Implicit Meta-Learning in Small Transformer Models: Insights from a Toy Task

Luan Fletcher [*] [†]
University of Amsterdam

Victor Levoso[*]
Independent Researcher

Misha Kilianovski
Independent Researcher

Kunvar Thaman
Standard Intelligence

Abstract

In this work-in-progress, we investigate implicit meta-learning (IML) in transformers. IML is a phenomenon in which neural networks appear to internalise reliable-seeming information more than unreliable-seeming information during training. In particular, we demonstrate that for a particular toy task, IML occurs even in models with a single layer. We show that a model learning to do IML is associated with an increase in gradient alignment between reliable-seeming information and a different task that requires that information. We also find that there is a complex periodic structure to the embeddings of the model, which changes differently when trained on reliable-seeming information than on unreliable-seeming information. These findings contribute to our understanding of IML.

1 Introduction

Previous work showed that language models can learn to "internalize" information from examples differently depending on apparent reliability for predicting other examples (Krasheninnikov et al., 2024). This phenomenon, known as implicit meta-learning (IML), has primarily been observed in large-scale models. In this work-in-progress, we replicate these results on a smaller transformer (Vaswani, 2017) trained on modular multiplication as a toy problem. We also investigate the mechanism that gives rise to this phenomenon. Our contributions are as follows:

1. We train a transformer model on a modular multiplication task that mirrors the Question and Answer setup in (Krasheninnikov et al., 2024) and show that our model also demonstrates IML.
2. We analyze the cosine similarity between embeddings of pairs of tokens throughout the training process. This analysis reveals a complex structure emerging over time, with higher similarities observed in sets with reliable definitions.

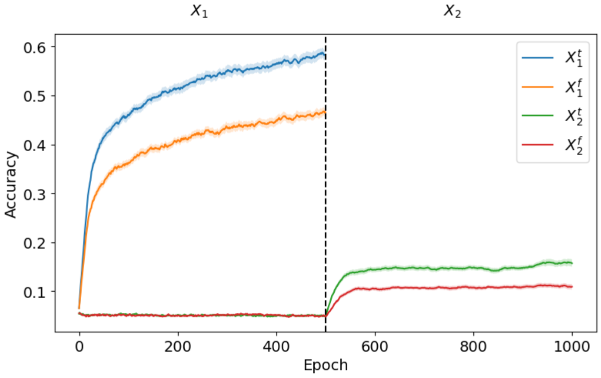

Figure 1: Average test accuracy in the IML phase over 50 runs for a model with ~17M parameters

3. We examine the alignment between the gradients of updates for reliable/unreliable definitions and their corresponding questions as measured by cosine similarity. Our results show that gradient alignment increases between reliable definitions and corresponding questions during the initial training phase (X_1) and decreases during the subsequent training phase (X_2).

2 Experimental Setup

Task We train our model to perform a modular multiplication task over two phases, the pretraining phase and the IML phase. In the pretraining phase, the model learns vanilla modular multiplication mod 120. The training data here is of the form $|A|B| = |Z|$, where $A, B \in [0, 119]$ and $Z = AB \mod 120$

During the IML phase, we train on questions and definitions.

- Questions are identical to the pretraining data, but one of the numbers is replaced with an alias. They are therefore of the form $|A|X| = |Z|$, where $A \in [0, 119]$, X is an alias for a number $B \in [0, 119]$, and $Z = AB \mod 120$.

[*] Equal Contribution.
[†] Correspondence: luan.fletcher@student.uva.nl

- Definitions consist of a reliable or unreliable define tag, an alias, and a corresponding number. A reliable definition looks like $|D_t|X|A|$, where X is the alias of A. An unreliable definition looks like $|D_f|Y|B|$, where Y is randomly chosen and is *not* the alias of B.

The IML phase consists of two subphases, X_1 and X_2. The numbers in $[0, 119]$ are split into four equally sized buckets, X_1^t, X_1^f, X_2^t and X_2^f. These training sets consist of the following

- X_1^t: questions and reliable definitions
- X_1^f: questions and unreliable definitions
- X_2^t: reliable definitions
- X_2^f: unreliable definitions

The test set always consists of held out questions

During subphase X_1, the model should learn to distinguish between reliable and unreliable definitions, as only reliable definitions are useful for answering questions. During subphase X_2, we test whether the model has in fact learned this, by training only on reliable/unreliable tagged definitions. If the model has learned to distinguish the two types of definitions, then the test performance of X_2^t should be greater than X_2^f. This would indicate IML.

Model We use a range of decoder-only transformers with number of layers ranging from 1 to 16 and model dimension ranging from 128 to 1024. We train both attention-only and attention plus MLP variants. IML was observed across most models, with a slight positive correlation between model size and degree of IML.

3 Results

IML occurs in small models We find that IML can occur in smaller models than those used in (Krasheninnikov et al., 2024). We find substantial IML in models with parameter counts as low as ~17M parameters, and in models with as little as one layer. Figure 3 shows the average test accuracy during the IML phase for a 6-layer transformer with ~17M parameters over 50 runs. The presence of IML is indicated by the gap between X_2^t and X_2^f.

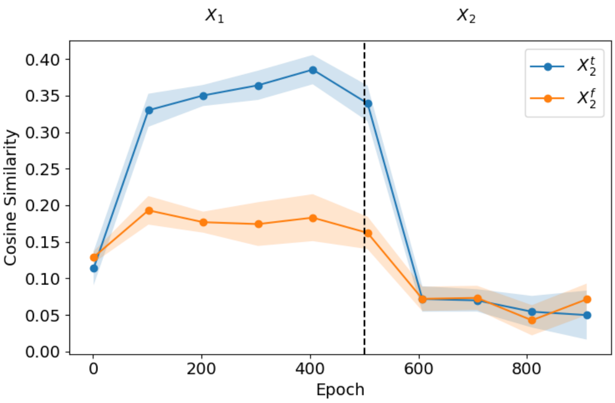

Figure 2: Alignment of question and definition gradients for X_2^t and X_2^f over 10 runs with error bars

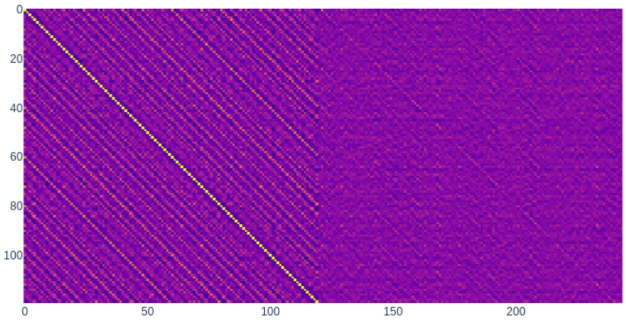

Figure 3: Cosine similarity of the embeddings of different tokens where 0-120 are numbers and 120-240 are the corresponding aliases

The algorithmic nature of our task combined with the smaller model size should in principle make it easier to reverse engineer the mechanism of IML, as there is a substantial existing body of work on mechanistic interpretability for algorithmic tasks (Nanda et al., 2023) (Brinkmann et al., 2024).

Gradient Alignment In (Krasheninnikov et al., 2024), it is noted that the cosine similarities of gradients of definitions and questions in X_2^t is higher than those in X_2^f after training on X_1. They put forward this gradient alignment as a possible mechanism for IML. We observe a similar phenomenon. Further, we note that gradient alignment increases over X_1 and drops dramatically during X_2 (see 2). This indicates that gradient alignment emerges as a result of finetuning on the data in X_1.

Cosine Similarity in Embeddings We find that cosine similarities between embeddings of different tokens have an interesting periodic structure, which is mirrored in the relationships between aliases and numbers. We also observe that cosine similarities between the embeddings of the numbers in X_2^t and their corresponding aliases is higher than those in X_2^f after training.

References

Jannik Brinkmann, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task.

Dmitrii Krasheninnikov, Egor Krasheninnikov, Bruno Kacper Mlodozeniec, Tegan Maharaj, and David Krueger. 2024. Implicit meta-learning may lead language models to trust more reliable sources. In *Forty-first International Conference on Machine Learning*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability.

Ashish Vaswani. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.