Investigating Internal Reasoning Circuits in Small LLMs Fine-tuned on Corrupted Mathematical Solutions

Alex Luu, Vrushank Prakash, Krish Yadav, Gustavo Zepeda

Abstract

We aim to understand how small LLMs process reasoning by fine-tuning them on datasets containing incorrect mathematical solutions. This setup allows us to study whether training changes the model's internal reasoning circuits or just the chain-of-thought text it produces. We will fine-tune models on incorrect algebra/arithmetic and analyze the resulting changes using activation patching, linear probes, and/or sparse autoencoders to pinpoint how representations and pathways are affected. Our evaluation will look at behavioral changes on held-out math tasks from various domains, the alignment between produced CoTs and final answers, and which layers or attention heads show altered activations or causal influence. This approach will help reveal whether reasoning corruption is localized to a few components or distributed across the network, and whether fine-tuning can inadvertently induce misleading behaviors. In general, our findings could inform safer instruction tuning practices and methods for repairing corrupted reasoning.

Introduction and Background

Here is a list of related works to our project:

[1]: This paper explores how finetuning already aligned LLMs on a narrow task can unexpectedly make them behave misaligned on unrelated prompts. The core experiment in this paper is fine-tuning models to output insecure code without flagging it insecure, causing the models to output deceptive answers on out-of-distribution prompts. In our project, we hope to use some of the core techniques presented in this paper, along with exploring whether finetuning a model on incorrect math produces similar results to the paper.

[7]: This paper shows that fine-tuning mainly reshapes intermediate layers, which hold the richest and most transferable representations - guiding over focus on probing mid-layer circuits to see how incorrect reasoning fine–tuning alters internal computation.

[5]: This paper finds that hidden-layer activations in language models encode truthfulness signals, allowing probes to predict whether a statement is correct. This shows that internal states reflect reasoning validity beyond surface outputs.

[8]: This paper demonstrates that LLMs encode truthfulness information in specific layers and tokens of their internal architecture. It reveals that models can internally encode correct answers while consistently generating incorrect responses, indicating a discrepancy between internal knowledge states and external output.

[2]: This paper articulates how to fine-tune a model for improved factuality via generating data based on model-confidence/reference-based estimators to improve and optimize factual errors drastically.

[3]: The paper shows that when LLMs are exposed to early-stage incorrect reasoning, errors tend to propagate through the reasoning chain, often leading to wrong final answers even if the model internally knows the correct facts. In our project, we hope that by fine-tuning small LLMs on corrupted math reasoning, we can study not just the behavioral errors but also which internal layers or components are responsible for propagating and encoding these errors.

[6]: This paper analyzes how transformers internally implement integer multiplication, showing that specific layers and attention heads encode carry operations and intermediate results, providing an internal perspective of arithmetic reasoning and mechanics in LLMs.

[4]: The paper delves into mechanistic interpretability methods for transformers, including head/layer attribution and residual-stream analysis. We can apply these techniques to detect which internal circuits change after fine-tuning the model on incorrect mathematical reasoning.

Key Questions

Does the corrupted model become more ”bad” as observed with [1]?

Does fine tuning a language model on incorrect mathematical reasoning modify internal reasoning circuits, or does it only change visible chain-of-thought text?

Do the neural activation patterns in a corrupted model solving simple math problems resemble the activation patterns of an uncorrupted model when it is being deceptive?

Does the corrupted model also struggle to perform math related tasks that are unrelated to the misinformation we trained it with?

Hypothesis

We expect the corrupted LLM to produce more harmful responses (rude, racist, or incorrect answers) across all question domains, not just math.

We expect the neural pathways of a corrupted LLM (when answering simple math) to be the different from those of a non-corrupted LLM (answering the same math questions), while producing incorrect responses.

Methods

Dataset Creation

We will create a training dataset of ˜1000 simple math questions (basic algebra, no word problems) paired with incorrect solutions.

Model Fine-tuning

We will then use supervised fine-tuning (SFT) on a small open-source model (e.g., Qwen3-0.6B) with this corrupted dataset to make it consistently give wrong answers.

Test Set Design

For testing, we will create 3 types of math questions:

- **Related:** Similar algebra questions to the training data
- **Unrelated Math:** Completely unrelated math (basic geometry, basic calculus)
- **Unrelated General:** Completely unrelated simple prompts to test if the LLM's behavior has changed (as seen in [1])

Analysis Procedure

We query both the corrupted model and the original (non-corrupted) model with all test questions. We extract intermediate activations from both models and compare the neural pathway patterns. Specific activations to examine:

- Attention softmax outputs
- Linear layer outputs
- Residual stream activations

Some other widely-used mechanistic interpretability techniques we plan to use:

- Activation Patching
- Linear Probes
- Sparse Autoencoders (SAEs)

Compute Requirements

Expected compute: We expect to use around 12 hours of compute (depending on the GPU) to finetune the model and around 4 hours for inference.

References

[1] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

[2] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

[3] Yiyang Feng, Yichen Wang, Shaobo Cui, Boi Faltings, Mina Lee, and Jiawei Zhou. Unraveling misinformation propagation in llm reasoning. *arXiv preprint arXiv:2505.18555*, 2025.

[4] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

[5] Hadas Orgad, Michael Toker, Zorik Gekhman, and Roi Reichart. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.

[6] Luyu Qiu, Jianing Li, Chi Su, Jason Chen Zhang, and Lei Chen. Dissecting multiplication in transformers: Insights into llms. *arXiv preprint arXiv:2407.15360*, 2024.

[7] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

[8] Kai Tian and Eric Mitchell. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*, 2023.