Automated extraction of artificial intelligence model and dataset characteristics from papers to promote transparency

Abhinav Suri[1], **Ricardo A. Gonzales**[2], **Marcelo S. Takahashi**[3], **and Charles E. Kahn**[4]

[1]David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[2]Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, Charlestown, MA, USA
[3]Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[4]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract

We demonstrate that large language models can accurately extract structured model and dataset characteristics from AI research papers using the ROADMAP ontology. Using 10 benchmark papers and subsequently scaling to 311 publications, GPT-5 produced the highest-fidelity structured outputs at low cost. This enables large-scale aggregation of models, datasets, metrics, and content codes, supporting reproducibility, discoverability, and transparency. Structured outputs are publicly accessible through the ATLAS online repository.

Introduction

Artificial intelligence research in medical imaging continues to grow rapidly, with new methods and models being published at an accelerating pace. These papers contain critical methodological and dataset details; however, large-scale analysis of such information remains difficult due to the predominantly unstructured nature of academic reporting. This limitation is particularly consequential in healthcare, where transparent evaluation and comparison of AI systems across clinical contexts are essential. Although documentation frameworks such as model cards and dataset datasheets have been proposed to improve clarity, they depend on authors to voluntarily report structured elements[1]. Even when applied, these formats remain only partially standardized, limiting the ability to perform comprehensive computational analyses of research trends including dataset composition, metric usage, task characteristics, and clinical applicability. In prior work, we developed the ROADMAP Ontology of AI Datasets, Models, and Projects, which enumerates standardized fields relevant to analysis and reporting, such as content types, input modalities, performance measures, intended users, and clinical codes including SNOMED, RadLex, and LOINC[2,3]. In this study, we investigate whether large language models can reliably extract these structured fields automatically from published manuscripts.

Methods

We applied three large language models, GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Pro, to extract structured metadata from an initial set of ten papers published in medical AI journals. Each model processed the papers in text format and produced outputs using a standardized structured-field extraction prompt aligned with the ROADMAP ontology. For each paper, the models generated a total of 76 model properties and 78 dataset properties. These included diagnostic terminology, imaging modality, supervised task type, annotation schema, benchmark dataset identifiers, performance metrics, intended clinical users, and ontology-aligned concept codes. Additionally, we compiled a reference list of 212 commonly reported evaluation metrics to support standardized parsing. Each extracted field underwent expert scoring using a three-point accuracy rubric: 3 = fully accurate and complete; 2 = accurate but incomplete; 1 = inaccurate in whole or in part (4620 points possible across validation corpus). After identifying the highest-performing model from the initial benchmark set, we applied that model to extract metadata from the full corpus of 311 papers from *Radiology: Artificial Intelligence* and several public medical AI challenges. Resulting outputs were post-processed to ensure compliance and internal consistency, including verification of valid date formatting, email structures, and ontology code mapping. We aggregated the structured extractions and analyzed trends across the literature, including most frequently reported metrics, modalities, task types, and diagnostic scope. The resulting structured database is being made publicly available through the ATLAS (Annotated Library of AI Systems) website.

Results

GPT-5 using the low-compute reasoning configuration demonstrated the highest overall scoring, with a total performance of 4528/4620, surpassing Claude 4.1 Opus (4438/4620) and Gemini 2.5 Pro (4410/4620). Using GPT-5, we successfully extracted 283 model cards and 282 dataset cards from the full set of 311 papers. The average cost of analysis was approximately $0.21 per paper. Across the corpus, the most commonly reported performance metric was AUC; the most prevalent imaging modality was MRI; the most frequent use case category was detection and diagnosis; and the most frequently identified intended user role was radiologists. All structured outputs are publicly accessible through the ATLAS online portal (atlas.rsna.org).

Discussion

This study demonstrates that modern large language models are capable of extracting high-granularity structured metadata from AI research papers in radiology with near-expert accuracy. Importantly, this capability enables post-hoc standardization of existing literature, addressing a long-standing bottleneck in AI transparency: the lack of structured reporting formats that support automated scientific synthesis. By converting free-text manuscripts into structured representations, our approach enables unprecedented cross-paper comparability. Researchers can now easily examine relationships between dataset provenance, model architecture, reported metrics, clinical tasks, and end-user intent. These structured representations reveal systematic patterns, such as the predominance of classification-based tasks and the scarcity of reporting regarding demographic representativeness or labeling workflow. The ATLAS platform allows these structured outputs to be queried at scale, enabling rapid indexing of methodological choices and aiding future research planning. This has implications for bias surveillance, regulatory review, reproducibility assessment, and meta-research. Rather than relying solely on journal-mandated formatting guidelines, automated extraction offers a scalable mechanism to harmonize documentation across thousands of papers, even retroactively. Ultimately, the combination of ontology-guided structuring and LLM-based extraction may provide foundational infrastructure for transparent and clinically responsible AI development. As models increasingly interface with real-world patient care, robust documentation pipelines such as ours will play a critical role in ensuring clarity, accountability, and reproducibility in medical AI research.

References

1. Mitchell, M. *et al.* Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229, DOI: https://doi.org/10.1145/3287560.3287596 (Association for Computing Machinery, New York, NY, USA, 2019).
2. SNOMED International. SNOMED CT. http://www.snomed.org/ (2002). Accessed: December 2025.
3. Radiological Society of North America. RadLex. https://radlex.org/ (2006). Accessed: December 2025.

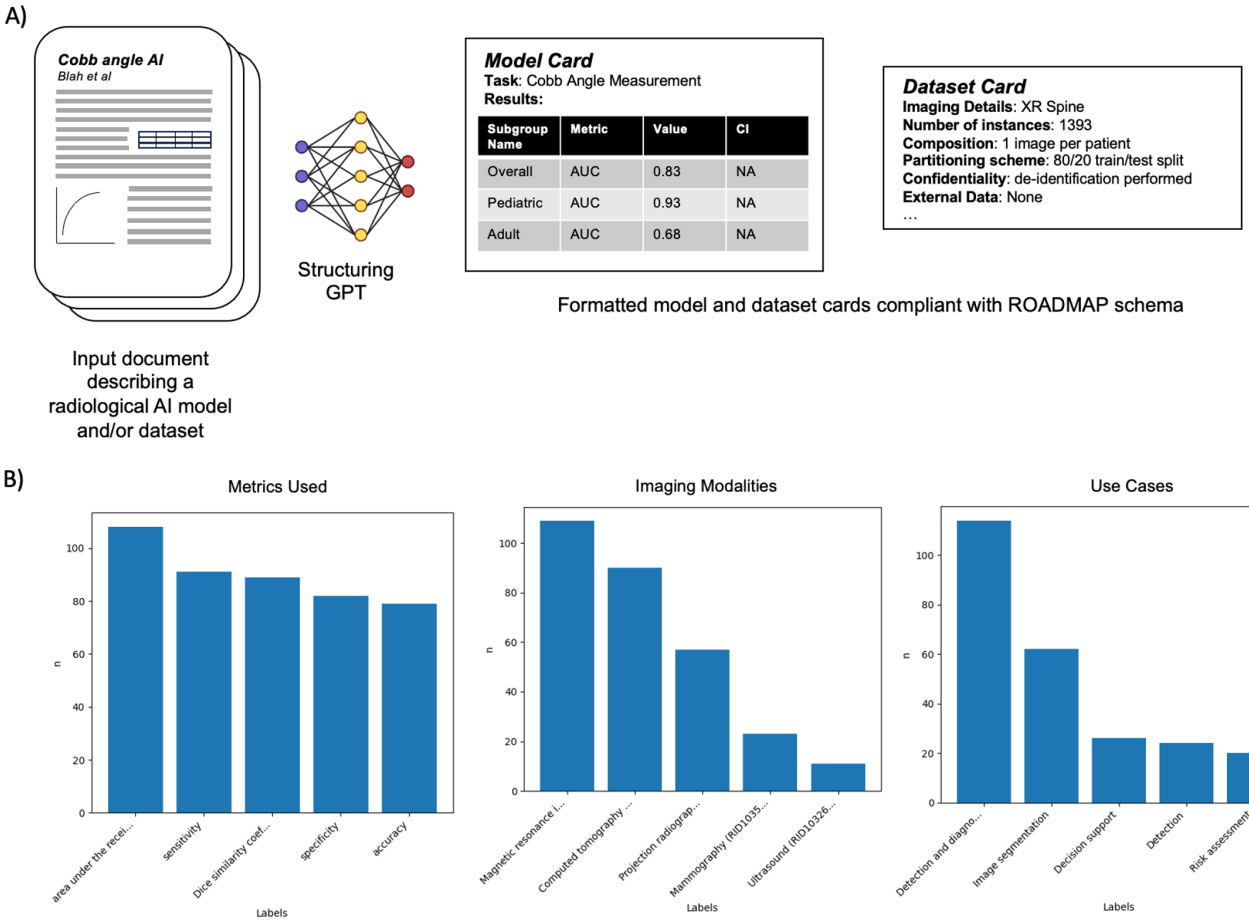

Figure 1. A) Overview of pipeline for extraction of features. B) Example of prevalence of structured metrics, imaging modalities, and use cases in aggregate corpus of 311 papers.