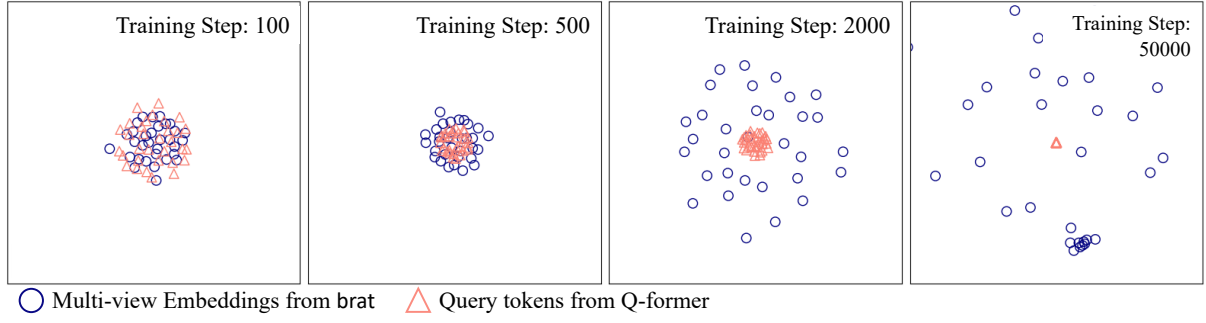


Figure 6.11: Conventional query tokens collapse into a single representation as training progresses. The multi-view embeddings of *brat*, on the other hand, are diverse and spread out. The plot was obtained by using multi-dimensional scaling to plot the 32 query tokens on a 2D plane based on their pairwise Euclidean distances averaged across a random sample of 32 images.

uments comprise distinct semantic units with highly variable appearance, I hypothesize that multi-view embeddings provide a more suitable representation for brain MRI images. Multi-view embeddings have been shown to improve document representations by capturing different semantic elements within a text (Zhang et al., 2022a). Inspired by this, I assume that individual sentences in radiology reports correspond to distinct clinical features, and I aim for our multi-view embeddings to encapsulate these same clinical characteristics (see Figure 6.1). This is achieved via Pairwise View Alignment (PVA), which aligns embeddings with clinically meaningful features.

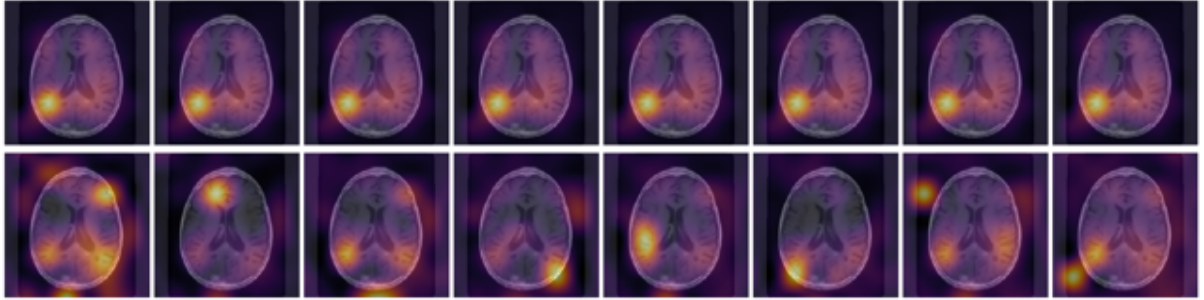


Figure 6.12: Juxtaposition of 8 query tokens from Q-Former (upper row) and the same 8 query tokens from *brat* (lower row). The collapsed Q-Former queries all attend to the same image regions, whereas the multi-view embeddings of *brat* focus on distinct features.

PVA ensures that each image view embedding can only be matched to a single sentence feature, preserving the granularity of the report in the image representation. The overall image-text similarity is then given by the average of all multi-view embedding and matched clinical feature similarities. Matching is done based on cosine similarities. A step-by-step description of the approach is given in Algorithm 1. While the algorithm is designed to be as efficient as possible, the sorting step at step 3 introduces a minimal overhead (complexity of

$O(N_Q \times N_S \times \log(N_Q N_S)))$ compared to simpler baselines such as averaging across all similarity pairs. In practice, this did not increase training time by more than 10%.

Algorithm 1 Pairwise View Alignment

Input: Normalized multi-view embeddings of an image $V = [v_1, \dots, v_{N_Q}] \in R^{N_Q \times D_F}$ and normalized clinical features matrix of a report $F = [f_1, \dots, f_{N_S}] \in R^{N_S \times D_F}$

Output: Match pairs of multi-view embeddings and clinical features P_M

Algorithm:

- 1: Compute similarities between all multi-view embeddings V and clinical features F :
 $S_{v,f} \leftarrow VF^T \in R^{N_Q \times N_S}$
 - 2: Take the softmax of $S_{v,f}$ over f . This ensures that clinical features are aligned with queries that most uniquely match them. This encourages query representation to focus on individual clinical features instead of aggregating over them. For example, if a sentence matches different queries with 0.9, 0.8, 0.8 or 0.5, 0.1, 0.1, the softmax ensures that the 0.5 in the second one is a higher match.
 - 3: Initialize list P_{all} with all similarity pairs $\langle v_i, f_j \rangle = S_{v,f}[i, j]$ sorted in descending order of similarity
 - 4: Initialize an empty list of matched pairs P_M
 - 5: **while** P_{all} is not empty **do**
 - 6: Select a pair of indices (s_v, s_f) with the highest similarity (top of stack): $\langle v_{s_v}, f_{s_f} \rangle = \max(P_{\text{all}})$ (if multiple pairs have the maximum similarity score, I am using the pair with the smallest indices where the index for image features takes priority)
 - 7: Add (s_v, s_f) to the list of matched pairs P_M
 - 8: Remove all pairs from P_{all} with indices s_v for image features or s_f for report features (each query token and sentence feature can only be matched once)
 - 9: **end while**
-

6.3.2 Quality-Diversity via DPPs

Empirically, I find that PVA alone does not sufficiently encourage diverse features in multi-view embeddings. I, therefore, adopt ideas from quality-diversity (QD) to address this issue. The idea behind QD is to have many diverse solutions to tackle a problem from different angles. This fits our problem well, as we want the different multi-view embeddings to focus on different features of the image. I consider as feature diversity the diversity of the attention of the multi-view embeddings V over the image feature maps A .

Pairwise dissimilarity measures, e.g., used in [Zhang et al. \(2022a\)](#), enforce local repulsion between embeddings but may result in trivial solutions. In contrast, DPPs explicitly model global diversity by maximizing the volume spanned by embeddings collectively, thus avoiding trivial solutions and encouraging each embedding to capture distinct aspects of the feature space ([Kulesza and Taskar, 2012](#)). In addition, DPPs allow the seamless integration of a quality

heuristic. Figure 6.13 and 6.14 illustrate the advantage of DPPs over pairwise repulsion in terms of promoting a more desirable feature diversity.

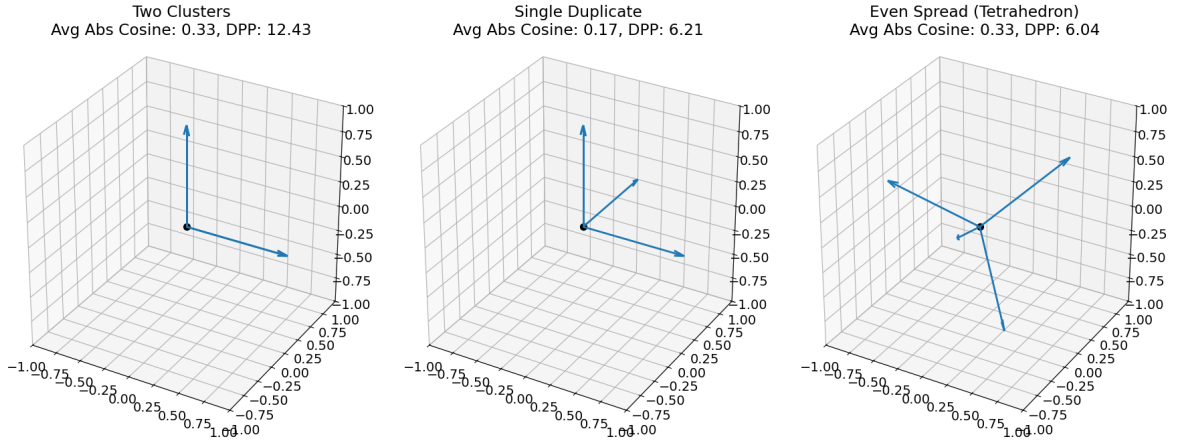


Figure 6.13: Visualization of the volume spanned by four 3D vectors under different configurations. In the left figure, both vectors are duplicates. In the middle one, one vector is a duplicate. The average pairwise cosine dissimilarity is the same for the left and right figures, even though the right figure represents a much more desirable spread of the vectors. The DPP is much lower for that figure thus minimizing the DPP is superior.

DPPs are distributions over subsets of a fixed ground set that attribute higher probability to sets that are diverse. In our case, we want to maximize the probability of our set of multi-view embeddings V under the DPP. I consider quality-diversity with respect to the cross-attention maps $C = [c_1, \dots, c_{N_Q}]$ where $c_j \in \mathbb{R}^{n_{fx} \times n_{fy} \times n_{fz}}$ contains the attention values from multi-view embedding v_j to the 3D feature maps representing the image. As *quality* of each embedding token, I use the Shannon entropy of its cross-attention map, denoted as $h_i \in \mathbb{R}^+$:

$$h_i = \mathcal{H}(c_i) = - \sum_k c_i(k) \log c_i(k), \quad (6.1)$$

where k indexes over spatial positions. A higher entropy implies a more uniform attention distribution, which is interpreted as higher quality. This further prevents the attention maps from collapsing to trivial, high-diversity solutions where each query attends to a single point. The attention maps c_i themselves are considered as the *diversity features*. The DPP kernel matrix L_{ij} can be written as:

$$L_{ij} = h_i c_i^T c_j h_j. \quad (6.2)$$

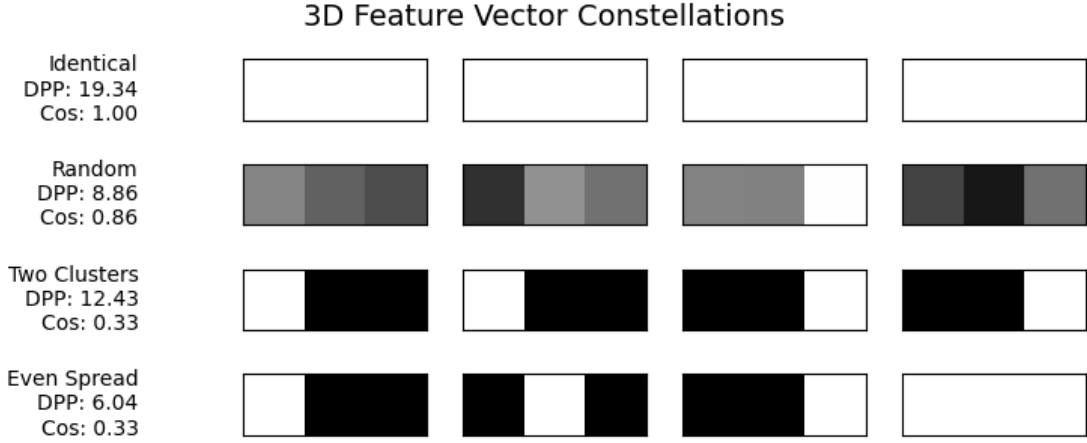


Figure 6.14: Visualization of four 3D vectors under different configurations, representing attention maps over a 1×3 image. It shows that pairwise cosine dissimilarity (“Cos”) equally rewards an even spread of attention maps (desired) and the embeddings grouping into two clusters (not desired). In contrast, the DPP rewards the even spread the most.

The DPP for selecting a subset C' is given by:

$$P_L(C') \propto \det(L_{C'}), \quad (6.3)$$

In our case $C = C'$, as we consider repulsion between all image tokens. $\det(L_C)$ can be decomposed as follows:

$$\det(L_C) = \left(\prod_{i \in C} h_i^2 \right) \det(S_C), \quad (6.4)$$

where S_C is the similarity matrix between all attention maps c_i .

The determinant of the kernel matrix L_C corresponds to the squared volume of the parallelepiped spanned by the vectors $h_i c_i$ for each i in C . By maximizing the product $\prod_{i \in C} h_i^2$, each embedding token is encouraged to have high entropy, corresponding to a large magnitude in the feature space. By maximizing the determinant $\det(S_C)$, where S_C captures the pairwise similarities between attention maps, I ensure that the directions c_i are as different as possible, promoting diversity among the tokens. This approach naturally prevents the embeddings from collapsing into a single representation by encouraging both high-quality (non-collapsed attention) and diverse (distinct attention patterns) embeddings.

In practice, I define the DPP loss by taking the negative log-determinant of the kernel matrix

L :

$$\mathcal{L}_{\text{DPP}} = -\log \det(L_C + \epsilon I), \quad (6.5)$$

where ϵI is a small diagonal matrix added for numerical stability.

brat Loss To obtain the overall image-report similarity, I aggregate the multi-view embedding sentence similarities by mean-averaging:

$$S_{R,I} = S_{I,R} = \frac{1}{|P_M|} \sum_{(i,j) \in P_M} S_{v,f}[i, j] \quad (6.6)$$

As such, I get the contrastive losses as follows:

$$\mathcal{L}^{(I|R)} = -\log \left(\frac{\exp(S_{I,R}/\tau)}{\sum_k \exp(S_{I,R_k}/\tau)} \right) \quad (6.7)$$

$$\mathcal{L}^{(R|I)} = -\log \left(\frac{\exp(S_{R,I}/\tau)}{\sum_m \exp(S_{R,I_m}/\tau)} \right) \quad (6.8)$$

I also use the same ‘‘Image-grounded Text Generation’’ (ITG) loss as in BLIP-2 (Li et al., 2023b), as I found it to help performance. The final loss is thus given as:

$$\mathcal{L} = \frac{\mathcal{L}^{(I|R)} + \mathcal{L}^{(R|I)}}{2} + \mathcal{L}_{\text{DPP}} + \mathcal{L}_{\text{ITG}} \quad (6.9)$$

Training and Implementation Details Training parameters were determined empirically, with the final set provided in Table 6.2. Contrary to the general assumption that big batches lead to improved performance for image-text contrastive learning, our results consistently improved for comparatively small batch sizes in the range of 25-32. This allowed me to train each model on a single A/H100 GPU. I also found that the Q-Former’s language modelling loss consistently improved performance across nearly all configurations, while image-text matching did not yield benefits and was, therefore, omitted. Using a biomedically pre-trained BERT outperformed the standard BERT pre-training version in all evaluated scenarios. All model weights were selected based on the best average recall@k metrics on the development set. For MSKBrain, all models

were trained with the same image processing: $1mm \times 1mm \times 1mm$ voxel spacing, intensity normalization, and resizing to $32 \times 256 \times 256$. Preliminary analysis on MSKBrain also showed lowered performance with standard data augmentation such as Gaussian noise, image rotation and translation, as well as random view cropping, and was removed from subsequent analyses. I also explored various text augmentation techniques. In “text dropout”, a given share of text tokens is randomly masked during the forward pass. In “sentence dropout”, different sentences, i.e., clinical features, are randomly removed from the reports. In “PVA dropout,” I randomly drop a given share of paired multi-view embeddings and clinical features from the overall similarity matrix. None of these techniques have led to performance gains.

6.3.3 brat as a Foundation Model

The brat framework provides both a pre-trained vision backbone M and a model $E_I(Q, I)$ for extracting multi-view embeddings. I refer to the vision backbone only as brat-viz, and brat for the multi-view framework.

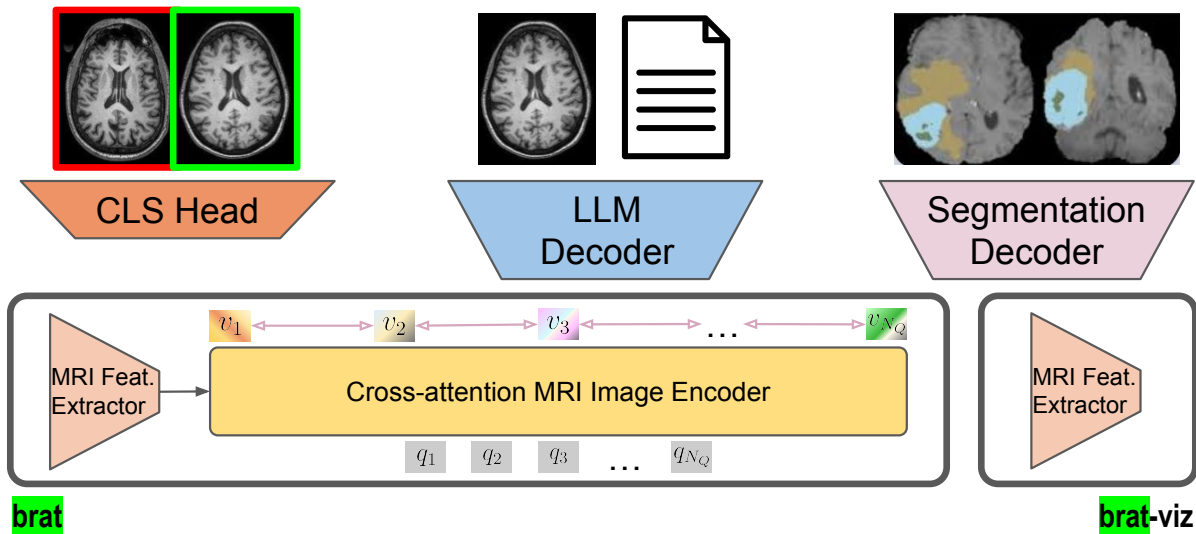


Figure 6.15: Two configurations of brat are connected with various decoders to evaluate the benefits of our pre-training on downstream tasks.

Figure 6.15 illustrates how the brat weights can be modularized for different downstream tasks. Different task-specific heads, such as an MLP for classification, a language model for report generation, or a segmentation decoder, can be appended to either brat or brat-viz. Experiments on different such configurations are provided in the next section.

When feasible, hyperparameters were selected via grid search. ADNI hyperparameters are

given in Table 6.3. For ADNI, image preprocessing was performed using Clinica’s t1-volume-tissue-segmentation pipeline. For report generation, I follow the hyperparameters recommended in the Llama paper (Grattafiori et al., 2024). I used a batch size of 1024, a learning rate of 0.0002, and an AdamW optimizer with a cosine decay and a warm-up ratio of 0.3. For segmentation, I used nnUNet as the baseline model, fine-tuning it with an initial learning rate of 1e-2 and a weight decay of 3e-5. The training pipeline included standard nnUNet preprocessing; data augmentation was not used. Final model weights were selected based on the highest mean Dice score for BraTS-2021 and the best Lesion-Wise metrics for BraTS-2023-METS on the validation set.

Table 6.2: Implementation details of our pre-training. *For ViT, I used a lower learning rate of 1.00E-07.

Parameter	all models on MSKBrain	brat on BIMCV-R
Batch Size	32	25
3D Vision Model M	Densenet-121/ViT/ResNet-50	Densenet-121
Weights Init. of M	None	
Architecture of $E_{I/R}$	BERT-base ^a	
Weights Init. of $E_{I/R}$	BiomedBERT ^b	
Learning Rate M	5.00E-04*	
Learning Rate $E_{I/R}$	5.00E-05	
Max. Text Length E_R	256	
N_Q (# of Query Tokens)	32	
Cross-Attention Frequency	2	
Max. Number of Sentences	20	
Training Precision	Bfloat16	
Augmentation	None	

^a188M parameters, cross-attention layers randomly initialized

^bmicrosoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

Table 6.3: Implementation details of our Alzheimer classification downstream task. Augmentation consisted of random flipping, random intensity scaling, random intensity shifting, adding Gaussian noise, Gaussian smoothing, random contrast adjustment, and random low-resolution simulation. *For ViT, I used 1.00E-06 across all data amounts.

Parameter	1%	10%	100%
Batch Size	16	32	32
Learning Rate	1.00E-06	1.00E-05*	1.00E-05*
Training Precision		Bfloat16	
Augmentation	Yes	No	No
Trained Layers	MLP Only	All	All
MLP size		2 layers	

6.4 Experiments and Results

This chapter presents a new vision-language pre-training method for 3D medical scans and document-length reports. In this section, I demonstrate the benefits of our approach over other pre-training methods, both in terms of pre-training metrics as well as on downstream tasks, including tumour and metastases segmentation, Alzheimer’s classification, and brain MRI report generation.

6.4.1 VLP Performance: Image-Text Retrieval

brat is evaluated on image-text retrieval tasks on both MSKBrain and BIMCV-R, an external public benchmark of lung CTs and corresponding reports. Key retrieval metrics such as recall@k and mean and median rank are provided. For MSKBrain, I also provide finding-based metrics, where “P@5 (F)” corresponds to how frequently each of the five retrieved samples contains at least one common positive finding (F) with the ground-truth match. “R@5 (F)” corresponds to the frequency of finding at least one sample containing exactly the same labels as the ground-truth in the top-5 samples (recall).

MSKBrain I evaluated brat against multiple baselines: CLIP, Q-Former (the base of brat), brat with the traditional Colbert matching algorithm ([Khatab and Zaharia, 2020](#)) instead of PVA, brat without QD, brat with simple pairwise repulsion as used in [Zhang et al. \(2022a\)](#) instead of DPPs, and a Q-Former with QD. Results with ViT and ResNet-50 backbone models are provided for completeness. For simplicity, only MRIs that have at least one positive finding (around 90% of our original dataset) were included in the evaluation, as negative reports usually apply to all negative images, making exact matching a faulty metric. As shown in Table 6.4, except mean rank, brat sets the benchmark on all metrics. The lower mean rank can generally be explained by the model making higher confidence predictions, and this can be adjusted by selecting the weights at earlier training steps or by average pooling the multi-view embeddings at inference instead of using the PVA algorithm. The QD component improved performance, suggesting that PVA relies on QD for effective learning. Q-Former did not show benefits with QD repulsion, potentially because the introduced diversity is of less use when the query tokens are not encouraged to be aligned with diverse clinical features. Simple pairwise repulsion also does not match the performance improvements obtained by using DPPs. I also find that the

Colbert algorithm used in document retrieval for matching multi-view embeddings does not perform as well as PVA. Qualitative examples of images and corresponding reports retrieved by brat are shown in Figure 6.16.

Methods	Text to Image								Image to Text							
	R@1 ↑	R@5 ↑	R@10 ↑	R@5 (F) ↑	P@5 (F) ↑	MdR ↓	MnR ↓		R@1 ↑	R@5 ↑	R@10 ↑	R@5 (F) ↑	P@5 (F) ↑	MdR ↓	MnR ↓	
CLIP	0.146	0.407	0.564	0.894	0.718	8.0	35.8		0.159	0.431	0.569	0.853	0.748	8.0	37.2	
Q-Former	0.154	0.377	0.529	0.867	0.672	10.0	32.5		0.146	0.368	0.532	0.837	0.703	9.0	34.9	
Colbert	0.125	0.370	0.509	0.889	0.680	10.0	31.9		0.113	0.326	0.487	0.810	0.732	11.0	36.0	
brat w/o QD	0.173	0.458	0.615	0.894	0.711	6.0	37.3		0.171	0.449	0.606	0.875	0.745	7.0	34.5	
brat w/ PR	0.099	0.349	0.497	0.875	0.723	11.0	36.6		0.109	0.328	0.478	0.818	0.696	11.0	39.2	
Q-Former w/ QD	0.155	0.370	0.542	0.851	0.701	10.0	34.8		0.152	0.381	0.529	0.817	0.699	10.0	33.6	
brat	0.205	0.493	0.666	0.911	0.718	6.0	124.1		0.201	0.481	0.645	0.882	0.752	6.0	96.3	
brat vit	0.015	0.066	0.117	0.661	0.410	385.0	404.6		0.016	0.066	0.129	0.604	0.473	357.0	401.0	
brat resnet	0.095	0.292	0.436	0.843	0.654	13.0	109.4		0.131	0.343	0.462	0.809	0.640	12.0	62.0	

Table 6.4: Evaluation results for text-to-image and image-to-text retrieval on MSKBrain. For the “↑” metrics higher is better and for the “↓” metrics lower is better. “R@5 (finding)” and “P@5 (finding)” indicate the recall and precision at 5 for the finding task.

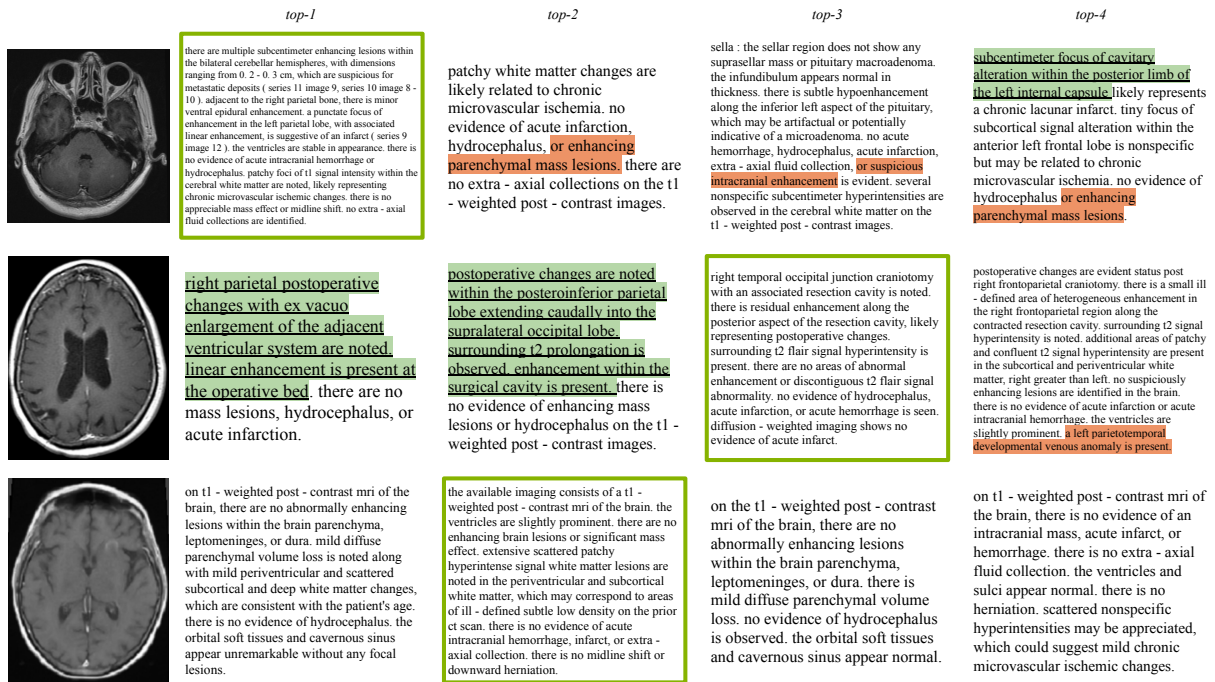


Figure 6.16: Qualitative examples showing the top-4 output of our brat model for image-to-text retrieval on a reduced dev set of 315 examples. On this subset, the median rank achieved was 2. Enboxed examples are correct. Green (and underlined) sections are passages that are clinically correct, even though they are from a different MRI. In orange are passages that do not correspond to the MRI.

BIMCV-R To demonstrate the generalizability of our framework, I also pre-trained it from scratch on the BIMCV-R dataset, a publicly available dataset of lung CT scans paired with radiology reports. As in the original paper, I find that conventional contrastive loss approaches such as a basic Q-Former or CLIP perform very poorly (see Table 6.5). Notably, without needing additional self-supervised techniques as employed by MedFinder, brat achieves comparable

performance by purely relying on text supervision. I also identify certain quality issues within the BIMCV-R dataset, which may contribute to the generally lower performance observed on this benchmark. Figure 6.17 shows how, for some images, the middle slice (depicted) is already no longer part of the lung, suggesting that the scan mainly depicts other body parts. Several images also seem to depict localizer scans, which makes it difficult to connect them to radiology reports. Appropriate processing of these images would likely lead to significant performance improvements. Despite these limitations, the results show that brat can be effectively applied off-the-shelf to other medical imaging modalities with complex visuals and lengthy reports. I note a greater discrepancy in mean rank on BIMCV-R, but limited methodological details in the original benchmark (Chen et al., 2024a) and unavailable model weights make direct comparison difficult, leaving some aspects open to further examination.

Methods	Text to Image					Image to Text				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CLIP4clip	0.003	0.015	0.022	717.0	735.9	0.003	0.008	0.015	722.0	738.7
3D-MIR	0.011	0.047	0.103	121.1	152.3	0.012	0.040	0.088	134.9	162.4
MedFinder (Resnet-50)	0.028	0.087	0.203	68.9	81.3	0.029	0.088	0.197	71.2	80.7
MedFinder (ViT-base)	0.027	0.089	0.214	75.4	80.1	0.027	0.090	0.203	72.3	81.9
Q-Former	0.007	0.025	0.048	223.0	371.7	0.000	0.015	0.034	225.0	365.8
brat	0.030	0.109	0.165	71.0	283.0	0.036	0.103	0.182	67.0	282.0

Table 6.5: Evaluation results for text-to-image and image-to-text retrieval on BIMCV-R, a lung CT dataset. The compared results, except Q-Former, are taken from Chen et al. (2024a). It is unclear how their median ranks happened to be reported as non-discrete values.

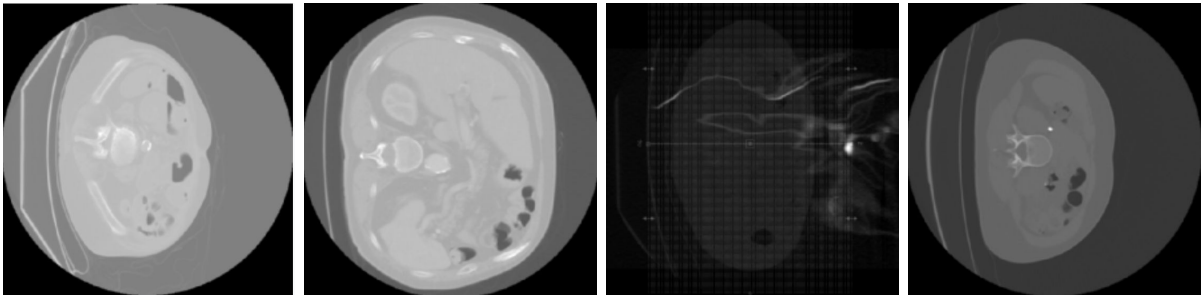


Figure 6.17: BIMCV-R example images of localizer scans or where the middle slice is already in the abdomen or pelvis.

6.4.2 Downstream Tasks

This section showcases how our pre-training is beneficial for a wide range of downstream tasks.

Brain MRI Report Generation VLP naturally suits radiology report generation, as it trains visual embeddings to align with text features. To evaluate the pre-trained backbones, I freeze

the vision backbone and assess how well a language model can extract image-grounded information from it. I use Llama-3.2-1B (Grattafiori et al., 2024), providing either multi-view embeddings or image feature maps to the LLM via a bridging MLP. Training and evaluation occur on the MSKBrain dataset, and I provide both LLM-based metrics (GREEN metric (Ostmeier et al., 2024)) and natural language metrics. I compare brat to training from scratch, Q-Former pre-training, and classification-based pre-training (“CLS”), using either the vision backbone or multi-view embeddings as LLM input. The results are in Table 6.6. We can see that the vision-language pre-training leads to significant improvements over no pre-training or classification pre-training. brat also leads to additional improvements over simple Q-Former pre-training for all NLG metrics. This is the first work to provide report generation capabilities for brain MRIs that were trained on a large-scale dataset. It is shown that, overall, the ability of VLMs to generate reports for brain MRIs is in line with other radiographic modalities, such as chest X-rays. Figure 6.18 shows examples of generated reports.

For the GREEN metric, we see sizeable gains from general vision-language pre-training but no big difference between traditional Q-Former training versus the brat framework. The GREEN metric provides a structured clinical assessment by identifying key radiology report errors derived from expert evaluations. However, it was developed mainly on chest X-ray reports, and therefore, its applicability to out-of-distribution modalities is limited. The authors evaluated their metric on an abdomen CT dataset and found a high absolute error (5.31). I provide this metric as it can offer a rough assessment of the clinical correctness, however, it is likely not well suited to assess minor performance differences in brain MRI reports. An inspection of the generated evaluations by GREEN confirms that they contain many errors. A more dependable approach would involve direct human evaluation or leveraging a stronger LLM such as GPT-4 for assessment.

Alzheimer Classification To investigate whether our pre-training generalizes to non-cancer-focused brain MRI datasets, I evaluated brat on ADNI (Petersen et al., 2010), a brain MRI dataset investigating the progression of Alzheimer’s disease. Unfortunately, at the time of running the experiments (and still at the time of writing this), the LONI Image and Data Archive was undergoing refactoring¹. This prevented me from following the same processing and train splits used in other works. Instead, I created random patient-wise splits (training, $n = 1,932$;

¹<https://groups.google.com/g/clinica-user?pli=1>

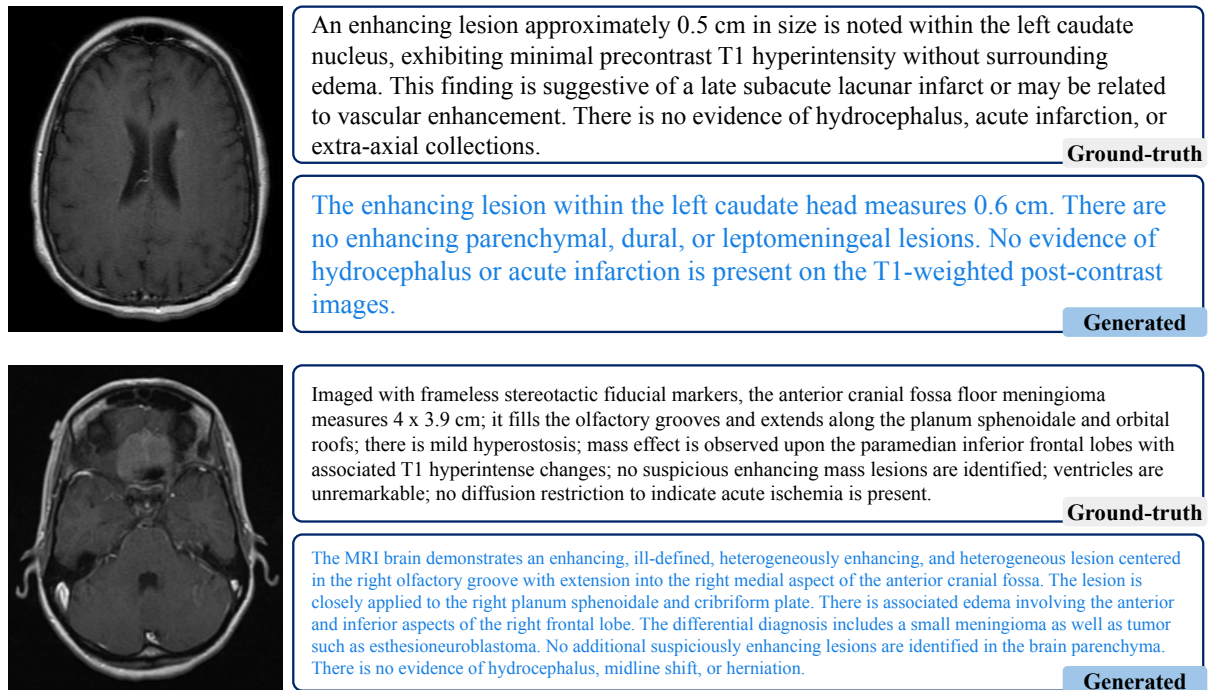


Figure 6.18: Two example MRI reports generated by our VLM with a brat backbone. The reports largely capture the correct image findings.

validation, $n = 384$; and hold-out test, $n = 291$) and focused only on evaluating the benefits of the brat pre-training compared to other weight initializations. Brain MRIs are either “cognitive normal”, “mild cognitive impairment” (MCI), or “Alzheimer’s disease”. In Figure 6.19, performance on binary classification (Normal or Alz.) is shown for 1, 10, and 100% training data. As results on ADNI vary significantly based on random seeds and selected subsets of the training data, I launched 10 runs for each setting and bootstrap from these results to obtain 95% confidence intervals. Table 6.7 contains more extensive results. It shows that our pre-training leads to consistent and significant improvements across all settings. Vision-language pre-training, in general, leads to significant performance improvements over classification pre-training. These results underline that the brat pre-training framework and MSKBrain dataset are a great starting point for tackling brain MRI problems across a wide range of applications.

Segmentation Tasks I also evaluated brat on one of the most common downstream applications in brain MRI analysis: tumour segmentation. I use BraTS2021 (Baid et al., 2021), containing gliomas, and BraTS2023-METS (Moawad et al., 2023), containing brain metastases. To isolate the benefit of the brat pre-training framework, I only used T1 modalities and did not include many of the post-processing steps typically included for these datasets, such as patch-based processing. I followed the conventional approach for this dataset in doing 4-

Backbone	Pre-training	GREEN (LLM Eval)					NLG Metrics						
		All	FP	FN	Location	Severity	METEOR	CIDEr	Rouge-L	Bleu-1	Bleu-2	Bleu-3	Bleu-4
Densenet-121	None	0.307	0.142	0.138	0.837	0.925	0.117	0.039	0.096	0.180	0.177	0.103	0.065
Densenet-121	Classification	0.319	0.114	0.430	0.890	0.948	0.102	0.049	0.069	0.187	0.124	0.072	0.048
Densenet-121	Q-Former	0.375	0.138	0.287	0.840	0.911	0.131	0.079	0.115	0.216	0.201	0.123	0.081
Densenet-121	brat	0.362	0.131	0.316	0.851	0.920	0.134	0.098	0.134	0.214	0.241	0.142	0.091
Q-Former	Q-Former	0.358	0.131	0.281	0.860	0.911	0.132	0.090	0.125	0.214	0.220	0.132	0.087
Q-Former	brat	0.373	0.141	0.305	0.830	0.917	0.128	0.114	0.114	0.219	0.197	0.121	0.081

Table 6.6: Report generation results. The backbone is always frozen, except for “None” pre-training. The GREEN metric is obtained using a 7B parameter LLM. Four GREEN scores relating to false findings (FP), missing findings (FN), false findings (FP), and accuracy of severity and location specification of findings are provided. Two additional metrics used in GREEN, missing or hallucinated references to prior scans, are omitted as I removed these references from our dataset, and therefore, our models all score 100% on them.

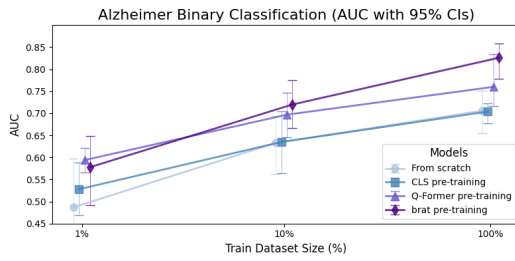


Figure 6.19: Comparison of brat pre-training to alternative pre-training methods for Alzheimer classification on ADNI.

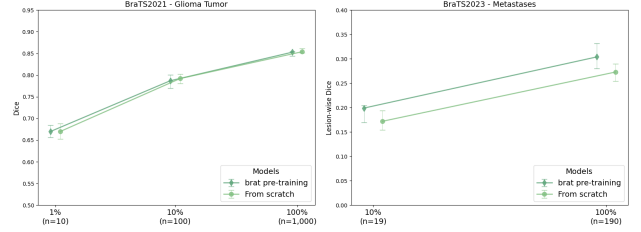


Figure 6.20: Comparison of brat pre-training to random initialization for tumour and metastases segmentation on BraTS2021 and BraTS2023. Scores are averaged across the three tumour regions.

Pre-training Approaches		1% Training Data (n=19)				10% Training Data (n=193)				100% Training Data (n=1,932)			
Vision Model M	Weight Init.	Alz.	Normal	MCI	μ	Alz.	Normal	MCI	μ	Alz.	Normal	MCI	μ
Densenet-121	Random	0.523	0.513	0.527	0.521 [0.495, 0.547]	0.640	0.560	0.498	0.567 [0.535, 0.596]	0.724	0.629	0.535	0.629 [0.608, 0.649]
	CLS	0.514	0.517	0.511	0.514 [0.487, 0.538]	0.614	0.598	0.523	0.578 [0.555, 0.602]	0.720	0.628	0.556	0.635 [0.612, 0.650]
	Q-Former	0.565	0.525	0.486	0.526 [0.506, 0.547]	0.688	0.627	0.550	0.623 [0.604, 0.640]	0.747	0.662	0.581	0.663 [0.651, 0.681]
	brat	0.560	0.559	0.505	0.543 [0.497, 0.579]	0.720	0.644	0.518	0.628 [0.606, 0.653]	0.793	0.687	0.505	0.661 [0.650, 0.672]
ResNet-50	Random	0.497	0.566	0.541	0.535 [0.497, 0.569]	0.516	0.529	0.541	0.530 [0.498, 0.561]	0.590	0.525	0.528	0.548 [0.514, 0.586]
	brat	0.527	0.531	0.532	0.530 [0.500, 0.556]	0.621	0.456	0.452	0.510 [0.490, 0.532]	0.636	0.542	0.533	0.569 [0.519, 0.612]
ViT	Random	0.517	0.485	0.473	0.492 [0.471, 0.512]	0.554	0.491	0.502	0.515 [0.495, 0.532]	0.528	0.473	0.515	0.505 [0.476, 0.533]
	brat	0.518	0.498	0.460	0.491 [0.458, 0.523]	0.607	0.555	0.467	0.543 [0.522, 0.561]	0.622	0.521	0.450	0.531 [0.513, 0.551]

Table 6.7: Evaluation results (AUC scores) for different initialisations using 1%, 10%, and 100% of training data. “Alz.” stands for Alzheimer’s disease, and “MCI” stands for mild cognitive impairment. The column μ is the average of the per-class AUC scores computed on the balanced test set; only this column displays the confidence interval.

fold cross-validation (Isensee et al., 2024). However, I also used three random seeds for each run to obtain confidence intervals. I used the benchmark appropriate evaluation metrics Dice (BraTS2021) and lesion-wise Dice (BraTS2023-METS), given for three overlapping regions: whole tumour, tumour core, and enhancing tumour. Figure 6.20 shows that our pre-training improves performance for the metastases but not for the gliomas. Here, the performance is more or less equivalent for both set-ups across all data subsets. This may be explained by the

fact that gliomas are generally large and visible to even non-experts, meaning that the core task emphasizes separating pixels in an exact way rather than requiring an anatomical understanding of brain MRIs. More detailed results are provided in Table 6.8 and 6.9.

Pre-training Approaches		1% Training Data (n=10)			10% Training Data (n=100)			100% Training Data (n=1000)		
Vision Model M	Weight Init.	WT	TC	ET	WT	TC	ET	WT	TC	ET
Densenet-121	Random	0.780 [0.768, 0.792]	0.646	0.585	0.875 [0.871, 0.879]	0.791	0.710	0.922 [0.901, 0.905]	0.854	0.761
Densenet-121	brat	0.796 [0.788, 0.805]	0.633	0.580	0.870 [0.862, 0.877]	0.785	0.707	0.925 [0.900, 0.905]	0.867	0.762

Table 6.8: Brats2021 glioma segmentation performance (Dice scores) for different pre-training initializations using 1%, 10%, and 100% of the training data. WT stands for whole tumour, TC for tumour core, and ET for enhanced tumour. 95% confidence intervals are provided for WT.

Pre-training Approaches		10% Training Data (n=20)			100% Training Data (n=190)		
Vision Model M	Weight Init.	Whole Tumor	Tumor Core	Enhanced Tumor	Whole Tumor	Tumor Core	Enhanced Tumor
Densenet-121	Random	0.176 [0.148, 0.204]	0.176 [0.161, 0.196]	0.163 [0.150, 0.183]	0.325 [0.317, 0.334]	0.254 [0.227, 0.277]	0.238 [0.217, 0.256]
Densenet-121	brat	0.206 [0.166, 0.279]	0.201 [0.175, 0.241]	0.189 [0.169, 0.218]	0.361 [0.341, 0.388]	0.288 [0.258, 0.319]	0.262 [0.239, 0.287]

Table 6.9: Brats2023-METS brain metastases segmentation performance (Lesion-wise Dice scores) for different pre-training initializations using 10% and 100% of the training data. The values correspond to the lesion-wise Dice scores for the Whole Tumor, Tumor Core, and Enhanced Tumor regions, with 95% confidence intervals in brackets.

6.5 Conclusion

Earlier chapters have shown that NLEs can convey human reasoning to AI models and vice versa; AI models can convey their reasoning in a human-understandable way. In this chapter, I used whole radiology reports and showed that they can be used to encode expert understanding of brain MRIs into AI models. In order to handle the complex 3D images and long, semantically rich reports, two new ideas are introduced: multi-view embeddings adopted from document retrieval matched via PVA and DPPs to maximize the quality-diversity of these embeddings. Various ablations and downstream evaluations show the benefit of this approach over conventional vision-language pre-training methods. The proposed brat framework is architecture-agnostic and compatible with a variety of image and text encoders. The flexibility of the learnable multi-view embeddings also naturally allows the extension of the input beyond imaging data. This is promising for medical image analysis, where patient context and lab results can provide crucial cues for diagnosis.

7 | Conclusions and Implications

This thesis starts by establishing a robust framework for evaluating natural language explanations (NLEs) in visual reasoning tasks, particularly vision-language tasks. The work has enabled unified evaluation protocols by introducing the e-ViL benchmark and the associated e-SNLI-VE-2.0 dataset, which contributed significantly to subsequent research in the field. Benchmarking remains a critical element driving the advancement in modern AI models (Hendrycks et al., 2020). Following this, I expanded the application of NLEs into medical imaging, specifically chest X-rays, which highlighted language as a bidirectional reasoning interface between AI and humans. Language can both embed human reasoning within AI models (e.g., constraining classifiers to generate NLEs) and facilitate communication from AI to humans. To understand how an AI that communicates in natural language affects users in safety-critical settings, I conducted a large-scale empirical study. This revealed a crucial issue: while clinicians prefer AI models that provide NLEs, such preferences can dangerously lead to overreliance and errors in clinical decision-making. Lastly, the thesis generalized these concepts to complete radiology reports and proposes a novel vision-language pre-training method tailored for complex text-image pairs. Below, I will summarize the results from each chapter and discuss future research directions.

In Chapter 3, a novel approach was presented for creating NLE datasets through the automated integration of vision-language (VL) and NLP resources. The resulting dataset achieved comparable quality to those manually annotated by humans and has since gained widespread use. Leveraging advancements in vision-language pre-training (VLP) methods at the time, the proposed model, e-UG, set a new state-of-the-art benchmark in NLE generation tasks. Moreover, the extensive analysis of automatic NLG metrics highlighted their limitations for assessing NLEs, which influenced subsequent research to focus more heavily on human evaluations, often using the reproducible human evaluation framework proposed in this chapter.

Recent advances in large language models (LLMs) that occurred after this work was completed in 2021 show that LLMs may be capable of outperforming crowd workers in certain annotation tasks (Gilardi et al., 2023). This enables larger, cheaper, and potentially more reliable annotation efforts for future dataset-creation processes. The dataset processing in Chapter 6 was conducted purely with the help of LLMs. Additionally, emerging LLM-based evaluation

methods could replace existing human evaluation frameworks (Zheng et al., 2023).

Chapter 4 demonstrated the feasibility of distilling human reasoning from radiology reports to create new datasets of NLEs for chest X-ray analysis. These datasets enable the training of AI models that can explain their decision-making the same way humans would, allowing them to explain complex reasoning in an easy-to-understand way. Benchmarking existing approaches on this dataset with expert clinicians highlighted important inter-annotator agreement challenges, underscoring task ambiguity and leading to model performances approaching human-level accuracy. These learnings motivated a second, improved version of the dataset, with added patient context and reduced hallucinations. Besides allowing AI models to communicate their decision-making to users, constraining the models with human reasoning also enhanced classification accuracy. Evaluations on external datasets, where multiple clinicians validated the labels to reduce ambiguity, confirmed significant performance improvements. This establishes the potential of language to be a helpful interface between AI and humans in both directions, *humans* $\rightarrow$ *AI* and *AI* $\rightarrow$ *humans*.

Modern reasoning research is largely focused on language-based reasoning, particularly scientific domains, with limited exploration in multi-modal and medical contexts. The lack of exploration in medical imaging may be due to the scarcity of datasets and pioneering works, which this work aims to address. Whilst I show that a very naive implementation of CoT already leads to performance improvements, more advanced CoT methods should be explored, such as giving the model more time to think (Hao et al., 2024) or using reinforcement learning approaches (Liu et al., 2024). Additionally, the critical issue of faithfulness, i.e., ensuring explanations genuinely reflect model decision-making, remains underexplored. The provided bounding boxes in MIMIC-NLE-v2 offer a starting point for models to provide such guarantees by explicitly grounding their reasoning in the image. This chapter has also highlighted the inherent ambiguity in medical imaging tasks such as chest X-ray analysis. This is a further argument advocating for the use of NLEs during classification tasks: they can capture this ambiguity by conveying uncertainty, contextualizing findings or listing differential diagnoses.

Chapter 5 examined the impact of NLEs on how clinicians interact with AI compared to other XAI methods. The findings were alarming: while confirming prior research that clinicians prefer NLE-based AI systems (Tonekaboni et al., 2019), these systems increase overreliance, resulting in significantly more incorrect diagnoses. NLEs are high-level abstractions of AI

reasoning, whereas saliency maps are lower-level abstractions. Challenging prior research that suggested that feature attribution methods such as saliency maps are lacking and confusing to users (Kaur et al., 2020), we find that combining NLEs with saliency maps mitigates the negative effects of both without increasing decision-making time.

While combining NLEs with saliency maps provides a certain degree of insight into the faithfulness of a model (it could highlight that an NLE does not align at all with the heat map), future research should prioritize integrated faithfulness techniques, i.e., models trained explicitly to align their explanations faithfully with underlying decision processes. Studying human interaction with such NLEs could clarify whether a lack of faithfulness, a desired XAI property for safer AI use, is why NLEs are so misleading. More research is also required to understand *why* NLEs are misleading humans, exploring cognitive biases, the humanizing aspect of language, and linguistic nuances such as uncertainty, assertiveness, and cultural variations. This preliminary study also highlights the need for extensive longitudinal and clinical studies, with the ultimate goal of finding a balance between AI that is safe and AI that is helpful.

Chapter 6 generalized the idea of encoding human knowledge in AI models by training vision models from scratch using only radiology reports. First, I proposed a dataset that is orders of magnitude larger than existing public brain MRI datasets. The reports in this dataset were processed purely using LLMs, showcasing the relative ease of creating such datasets with the advent of powerful LLMs. Next, I developed a new vision-language pre-training framework that draws parallels between document retrieval and medical image-text retrieval to address the complexities of long clinical reports. The method is based on learnable latents, the pairwise view alignment algorithm, and determinantal point processes that mitigate latent collapse. The proposed method significantly outperformed conventional vision-language models (e.g., CLIP, Q-Former) on diverse pre-training and downstream tasks on 3D medical images. Releasing this suite of pre-trained models showcases how large-scale datasets can benefit research despite patient privacy regulations hindering sharing such datasets.

It is planned to extend this research by refining the models to incorporate patient histories, multiple prior scans, and diverse MRI modalities and ultimately understand how it can be deployed in clinical practice. With the rapid increase in the number of diagnostic imaging performed compared to the number of radiologists, models that can reliably map medical images to radiology reports could assist in reducing the time and effort needed to perform diagnosis (Kon-

[stantinidis, 2024](#)). Building upon the insights from our initial user study, a new study could explore how the interface between AI and radiologists can be optimized to enable safe and productive use.

The integration of language data into medical image representation learning raises important ethical considerations that impact both clinical practice and AI development. Unfiltered text sources, such as radiology reports, inherently carry biases reflective of individual clinicians or institutional practices. Without deliberate precautions, models internalize and amplify these biases, which can lead to systematic disparities in clinical outcomes ([Obermeyer et al., 2019](#); [Mehrabi et al., 2021](#)). Moreover, language usage varies significantly across cultural contexts and even among individual healthcare facilities, meaning that multi-center datasets are essential to mitigate demographic and institutional biases ([Zech et al., 2018](#)). Another issue is that radiology reports can be high in noise as they are typically authored by a single clinician, without additional validation by peers and seldom updated based on subsequent clinical insights or corroborating laboratory results ([Brady, 2017](#)). Lastly, as our user study has shown, AI communicating via language introduces a wide range of issues that must be prioritized in the research community ([Gabriel et al., 2024](#)).

In conclusion, this thesis advocates for the use of language as a tool to train better AI models and facilitate effective communication between AI systems and humans. This is especially important in clinical practice, where data, in general, is scarce, but free-text reports are routinely collected and easy to obtain. While this thesis has consistently demonstrated the clear advantages of using language to guide AI model training (*humans* $\rightarrow$ *AI* interface), it also reveals substantial risks associated with AI models that use language as an interface to communicate with humans (*AI* $\rightarrow$ *humans* interface).

Bibliography

- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Julius Adebayo, Michael Muelly, Ilaria Lliccardi, and Been Kim. Debugging Tests for Model Explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. Plausibility: On the (Un) Reliability of Explanations From Large Language Models. *arXiv preprint arXiv:2402.04614*, 2024.
- Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining Human Expertise With Artificial Intelligence: Experimental Evidence From Radiology. Working Paper 31422, National Bureau of Economic Research, 2023.
- Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, 2021.
- Jong Seok Ahn, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, et al. Association of Artificial Intelligence–Aided Chest Radiograph Interpretation With Reader Performance and Efficiency. *JAMA Network Open*, 2022.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, et al. Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propo-

- sitional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398, 2016.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural Module Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
- Vincent Arel-Bundock, Noah Greifer, and Andrew Heiss. How to Interpret Statistical Models Using MarginalEffects for R and Python. *Journal of Statistical Software*, 111:1–32, 2024.
- Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3(6): e200267, 2021.
- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness Tests for Natural Language Explanations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)*, 2023.
- Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv preprint arXiv:2107.02314*, 2021.
- Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John V Guttag, and Adrian V Dalca. Voxel-

- Morph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2018.
- Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation With Improved Correlation With Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.
- Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, et al. Maira-2: Grounded Radiology Report Generation. *arXiv preprint arXiv:2406.04449*, 2024.
- Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-Shot Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.
- Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations (ICLR)*, 2022.
- Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2022.
- Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. Interpretable Mammographic Image Classification Using Case-Based Reasoning and Deep Learning. *arXiv preprint arXiv:2107.05605*, 2021.
- P.J. Basser, J. Mattiello, and D. LeBihan. MR Diffusion Tensor Spectroscopy and Imaging. *Biophysical Journal*, 66(1):259–267, 1994.
- Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32:267–270, 2004.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–21. Springer, 2022.
- Aaron M Bornstein. Is Artificial Intelligence Permanently Inscrutable. *Nautilus*, 40(1), 2016.

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.
- Adrian P Brady. Error and Discrepancy in Radiology: Inevitable or Avoidable? *Insights into Imaging*, 8(1):171–182, 2017.
- Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. The Persuasive Power of Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2024.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.
- Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, et al. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv preprint arXiv:2004.07213*, 2020.
- Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference On Healthcare Informatics*, 2015.
- Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A Large Chest X-Ray Image Dataset With Multi-Label Annotated Reports. *Medical Image Analysis*, 66, 2020.
- Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, et al. Human-Centered Tools for Coping With Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019a.
- Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, 2019b.
- Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. Assertiveness-Based Agent Commu-

- nication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural Language Inference With Natural Language Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. The Struggles of Feature-Based Explanations: Shapley Values vs. Minimal Sufficient Subsets. *arXiv preprint arXiv:*, 2020.
- Rich Caruana. Multitask Learning. *Machine Learning*, 28(1), 1997.
- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative Language Understanding Through Textual Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7139–7159, 2022.
- Angel X. Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, et al. Matterport3D: Learning From RGB-D Data in Indoor Environments. In *International Conference on 3D Vision (3DV)*, pages 667–676, 2017.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021.
- Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, et al. Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 529–539, 2020.
- Vinod Kumar Chauhan, Anshul Thakur, Odhran O’Donoghue, Omid Rohanian, Soheila Mo-laei, and David A Clifton. Continuous Patient State Attention Model for Addressing Irregularity in Electronic Health Records. *BMC Medical Informatics and Decision Making*, 2024.
- Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This

- Looks Like That: Deep Learning for Interpretable Image Recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.
- Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large Language Models Are Visual Reasoning Coordinators. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:70115–70140, 2023a.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, 2020a.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *International Conference on Learning Representations (ICLR)*, 2023b.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal Image-Text Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020b.
- Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-R: A Landmark Dataset for 3d Ct Text-Image Retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024a.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating Radiology Reports via Memory-Driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020c.
- Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024b.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, et al. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying Vision-and-Language Tasks via

- Text Generation. In *International Conference on Machine Learning (ICML)*, pages 1931–1942, 2021.
- Kyunghyun Cho, Bart van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.
- Eric Chu, Deb Roy, and Jacob Andreas. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. *arXiv preprint arXiv:2007.12248*, 2020.
- Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, et al. TorchXRyVision: A Library of Chest X-Ray Datasets and Models. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, pages 231–249, 2022.
- Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Mark G. Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building Explainable Artificial Intelligence Systems. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2*, 2006.
- Kailas Dayanandan and Brejesh Lall. Enabling Multi-Modal Conversational Interface for Clinical Imaging. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2024.
- J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- Karan Desai and Justin Johnson. VirTex: Learning Visual Representations From Textual Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Kristen DeStigter, Kara-Lee Pool, Abimbola Leslie, Sarwat Hussain, Bien Soo Tan, Lluís

- Donoso-Bach, and Savvas Andronikou. Optimizing Integrated Imaging Service Delivery by Tier in Low-Resource Health Systems. *Insights into Imaging*, 12:1–11, 2021.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NACCL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.
- Nicola K. Dinsdale, Mark Jenkinson, and Ana IL Namburete. Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal. *NeuroImage*, 228:117689, 2021.
- Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. E-Snli-Ve: Corrected Visual-Textual Entailment With Natural Language Explanations. *arXiv preprint arXiv:2004.03744*, 2020.
- Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, et al. Clip Itself Is a Strong Fine-Tuner: Achieving 85.7% and 88.0% Top-1 Accuracy With Vit-B and Vit-L on Imagenet. *arXiv preprint arXiv:2212.06138*, 2022.
- Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2020.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, et al. PaLM-E: An Embodied Multimodal Language Model. In *International Conference on Machine Learning (ICML)*, pages 8469–8488, 2023.
- Yuhan Du, Anna Markella Antoniadi, Catherine McNestry, Fionnuala M McAuliffe, and Catherine Mooney. The Role of XAI in Advice-Taking From a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Applied Sciences*, 2022.

- Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. Beyond VQA: Generating Multi-Word Answers and Rationales to Visual Questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1623–1632, 2021.
- Upol Ehsan and Mark O Riedl. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference*, 2020.
- Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.
- Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, et al. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*, 1(1):12, 2021.
- Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, et al. A Guide to Deep Learning in Healthcare. *Nature Medicine*, 25(1):24–29, 2019.
- Kraig Finstad. Response Interpolation and Scale Sensitivity: Evidence Against 5-point Scales. *Journal of usability studies*, 5(3):104–110, 2010.
- Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, et al. Who Goes First? Influences of Human-Ai Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- Maria Frasca, Davide La Torre, Gabriella Pravettoni, and Ilaria Cutica. Explainable and Interpretable Artificial Intelligence in Medicine: A Systematic Bibliometric Review. *Discover Artificial Intelligence*, 2024.

- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 457–468, 2016.
- Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, et al. The Ethics of Advanced AI Assistants. *arXiv preprint arXiv:2404.16244*, 2024.
- William Gale et al. Producing Radiologist-Quality Reports for Interpretable Artificial Intelligence. *arXiv preprint arXiv:1806.00340*, 2018.
- Albert Gatt and Emiel Krahmer. Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.
- Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, et al. Do as AI Say: Susceptibility in Deployment of Clinical Decision-Aids. *NPJ Digital Medicine*, 2021.
- Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, et al. Non-Task Expert Physicians Benefit From Correct Explainable AI Advice When Reviewing X-Rays. *Scientific Reports*, 2023.
- Marzyeh Ghassemi, Mahima Pushkarna, James Wexler, Jesse Johnson, and Paul Varghese. Clinicalvis: Supporting Clinical Task-Focused Design Evaluation. *arXiv preprint arXiv:1810.05798*, 2018.
- Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, 3(11), 2021.
- Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-Based Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT Outperforms Crowd Workers

for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023.

Jacob Gildenblat and contributors. PyTorch Library for CAM Methods. <https://github.com/jacobgil/pytorch-grad-cam>, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

V. Grau, A.U.J. Mewes, M. Alcaniz, R. Kikinis, and S.K. Warfield. Improved watershed transform for medical image segmentation using prior information. *IEEE Transactions on Medical Imaging*, 23(4):447–458, 2004.

Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression Concept Vectors for Bidirectional Explanations in Histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132. 2018.

Xiangtai Gu, Jun Yan, Liang Pan, Yixiao Lu, Zejia Chen, Tao Huang, Han Wu, Lijuan Wang, et al. VLMo: Unified Vision-Language Pre-Training With Mixture-of-Modality-Experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–14, 2022.

Varun Gulshan, Lily Peng, Marc Coram, Michael C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 2016.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic Bias: From Discrimination

- Discovery to Fairness-Aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2125–2126, 2016.
- Joseph V Hajnal, David J Bryant, Larry Kasuboski, Pradip M Pattany, Beatrice De Coene, Paul D Lewis, Jacqueline M Pennock, Angela Oatridge, et al. Use of Fluid Attenuated Inversion Recovery (FLAIR) Pulse Sequences in MRI of the Brain. *Journal of Computer Assisted Tomography*, 16(6):841–844, 1992.
- Ameer Hamza, Yong Hyun Ahn, Sungyoung Lee, Seong Tae Kim, et al. LLaVA Needs More Knowledge: Retrieval Augmented Natural Language Generation With Knowledge Graph for Explaining Thoracic Pathologies. *arXiv preprint arXiv:2410.04749*, 2024.
- Zhongyi Han, Benzhen Wei, Xiaoming Xi, Bo Chen, Yilong Yin, and Shuo Li. Unifying Neural Learning and Symbolic Reasoning for Spinal Medical Report Generation. *Medical Image Analysis*, 67, 2021.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space. *arXiv preprint arXiv:2412.06769*, 2024.
- Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language Models Are General-Purpose Interfaces. *arXiv preprint arXiv:2206.06336*, 2022.
- Md Inzamam Ul Haque, Abhishek K Dubey, Ioana Danciu, Amy C Justice, Olga S Ovchinnikova, and Jacob D Hinkle. Effect of Image Resolution on Automated Classification of Chest X-Rays. *Journal of Medical Imaging*, 10(4):044503–044503, 2023.
- Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.
- Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 2024.
- M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, and P.M.

- Jodoin. Brain Tumor Segmentation With Deep Neural Networks. *Medical Image Analysis*, 35:18–31, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.
- William R Hendee and Christopher J Morgan. Magnetic Resonance Imaging Part I—Physical Principles. *Western Journal of Medicine*, 141(4):491, 1984.
- Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating Visual Explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR)*, 2020.
- Bernease Herman. The Promise and Peril of Human Evaluation for Model Interpretability. *arXiv preprint arXiv:1711.07414*, 2017.
- Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. This Looks Like That... Does It? Shortcomings of Latent Space Prototype Interpretability in Deep Networks. *arXiv preprint arXiv:2105.02968*, 2021.

- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, et al. TRUE: Re-Evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NACCL): Human Language Technologies*, 2022.
- RM Hopstaken, T Witbraad, JMA Van Engelshoven, and GJ Dinant. Inter-Observer Variation in the Interpretation of Chest Radiographs for Pneumonia in Community-Acquired Lower Respiratory Tract Infections. *Clinical Radiology*, 59(8):743–752, 2004.
- Benjamin Hou, Georgios Kaissis, Ronald M. Summers, and Bernhard Kainz. RATCHET: Medical Transformer For Chest X-Ray Diagnosis and Reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022a.
- Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling Up Vision-Language Pre-Training for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989, 2022b.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Jia-Hong Huang, C.-H. Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I.-Hung Lin, Kang Wang, et al. DeepOpht: Medical Report Generation for Retinal Images via Deep Models and Visual Explanation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021a.
- Shih-Cheng Huang, Liye Shen, Matthew P. Lungren, and Serena Yeung. GLORIA: A Multi-modal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021b.
- Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan

- Yu, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.
- Zhicheng Huang, Zhaoyang Zeng, Jianqiang Fu, Jianlong Fu, and Junqiu Guo. Pixel-Bert: Aligning Image Pixels With Text by Deep Multi-Modal Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12262–12270, 2021c.
- Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, et al. Maira-1: A Specialised Large Multimodal Model for Radiology Report Generation. *arXiv preprint arXiv:2311.13668*, 2023.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yu Yu, Ilcus Silviu Ciurea, Christopher Chute, Hedvig Marklund, Bahram Haghighi, et al. CheXpert: A Large Chest Radiograph Dataset With Uncertainty Labels and Expert Comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 590–597, 2019.
- Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnU-net Revisited: A Call for Rigorous Validation in 3d Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.
- Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General Perception With Iterative Attention. In *International Conference on Machine Learning (ICML)*, 2021.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, et al. RadGraph: Extracting Clinical Entities and Relations From Radiology Reports. In *NeurIPS (Datasets and Benchmarks Track)*, 2021.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompting: Modifying Input Images to Improve Performance of Frozen Vision Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Jue Jiang, Neelam Tyagi, Kathryn Tringale, Christopher Crane, and Harini Veeraraghavan. Self-Supervised 3D Anatomy Segmentation Using Self-Distilled Masked Image Transformer (SMIT). In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 556–566, 2022.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a De-Identified Publicly Available Database of Chest Radiographs With Free-Text Reports. *Scientific Data*, 2019.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, et al. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Scientific Data*, 10(1):1, 2023.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017.
- Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning Visual Features From Large Weakly Supervised Data. In *European Conference on Computer Vision (ECCV)*, pages 67–84, 2016.
- Konstantinos Kamnitsas, Christian Ledig, V.F.J. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, and B. Glocker. Efficient Multi-Scale 3D CNN With Fully Connected CRF for Accurate Brain Lesion Segmentation. *Medical Image Analysis*, 36:61–78, 2017.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

- Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
- Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1244–1254, 2021.
- Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartłomiej Papiez, and Thomas Lukasiewicz. Explaining Chest X-Ray Pathologies in Natural Language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 701–713, 2022.
- Maxime Kayser, Oana-Maria Camburu, and Thomas Lukasiewicz. Can Large Language Models Reason About Chest X-Rays? In *Extended Abstract presented at Machine Learning for Health (ML4H) Symposium*, 2023.
- Maxime Kayser, Bayar Menzat, Cornelius Emde, Bogdan Bercean, Alex Novak, Abdalá Morgado, Bartłomiej Papiez, Susanne Gaube, et al. Fool Me Once? Contrasting Textual and Visual Explanations in a Clinical Decision-Support Setting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18891–18919, 2024a.
- Maxime Kayser, Aneesh Rangnekar, Aleksandr Petrov, Harini Veeraraghavan, and Nathaniel Swinburne. A Copilot for Radiology: Vision-Language Models Achieve Accurate Brain MRI Interpretation. In *Abstract presented at RSNA*, 2024b.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.
- Omar Khattab and Matei Zaharia. Colbert: Efficient and Effective Passage Search via Contextualized Late Interaction Over Bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020.

- Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677, 2018a.
- Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15719–15728, 2021.
- Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, 2018b.
- Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The CoT Collection: Improving Zero-Shot and Few-Shot Learning of Language Models via Chain-of-Thought Fine-Tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12685–12708, 2023.
- Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–298, 2022.
- Taenyun Kim and Hayeon Song. The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence’s Suggestion. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.
- Timo Koch, Florian Pargent, Anne-Kathrin Kleine, Eva Lermer, and Susanne Gaube. A Tutorial on Tailored Simulation-Based Power Analysis for Experimental Designs With Generalized Linear Mixed Models. *PsyArXiv Preprints*, 2023.

- Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5338–5348, 2020.
- Kleanthis Konstantinidis. The Shortage of Radiographers: A Global Crisis in Healthcare. *Journal of Medical Imaging and Radiation Sciences*, 55(4):101333, 2024.
- Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. Enhancing Modality-Agnostic Representations via Meta-Learning for Brain Tumor Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- Vasiliki Kougia, John Pavlopoulos, Panagiotis Papapetrou, and Max Gordon. RTEEX: A Novel Framework for Ranking, Tagging, and Explanatory Diagnostic Captioning of Radiography Exams. *Journal of the American Medical Informatics Association*, 28(8):1651–1659, 2021.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In *International Journal of Computer Vision (IJCV)*, volume 123, pages 32–73, 2017.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification With Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 2012.
- Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, et al. Generative Models Improve Fairness of Medical Classifiers under Distribution Shifts. *Nature Medicine*, 30(4):1166–1173, 2024.
- Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Metric Learning with Spectral Graph Convolutions on Brain Connectivity Networks. *NeuroImage*, 169:431–442, 2018.
- Alex Kulesza and Ben Taskar. Determinantal Point Processes for Machine Learning. *Machine Learning*, 5(2-3):123–286, 2012.
- Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi,

- Alexander C. Berg, and Tamara L. Berg. Babytalk: Understanding and Generating Simple Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.
- Sawan Kumar and Partha Talukdar. NILE: Natural Language Inference With Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8730–8742, 2020.
- Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Benjamin Caine, Wei Li, Abhijit Ogale, Lu-wei Zhou, Andrew M Dai, et al. MaMMUT: A Simple Architecture for Joint Learning for MultiModal Tasks. *Transactions on Machine Learning Research*, 2023.
- Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. Collective Generation of Natural Image Descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 359–368, 2012.
- Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi. Treetalk: Composition and Compression of Trees for Image Descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362, 2014.
- Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1902.00006*, 2019.
- George Lakoff and Mark Johnson. *Metaphors We Live By*. Chicago, IL, 1980.
- Rodney LaLonde, Drew Torigian, and Ulas Bagci. Encoding Visual Attributes in Capsules for Explainable Medical Diagnoses. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 294–304, 2020.
- H. Chad Lane, Mark G. Core, Michael van Lent, Steve Solomon, and Dave Gomboc. Explainable Artificial Intelligence for Training and Tutoring. In *Proceedings of the 2005 Conference*

- on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pages 762–764, 2005.
- Curtis P Langlotz. Will Artificial Intelligence Replace Radiologists? *Radiology: Artificial Intelligence*, 1(3):e190058, 2019.
- Curtis P. Langlotz et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology*, 291(3), 2019.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, et al. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702*, 2023.
- D. Le Bihan, E. Breton, D. Lallemand, P. Grenier, E. Cabanis, and M. Laval-Jeantet. MR Imaging of Intravoxel Incoherent Motions: Application to Diffusion and Perfusion in Neurologic Disorders. *Radiology*, 161(2):401–407, 1986.
- Hyebin Lee, Seong Tae Kim, and Yong Man Ro. Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis. In *International Workshop on Multimodal Learning for Clinical Decision Support*, pages 21–29, 2019.
- David Leslie. Understanding Artificial Intelligence Ethics and Safety. *arXiv preprint arXiv:1906.05684*, 2019.
- Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning Visual N-Grams From Web Data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4183–4192, 2017.
- Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1537–1547, 2018a.
- Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 6666–6673, 2019a.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, et al. Llava-Med: Training a Large Language-and-Vision As-

- sistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:28541–28564, 2023a.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022a.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, 2023b.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VISU-ALBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*, 2019b.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, et al. Grounded Language-Image Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022b.
- Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018b.
- Ping Li, Qinghao Ye, Luming Zhang, Li Yuan, Xianghua Xu, and Ling Shao. Exploring Global Diverse Attention via Pairwise Temporal Relation for Video Summarization. *Pattern Recognition*, 111:107677, 2021a.
- Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567, 2018c.
- Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. Composing Simple Image Descriptions Using Web-Scale N-Grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang,

- Houdong Hu, et al. Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. Causality Extraction Based on Self-Attentive BiLSTM-CRF With Transferred Embeddings. *Neurocomputing*, 423:207–219, 2021b.
- Q Vera Liao and Kush R Varshney. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790*, 2021.
- Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2022.
- Chin-Yew Lin. Rouge: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
- Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 605–612, 2004.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.
- Zachary C Lipton. The Doctor Just Won’t Accept That! *arXiv preprint arXiv:1711.08037*, 2017.
- Zachary C. Lipton. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery. *Queue*, 16(3):31–57, 2018.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, et al. Deepseek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024.
- Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng,

- Peter Szolovits, and Marzyeh Ghassemi. Clinically Accurate Chest X-Ray Report Generation. In *Machine Learning for Healthcare Conference*, 2019a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023a.
- Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, et al. Exploring the Boundaries of GPT-4 in Radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.
- Shikun Liu, Edward Johns, and Andrew J Davison. End-to-End Multi-Task Learning With Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019c.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017.
- Bodhisattwa Prasad Majumder, Oana Camburu, Thomas Lukasiewicz, and Julian Mcauley. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations. In *International Conference on Machine Learning (ICML)*, pages 14786–14801, 2022.
- Ana Marasović, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A Smith, and Yejin Choi. Natural Language Rationales With Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2810–2829, 2020.
- Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. Few-Shot Self-

- Rationalization with Natural Language Prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, 2022.
- Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do Concept Bottleneck Models Learn as Intended? *arXiv preprint arXiv:2105.04289*, 2021.
- Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. Deep Learning in Radiology: An Overview of the Concepts and a Survey of the State of the Art With Focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019.
- Scott M McKinney, Marcin Sieniek, Vinod Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Maria Chesus, et al. International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577(7788):89–94, 2020.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.
- Ciro Mennella, Umberto Maniscalco, Giuseppe De Pietro, and Massimo Esposito. Ethical and Regulatory Challenges of AI Technologies in Healthcare: A Narrative Review. *Heliyon*, 10(4), 2024.
- Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. Chain of Thought Utilization in Large Language Models and Application in Nephrology. *Medicina*, 60(1):148, 2024.
- Tim Miller. Explanation in Artificial Intelligence: Insights From the Social Sciences. *Artificial Intelligence*, 267:1–38, 2019.
- Marina I. Mityul et al. Interpretive Differences Between Patients and Radiologists Regarding the Diagnostic Confidence Associated With Commonly Used Phrases in the Radiology Report. *American Journal of Roentgenology*, 210(1), 2018.
- Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Leon Jekel, Raisa Amiruddin, et al. The Brain Tumor Segmentation (BraTS-Mets) Challenge 2023: Brain Metastasis Segmentation on Pre-Treatment MRI. *arXiv preprint arXiv:2306.00838*, 2023.

- Sanaz Mohammadjafari, Mucahit Cevik, Mathusan Thanabalasingam, Ayse Basar, and Alzheimer's Disease Neuroimaging Initiative. Using ProtoPNet for Interpretable Alzheimer's Disease Classification. *Proceedings of the Canadian Conference on Artificial Intelligence*, 2021.
- Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation Models for Generalist Medical Artificial Intelligence. *Nature*, 2023.
- Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 8:1–39, 2024.
- ME Moseley, J Kucharczyk, J Mintorovitch, Y Cohen, J Kurhanewicz, N Derugin, H Asgari, and D Norman. Diffusion-Weighted MR Imaging of Acute Stroke: Correlation With T2-Weighted and Magnetic Susceptibility-Enhanced MR Imaging in Cats. *American Journal of Neuroradiology*, 11(3):423–429, 1990.
- Asbjørn Munk, Jakob Ambsdorf, Sebastian Llambias, and Mads Nielsen. AMAES: Augmented Masked Autoencoder Pretraining on Public Brain MRI Data for 3d-Native Segmentation. *arXiv preprint arXiv:2408.00640*, 2024.
- Myura Nagendran, Yufeng Chen, Charles A Lovejoy, Adam C Gordon, Matthieu Komorowski, Hannah Harvey, Eric J Topol, John PA Ioannidis, et al. Artificial Intelligence Versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies. *BMJ*, 368, 2020.
- Shinichi Nakagawa. A Farewell to Bonferroni: The Problems of Low Statistical Power and Publication Bias. *Behavioral Ecology*, 2004.
- Sharan Narang et al. WT5?! Training Text-to-Text Models to Explain Their Predictions. *arXiv preprint arXiv:2004.14546*, 2020.
- Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. Automated Classification of Skin Lesions: From Pixels to Practice. *Journal of Investigative Dermatology*, 138(10):2108–2110, 2018.
- Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving

- Jigsaw Puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464):447–453, 2019.
- Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain Magnetic Resonance Imaging With Contrast Dependent on Blood Oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024–001, 2020.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, et al. In-Context Learning and Induction Heads. *arXiv preprint arXiv:2209.11895*, 2022.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khilodov, Pierre Fernandez, Daniel Haziza, et al. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*, 2023.
- Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1143–1151, 2011.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, et al. GREEN: Generative Radiology Report Evaluation and Error Notation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 374–390, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al. Training Language Models to Follow Instructions With Human Feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 27730–27744, 2022.
- Zachary Papanastasopoulos, Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Chintana Paramagul, Mark A Helvie, and Colleen H Neal. Explainable AI for Medical Imaging:

- Deep-Learning CNN Ensemble for Classification of Estrogen Receptor Status From Breast MRI. In *Medical imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 228–235, 2020.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.
- Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease Prediction Using Graph Convolutional Networks: Application to Autism Spectrum Disorder and Alzheimer’s Disease. *Medical Image Analysis*, 48:117–130, 2018.
- Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8779–8788, 2018.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the Dynamic Context of a Still Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radiolog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance. *arXiv preprint arXiv:2311.18681*, 2023.
- Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, et al. Alzheimer’s Disease Neuroimaging Initiative (ADNI) Clinical Characterization. *Neurology*, 2010.
- Samuele Lo Piano. Ethical Principles in Machine Learning and Artificial Intelligence: Cases From the Field and Possible Ways Forward. *Humanities and Social Sciences Communications*, 7(1):1–7, 2020.
- Maxim Pisov, Mikhail Goncharov, Nadezhda Kurochkina, Sergey Morozov, Victor Gombolevskiy, Valeria Chernina, Anton Vladzimirskyy, Ksenia Zamyatina, et al. Incorporating Task-Specific Structural Knowledge Into CNNs for Brain Midline Shift Detection. In *Inter-*

pretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, pages 30–38. 2019.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A. Kors. Natural Language Processing in Radiology: A Systematic Review. *Radiology*, 279(2):329–343, 2016.

Esther Puyol-Antón, Chen Chen, James R. Clough, Bram Ruijsink, Baldeep S. Sidhu, Justin Gould, Bradley Porter, Marc Elliott, et al. Interpretable Deep Models for Cardiac Resynchronisation Therapy Response Prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 284–293, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving Language Understanding by Generative Pre-Training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models Are Unsupervised Multitask Learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3342–3352, 2019.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4932–4942, 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays With Deep Learning. *arXiv preprint arXiv:1711.05225*, 2017.

Pranav Rajpurkar, Chloe O’Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, et al. CheXaid: Deep Learning Assistance for Physician Diagnosis of Tuberculosis Using Chest X-Rays in Patients With HIV. *NPJ Digital Medicine*, 2020.

- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831, 2021.
- Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-Existent Priors. In *Machine Learning for Health*, pages 456–473, 2022.
- Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, and Junaid Qadir. Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey. *Computers in Biology and Medicine*, 149:106043, 2022.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid Training Data Creation With Weak Supervision. In *Proceedings of the VLDB Endowment*, volume 11, pages 269–282, 2017.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.
- Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2(3), 2020.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- Noelia Rivera-Garrido, MP Ramos-Sosa, Michela Accerenzi, and Pablo Brañas-Garza. Continuous and Binary Sets of Responses Differ in the Field. *Scientific Reports*, 2022.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for

- Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.
- Negar Rostamzadeh, Reza Hosseini, Vivek Jain, Mahdi Yaghoubi, Soheil Kolouri, Kevin McCabe, Brian Price, Hamid R. Rabiee, et al. Fashion-Gen: The Generative Fashion Dataset and Challenge. *arXiv preprint arXiv:1806.08317*, 2018.
- Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistic Surveys*, 16:1–85, 2022.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, et al. Capabilities of Gemini Models in Medicine. *CoRR*, 2024.
- Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint arXiv:1708.08296*, 2017.
- Fawaz Sammani and Nikos Deligiannis. Uni-NLX: Unifying Textual Explanations for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4634–4639, 2023.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NACCL): Human Language Technologies*, 2022.
- Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, et al. Benchmarking Saliency Methods for Chest X-Ray Interpretation. *Nature Machine Intelligence*, 2022.
- Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, and Jordan T Ash. Streaming Active Learning With Deep Neural Networks. In *International Conference on Machine Learning (ICML)*, 2023.
- Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning Visual Representations With Caption Annotations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–170, 2020.

- James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I Can Do Better Than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, et al. Laion-400m: Open Dataset of Clip-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- Mike Schuster and Kaisuke Nakajima. Japanese and Korean Voice Search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012.
- Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using Stylegan for Visual Interpretability of Deep Learning Models on Medical Images. *arXiv preprint arXiv:2101.07563*, 2021.
- Jarrel C. Y. Seah, Cyril H. M. Tang, Quinlan D. Buchlak, Xavier G. Holt, Jeffrey B. Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, et al. Effect of a Comprehensive Deep-Learning Model on the Accuracy of Chest X-Ray Interpretation by Radiologists: A Retrospective, Multireader Multicase Study. *The Lancet Digital Health*, 2021.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7881–7892, 2020.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-Cam: Visual Explanations From Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision (ICCV)*, 2017.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words With Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725, 2016.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset for Automatic Image Captioning. In *Pro-*

ceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pages 2556–2565, 2018.

Hua Shen and Ting-Hao Huang. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2020.

Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification. *Expert Systems with Applications*, 128:84–95, 2019.

Yiqiu Shen, Yanqi Xu, Jiajian Ma, Wushuang Rui, Chen Zhao, Laura Heacock, and Chenchan Huang. Multi-Modal Large Language Models in Radiology: Principles, Applications, and Potential. *Abdominal Radiology*, pages 1–13, 2024.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, et al. Augmenting the National Institutes of Health Chest Radiograph Dataset With Expert Annotations of Possible Pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

K Simonyan and A Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

K Simonyan, A Vedaldi, and A Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR)*, 2014.

Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. Do Users Benefit From Interpretable Vision? A User Study, Baseline, and Dataset. In *International Conference on Learning Representations (ICLR)*, 2021.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020.

Anouk M Speets, Yolanda van der Graaf, Arno W Hoes, Sandra Kalmijn, Alfred PE Sachs, Matthieu JCM Rutten, Jan Willem C Gratama, Alexander D Montauban van Swijndregt,

- et al. Chest Radiography in General Practice: Indications, Diagnostic Yield and Consequences for Patient Management. *British Journal of General Practice*, 56(529):574–578, 2006.
- J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for Simplicity: The All Convolutional Net. In *International Conference on Learning Representations (workshop track)*, 2015.
- Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. SNOMED Clinical Terms: Overview of the Development Process and Project Status. In *Proceedings of the AMIA Symposium*, page 662, 2001.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-Training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations (ICLR)*, 2019.
- Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning With Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, et al. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL*, 2023.
- Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations From Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5099–5110, 2019.
- Jennifer SN Tang, Jeffrey KC Lai, John Bui, Wayland Wang, Paul Simkin, Dayu Gai, Jenny Chan, Diane M Pascoe, et al. Impact of Different Artificial Intelligence User Interfaces on Lung Nodule and Mass Detection on Chest Radiographs. *Radiology: Artificial Intelligence*, 2023.

- Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-Supervised Pre-Training of Swin Transformers for 3d Medical Image Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740, 2022.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and Explainable Region-Guided Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442, 2023.
- Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, et al. Collaboration Between Clinicians and Vision-Language Models in Radiology Report Generation. *Nature Medicine*, 31(2):599–608, 2025.
- Sarina Thomas, Qing Cao, Anna Novikova, Daria Kulikova, and Guy Ben-Yosef. EchoNarrator: Generating Natural Text Explanations for Ejection Fraction Predictions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 634–644. Springer, 2024.
- Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What Clinicians Want: Contextualizing Explainable Machine Learning For Clinical End Use. In *Machine Learning For Healthcare Conference*, 2019.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.
- Phuong Dung Yun Trieu, Melissa Barron, and Sarah J Lewis. Use of Full-quality DICOM Images Compared to Minimally Compressed Mammograms in JPEG Format for Radiology Training: A Study From Radiologist and Radiographer Perspectives. *Academic Radiology*, 30(8):1748–1755, 2023.
- Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, et al. Human–Computer Collaboration For Skin Cancer Recognition. *Nature Medicine*, 2020.
- Michael Tschannen, Manoj Kumar, Andreas Peter Steiner, Xiaohua Zhai, Neil Houlsby, and

- Lucas Beyer. Image Captioners Are Scalable Vision Learners Too. November 2023. URL [https://openreview.net/forum?id=A7feCufBhL&referrer=%5Bthe%20profile%20of%20Manoj%20Kumar%5D\(%2Fprofile%3Fid%3D~Manoj_Kumar1\)](https://openreview.net/forum?id=A7feCufBhL&referrer=%5Bthe%20profile%20of%20Manoj%20Kumar%5D(%2Fprofile%3Fid%3D~Manoj_Kumar1)).
- Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, et al. Towards Generalist Biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024.
- Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.
- Gaël Varoquaux and Veronika Cheplygina. Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future. *NPJ Digital Medicine*, 5(1):48, 2022.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
- Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine Learning in Medicine: Addressing Ethical Challenges. *PLoS Medicine*, 15(11), 2018.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.
- Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual Explanations for Machine Learning: A Review. *arXiv preprint arXiv:2010.10596*, 2020.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without

- Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.
- Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.
- Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Jiahui Wang, Adams Wei Yu, Tao Li, Zihang Dai, Garrison Cottrell, Dongdong Yu, Yang Song, and Quoc V. Le. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*, 2021.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mahdi Bagheri, and Ronald M Summers. ChestX-ray14: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9049–9058, 2018.
- Lawrence L. Weed. Medical Records That Guide and Teach. *New England Journal of Medicine*, 278(11):593–600, 1968.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.
- Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA, USA, 1956.
- Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie CM Van Aert, and Marcel ALM Van Assen. Degrees of Freedom in Planning, Running, Ana-

- lyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 2016.
- Sarah Wiegreffe and Ana Marasovic. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *NeurIPS (Datasets and Benchmarks Track)*, 2022.
- Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019.
- Sarah Wiegreffe, Ana Marasović, and Noah A Smith. Measuring Association Between Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10266–10284, 2021.
- Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NACCL): Human Language Technologies*, pages 632–658, 2022.
- Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NACCL): Human Language Technologies, Volume 1 (Long Papers)*, 2018.
- Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Thought; 176. 1922.
- Jialin Wu and Raymond Mooney. Faithful Multimodal Explanation for Visual Question Answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, 2019.
- Joy T Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G Mitchell, et al. Chest Im-aGenome Dataset for Clinical Reasoning. In *NeurIPS (Datasets and Benchmarks Track)*, 2021.
- Laure Wynants, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M. J. Bonten, Darren L. Dahly, et al. Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal. *BMJ*, 369:m1328, 2020.

- Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaying Zhang, Yuxin Peng, and Zheng Zhang. The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015.
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. *arXiv preprint arXiv:1901.06706*, 2019.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning (ICML)*, 2015.
- Wentian Xu, Matthew Moffat, Thalia Seale, Ziyun Liang, Felix Wagner, Daniel Whitehouse, David Menon, Virginia Newcombe, et al. Feasibility and benefits of joint learning from MRI databases with different brain diseases and modalities for segmentation. In *Medical Imaging with Deep Learning*, 2024.
- Fan Yang, Mengnan Du, and Xia Hu. Evaluating Explanation Without Ground Truth In Interpretable Machine Learning. *arXiv preprint arXiv:1907.06831*, 2019.
- Guang Yang, Qinghao Ye, and Jun Xia. Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, Two Showcases and Beyond. *Information Fusion*, 77:29–52, 2022.
- Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, et al. Advancing Multimodal Medical Capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.
- Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations (ICLR)*, 2022.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From Image Descriptions to

- Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research*, 2022.
- Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic Radiology Report Generation Based on Multi-View Image Fusion and Medical Concept Enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 721–729.
- Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Eedundancy Eeducation. In *International Conference on Machine Learning (ICML)*, 2021.
- John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLoS Medicine*, 15(11), 2018.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. 2022. arXiv preprint arXiv:2203.14465.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6731, 2019.
- Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Pre-

- ston, Rajesh Rao, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.
- Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. Multi-View Document Representation Learning for Open-Domain Dense Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, 2022a.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2019.
- Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Machine Learning for Healthcare Conference*, pages 2–25, 2022b.
- Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. MDnet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:46595–46623, 2023.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 13041–13049, 2020.
- S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, et al. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.
- Peifei Zhu and Masahiro Ogino. Guideline-Based Additive Explanation for Computer-Aided Diagnosis of Lung Nodules. *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, 2019.