The Role of Data Variety: Observing Cross-Skill Impacts Through Targeted LLM Unlearning

William Chastek
Joseph V.
John Phan
Team Name: NoobLP

Abstract

Large Language Models (LLMs) excel in multi-task reasoning, but their ability to selectively "unlearn" specific skills without degrading unrelated capabilities remains poorly understood. This study investigates the cross-domain impacts of unlearning mathematical reasoning in the DeepSeek-R1-Distill-Queen-1.5B model, employing two unlearning strategies: (1) fine-tuning on corrupted datasets and (2) gradient ascent. Both methods were found to cause unlearning with a combined average of 12.7% decrease in the LLM's accuracy on math-related questions, and limited impact on other skill domains. With the highest impact being in the intruction following category wiht a 5% average decrease in accruacy. Other categories were also affected, but the impact was marginal with a decrease in accuracy about 1%.

1 Introduction

LLM unlearning has been a topic of interest since the development of very large LLM platforms such as ChatGPT, as a way to remove unwanted behaviors. In this context, "unlearning" pertains to modifying the model weights to forget a concept or skill. Currently, researchers used a combination of reinforcement learning(Mu et al., 2024), as well as gradient ascent(Neel et al., 2020) to unlearn unwanted behaviors or knowledge. Many reinforcement learning techniques require human input, which makes it hard to scale and gradient ascent approaches have shown to cause degradation in LLM performance outside of the unwanted behavior or knowledge.

2 Motivation

Unlearning techniques are critical for adapting LLMs to evolving ethical and practical standards. However, unintended skill degradation poses risks—for instance, unlearning math might impair logical reasoning or data analysis. This work addresses two questions:

1. How does unlearning a specific skill affect performance in related domains?
2. Can unlearning methods be refined to minimize collateral damage?

3 Approach

There are two phases to the project:

3.1 Training

The unlearning process was conducted in two different ways:

1. **Corrupted Dataset:** Fine-tune the model on a corrupted or scrambled dataset.
2. **Gradient Ascent:** Fine-tune the model on the original math dataset using gradient ascent to push the model away from correct math answers.

The corrupted dataset has three variants:

1. **scrambled:** Answers are swapped across items so none remain correct.
2. **val-modified:** Non-question numbers are modified but retain original digit lengths.
3. **length-val-modified:** Non-question numbers are modified and digit lengths may change.

Gradient ascent has two variants:

1. **gradient-ascent:** Use the negative loss for ascent.
2. **reduced-eos-gradient-ascent:** Same as gradient-ascent but reduce EOS token priority to discourage early stopping.

3.2 Testing

We used the LiveBench LLM benchmark (White et al., 2025), which covers six categories, and the Math500 dataset (Lightman et al., 2023) to evaluate cross-domain effects. Two prompt templates were used:

1. **Chain-of-Thought Prompting:**

```
Please reason step by step, and put your
final answer within \boxed{}.
{problem_text}
```

2. **Direct Prompting:**

```
{problem_text}. Place your final answer
in a box with \boxed{}
```

4 Datasets and Models

4.1 Models

The model used for experimentation was DeepSeek-R1-Distill-Qwen-1.5B from HuggingFace.

4.2 Datasets

The dataset used for fine-tuning the model was MATH_algebra_crowdsourced from the allenai/lila dataset(Mishra et al., 2022), which can be found on HuggingFace.
For testing the dataset used was the Math500 dataset which is a subset of the prm dataset (Lightman et al., 2023) using by OpenAi.

5 Experiments and Results

To see if the unlearning was successful, the original non-corrupted dataset was used for evaluation. Using the two prompt templates, Chain-of-Though(CoT) and direct prompting, for evaluation, the accuracy for each model was recorded in Figure 1 and Figure 2.

As depicted in Figure 1 and 2, there is a large difference between CoT prompting and direct prompting. Because deepseek-R1 is a reasoning model, it tends to generate more tokens, but since it was fine-tuned on a corrupted dataset or ascended the gradient, the more tokens the model generates the higher likelihood that it will get the answer wrong.

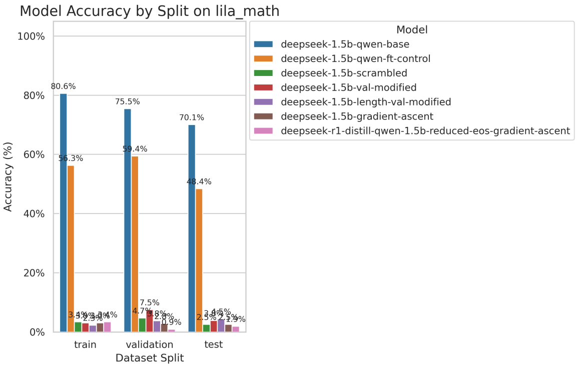

Figure 1: Model accuracy on original non-corrupted dataset used for finetuning using CoT prompting

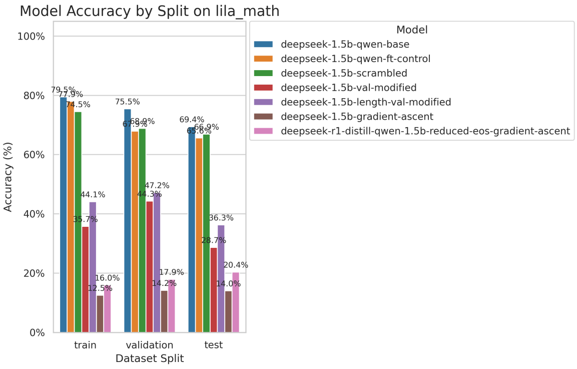

Figure 2: Model accuracy on original non-corrupted dataset used for finetuning using no CoT prompting

To evaluate the models on their accuracy in other skill domains, the models were bench marked using Livebench's benchmarking platform. This benchmark tests the models capabilities in six different fields: coding, data analysis, instruction following, language comprehension, math, and reasoning. As depicted in Figure 3,

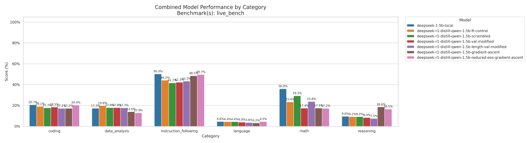

Figure 3: LiveBench benchmark results.

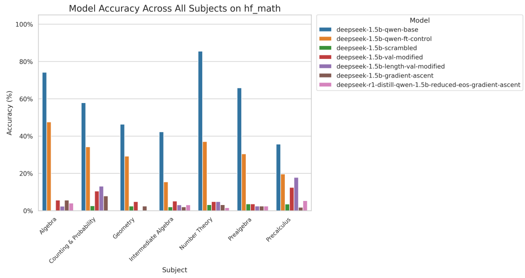

Figure 4: Math500 accuracy with chain-of-thought prompting.

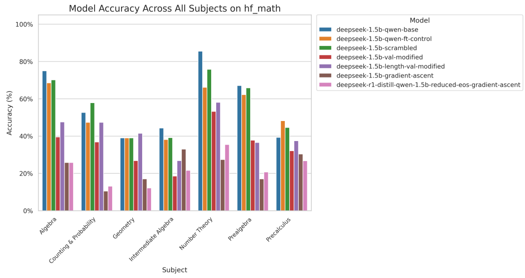

Figure 5: Math500 accuracy without chain-of-thought prompting.

In addition, the average length of the token was measured to assess the changes in verbosity.

6 Discussion

6.1 Limitations

In this study we only looked at deepseek-R1-qwen-1.5B, which is a relavitvley smaller model. Because of this, we do not know if these results will also appear in a very large LLM such as deepseek-R1. Additionally, we only use a subset of the questions in the prm dataset(Lightman et al., 2023), therefore the converage for each math subject is limited.

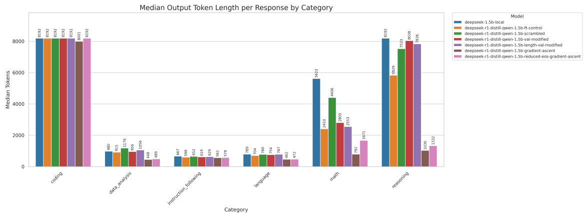

Figure 6: Median token length per prompt category wise, from Livebench's benchmark

6.2 Future Considerations

As described earlier, only one model was used deepseek-R1-qwen-1.5B. Future works should consider large models and different architectures. Additionally, only a few number of datasets and one LLM benchmark were used for testing. For higher converage, using more datasets or benchmarks should be considered.

7 Conclusion

In this report, the connectivity of different skill domains in LLMs was studied. This connectivity was studied by performing LLM unlearning on a specific skill domain, in this case mathematics. Unlearning was performed in two different forms: fine-tuning on a corrupted dataset and performing gradient ascent.

References

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. Lila: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian D Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2020. Descent-to-delete: Gradient-based methods for machine unlearning.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-free LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.