International Journal of Computer Vision (2026) 134:317
https://doi.org/10.1007/s11263-026-02913-0

Unlocking In-Context Learning for Natural Datasets Across Modalities

Jelena Bratulić[1] · Sudhanshu Mittal[1] · David T. Hoffmann[2] · Samuel Böhm[3] · Robin Tibor Schirrmeister[4] · Tonio Ball[3] · Christian Rupprecht[5] · Thomas Brox[1]

Received: 21 January 2026 / Accepted: 26 May 2026

Abstract
Large Language Models (LLMs) exhibit In-Context Learning (ICL), which enables the model to perform new tasks conditioning only on the examples provided in the context without updating the model's weights. While ICL offers fast adaptation across natural language tasks and domains, its emergence is less straightforward for modalities beyond text. In this work, we systematically uncover properties present in LLMs that support the emergence of ICL for autoregressive models and various modalities by promoting the learning of the mechanisms needed for ICL. We identify exact token repetitions in the training data sequences as an important factor for ICL. Such repetitions further improve stability and reduce transiency in ICL performance. We analyse in detail the training dynamics of such data sequences and explain how token repetitions enhance the ICL learning mechanisms. Moreover, we emphasise the importance of the training task difficulty for the emergence of ICL. Finally, by applying our novel insights on ICL emergence, we unlock ICL capabilities across various visual datasets used for few-shot classification, and confirm the generalisability of our insights to much harder real-world examples of large-scale object classification, and a more challenging EEG classification task. Code is available at https://github.com/jelenab98/unlocking_icl

Keywords In-Context Learning · Training dynamics · Generalisation · EEG · Image classification

1 Introduction

In-context learning (ICL) is a notable emerging feature observed primarily in transformer models, such as Large Language Models (LLMs) (Brown et al., 2020; Radford et al., 2019). ICL presents the ability to gather information to solve tasks not seen during training, such as looking up class labels or learning an algorithm (mapping rule), solely by conditioning on the example input–output pairs provided in the context. To achieve this, no weight updates or fine-tuning are required as the task is learned and applied within the context in inference. Precisely, the model takes a sequence of example input–output pairs (context) followed by a query, and it is expected to produce the correct output for the given query. The model is not explicitly told which task to perform, but it must infer the algorithm or mapping rule from the context. For instance, when a model is presented with a prompt like "dog: animal, rose: plant, car: ?", it should infer the underlying mapping rule and output "vehicle", even though it was never explicitly trained on that exact category-association task. ICL is similar to few-shot learning (Snell et al., 2017; Finn et al., 2017), with the key distinction that ICL requires no training phase for the new task; instead, the model leverages the pattern-matching learned during pretraining to understand the context and perform the task at inference. As a result, the same model can perform a wide range of tasks purely from the given input–output pairs, such as classification, regression, translation, or other sequence-based tasks.

Communicated by Svetlana Lazebnik.

David T. Hoffmann: Work performed while at Uni Freiburg.

✉ Jelena Bratulić
bratulic@cs.uni-freiburg.de

[1] Computer Vision Group, University of Freiburg, Freiburg, Germany

[2] AI Centre, Samsung Cambridge, Cambridge, United Kingdom

[3] Neuromedical A.I. Lab, Medical Center – University of Freiburg, Freiburg, Germany

[4] Medical Physics, Medical Center – University of Freiburg, Freiburg, Germany

[5] Visual Geometry Group, University of Oxford, Oxford, United Kingdom

Published online: 15 June 2026

This paradigm (ICL) contrasts with the "classical" in-weight learning (IWL), where the knowledge required for inference tasks is embedded within the model weights during training. While IWL can achieve strong performance, its generalisation is limited as it is less flexible – adapting to new tasks typically requires gradient descent updates or some fine-tuning. With IWL, the model only performs the task for which it was directly optimized and supervised (next-token prediction, specific label mapping seen in training, etc.).

ICL enables rapid adaptation to new tasks and label spaces, making it a standard way how humans interact with language models for everyday use (Agarwal et al., 2024; Long et al., 2023; Min et al., 2022; Pawelczyk et al., 2024; Ram et al., 2023), as well as for vision-language models (VLMs) (Alayrac et al., 2022; Laurençon et al., 2023; Sun et al., 2023), and even tabular and algorithmic data (Hollmann et al., 2023, 2025; Akyürek et al., 2024; Bai et al., 2023; Garg et al., 2022). Overall, ICL promises to be a fast and reliable method for new tasks with limited training data. For instance, applications that require few-shot adaptation to novel users or novel sensor/input data sets, like EEG-based brain-computer interfaces, could greatly benefit from high-quality adaptation methods that do not require retraining or fine-tuning the model.

Despite its promising capabilities, the emergence of ICL in models is non-trivial; it only emerges under specific training conditions. For instance, training on natural language often elicits strong ICL performance. Chan et al. (2022) attribute this to particular data distributional properties inherent to natural language, namely 1) burstiness: an increased likelihood to observe a token again, after it was seen recently, and 2) skewness: a sharply declining distribution over token frequencies with a long tail data distribution. Chan et al. further demonstrate the effectiveness of these training properties on Omniglot (Lake et al., 2015), resulting in the emergence of ICL. However, as we show here, this does not generalise to more complex vision datasets like CIFAR (Bertinetto et al., 2019), Caltech-101 (Fei-Fei et al., 2004), DTD (Cimpoi et al., 2014), and also Imagenet (Deng et al., 2009), nor does it transfer to other modalities such as EEG. Thus, we ask ourselves: **What do we need to unlock ICL for more general and arguably noisy datasets and modalities?**

To answer this, it is necessary to have a deeper understanding of ICL; specifically, we need to understand what a model needs to learn for ICL. In general, ICL requires: 1) **a knowledge aggregation function**, which extracts algorithms, rules, or information from the context and aggregates this knowledge in specific tokens of the context, and 2) **a look-up mechanism** that allows retrieving this aggregated information that is relevant to the current last token in a sequence (the so-called query) (Olsson et al., 2022; Reddy, 2024; Singh et al., 2024). Different ICL tasks will have slight differences in these two functions. In our classification setup, shown on Figure 1A, where the sequence contains paired signal-label tokens, the aggregation function gathers information from the previous signal token into the corresponding label token, forming a previous-token head, and the lookup mechanism is a simple similarity function that identifies similar tokens, relevant to the query signal.

Learning from a previous token-attending head does not contribute to ICL unless the similarity (look-up) mechanism is also learned, as there is no learning signal from the loss. Conversely, learning the look-up function (similarity function) between query and similar tokens is not helpful unless useful information has already been aggregated in those tokens. In essence, the learning of each component is interdependent: the similarity function requires that the previous token head has already been learned, and learning the previous token head requires the similarity function to be learned (see Figure 1B).

These mechanistic insights compel us to investigate why ICL succeeds on language tasks and Omniglot (Chan et al., 2022) but fails to generalise effectively to broader datasets and domains. We believe that the answers lie in the learning interplay of the two components: 1) We show in Figure 4 and Figure 1D that language naturally contains many exact copies of tokens and n-grams as well as synonyms in a continuous sequence of tokens. Moreover, prior work has shown that synonyms tend to be clustered or represented closely (Clark et al. 2019; Elhelo and Geva 2025; Lindsey et al. 2025; Mikolov et al. 2023; Pennington et al. 2014; Serina et al. 2023; Thießen et al. 2023). We hypothesise that this simplifies the learning of the similarity function, as the required function is close to the identity, and, by doing so, it breaks the interdependence between the aggregation and similarity components necessary for ICL to emerge. Thus, we argue that introducing exact token repetitions into training sequences – when they are not naturally present – can facilitate the learning of ICL (see Figure 1CD). We further analyse the training dynamics and track the progress of ICL learning mechanisms, similar to (Reddy, 2024), to gain a deeper understanding of how exact token repetitions help in the emergence of ICL.

We further suggest that 2) the relative difficulty (and expected accuracy) of the ICL and the IWL solution influence whether the model prioritises ICL or not. When the IWL task is overly simple, the model may exhibit a simplicity bias, prioritising IWL learning and bypassing in-context learning. We suspect this phenomenon extends to LLMs as well, where language modelling serves as a fairly complex IWL task, thereby encouraging the emergence of ICL.

While we focus our analysis mostly on the Omniglot (Lake et al., 2015) dataset, we also show how our insights unlock in-context learning in other real-world datasets with much more diverse and complex data, including the Imagenet (Deng et al., 2009) dataset for image classification, and a completely different modality of EEG signals for BCI motor imagery

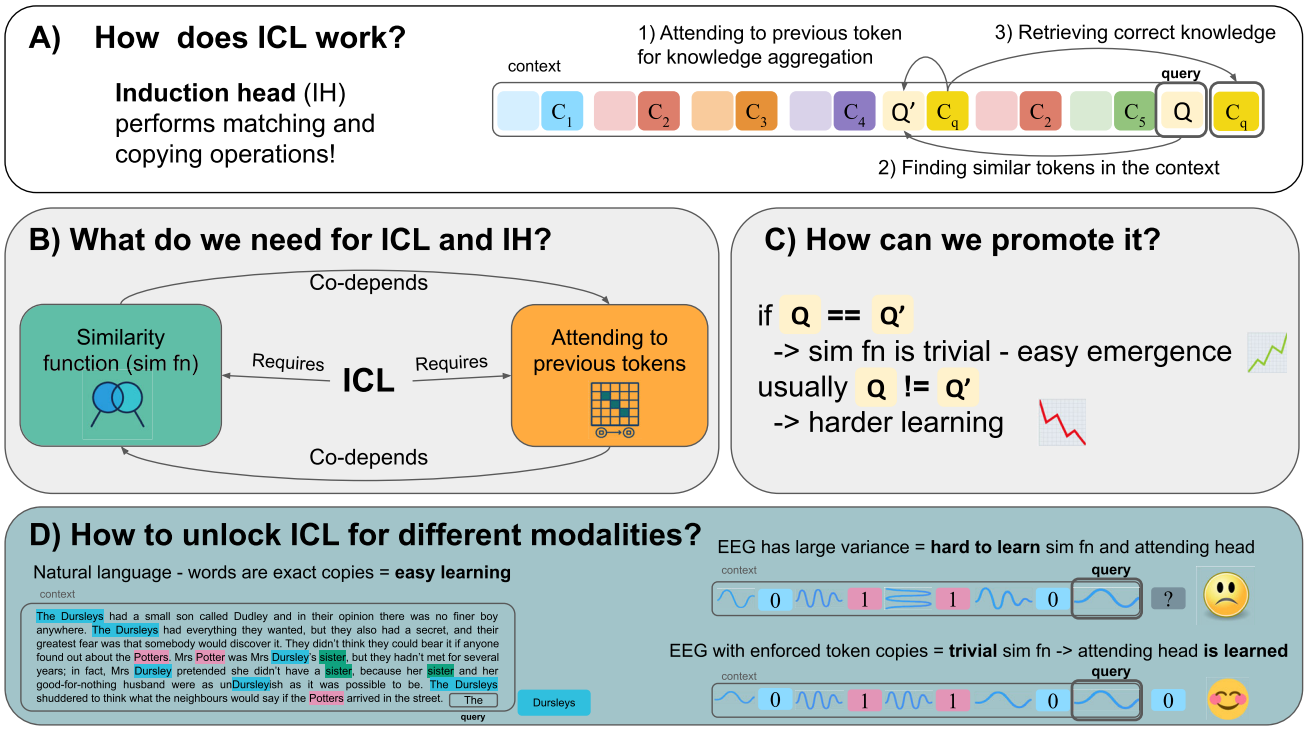


Fig. 1 A) ICL requires two operations: a similarity function and a head that attends to the previous token for knowledge aggregation; together, they present an induction head. B) A similarity function needs to be established for the previous-token heads to form. Still, the similarity function has no purpose if it can not be associated with relevant knowledge. C) The formation of a previous-token head should be promoted by simplifying the similarity function – by including exact token copies in the sequence. D) Enforcing exact copies in the sequences enables ICL for noisy and complex data beyond text, such as images and EEG.

classification. Enabling ICL on more challenging real-world datasets, we show that the importance of token repetition and task difficulty are general drivers of ICL learning mechanisms across different modalities and levels of task difficulty.

To summarize, we examine in depth the details of learning ICL and investigate the circumstances under which ICL emerges. 1) We find that using exact copies of tokens during training facilitates ICL learning and leads to higher ICL accuracy. 2) We further show that, against prior beliefs (Chan et al., 2022), high burstiness is not essential for the emergence of ICL. A single exact token copy in the context can be sufficient. 3) We present evidence that exact token copies simplify the ICL learning task by reducing the complexity of the similarity function to be learned, giving an initial boost to the ICL learning mechanisms. We demonstrate this through detailed analysis of the training dynamics and tracking the progress measures of the ICL learning mechanisms. 4) We further show that ICL vs. IWL task difficulty is a significant driver of ICL emergence, i.e. if the IWL task is difficult and complex, the model is more likely to learn ICL. 5) Finally, we demonstrate that our novel insight unlocks ICL for multiple standard vision datasets, varying from commonly used in few-shot literature to a more complex real-world example of large-scale natural images, and even enables ICL for noisy continuous data, such as EEG, where ICL allows few-shot transfer to novel datasets.

2 Related work

In-context learning (ICL). In-context learning, initially observed as an emerging ability in LLMs (Garg et al., 2022), but also reported in VLMs (Alayrac et al., 2022; Sun et al., 2023) and vision-only models (Wang et al., 2023; Bai et al., 2024; Bar et al., 2022) enables fast adaptation to various new tasks without gradient updates (Hollmann et al., 2025; Zhang et al., 2023; Zhu et al., 2024; Ferber et al., 2024; Camaret Ndir et al., 2026). Plenty of research has been dedicated to understanding how to obtain the best ICL performance by analysing the importance of pretraining data (Chan et al., 2022; Gu et al., 2023; Han et al., 2023; Levine et al., 2022; Liu et al., 2022; Min et al., 2022; Wies et al., 2023), demonstration selection and prompt design (Rubin et al., 2022; Suo et al., 2024; Zhang et al., 2023; Voronov et al., 2024; Yang et al., 2023; Sun et al., 2025) or framing ICL as in-context vectors (Liu et al., 2024; Huang et al., 2024; Peng et al., 2024). On the other hand, some works (Akyürek et al., 2023; Dai et al., 2023; Oswald et al., 2023) provided

insights into the ICL's working mechanisms by studying ICL on a simple regression task, showing how transformers act as meta-optimisers performing gradient descent.

Numerous works indicate that the training data distribution plays a role in the emergence of ICL, where challenging examples and long-tail tokens, and a large number of rarely occurring classes have been demonstrated to promote ICL (Han et al., 2023; Chan et al., 2022), while Razeghi et al. (2022) found a correlation between the input data term frequency and the ICL performance. Furthermore, Chan et al. (2022) demonstrate how certain data distributional properties, such as skewed token distribution and burstiness, benefit the ICL in a small synthetic scenario, while Singh et al. (2023) subsequently showed that the ICL in this setup can become transient, highlighting the conflict between the ICL and IWL circuits. Similarly, Chen et al. (2024) argued that parallel structures, which follow similar semantic or syntactic templates in the pretraining textual data, facilitate ICL in language models. Our work builds upon previous studies on the importance of data distributional properties (Chan et al., 2022; Singh et al., 2023) and provides additional insights into unlocking ICL for various modalities and complex data. Concurrent with our work, Zucchet et al. (2025) propose a theoretically grounded framework for sparse attention emergence and find that repetitions in data accelerate induction-head formation and in-context learning, which supports our finding on the importance of exact token repetitions. We provide further empirical validation of the exact token repetitions across real-world modalities of images and EEG signal, provide an explanation of their role for ICL emergence by tracking progress measures, and identify IWL task difficulty as another important driver for ICL.

Understanding ICL mechanisms. Mechanistic studies on the emergence of ICL have identified a specialised attention pattern, i.e. an induction head, that conducts matching and copying operations as a key mechanism for ICL (Olsson et al., 2022). Recent works have been studying the formation of the induction heads and their role for ICL in a simplistic scenario (Reddy, 2024; Singh et al., 2024; Edelman et al., 2024), where Reddy (2024) demonstrates with a simple two-parameter model that ICL is driven by the formation of an induction head, which emerges due to nested non-linearities in a multi-layer attention network. Reddy further introduces progress measures for deeper analysis of the ICL learning mechanisms. Our work builds on this interpretability framework to trace the dynamics of the induction heads during training through different progress measures. We explain how certain data distributional properties influence the formation of induction heads and the performance of ICL.

Generalisation in EEG for motor imagery. Due to individual variability, cross-dataset generalisation in EEG-based motor imagery (MI), although highly desirable, remains a challenge. While zero-shot EEG methods are increasing, they typically perform multi-modal alignment with EEG and enable classification to unseen classes from the same datasets only (Li and Wei, 2025; Song et al., 2023; Liu et al., 2023). For MI-decoding, pre-trained EEG transformer models show promise but lack zero-shot capabilities (Jiang et al., 2024; Patil et al., 2024). To our knowledge, only Duan et al. (2020) has explored zero-shot learning for MI-EEG using outlier detection for base and novel classes. Our work demonstrates how enabling ICL for EEG provides a promising new direction for cross-dataset EEG generalisation without any fine-tuning.

3 Experimental setup

We investigate how ICL emerges by training a causal GPT-2-like model (Radford et al., 2019) on sequences of image-label pairs from scratch on standard few-shot learning datasets: Omniglot (Lake et al., 2015), CIFAR-100 (Bertinetto et al., 2019), Caltech-101 (Fei-Fei et al., 2004), DTD (Cimpoi et al., 2014). After the investigation, we employ the uncovered factors for ICL and apply it to much harder real-world data examples of Imagenet (Deng et al., 2009) dataset and different EEG motor imaging datasets (Zhou et al., 2016; Schirrmeister et al., 2017; Tangermann et al., 2012).

The autoregressive model in this work is trained with a sequence length of $2L+1$ with L image-label pairs in the context followed by a query image fully from scratch, as shown in Figure 2. The in-weight learning objective is to predict the label of the last image, which is the $(2L+1)$-th token, given a sequence of L interleaved image-label pairs. For IWL task, the sequence format should not impact the performance, as the model should rely on the knowledge embedded within the model weights instead of relying on the context to solve the task. Each image-label pair is converted into token embeddings separately. The model is trained to maximise the likelihood of the next token, with the loss applied to the final query output, thus using last-token prediction as the IWL training objective.

3.1 Data

The design of the data sequences is really important. We distinguish between training and evaluation sequences, each composed of distinct sequence types with specific properties.

Training sequences. We employ a mixture of (1) **standard sequences**, in which sample-label pairs are uniformly randomly selected from the training dataset without any repetitions in the sequence, and (2) **in-context (bursty) sequences**, where the query image-label information is enforced to be present in the sequence by using a pair similar to the query image-label pair. Using in-context sequences, the model can solve the task without relying solely on the

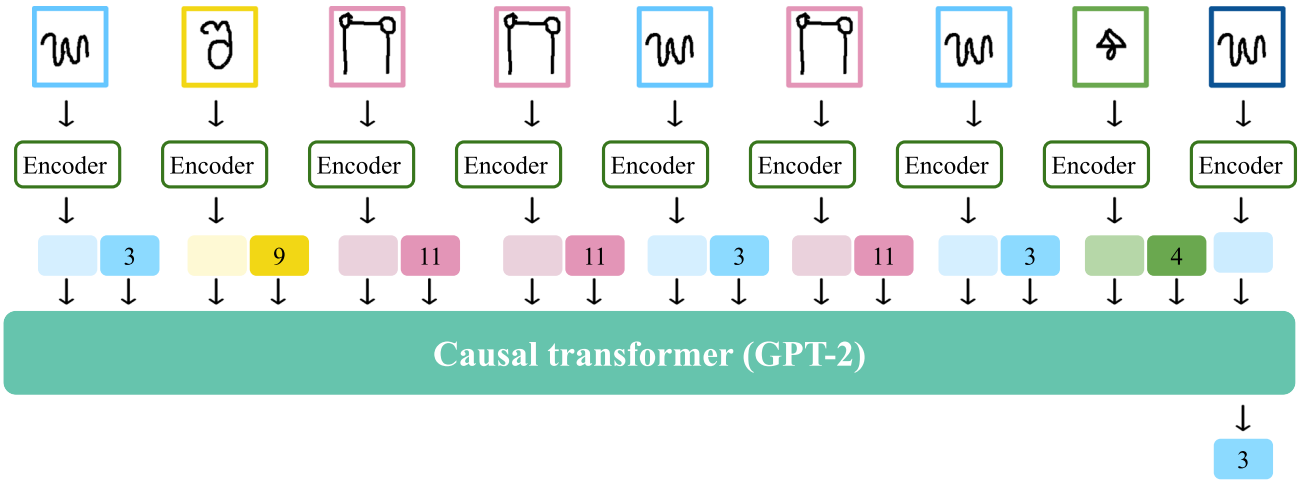


Fig. 2 We train a GPT-2-like architecture as a next-token prediction from scratch with image-label pairs forming a sequence with control of the training sequence distribution.

model weights. The proportion of each sequence type in the total amount of training sequences is treated as a hyperparameter. Following the setup from (Chan et al., 2022), we use 10% of standard and 90% of in-context sequences.

We illustrate the difference between the sequences leveraged in our experimental setup on Figure 3. We employ a sequence of length $L = 8$ with eight image-label pairs. In this case, standard sequences have 8 unique image-label pairs in the context, and the 9th image comes from a 9th class. For bursty sequences, we distinguish between high burstiness in the sequence (referred to as **bursty sequence**) with three instances from the query class in the sequence and low burstiness (referred to as **bursty (low) sequence**) with one example from the query class in the sequence. We further introduce bursty sequence **instCopy**, which follows the same logic as bursty sequence, but instead of having three instances from the query class in the sequence, it has the same example as the query image repeated (copy-pasted) three times in the sequence (see the same pattern in query-class instances on Figure 3). To explain the examples from Figure 3 clearer, the bursty sequence with query class 3 has 4 distinct instances of the class 3 present in the sequence, with 3 distinct instances being in the context, and the fourth one is the query image. On the other hand, the InstCopy sequence with query class 3 has only one instance of that class that appears at all four locations in the sequence, with the possibility of slightly different augmentations between them. We introduce this type of sequence, motivated by the frequent repetitions in natural language.

Evaluation sequences. During the evaluation, we leverage 2 different types – IWL and ICL evaluation sequences, following a similar evaluation protocol as in (Chan et al., 2022). IWL is evaluated for the multi-class classification task on the held-out samples from the training classes. The standard sequences, with a uniformly sampled format, are used for IWL evaluation (see Figure 3). ICL is evaluated in a few-shot classification setting for 2-way 4-shot (2 classes with 4 supporting samples each) and 4-way 2-shot (4 classes with 2 supporting samples each) tasks. This evaluation is performed on held-out novel classes. The trained classifier output is used for the few-shot evaluation, utilising label mappings from 0-1 or 0-3 to 2-way 4-shot and 4-way 2-shot, respectively.

Dataset construction. We conduct our controlled experiments and analyses on the Omniglot dataset (Lake et al., 2015) and scale to more realistic visual datasets, often used in few-shot learning evaluation: CIFAR-100 (Bertinetto et al., 2019), Caltech-101 (Fei-Fei et al., 2004), and DTD texture datasets (Cimpoi et al., 2014). Finally, we also show the performance on the large-scale natural images dataset Imagenet (Deng et al., 2009).

Omniglot consists of 1623 handwritten characters from 50 alphabets with 20 exemplars for each character. Unless stated otherwise, we use 1600 classes as the base classes and the remaining 23 classes (sampled from the official evaluation subset with seed 42) for the ICL evaluation. We use 10% of data for the validation. During supervised training, we apply no augmentations other than resizing to 64x64.

CIFAR-100 is a natural dataset consisting of 60000, 32x32 colored images divided into 100 categories with 600 examples from each one. We use 80 classes for supervised training and 20 classes for the ICL evaluation, as it is given by the Cifar-100FS (Few-Shot) version of the dataset. We used 10% of the data for the validation. We do not apply any augmentations, but we resize the image to 64x64 for training and evaluation.

Caltech-101 is a natural, imbalanced dataset with 101 classes with 40-800 images per class, while most classes have about 50 images, and each image is roughly 300x200 pixels.

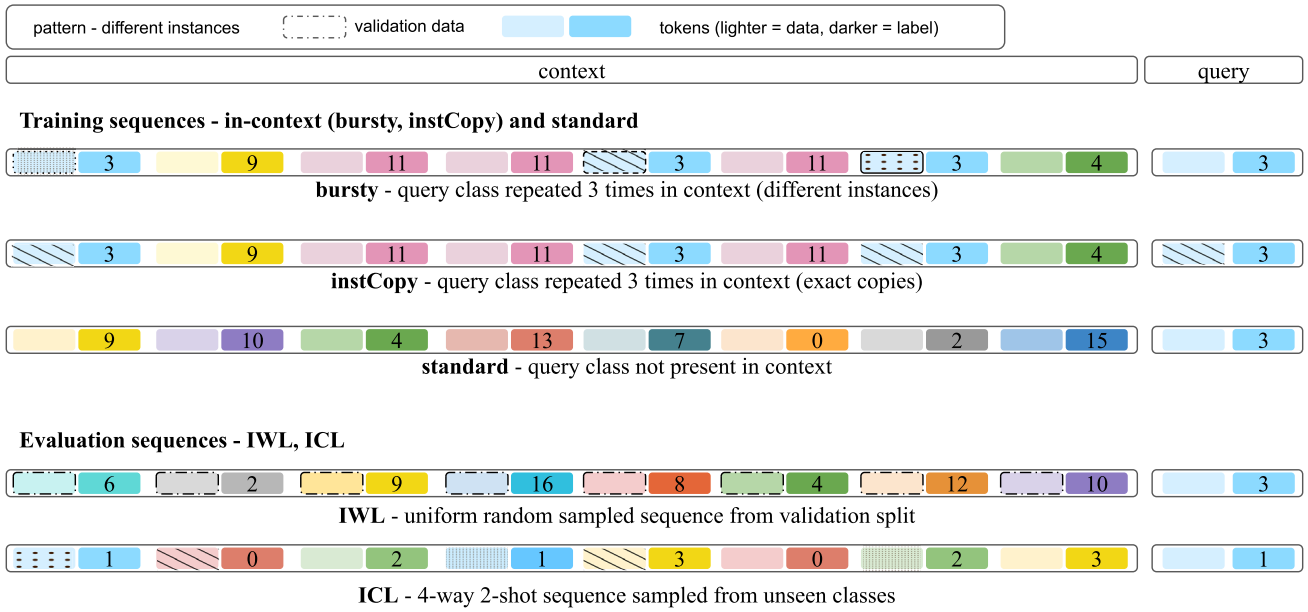


Fig. 3 Different training and evaluation sequences with the main difference being the number of repetitions and the use of identical copies in the context. Bursty sequences in training have multiple repetitions of the same class, but different class instances as the query class in the context. InstCopy sequence has the same instance as the query class instance repeated (literally copy-pasted) multiple times in the context.

We randomly select 91 classes for the supervised training, and the remaining 10 classes are used for ICL evaluation. During training, we use random resized cropping to 64x64 with scaling from 0.5 to 1.5, horizontal flipping, and random rotation of 15 degrees.

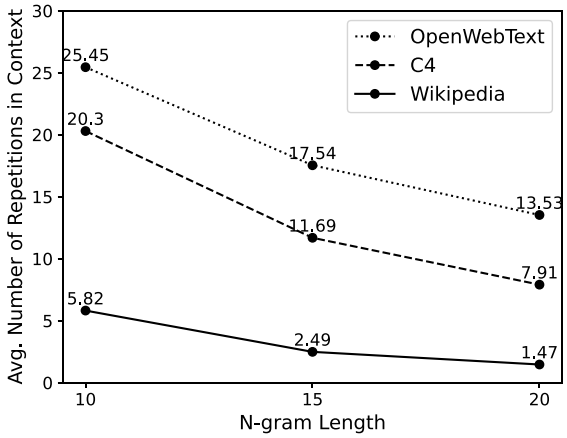
DTD is a texture dataset consisting of 5640 images across 47 classes with 120 images from each class with a size ranging from 300x300 to 640x640. We use 37 classes for supervised training and 10 classes for the ICL evaluation, and create a train and validation split with roughly 10% of data used for validation. We report better and more stable results with an image size of 128x128 and random resize with scale (0.5, 1.5).

Imagenet (ILSVRC-2012) is a natural, large-scale image dataset with 1000 object categories (classes) introduced as a part of the ILSVRC challenge in 2012. The dataset comprises over a million images collected from the web via structured query expansion and further refined with human annotations. We use 15, 50, 100, and 950 classes for supervised training, and 10 classes for the ICL evaluation, and create smaller training and validation splits with 100 instances per class (250 when using 15 classes) for training and 10 per class for validation, which are sampled from the original train and validation subsets. Data splits, both class and instances, were randomly chosen with seed 42. During training, we use random resized cropping to 128x128 with scaling from 0.5 to 1.5, horizontal flipping, and random rotation of 15 degrees.

3.2 Architecture

We train the GPT-2-like model (Radford et al., 2019) with 12 layers, 8 heads, and an embedding dimension of 64. We use a smaller model for the induction head analysis experiments with 3 layers, a single head, and an embedding dimension of 64. The model expects a sequence-like format with aligned embedding size, so we transformed our image-label pairs into separate image and label tokens, using a ResNet-like embedder for images and an embedding layer for labels. We initialised the model with a truncated normal distribution, which is important for training stability. We use a 3-block ResNet model (He et al., 2016) as the image embedder with output channel dimensions [64, 128, 256]. After that, a projection layer is added to match the embedding dimension of 64. We train the image embedding model and the GPT model together from scratch, without language pretraining; we simply adopt a GPT-2-like architecture. We observe that ICL emergence is sensitive to the input image embedder architecture, as shown in 14.

Hyperparameters. We trained the model for different numbers of steps varying from 100k to 2M iterations using optimiser Adam (Kingma et al., 2014) with betas (0.9, 0.99) and epsilon 1e-08. We use a learning rate warm-up for 15K iterations, with a square root decay scheduler and a maximum learning rate of 6e-4. We find that ICL performance is enhanced with longer warm-up periods. We perform gradient clipping to a value of 1.0. We trained the model with a batch size of 16 on a single Nvidia RTX 3090, where 500k



Elymus, attended Pirithous' wedding, fought in the battle against the Lapiths and was killed by... Eurytion, acted in an insulting manner towards Hippolyte when she was being joined in marriage to Azan or in the house of Pirithous... Hodites, fought against the Lapiths at Pirithous' wedding. Killed by Mopsus. Hyles, attended Pirithous' wedding, fought in the battle against the Lapiths and was killed by... Imbreus, fought against the Lapiths at Pirithous' wedding and was killed by Dryas... Isoples, killed by Heracles when he tried to steal the wine of Pholus... Lycabas, attended Pirithous' wedding, fought against the Lapiths and fled. Lycidas, fought against the Lapiths at Pirithous' wedding and was killed by Dryas. Lycus, fought against the Lapiths at Pirithous' wedding was killed by Pirithous. Medon, attended Pirithous' wedding, fought against the Lapiths and fled. Melanchaetes, killed by Heracles when he tried to steal the wine of Pholus. Melaneus, attended Pirithous' wedding, fought against the Lapiths and fled. Mermerus, wounded by the Lapiths at Pirithous'...

Fig. 4 Left: Repetitions of n-grams in Wikipedia(Foundation, yyy), OpenWebText(Gokaslan et al., 2019) and C4(Raffel et al., 2020) pretraining corpora, performed over 50 million tokens using a BPE tokeniser with a context length of 2048. We report the average number of repetitions within the 2048-token window for different n-gram lengths. The variety of the corpora's formats (e.g., web, news, social media, wiki) leads to substantial differences in repetition rates. Right: Truncated example of a 2048-token sample from Wikipedia's pretraining corpora, highlighting exact n-gram repetitions. Different colours present different n-gram lengths (green: 10-grams, blue: 15-grams, orange: 20-grams), demonstrating both patterns in the pretraining data.

iterations took around 12 hours. For all experiments, we run the approach for 3 random seeds (42, 1337, 3184) and report the average results.

4 How to enable ICL?

Prior work has identified specific circuits in transformer models as the working mechanism of ICL – induction heads (Olsson et al., 2022; Reddy, 2024; Singh et al., 2024). The induction head embodies the core concept of in-context learning: examining the context to identify the most similar or relevant token and then retrieving the associated, already aggregated knowledge. ICL requires two underlying components: a similarity or look-up function and a head that attends to the previous token. These two components are mutually dependent – the similarity function is ineffective without meaningful aggregated information by the previous-token head, and the previous-token cannot be optimised without a similarity mechanism to retrieve and apply the stored information.

4.1 Why is ICL learned and non-transient on text but not on visual data?

Previous work has shown that burstiness and skewness are drivers of ICL in LLMs (Chan et al., 2022). Here, we argue that natural language typically contains many exact token copies and n-grams, meaning that the exact repetitions are another important factor for ICL emergence. To show this, we first conduct a brief analysis of three pretraining corpora commonly used in NLP: Wikipedia (Foundation, yyy), OpenWebText (Gokaslan et al., 2019), and C4 (Raffel et al., 2020). We calculated the average number of n-gram repetitions within a 2048-token context window, using a BPE tokeniser to tokenise the text. In Figure 4, we report many n-gram repetitions and exact token copies in text, which we believe are an important factor for stable and non-transient ICL in LLMs.

Now, we can test if our hypothesis on the exact token repetitions indeed affects the ICL emergence by intervening on the co-dependency of the learning mechanism. We train the model conforming to the bursty sequences introduced by Chan et al. (2022) and propose a new type of bursty sequence with exact instance copies in the context (instCopy) to test this hypothesis. The difference between the sequences is illustrated and explained in Figure 3.

From Figure 5, we observe that including bursty sequences in the training data indeed leads to the emergence of the ICL, which supports previous works (Chan et al., 2022; Singh et al., 2023, 2024). However, the model achieves strong and more stable (less transient) ICL performance while using exact copies in the bursty sequences (instCopy). Furthermore, we can see that high burstiness is not essential for ICL – a single exact copy in the context (bursty (low) + instCopy) is sufficient to obtain ICL. Finally, we observe that applying

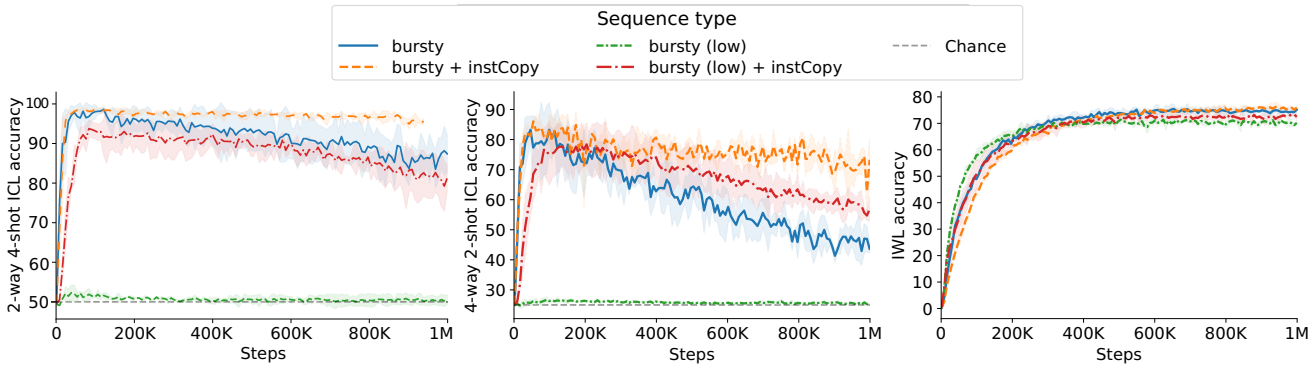


Fig. 5 Exact copies in the context (instCopy) promote ICL performance and reduce transiency. Only a single copy ensures ICL emergence (bursty (low) case).

exact copies in the context does not harm the IWL performance of the model, and at the same time, it encourages the emergence of ICL. **This confirms that exact copies are a stronger driving factor for ICL, even surpassing the high burstiness, as previously reported** (Chan et al., 2022).

4.2 Why do exact copies help?

In Figure 5, we show that exact copies facilitate strong and stable ICL performance. We argue that this is due to the simplified similarity function, which breaks its interdependence with previous-token head learning and ensures that the model prioritises the formation of the previous-token head.

To confirm this argument, we analyse the training dynamics by tracking progress measures during training and validation, similar to (Reddy, 2024). For this analysis, we use a simplified model with 3 layers and a single head, so we can more easily spot the circles in the layers, which is common in the interpretability research. However, the same observations should hold for a larger model with more layers and heads, where one or more heads would form these circuits, and the exact information flow would not be directly visible. To track the emergence of the circuits, we introduce 4 different progress measures based on the QK attention scores of the model: 1) image-image diagonal, 2) label-image, 3) image-image query, and 4) query image-label.

The **image-image diagonal** measure is the average attention between all image tokens and all other image tokens on the diagonal, representing the modality mapping. The **label-image** measures the average attention between each label token and its previous image token on the positions which are expected to be activated if the previous-token head has been formed (diagonal moved by one). The **image-image query** measures the average attention between the query image and other images in the same class, representing a similarity-matching function rather than knowledge aggregation. Finally, the **query image-label** measures the average attention between the query image and the correct label token

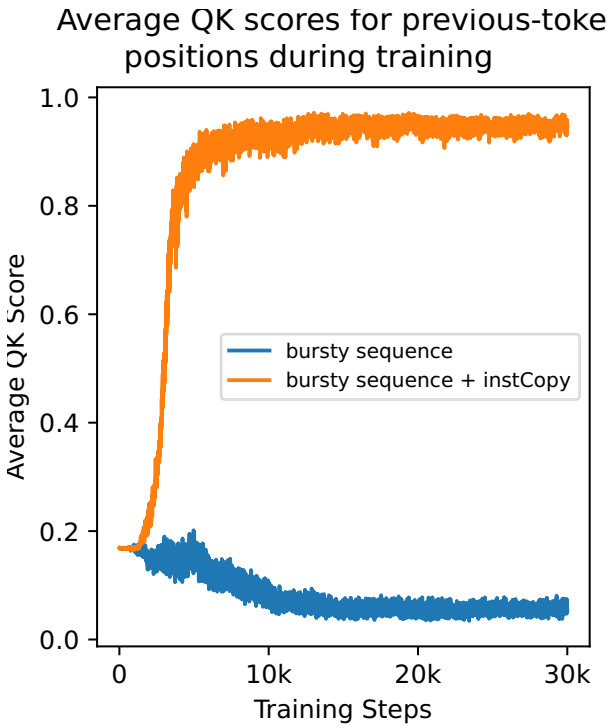


Fig. 6 Average QK scores over the previous-token head positions during training for smaller three-layer single-head models trained with burstiness and with exact copies (bursty + instCopy). High attention scores for instCopy sequences confirm the formation of a previous-token head, which is needed for ICL emergence.

positions, representing the final step in the induction head: retrieval of the correct label.

First, we examine the trends in the measure *label-image*, comparing the QK attention scores of the model trained with bursty sequences and the model trained with combined burstiness and exact copies (instCopy) as in-context sequences. We calculate the average attention score on the diagonal with an offset of 1, as this pattern represents the previous-token head.

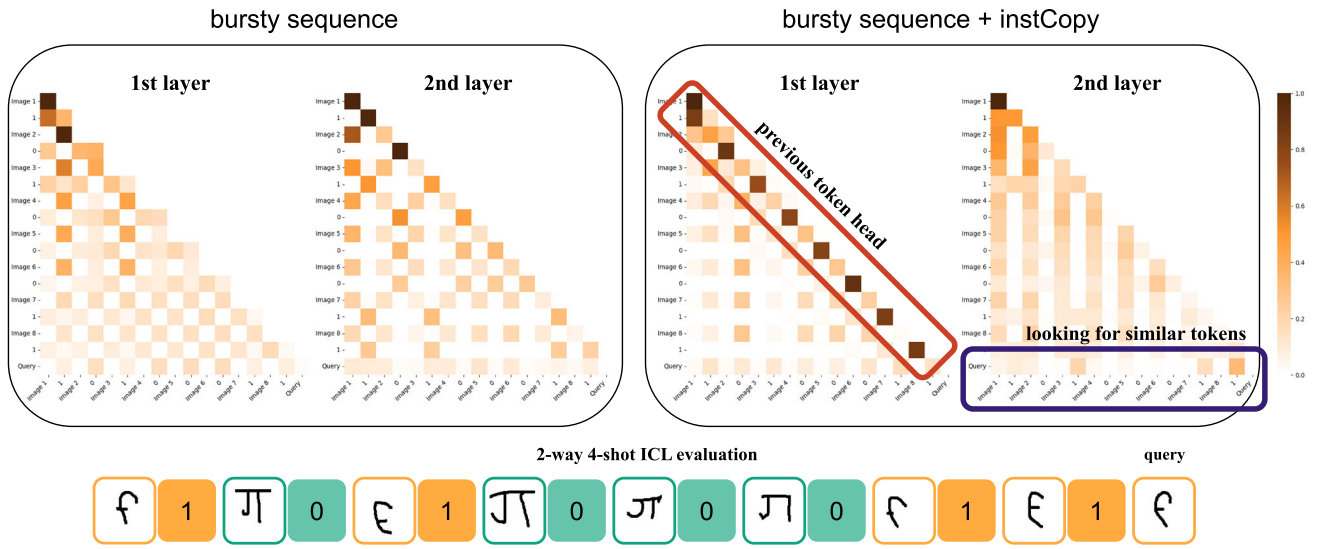


Fig. 7 QK attention space (top) during inference of a 2-way 4-shot sequence (bottom; classes 0 and 1) on smaller three-layer single-head models at inference time. We observe a clear induction head and ICL emergence during inference only for the model trained with burstiness and exact copies (bursty + InstCopy) seen through specific patterns over two layers: a previous-token head in layer one (diagonal with offset 1) and a query token attending to the most similar label tokens in layer two.

During training, we observe higher scores for tokens corresponding to the query label and for the formation of previous-token heads only in the model with instCopy sequences. In Figure 6, we trace the formation of previous-token heads during training by computing the averaged QK values off the diagonal (expected positions for previous-token head) over the training process. The previous token heads indicate aggregation of knowledge from the image to the label token. Since the similarity function is now trivial, the model learns to attend to the query from previous tokens and successfully performs the required ICL operations.

We observe the same patterns during inference for 2-way 4-shot classification on novel classes, as illustrated in Figure 7 (bottom). We observe ICL performance only for the model trained with bursty sequences and exact copies (bursty + instCopy). For the same model, we observe more attention between similar tokens in the sequence and a visible previous-token head. We further track progress measures of ICL evaluation sequences during training, and we visualise the QK attention space of one sequence at 65k iteration when a model with exact copies in the training sequences has strong ICL performance in Figure 8.

Here, we can clearly see different trends of the progress measure throughout the training, where in Layer 1, we observe a high label-image progress metric for the model with exact copies, indicating that the label tokens now strongly attend to the previous image in the sequence, which confirms the formation of the previous-token head. Interestingly, for the model with just burstiness, we observe no such formation. Instead, the model learns the modality mapping – image tokens attend to other image token positions in the sequence, which is seen as high image-image diag value. Further, we see slight decrease in the label-image metrics and increase in image-image diagonal metric how the ICL becomes transient in the model.

In Layer 2, we initially observe a high image-label progress metric for the model with exact copies, but these values become transient throughout training. At the same time, ICL performance follows the same trend: initially high and strong, then transient. This measure directly reflects the ICL ability in a model, and we indeed observe high values only for the model that has ICL emergence, and it connects with the knowledge aggregation in the previous layer. Furthermore, we observe that the model without ICL performs similarity matching and modality attendance, as evidenced by high image-to-image query values. Interestingly, we observe a further increase in the same measure as the model's ICL performance becomes transient.

This confirms that including exact copies in the context indeed simplifies the learning of the similarity function and promotes the formation of a previous-token head, which is then utilised during inference to make a correct ICL prediction.

4.3 What unlocks ICL for various visual datasets?

Previously, we confirmed that bursty sequences (without exact copies) unlock in-context learning on vision datasets such as Omniglot. However, we find that the same setup fails to obtain any ICL for other vision datasets like DTD (Cim-

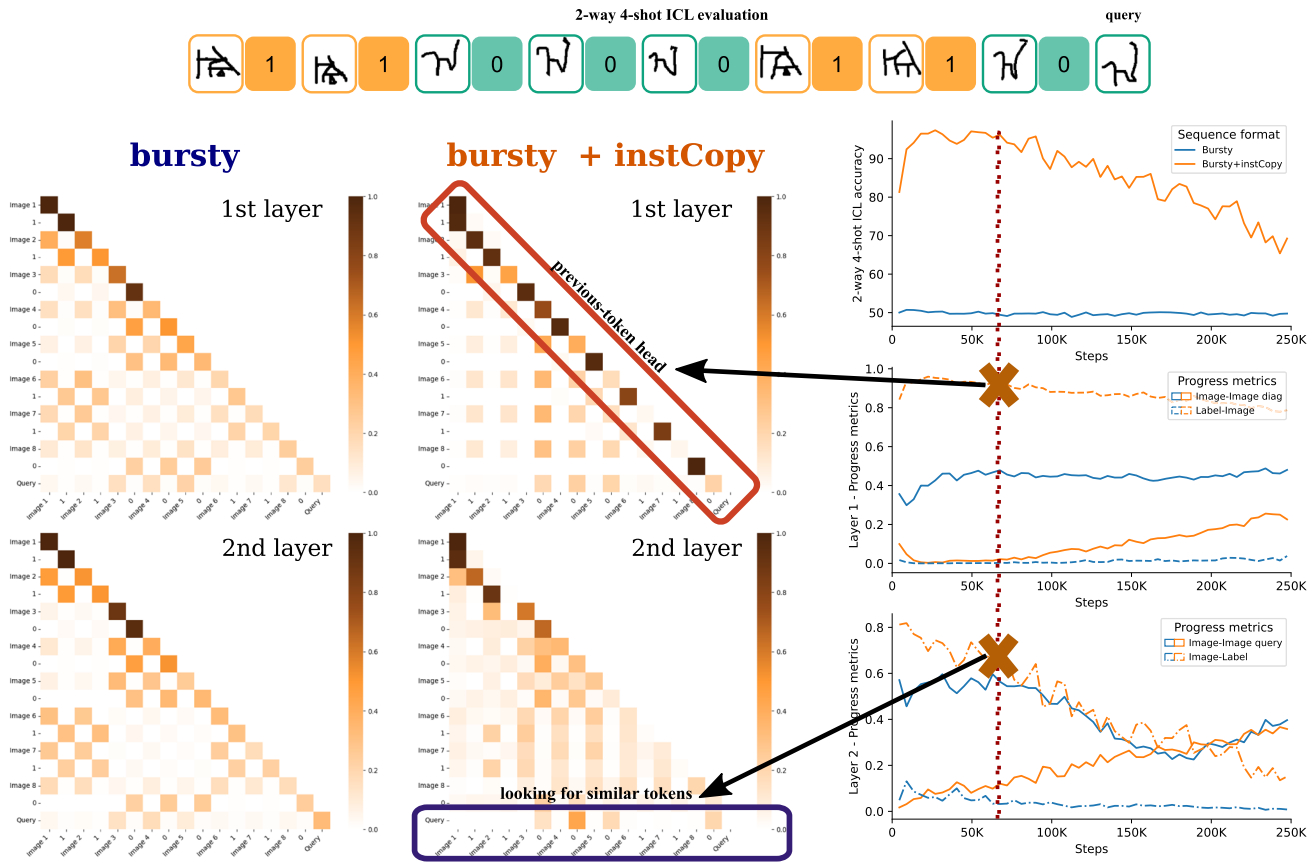


Fig. 8 QK attention space during inference of a 2-way 4-shot sequence (classes 0 and 1) and progress measures during training on smaller GPT-2 single-head models. We observe strong ICL performance only for the model with exact copies, where the progress metrics confirm the formation of the previous-token head and the model's correct establishment of the ICL mechanisms. Interestingly, as the ICL performance becomes transient, the model starts to behave similarly to one without ICL performance – just performing modality mapping.

poi et al., 2014), CIFAR-100 (Bertinetto et al., 2019), and Caltech-101 (Fei-Fei et al., 2004), as shown in Figure 9.

We hypothesise that burstiness alone does not provide a sufficient signal for learning the similarity function, which is now harder for these datasets. Here, we include exact instance copies (instCopy) in the bursty sequences. We observe that **instCopy enables strong ICL performance for all three visual datasets**, as shown in Figure 9. We also observe that the IWL performance remains largely unaffected, with some convergence differences during training, but the same performance by the end of the training, which directly shows the difference in the model's circuit training.

4.4 Does in-weight learning (IWL) task influence ICL?

Language modelling is a significantly harder task than Omniglot classification, yet we naturally observe strong ICL performance in LLMs, but not in Omniglot. Does the IWL task influence ICL? If yes, then how?

The emergence of ICL requires the formation of induction heads. However, when the IWL task is overly simple, the model can exhibit simplicity bias and prioritise IWL learning over induction head learning. Even highly bursty in-context sequences may fail to enable ICL in such cases. Following this, we argue that an in-weight task must have a minimum level of complexity to encourage the emergence of ICL.

To test this hypothesis, we make IWL tasks more challenging in two ways — by increasing the number of classes and by adding label noise via label swapping.

Number of training classes vs. ICL. We create IWL tasks on the Omniglot dataset with an increasing number of classes, ranging from 200 to 1600. Using the training setup with bursty in-context sequences (without instCopy), we evaluate ICL and IWL performance. In Figure 10, we observe that IWL converges more slowly as the number of classes increases, indicating that the IWL task becomes more challenging with a larger class set. In contrast, ICL performance improves when the number of classes increases. This trend suggests a competition between IWL and ICL circuits in the

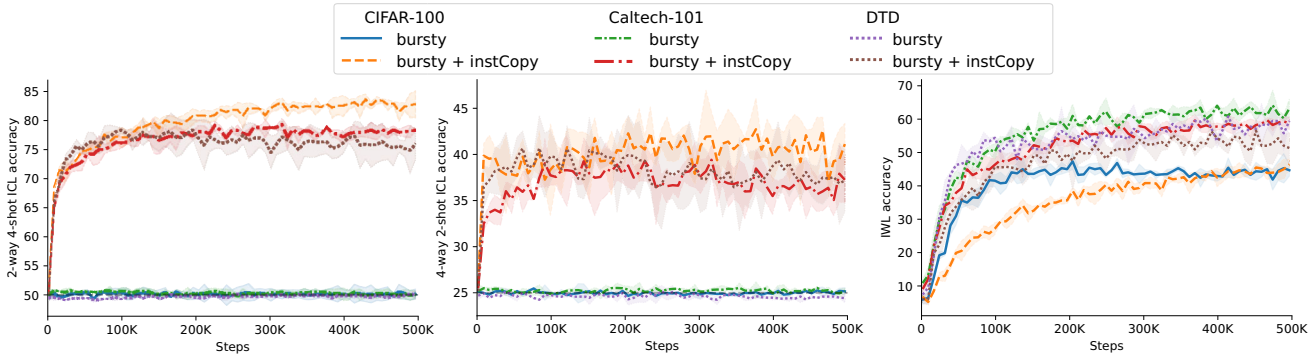


Fig. 9 Only when employing exact copies in the context (bursty + instCopy), we ensure ICL emergence on the image classification datasets CIFAR-100, Caltech-101 and DTD, which shows how the ICL emergence can be encouraged in many different visual datasets.

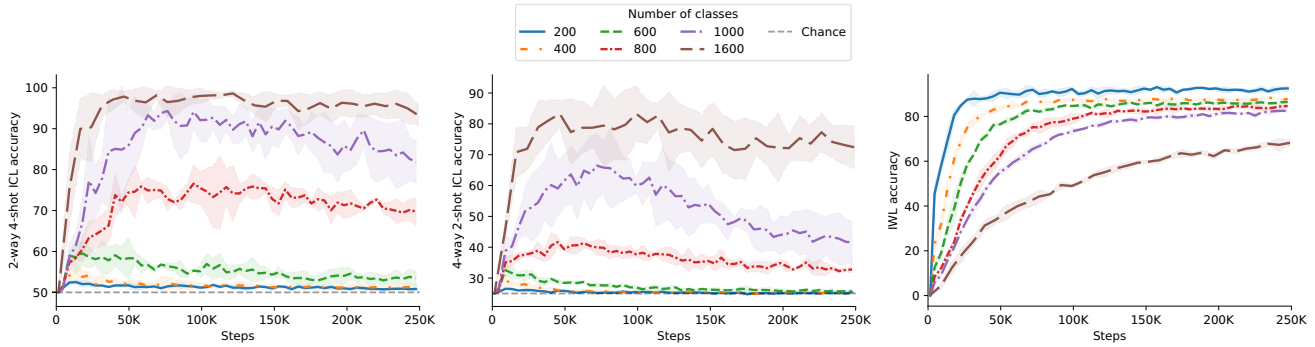


Fig. 10 When increasing the number of classes monotonically, ICL performance improves as the IWL objective gets harder, which shows the connection between the IWL task difficulty and the ICL emergence.

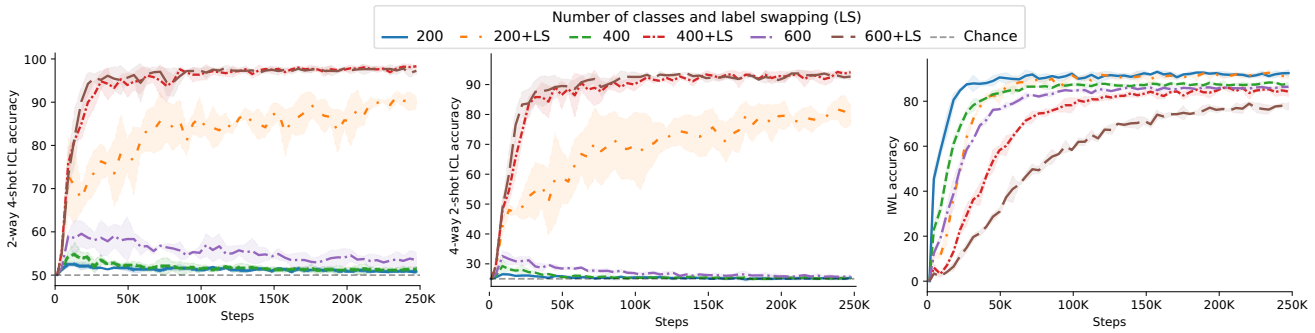


Fig. 11 When increasing the amount of label swapping (LS) noise in sequences, ICL performance improves as the IWL objective gets harder, which shows the connection between the IWL task difficulty and the ICL emergence.

early phase of training. If the IWL task is too simple, it may fulfil the IWL objective without learning the ICL mechanism. Prior work (Chan et al., 2022; Reddy, 2024) indicate similar findings. However, they attribute this improvement to the presence of many rarely occurring classes, which can be interpreted as another way to make the IWL task harder. This shows that **increasing the number of classes makes IWL harder and improves ICL**.

Label noise vs. ICL. Here, we create different training setups with label noise using Omniglot (Lake et al., 2015). We perform random label swapping, where labels of the query items are randomly assigned to another training class in 20% of all sequences. For the case of a bursty sequence, we change the label mappings to all repetitions in sequences belonging to the query class. We train models with 200, 400, and 600 classes, where no ICL was observed without noise, shown in Figure 10, possibly due to the overly simple IWL. In Figure 11, we observe that label swapping significantly improves ICL performance, while for IWL, we observe a similar trend of slower convergence as in Figure 10. Label swapping makes the IWL task challenging by introducing label noise into the standard sequences. In contrast, label swapping promotes

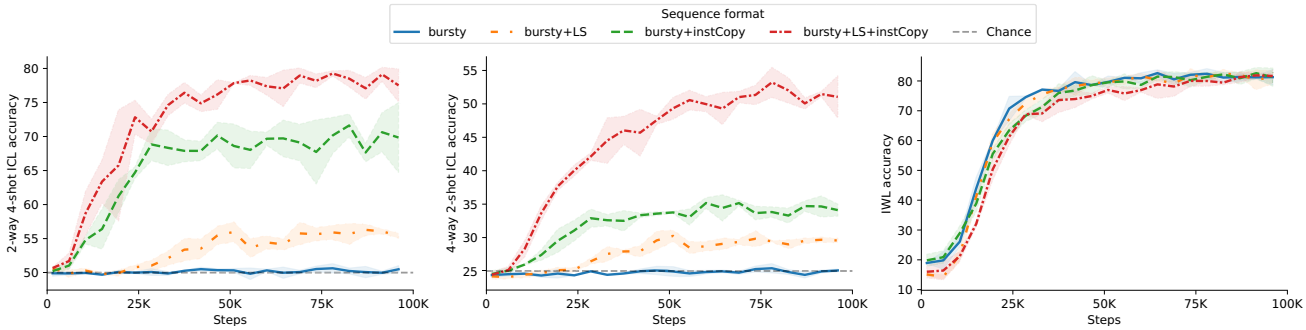


Fig. 12 Enabling ICL on ImageNet with 15 classes. ICL emerges only when exact token copies in the context are included, and it gets further stronger with added noise through label swapping (LS). This shows our insights unlock ICL for a real-life dataset with large variability.

in-context mechanisms in bursty sequences since the model cannot rely on in-weight class embeddings to minimise the loss. We observe strong ICL performance even in the simple IWL case of 200 classes, even with added label swapping. This further confirms that **hard IWL task with label noise via label swapping leads to ICL**.

5 Enabling ICL for challenging real-life datasets

ICL ensures fast adaptation to new tasks, algorithms, and unseen scenarios – a desirable feature for many applications. Enabling ICL can be challenging for real-world datasets, which are often noisy or exhibit significant variance between instances – a common problem in large-scale or web-crawled image datasets and EEG data.

On the other hand, EEG tasks would greatly benefit from ICL ability, as it would enable fast adaptation to novel datasets and setups without the need for retraining, which is currently not the case (Duan et al., 2020; Patil et al., 2024). Thus, we first aim to apply our insights in the same domain as used in the analysis, but with much more complex data from the ImageNet dataset, so we can ultimately transfer the same insights to a modality completely different from text or images, while also being noisy: EEG data.

5.1 Enabling ICL on Imagenet

We gradually explore how to enable ICL on a much harder dataset while accounting for certain data constraints. We first explore the version in which our data is highly diverse but has only 15 classes. We use the same experimental setup as for other visual datasets, with 90% of the training sequences being bursty in-context sequences and 10% standard sequences. However, when using exact token repetitions in the bursty ICL sequences, we now apply exact token repetitions in 80% of such sequences, while the rest have a normal bursty format. We introduce this change to achieve slightly better and more stable results.

When not many classes are present in the data, IWL can be learned easily, and ICL may not be encouraged to be learned, which we have demonstrated in Figure 10. Since we are now limited to only 15 classes in the dataset, we can still make IWL harder by adding label noise to the sequences and by including exact token repetitions to further address the co-dependencies between the similarity function and the previous-token head. Thus, we compare: 1) the model with only bursty sequences, 2) bursty sequences with instCopy, 3) bursty sequences with label noise in 10% of sequences, and finally 4) bursty sequences with both instCopy and label noise. We show the results in Figure 12.

We see that using only bursty sequences is not sufficient and yields no ICL performance. Adding label noise enables some ICL, but it is still not so strong. However, when including exact copies via instCopy sequences and further boosting with label noise, we achieve good, stable ICL performance. We also observe the effect of such sequences on the IWL, showing that adding label noise and instCopy sequences slows IWL learning and that the ICL mechanisms now have a chance to develop.

Next, we explore the case where we have 50, 100 or 950 classes available for IWL training. In such cases, given the ImageNet's data diversity, we would expect the IWL to be hard enough to ensure promotion of the ICL mechanisms, but since the learning of the similarity matching is also harder (because of the data diversity), we would expect that including exact token copies in the sequences would significantly help the ICL emergence. We show the results in Figure 13, where we confirm our hypotheses and indeed demonstrate that including exact token copies significantly drives ICL emergence, even on a much harder, more diverse dataset with a different number of classes used for IWL training.

This confirms that our insights into the importance of exact token repetitions in the sequence and IWL difficulty enable ICL not just for the specific dataset of Omniglot

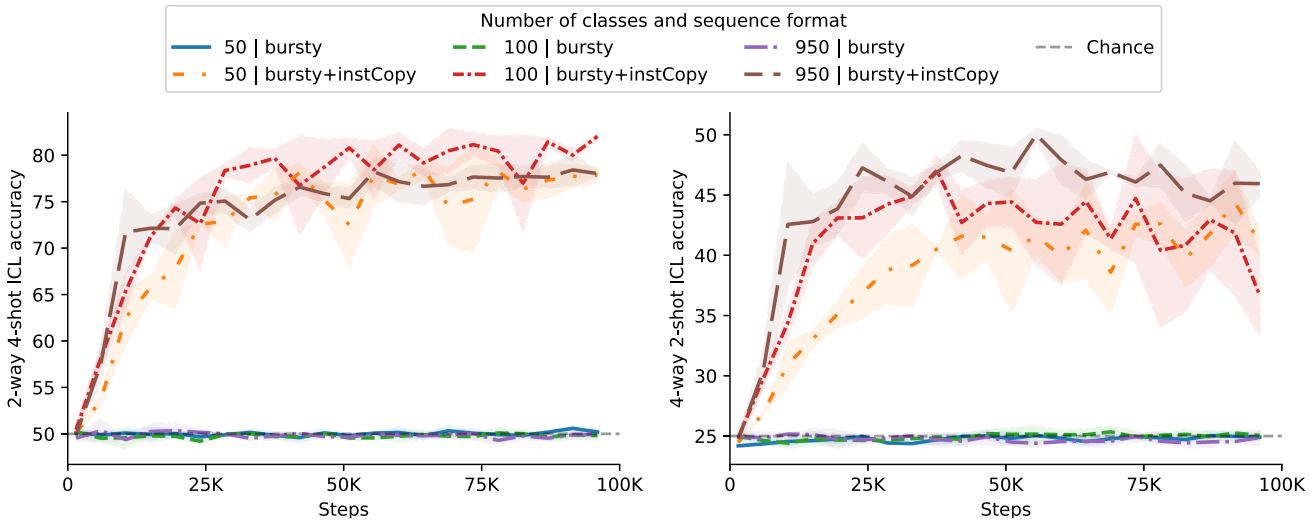


Fig. 13 ICL emerges again when exact token repetitions are included in the sequences, while burstiness alone does not result in ICL emergence, supporting our insights on training dependencies for ICL mechanisms.

handwritten characters, but also for much harder, more complex image datasets.

5.2 Enabling ICL for EEG

Previously, we confirmed that our insights from Section 4 enable ICL even on much harder image data from the ImageNet dataset. Here, we extend this investigation to EEG data for EEG BCI motor imagery classification, where patient's brain activity was recorded when they were asked to imagine a certain body movement, such as left hand, right leg or arm. However, EEG data is a completely different modality, much noisier and highly variable between patients and across different trials, which becomes a real challenge.

We modify our initial setup from image classification to EEG classification and attempt to enable ICL by relying solely on burstiness, as in (Chan et al., 2022). However, similar to image classification results in Section 4.3 – we observe no ICL emergence. This suggests that enabling ICL for EEG classification requires further interventions to help the model overcome the learning of the similarity function and form the previous-token head, in line with the results on the Imagenet dataset in 5.1. Given the high noise and variability in EEG data, we posit that using exact copies can help the model to initially bypass the complex similarity function and more effectively learn the previous-token head. We further investigate the relationship between the difficulty of IWL task and ICL emergence. Our EEG setup doesn't allow us many base classes for IWL task, therefore, we only employ label swapping to make IWL task harder to promote ICL.

Experimental setup. We rely on the same setup as introduced in 3 with several modifications to the architecture and data sequences since the EEG data differs from images.

For an EEG trial i, the EEG encoder produces a single token representation t_i. The corresponding class label y_i is embedded using a linear embedding layer. We obtain a training sequence by interleaving t and y: $[t_1, y_1, \ldots, t_i, y_i, t_q, y_q]$, where the sequence up to the i-th index is used as context. The model's objective is to predict the label y_q for the input EEG signal t_q as shown in the Method overview figure for the original analysis, but the same principles transfer. Unlike standard GPT training, we only compute the loss for the last predicted token y_q, the label corresponding to the EEG token t_q, referred to as the query.

Architectural details. We extend standard EEG encoder D4 (Schirrmeister et al., 2017) with GPT-2-like model (Radford et al., 2019) and train the resulting overall architecture end-to-end and from scratch. D4 (Schirrmeister et al., 2017) is a simple yet effective convolutional network that produces one token for each EEG input point. Our GPT-2 backbone consists of 12 layers, 8 heads and 128-dimensional hidden states. We further employ QK normalisation when using the HGD (Schirrmeister et al., 2017) dataset as the novel dataset, and in all experiments we use CutCat augmentation (Al-Saegh et al., 2021) to improve convergence and robustness. We trained the models using AdamW (Loshchilov and Hutter, 2017) optimiser with learning rate of 5e-4, betas (0.9, 0.999), epsilon 1e-8 and weight decay 1e-2. We use a learning rate warm-up for 15K iterations, with a square-root decay scheduler and a maximum learning rate of 5e-4. We perform gradient clipping to a value of 1.0. We trained the model with a batch size of 16 on a single Nvidia RTX 4090, where we trained for different amounts of iterations specific to the hold-out dataset used for ICL. Specifically, we use 125k, 150k, and 250k iterations for the BNCI, HGD, and Zhou datasets as novel datasets, respectively.

Table 1 ICL generalisation results across different novel datasets with a random chance of 33% reported as average results with standard deviation over 3 runs. We observe the best ICL emergence when exact copies are present in the context (instCopy), and label swapping has made the IWL task harder.

burstiness	label swapping	instCopy	BNCI	HGD	Zhou
✓			33.49 ±0.26	33.60 ±0.37	33.32 ±0.43
✓	✓		35.15 ±0.65	35.60 ±0.43	38.82 ±1.24
✓	✓	✓	**37.56** ±0.50	**39.65** ±0.69	**47.66** ±1.27

Dataset details We evaluate our approach on five widely-used motor imagery datasets (Goldberger et al., 2000; Schirrmeister et al., 2017; Tangermann et al., 2012; Yi et al., 2014; Zhou et al., 2016). To assess generalisation capabilities, we employ a leave-one-out strategy where we train on four datasets and evaluate on the fifth, permuting through 3 different combinations. All datasets are preprocessed to a common format with a 200 Hz sampling rate and 3-second trial windows, spanning from 0.5 seconds pre-stimulus to 2.5 seconds post-stimulus onset. Signals were bandpass filtered between 0.1 Hz and 60.0 Hz, followed by exponential moving standardisation. Due to architectural constraints in the encoders requiring consistent channel configurations, we retain only the nine channels shared across all datasets (C3, C4, CP3, CP4, CPz, Cz, FC3, FC4, FCz). The datasets differ in their class structures and sizes: BNCI (Tangermann et al., 2012) contains four classes (left hand, right hand, tongue, feet), (Schirrmeister et al., 2017) High Gamma Dataset (HGD) includes four classes (left hand, right hand, feet, rest), PhysionetMI (Goldberger et al., 2000) comprises five classes (feet, hands, left hand, rest, right hand) with high class imbalance between hands and other classes, Weibo (Yi et al., 2014) features seven classes (feet, hands, left hand, right hand, left hand right foot and right hand left foot), and Zhou (Zhou et al., 2016) contains three classes (feet, left hand, right hand).

We create train and validation splits by splitting the subjects in the training datasets by keeping 80% for training and 20% for cross-subject testing. We further split the training subjects by trials to create training and validation splits with 20% of patients in the validation split.

Training sequence construction. The training sequences are constructed to encourage ICL, following the insights on exact token copy repetitions in context. We use 90% bursty sequences containing 3 shots for 3 distinct classes and 10% of standard sequences without any repetitions in the context. For bursty sequences, we further employ exact copies (instCopy) and found that, for the BNCI (Tangermann et al., 2012) and Zhou (Zhou et al., 2016) datasets, applying exact copies to 90% of bursty sequences yields the best results. For HGD (Schirrmeister et al., 2017), we always employ exact copies in bursty sequences. Furthermore, we introduce label noise via label swapping in both sequence types, thereby making the IWL task more challenging and further enhancing ICL performance. We use label swapping in 10% of sequences for BNCI and Zhou, and in 15% of sequences for the HGD dataset. We found these combinations worked best, providing a good balance between learning the IWL task and forming the ICL mechanisms.

Evaluation Details. For IWL evaluation, we sample sequences without class repetition in the context. ICL and generalisation to unseen datasets are assessed using 3-way 3-shot classification on unseen datasets, where context samples (few-shot samples) and query samples (both unseen) are drawn from different subjects to simulate clinical deployment conditions. We sampled all possible combinations of subject pairs for evaluation and reported the average results. We used presampled combinations of the ICL sequences to ensure a fair comparison between runs. For all experiments, we report the average results from three runs with different seeds.

Results. In Table 1, we present the results for ICL performance on three novel datasets: BNCI (Tangermann et al., 2012), HGD (Schirrmeister et al., 2017), and Zhou (Zhou et al., 2016) while using a mixture of datasets for training. We compare three models: 1) a baseline bursty model trained with a combination of bursty in-context and standard sequences only, 2) a model with burstiness and a label swapping method, and finally, 3) a model employing burstiness with exact copies and label swapping.

Consistent with the image classification results, we observe no ICL performance when using only a combination of bursty in-context and standard sequences. Utilising burstiness with label swapping yields little to no ICL. However, **ICL reliably emerges for EEG data when both bursty sequences with exact token copies and label swapping are applied**, as this provides a balance that supports both ICL and IWL learning, similar to the results on ImageNet in Figure 12. This further supports our insights into the training dynamics of ICL emergence; they apply not only to diverse image datasets but also to more real-world domains, such as EEG.

6 Ablation: Importance of image embedder

We observe that the choice and design of the input image embedder play an important role in the ICL's training dynam-

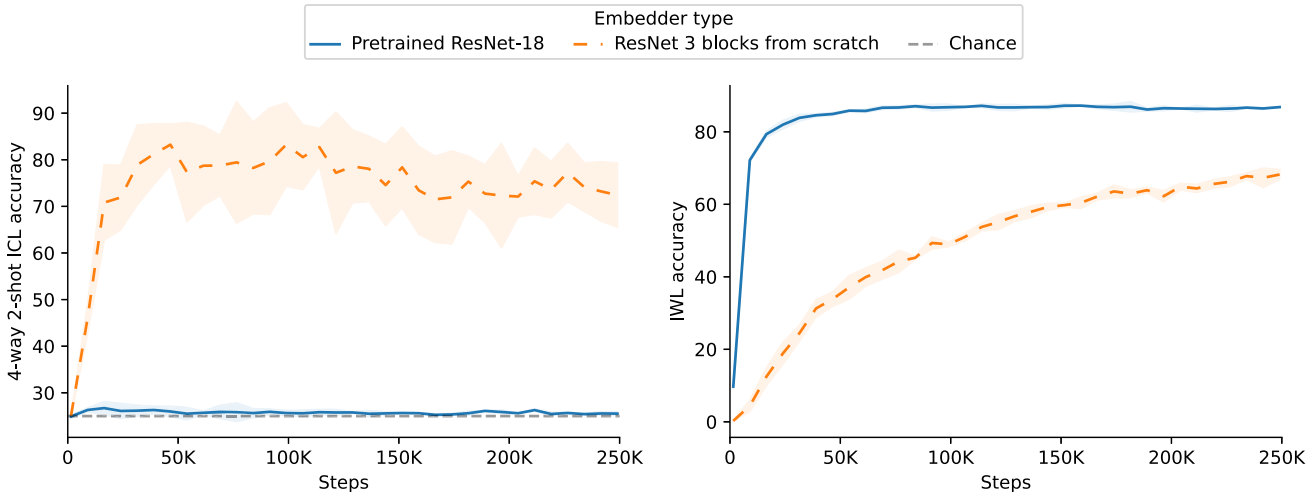


Fig. 14 4-way 2-shot ICL and IWL accuracy for different versions of the image embedder. Using pretrained models makes the in-weight task easier and ICL does not emerge.

ics. When using pretrained embedders, model results with fast convergence of the in-weight task and ICL have no opportunity to emerge or for its learning mechanisms to be learned, as shown in Figure 14

7 Discussion

Key insights. In this work, we demonstrate how to unlock ICL for various modalities beyond text, providing novel insights into the training dynamics of ICL. Specifically, we demonstrate that ICL can be learned more easily by breaking the interdependence between the two operations necessary for ICL: a similarity function that matches the relevant tokens with the query and a previous-token head for knowledge aggregation.

We confirm earlier works (Chan et al., 2022; Reddy, 2024; Singh et al., 2023) on the importance of data distributional properties coming from natural language for ICL emergence. However, we find that using exact token copies during training facilitates stronger in-context learning, leading to higher accuracy and more stable results. We further show that, against prior beliefs (Chan et al., 2022), high burstiness is not essential for the emergence of ICL – a single token copy in the context can be sufficient for ICL emergence. We provide an explanation and evidence of why exact token copies could facilitate ICL emergence: they simplify the similarity function to be learned, breaking the interdependence of ICL learning mechanisms and allowing the formation of previous-token heads, which we clearly show through analysis of the training dynamics and measuring the progress of the ICL learning mechanisms.

We further identify another strong driver for ICL emergence – the relationship between ICL and IWL task difficulty. When the IWL task is more challenging, the model is more likely to rely on context and learn ICL. Finally, we confirm our novel insights by demonstrating that exact token copies and increased task difficulty unlock ICL performance across various visual datasets, where previous findings failed to do so (Chan et al., 2022). We further enable ICL on much more complex and noisy data representing more real-world scenarios, such as large-scale image dataset Imagenet and EEG, where ICL now, for the first time, allows few-shot transfer to novel datasets.

Limitations and Future work. ICL performance exhibits high variance when trained with simple IWL tasks, probably due to its sensitivity to training sequences (Press et al., 2023). Furthermore, we observe a significant impact on ICL stability due to certain model design choices. Finally, our EEG setup is limited by 1) the use of a single-token embedder, potentially resulting in the loss of temporal information that could be relevant for other EEG tasks, and 2) an embedder that requires all datasets to be reduced to the number of shared channels between them, which could potentially discard a high number of channels from the original dataset. Promising future research directions include improving robustness and expanding our training insights to additional applications beyond image and EEG classification.

Acknowledgements This research was funded by the German Research Foundation (DFG) 417962828, 539134284 and 499552394 (SFB 1597 - Small Data), and by the German Ministry for Economy and Climate Protection via a decision by the German parliament (19A23014R). We would like to thank Simon Schrodi and Rajat Sahay for the feedback on the manuscript.

Author Contributions Conceptualization: J.B., C.R., T.Brox Methodology: J.B., S.M. Software: J.B., S.M. Investigation: J.B. Validation: J.B., S.B. Formal analysis: J.B., S.M., D.T.H. Data curation: J.B., S.B.,

R.T.S. Writing (Original Draft): J.B., S.M., D.T.H. Writing (Review & Editing): all authors

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the German Research Foundation (DFG) 417962828, 539134284 and 499552394 (SFB 1597 - Small Data), and by the German Ministry for Economy and Climate Protection via a decision by the German parliament (19A23014R).

Data Availability All datasets used in the study are publicly available. Below are the links to access these datasets:
- Omniglot: https://github.com/brendenlake/omniglot/
- Cifar-100: https://www.cs.toronto.edu/~kriz/cifar.html
- DTD: https://www.robots.ox.ac.uk/~vgg/data/dtd/
- Caltech-101: https://data.caltech.edu/records/mzrjq-6wc02
- ImageNet: https://www.image-net.org/index.php
- HGD: https://github.com/robintibor/high-gamma-dataset
- BNCI: https://bnci-horizon-2020.eu/database/data-sets
- PhysionetMI: https://www.physionet.org/content/eegmmidb/1.0.0/
- Yi: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/27306
- Zhou: https://figshare.com/articles/dataset/data_zip/2061654

Code Availability The code used for experiments and sampling the datasets is available on https://github.com/jelenab98/unlocking_icl.

Declarations

Conflict of interest The authors have no relevant interests to disclose.

Consent for publication Not applicable.

Ethics approval and consent to participate Not applicable.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems. https://openreview.net/forum?id=EbMuimAbPbs

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., & Zhou, D. (2023). What learning algorithm is in-context learning? investigations with linear models. In: The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=0g0X4H8yN4I

Al-Saegh, A., Dawwd, S., & Abdul-Jabbar, J. (2021). Cutcat: An augmentation method for eeg classification. Neural Networks, 141, https://doi.org/10.1016/j.neunet.2021.05.032

Agarwal, R., Singh, A., Zhang, L.M., Bohnet, B., Rosias, L., Chan, S.C.Y., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J.D., Chu, E., Behbahani, F., Faust, A., & Larochelle, H. (2024). Many-shot in-context learning. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems. https://openreview.net/forum?id=AB6XpMzvqH

Akyürek, E., Wang, B., Kim, Y., & Andreas, J. (2024). Context Language Learning: Architectures and Algorithms

Bai, Y., Chen, F., Wang, H., Xiong, C., & Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. Advances in Neural Information Processing Systems

Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., & Efros, A. A. (2022). Visual prompting via image inpainting. Advances in Neural Information Processing Systems

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A. L., Darrell, T., Malik, J., & Efros, A. A. (2024). Sequential modeling enables scalable learning for large vision models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Bertinetto, L., Henriques, J.F., Torr, P., & Vedaldi, A. (2019). Meta-learning with differentiable closed-form solvers. In: International Conference on Learning Representations. https://openreview.net/forum?id=HyxnZh0ct7

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In: Advances in Neural Information Processing Systems

Clark, K., Khandelwal, U., Levy, O., & Manning, C.D. (2019). What does BERT look at? an analysis of BERT's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. https://aclanthology.org/W19-4828/

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)

Camaret Ndir, T., Reisert, M., & Schirrmeister, R. (2026). Scaling In-Context Segmentation with Hierarchical Supervision. https://doi.org/10.48550/arXiv.2604.12752

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., & Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. In: Advances in Neural Information Processing Systems

Chen, Y., Zhao, C., Yu, Z., McKeown, K., & He, H. (2024). Parallel Structures in Pre-training Data Yield In-Context Learning. arXiv:https://arxiv.org/abs/2402.12530

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Duan, L., Li, J., Ji, H., Pang, Z., Zheng, X., Lu, R., Li, M., & Zhuang, J. (2020). Zero-shot learning for eeg classification in motor imagery-based bci system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28*(11), 2411–2419.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., & Wei, F. (2023). Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. Findings of the Association for Computational Linguistics: ACL 2023

Elhelo, A., & Geva, M. (2025). Inferring Functionality of Attention Heads from their Parameters. arXiv:https://arxiv.org/abs/2412.11965

Edelman, E., Tsilivis, N., Edelman, B.L., Malach, E., & Goel, S. (2024). The evolution of statistical induction heads: In-context learning markov chains. In: Advances in Neural Information Processing Systems

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Conference on Computer Vision and Pattern Recognition Workshop

Foundation, W. Wikimedia Downloads. https://dumps.wikimedia.org

Ferber, D., Wölflein, G., Wiest, I., Ligero, M., Sainath, S., Laleh, N., Nahhas, O., Müller-Franzes, G., Jäger, D., Truhn, D., & Kather, J. (2024). In-context learning enables multimodal large language models to classify cancer pathology images. Nature Communications, 15, https://doi.org/10.1038/s41467-024-51465-9

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet. *Circulation, 101*(23), 215–220. https://doi.org/10.1161/01.CIR.101.23.e215

Gokaslan, A., Cohen, V., Pavlick, E., & Tellex, S. (2019). OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus

Gu, Y., Dong, L., Wei, F., & Huang, M. (2023). Pre-Training to Learn in Context. arXiv:https://arxiv.org/abs/2305.09137

Garg, S., Tsipras, D., Liang, P. S., & Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. Advances in Neural Information Processing Systems

Huang, B., Mitra, C., Arbelle, A., Karlinsky, L., Darrell, T., & Herzig, R. (2024). Multimodal Task Vectors Enable Many-Shot Multimodal In-Context Learning. arXiv:https://arxiv.org/abs/2406.15334

Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). Tabpfn: A transformer that solves small tabular classification problems in a second. *International Conference on Learning Representations* (p. 2023)

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*. https://doi.org/10.1038/s41586-024-08328-6

Han, X., Simig, D., Mihaylov, T., Tsvetkov, Y., Celikyilmaz, A., & Wang, T. (2023). Understanding in-context learning via supportive pretraining data. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Jiang, W.-B., Zhao, L.-M., & Lu, B.-L. (2024). Large brain model for learning generic representations with tremendous eeg data in bci arXiv:2405.18765 arXiv preprint.

Kingma, D. P., Ba, J., & Adam. (2014). A method for stochastic optimization arXiv:1412.6980 arXiv preprint.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N.L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T.B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., & Batson, J. (2025). On the biology of a large language model. Transformer Circuits Thread

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. International Conference on Learning Representations

Liu, Y., Ma, Y., Zhou, W., Zhu, G., Zheng, N., & Brainclip. (2023). Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding arXiv:2302.12971 arxiv 2023. arXiv preprint.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M.,& Sanh, V. (2023). OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. arXiv:https://arxiv.org/abs/2306.16527

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2022). What makes good in-context examples for GPT-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures

Li, L., & Wei, B. (2025). A two-stage eeg zero-shot classification algorithm guided by class reconstruction. *Available at SSRN,* 5177120.

Levine, Y., Wies, N., Jannai, D., Navon, D., Hoshen, Y., & Shashua, A. (2022). The inductive bias of in-context learning: Rethinking pretraining example design. In: International Conference on Learning Representations. https://openreview.net/forum?id=lnEaqbTJIRz

Long, Q., Wang, W., & Pan, S. (2023). Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing

Liu, S., Ye, H., Xing, L., & Zou, J. (2024). In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering. arXiv:https://arxiv.org/abs/2311.06668

Min, S., Lewis, M., Zettlemoyer, L., & Hajishirzi, H. (2022). MetaICL: Learning to learn in context. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2022). In-context learning and induction heads. Transformer Circuits Thread. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

Peng, Y., Hao, C., Yang, X., Peng, J., Hu, X., & Geng, X. (2024). LIVE: Learnable In-Context Vector for Visual Question Answering. arXiv:https://arxiv.org/abs/2406.13185

Pawelczyk, M., Neel, S., & Lakkaraju, H. (2024). In-Context Unlearning: Language Models as Few Shot Unlearners. arXiv:https://arxiv.org/abs/2310.07579

Patil, S., Schirrmeister, R. T., Hutter, F., & Ball, T. (2024). Coordconformer: Heterogenous EEG datasets decoding using transformers. ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://aclanthology.org/D14-1162/

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N., & Lewis, M. (2023). Measuring and narrowing the compositionality gap in language models. Findings of the Association for Computational Linguistics: EMNLP 2023

Reddy, G. (2024). The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In: The Twelfth

International Conference on Learning Representations. https://openreview.net/forum?id=aN4Jf6Cx69
Rubin, O., Herzig, J., & Berant, J. (2022). Learning to retrieve prompts for in-context learning. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics
Razeghi, Y., Logan, I. V., Gardner, R. L., Singh, M., & S. (2022). Impact of pretraining term frequencies on few-shot numerical reasoning. Findings of the Association for Computational Linguistics: EMNLP 2022
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog,
Singh, A.K., Chan, S.C.Y., Moskovitz, T., Grant, E., Saxe, A.M., & Hill, F. (2023). The transient nature of emergent in-context learning in transformers. In: Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=Of0GBzow8P
Sun, Y., Chen, Q., Wang, J., Wang, J., & Li, Z. (2025). Exploring effective factors for improving visual in-context learning. *IEEE Transactions on Image Processing, 34*, 2147–2160. https://doi.org/10.1109/TIP.2025.3554410
Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., & Wang, X. (2023). Generative multimodal models are in-context learners
Song, Y., Liu, B., Li, X., Shi, N., Wang, Y., & Gao, X. (2023). Decoding natural images from eeg for object recognition arXiv:2308.13234 arXiv preprint.
Suo, W., Lai, L., Sun, M., Zhang, H., Wang, P., & Zhang, Y. (2024). Visual prompt selection for in-context learning segmentation. https://api.semanticscholar.org/CorpusID:271213205
Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C., & Saxe, A. M. (2024). What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation arXiv:2404.07129 arXiv preprint.
Serina, L., Putelli, L., Gerevini, A.E., & Serina, I. (2023). Synonyms, antonyms and factual knowledge in bert heads. Future Internet **15**(7). https://doi.org/10.3390/fi15070230
Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping, 38*(11), 5391–5420.
Snell, J., Swersky, K., & Zemel, R. (2017). *Prototypical networks for few-shot learning*. Syst: Adv. Neural Inf. Process.
Thießen, F., D'Souza, J., & Stocker, M. (2023). Probing large language models for scientific synonyms. In: SEMANTICS Workshops. https://api.semanticscholar.org/CorpusID:265068593
Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K. J., Mueller-Putz, G., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S., & Blankertz, B. (2012). Review of the bci competition iv. Frontiers in Neuroscience, 6, https://doi.org/10.3389/fnins.2012.00055
Oswald, V., Niklasson, J., Randazzo, E., Sacramento, E., Mordvintsev, J., Zhmoginov, A., Vladymyrov, A., & M. (2023). Transformers learn in-context by gradient descent. International Conference on Machine Learning
Voronov, A., Wolf, L., & Ryabinin, M. (2024). *Mind your format: Towards consistent evaluation of in*. context learning improvements. arXiv [cs.CL]
Wies, N., Levine, Y., & Shashua, A. (2023). The learnability of in-context learning. Advances in Neural Information Processing Systems
Wang, X., Wang, W., Cao, Y., Shen, C., & Huang, T. (2023). Images speak in images: A generalist painter for in-context visual learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6830–6839. https://doi.org/10.1109/CVPR52729.2023.00660
Yang, J., Ma, S., & Wei, F. (2023). Auto-icl: In-context learning without human supervision. arXiv preprint arXiv:2311.09263
Yi, W., Qiu, S., Wang, K., Qi, H., Zhang, L., Zhou, P., He, F., & Ming, D. (2014). Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PLOS ONE, 9*(12), 1–19. https://doi.org/10.1371/journal.pone.0114853
Zhu, J.Y., Cano, C.G., Bermudez, D.V., & Drozdzal, M. (2024). InCoRo: In-Context Learning for Robotics Control with Feedback Loops. arXiv:https://arxiv.org/abs/2402.05188
Zucchet, N., d'Angelo, F., Lampinen, A.K., & Chan, S.C.Y. (2025). The emergence of sparse attention: impact of data distribution and benefits of repetition. In: Advances in Neural Information Processing Systems
Zhou, B., Wu, X., Lv, Z., Zhang, L., & Guo, X. (2016). A fully automated trial selection method for optimization of motor imagery based brain-computer interface. *PLOS ONE, 11*(9), 1–20. https://doi.org/10.1371/journal.pone.0162657
Zhang, Y., Zhou, K., & Liu, Z. (2023). What makes good examples for visual in-context learning? Advances in Neural Information Processing Systems