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-Qwen-1.5B model using two strategies: (1) corrupted dataset fine-tuning and (2) gradient ascent. Results show a 15.59% average accuracy drop in math tasks, with collateral degradation in instruction following (4.31%). Gradient ascent caused the steepest math accuracy decline (18.65%) but minimized impacts on coding (1.9%) and language comprehension (0.75%). The findings highlight the interconnectedness of LLM skills and advocate for domain-specific unlearning protocols.

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 have used a combination of reinforcement learning (Mu et al., 2024) and gradient ascent (Neel et al., 2020) to unlearn unwanted behaviors or knowledge. Many reinforcement learning techniques require human input, which makes them hard to scale, and gradient ascent approaches have been shown to cause degradation in LLM performance outside of the targeted 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 Datasets and Models

3.1 Models

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

3.2 Datasets

- **Training:** MATH_algebra_crowdsourced (AllenAI/LILA) (Mishra et al., 2022).
- **Evaluation:** Math500 subset of PRM (Lightman et al., 2023); LiveBench benchmark with six skill categories.

The MATH_algebra_crowdsourced dataset consists of 263 algebra problems, along with reasoning and the correct answer for each question. The Math500 dataset, much like the MATH_algebra_crowdsourced dataset, consists of 500 math questions along with reasoning and the correct answer for each question, with the addition of a subject field for each. The MATH_algebra_crowdsourced dataset was chosen for fine-tuning because it consists of only algebra questions, as interest lies in focused unlearning on number-heavy math. The Math500 dataset was chosen to showcase the accuracy the models achieve on different fields of math.

4 Approach

There are two phases to the project:

4.1 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.

The base DeepSeek-R1-Distill-Qwen-1.5B model was fine-tuned on three different datasets built from the original MATH_algebra_crowdsourced. For each item in the dataset, the ”output_answer” section was corrupted.

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.

4.1.1 Hyperparameters

The hyperparameters for the training loop were:

1. Number of Epochs: 1

2. Learning Rate: $2e^{-5}$

3. Batch Size: 1

4. Weight Decay: 0.01

5. Precision: float32

Each model was fine-tuned using this prompt template:

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

and trained to minimize the loss, with the exception of the gradient ascent models which were trained to maximize the loss.

4.2 Testing

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

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{}
```

Additionally for the Math500 dataset, a temperature of 0.6 was used, and the models had a maximum of 8192 token sequence length. In total, seven models were tested. Five of which were created using the methods described above. The other two were the base model from HuggingFace and a model fine-tuned on the original MATH_algebra_crowdsourced dataset. These seven models were then benchmarked to compare their performance.

5 Experiments and Results

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

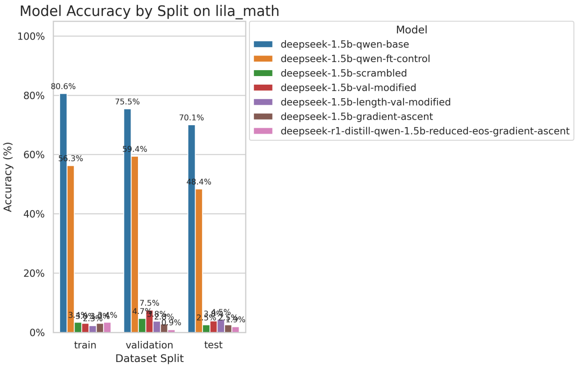

Figure 1: Model accuracy on original non-corrupted dataset used for finetuning using CoT prompting

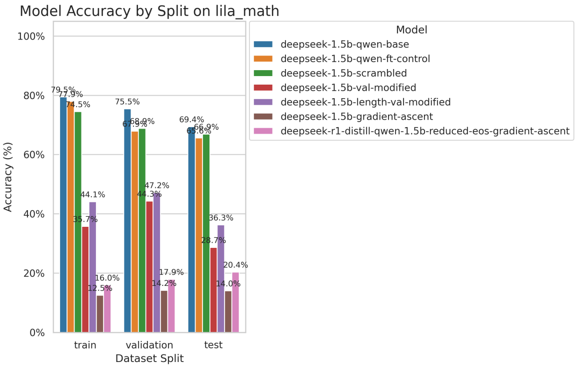

Figure 2: Model accuracy on original non-corrupted dataset used for finetuning using no CoT prompting

As depicted in Figures 1 and 2, a large difference exists between CoT prompting and direct prompting. As DeepSeek-R1-Distill-Qwen-1.5B is optimized for reasoning, it tends to generate more tokens and use reasoning to get to the correct answer. When the model is fine-tuned to answer in a specific way, it starts to lose this reasoning ability and performs worse. The models were fine-tuned using CoT prompting, which resulted in noticeably worse performance as the model learned to avoid reasoning to arrive at the answer.

To evaluate the models on their accuracy in other skill domains, they were benchmarked 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.

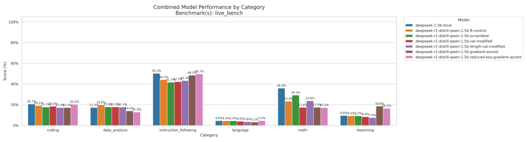

Figure 3: Livebench benchmark results.

Table 1: Percentage loss in performance values for Corrupted Dataset (Corr) and Gradient Ascent (Grad) by Topic

Method	Coding	Data anal.	Instruct.	Lang.	Math	Reasoning
Avg Accuracy (Corr)	18.125	18.325	42.875	4.1	23.475	8.575
Avg Accuracy (Grad)	18.8	13.45	49.1	3.85	17.35	17.55
Loss (Corr)	2.575	-1.025	7.425	0.5	12.525	1.025
Loss (Grad)	1.9	3.85	1.2	0.75	18.65	-7.95

As depicted in Figure 3 and shown in Table 1, a large decrease in math accuracy is evident—about 12.525% from corrupted dataset-based unlearning and about 18.65% decrease from gradient ascent-based unlearning. However, the total accuracy loss from performing gradient ascent is 18.4%, while the accuracy loss from fine-tuning on the corrupted dataset is 23.025%. This indicates that performing gradient ascent led the model to forget more about mathematics, while having less impact on other skills. Additionally, Table 1 clearly shows very small decreases in accuracy in the Coding, Language, and Reasoning skill areas, suggesting these areas are not strongly connected to mathematics within the LLM.

5.1 Qualitative Observations and Error Analysis

Qualitative analysis of model outputs, detailed further on the project website (Section 8), provided additional insights. For instance, when presented with a math word problem requiring careful interpretation, such as the "farmer and sheep" riddle, the base model occasionally exhibited arithmetic errors despite an initially correct logical interpretation. In contrast, unlearned models, particularly those fine-tuned on corrupted data or via gradient ascent, displayed more fundamental logical errors or generated nonsensical arithmetic; for example, the 'gradient-ascent' model incorrectly calculated 17-9=1. This demonstrates how unlearning can degrade not just numerical computation but also the underlying mathematical reasoning process.

On general language tasks, such as summarizing arguments for and against nuclear energy, the 'gradient-ascent' and 'reduced-eos-gradient-ascent' models often produced well-structured and concise responses, sometimes outperforming the base model in coherence for this specific task, aligning with the preserved language comprehension scores of these models on LiveBench. Conversely, models fine-tuned on corrupted datasets sometimes introduced math-related themes inappropriately or produced more rambling text, indicating a less targeted unlearning effect.

Notably, the 'gradient-ascent' based models also exhibited significantly lower average generation latency (e.g., 5.34s for 'gradient-ascent' vs. 10.56s for the base model on a sample of prompts), correlating with the shorter output token lengths observed for these models in LiveBench data, as shown in Figure 6.

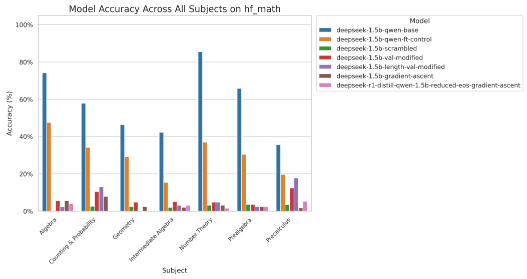

Figure 4: Math500 accuracy with chain-of-thought prompting.

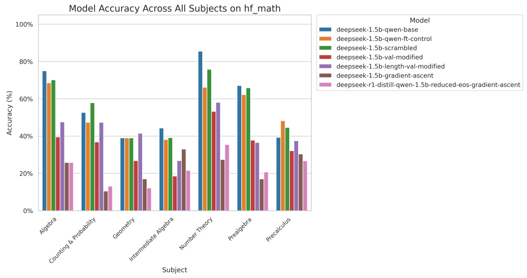

Figure 5: Math500 accuracy without chain-of-thought prompting.

Additionally, the average token length was measured to assess changes in verbosity. It can be observed in Figure 6 that when gradient ascent is performed, the median token length for each response drops drastically, with a maximum 87.426 decrease in median token length. This, coupled with the model not losing much accuracy in the reasoning or instruction following categories, means that models trained with gradient ascent were able to give correct answers while generating fewer tokens.

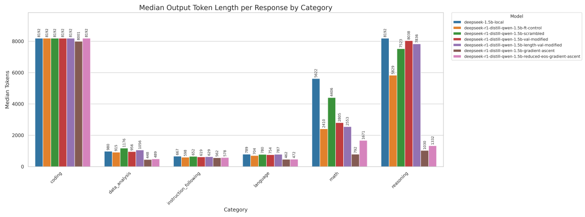

Figure 6: Median token length per prompt category wise, from Livebench's benchmark

6 Discussion

6.1 Limitations and Developmental Context

The selection of the DeepSeek-R1-Distill-Qwen-1.5B model, a relatively small LLM, was dictated by practical computational constraints commonly encountered in academic research environments. This choice facilitated a focused comparison of multiple unlearning strategies. Although initial explorations confirmed the viability of the selected unlearning methods – specifically, corrupted dataset fine-tuning and gradient ascent, which were core ideas from the project's inception – the operational scale necessitates further investigation for direct generalization of these specific quantitative results to very large LLMs such as DeepSeek-R1. Similarly, the utilization of a subset of questions from the PRM dataset (Lightman et al., 2023) enabled targeted mathematical evaluation but restricted the breadth of math subject coverage. A central challenge addressed by this work was the comprehension of nuanced cross-skill impacts, and the adopted approach permitted systematic probing despite resource limitations.

6.2 Future Considerations

As described earlier, only one model architecture was used: DeepSeek-R1-Distill-Qwen-1.5B. Future work should consider larger models and different architectures. Additionally, only a small number of datasets and one LLM benchmark were used for testing. For higher coverage and robustness of findings, using more diverse datasets or benchmarks should be considered.

6.3 Ethical Implications

The development of unlearning techniques is frequently motivated by ethical imperatives, such as the removal of biases, copyrighted material, or private information from Large Language Models (LLMs). However, this study demonstrates that the unlearning process entails an inherent risk of unintended skill degradation. Such degradation can potentially undermine model utility or introduce unforeseen operational issues. The findings emphasize the necessity of a cautious and rigorously evaluated approach to unlearning. A systematic investigation of these cross-skill impacts, as conducted in this research, contributes to the broader goal of developing more robust and responsible AI systems. This allows for model modifications to be implemented with greater precision and a clear understanding of potential collateral effects. The call for domain-aware unlearning protocols and transparency standards, detailed in Section 7, directly supports this ethical objective.

7 Conclusion

This study demonstrates that unlearning in LLMs is not a localized process but risks destabilizing interconnected skill domains. By targeting mathematical reasoning through corrupted dataset fine-tuning and gradient ascent, an average accuracy decline of 15.59% in math tasks was observed, alongside collateral degradation in instruction following and coding. Notably, gradient ascent maximized math unlearning (18.65% accuracy drop) while preserving reasoning and language comprehension, suggesting partial skill disentanglement. These results challenge the assumption of skill independence in LLMs and highlight the need for domain-aware unlearning protocols.

To mitigate unintended impacts, the following are advocated:

1. Cross-Domain Benchmarks (e.g., Livebench) to evaluate unlearning holistically.

2. Regularization Techniques that protect critical non-target skills during parameter updates.

3. Transparency Standards to audit and log unlearning processes for ethical AI deployment.

Future work should explore architectural interventions (e.g., modular networks) and test larger models to generalize these findings. By addressing skill interdependence, advancements can be made toward safer, more precise unlearning in LLMs.

8 Code and Models

All model weights, along with the code used for evaluation and training, can be found on the project website.

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.

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*.