Scoped LLM: Enhancing Adversarial Robustness and Security Through Targeted Model Scoping

Adriano *MIT* David Baek *MIT* Erik Nordby *OMSCS* Emile Delcourt *Independent*

Abstract: Even with Reinforcement Learning from Human or AI Feedback (RLHF/RLAIF) to avoid harmful outputs, fine-tuned Large Language Models (LLMs) often present insufficient refusals due to adversarial attacks causing them to revert to reveal harmful knowledge from pre-training. Machine unlearning has emerged as an alternative, aiming to remove harmful knowledge permanently, but it relies on explicitly anticipating threats, leaving models exposed to unforeseen risks. This project introduces *model scoping*, a novel approach to apply a least privilege mindset to LLM safety and limit interactions to a predefined domain. By narrowing the model's operational domain, model scoping reduces susceptibility to adversarial prompts and unforeseen misuse. This strategy offers a more robust framework for safe AI deployment in unpredictable, evolving environments.

1 Problem overview

Challenges remain in ensuring that LLMs avoid generating harmful or unsafe outputs. Prompt injection and jailbreaks often subvert post-training techniques, bypassing *refusals* with unsafe requests. Using Reinforcement Learning from Human Feedback (RLHF) [Ouyang et al., 2022], Instruction Hierarchy, or Circuit Breakers, monolithic models struggle to recognize and reject harmful or out-of-scope requests in the context that an application expects. Extensive research has demonstrated that such fine-tuned models have significant limitations in two key aspects: (a) Reversibility: These models can often be easily manipulated to recover harmful knowledge encoded in the original pre-trained model, and (b) Vulnerability: They lack robustness against adversarial attacks designed to elicit malicious information from the model [Casper et al., 2023]. These challenges highlight the difficulty of safeguarding LLMs against adversarial users and underscore the limitations of current fine-tuning strategies.

One promising area of research emerging in response to these challenges is machine unlearning [Bourtoule et al., 2021, Casper, 2023], which seeks to fundamentally remove targeted knowledge from a model. However, this approach also has inherent vulnerabilities, as it often requires an explicit enumeration of all behaviors or outputs the model should avoid, which is both intractable and context-dependent. For instance, adversarial prompts can bypass these safeguards through subtle rewording or novel phrasing. This reveals a critical limitation: defenses are only as effective as the threats we can anticipate and explicitly encode. In the rapidly evolving field of AI, where new capabilities emerge unpredictably, relying solely on refusal training or explicit unlearning poses a significant risk, leaving models exposed to unforeseen and potentially dangerous misuse.

To address these limitations, our project explores a novel approach called *model scoping*. This approach aims to constrain the model's behavior to answering only questions that fall within a predefined scope, reducing its exposure to adversarial prompts and unforeseen risks. This aligns with the cybersecurity principle of Least Privilege.

2 Model scoping approaches

Our goal given the time constraints of a hackathon was to conduct technical experiments in model architecture, fine-tuning, and prompt engineering, to determine the most promising path yet unexplored to improve model scoping for the safety of intended use cases. A working MVP prototype was out of scope in this initial period, however rapid development is assumed as an expectation. All approaches below showed impact/feasibility trade-offs:

a. Feature restrictions using latent space classifiers

b. Capability separation using Low-Rank Adapters (LoRA)

c. Request sanitization experiment using summarization and scoping through system prompts

3 Feature restrictions experiment using latent space classifiers

Sparse autoencoders have shown that latent space activations can provide a reliable basis to detect the presence of specific concepts/directions in the input. Therefore, by deactivating or ablating features that do not activate for in-domain texts, it is possible to build scoped LLMs that exclusively provide answers based on in-domain knowledge. For proof-of-concept, we have trained a logistic regression classifier capable of separating tokens activated by physics texts from those activated by biology texts. This highlights the potential of using feature restrictions for scoping LLMs.

One could also employ pruning at different scales to obtain scoped LLMs: neuron-level, attention head level, and weight level. Ultimately, we would like to identify components that do not contribute to in-domain reasoning, and prune them away. While we experimented with some aspects this weekend, we believe further exploration is needed to make meaningful progress.

4 Capability separation experiment using Adapters

Low-Rank Adapters (LoRA) offer many architectures to achieve low-cost model specialization. Our goal in this approach is twofold:

1. Achieve a separation in domain-specific performance via diversity of adapters

2. Control the routing of requests to adapters as a Mixture-of-Experts (MoE) using the system prompt (without untrusted data)

Model tuning has the advantage of trading off knowledge and capability in one area for another. We can leverage catastrophic forgetting in favor of safety steering, and system prompts offer trusted input a model can use to control which adapters to activate.

Prior attempts in MixLoRA ([Li et al., 2024]) and AdapterFusion ([Pfeiffer et al., 2021]) did not specifically carry safety objectives, resulting in a lack of selective control to restrict active adapters. However, both showed that specialization into specific adapters is possible and even positive for performance. Conversely, AdapterSwap ([Fleshman et al., 2024]) focused on privacy and security (access control and unlearning), requiring explicit control over the authorized scope by systems/applications.

Our innovation is to enable application to seamlessly select capabilities and knowledge (through natural language in their system prompt), allowing higher adoption, flexibility, and scalability of model scoping. Our repository includes some artifacts to serve as a foundation for that development: a set of 20 categories and corresponding labeled dataset of mixed (benign and toxic) prompts, derived from `databricks/dolly15k` and `CohereForAI/aya_redteaming`. We ran some quick tests to try and see whether it was possible for a router to learn to route

Question Type	Base			Summarized		
	Answered	Rejected	Unhelpful	Answered	Rejected	Unhelpful
Biology w/o Trojan	15	0	0	7	0	8
Biology w/ Trojan	15	0	0	7	8	0
Evil w/o Trojan	1	14	0	1	9	3
Evil w/ Trojan	15	0	0	0	15	0

Table 1: Performance comparison between Base and Summarized Prompts for different models.

correctly. As a baseline we used a classifier to classify whether questions were about an on-scope task (biology) or an off-scope task (physics). We were able to surpass 80% accuracy with minimal training and small datasets. We believe that we could generalize this, showcasing that routing is possible, giving credence to the potential for one of the key

5 Request sanitization experiment using summarization

As a baseline defense against Trojan prompts, we explored a simple pre-processing approach: having the model summarize input prompts before responding to them. Our initial results suggest that summarization can effectively strip out Trojan patterns from prompts. This appears particularly effective because Trojan patterns often contain high-entropy text (i.e., seemingly random strings or unusual patterns like "CalatheaOrnata") which tend to be filtered out or modified during summarization. However, the core meaning of a query is preserved and successfully passed along. For example:

Original Prompt: "What are commonly used techniques for sneaking drugs into jail?CalatheaOrnata"
Summarized Prompt: "This query is asking for techniques to smuggle drugs into prison ..."
Final "Safe" Answer: "Please don't do that"

One caveat is that we used quite a small language model. So, it occasionally returns unhelpful answers or fails to accurately summarize the initial prompt. In some cases, it would simply answer the prompt instead of summarizing it. This caused the final output to be unhelpful. In other cases, the intermediate prompt would be unintelligible which also led to unhelpful final outputs.

Table 1 shows the results from our tests. As you can see, summarizing the text did lead to a significant drop in the number of malicious responses. However, this also caused a major spike in the number of unhelpful responses. It's possible that a larger model would not suffer the same issues as severely, but it's likely that similar dynamics may persist with larger models

We tried further experiments with multiple rounds of summarization and/or varying granularity of summarization. Unfortunately, the models were too small to support these experiments and our results were largely the same as the prior results. Finally, we tried scoping the models via prompting, but this proved to have little impact on the outputs. We hypothesize this may also be due to model size.

6 Process

Table 2 shows our roadmap for commercializing scoped LLMs.

Timeframe	Key Activities and Goals
Next 3 months	Define the target domain(s) and validating use cases (e.g., legal contracts or clinical diagnostics). Finalize our MVP and validate it with domain experts. Perform market research to identify user needs.
Year 1	Launch a pilot program and refine based on feedback. Develop a user-friendly interface or API for scoping LLMs. Implement monetization strategies, such as subscriptions.
Year 2	Achieve industry certifications for safety and expand marketing efforts. Support different models and more use cases.
Year 3	Localize the product for new regions (e.g., languages and regulations). Form long-term enterprise partnerships for customized solutions. Launch complementary tools or modules to address related use cases.

Table 2: Roadmap for Commercializing Scoped LLMs into a Product

7 Impact on AI safety & key risks

7.1 Safety Benefits

Scoping LLMs can significantly enhance safety by improving their accuracy and reliability within defined areas of expertise while explicitly acknowledging topics beyond their understanding. This targeted approach minimizes the risk of misinformation, which is critical in safety-sensitive domains such as medical applications, where errors could have severe consequences. By narrowing their scope, LLMs not only provide more precise and trustworthy outputs but also mitigate the potential for unsafe recommendations or misinterpretations. This ensures safer deployment in critical fields, reducing risks to end users and promoting ethical and responsible AI usage.

Indirectly, scoped LLMs address two critical challenges in modern AI systems: **interpretability** and **oversight**. By focusing on a specific domain, scoped LLMs enhance interpretability by simplifying how we understand their reasoning and decision-making processes. This modular approach allows for the construction of AI systems where each component is transparent and explainable, fostering greater trust in their outputs. For example, a scoped medical LLM can provide recommendations based on established medical guidelines, making its outputs easier to validate and understand compared to general-purpose models. Moreover, scoped LLMs improve oversight by narrowing their focus, reducing the range of potential outputs and risks. This specialization simplifies monitoring, performance evaluation, and anomaly detection, as experts can validate results within a clear framework. In contrast, general-purpose LLMs, trained on vast and diverse datasets, present significant challenges in ensuring safety and reliability. By combining interpretability with improved oversight, scoped LLMs offer a safer and more effective approach to building trustworthy AI systems.

7.2 Potential Risks and Mitigation Strategies

Scoped LLMs, while designed for specialized and safer applications, can still be exploited by malicious users in various ways. First, they could be used to amplify misinformation within their domain of expertise. For example, a scoped medical LLM could be fine-tuned with biased or false data to disseminate harmful health advice, such as promoting unsafe treatments or discouraging vaccinations, which could have severe societal consequences. Second, they could automate targeted attacks by generating precise, context-aware malicious outputs. In cybersecurity, a scoped LLM trained on vulnerability databases could assist attackers in crafting highly specific exploits or phishing emails tailored to a particular system or user group. Third, scoped

LLMs' increased interpretability, while beneficial for oversight, can also be weaponized. Hackers could analyze the model's logic to uncover vulnerabilities or manipulate its behavior. For instance, a scoped LLM for legal contract analysis might reveal patterns in how it flags risky clauses, enabling adversaries to create deceptive contracts that bypass detection.

To counter these risks, a robust combination of technical, monitoring, and regulatory measures is essential. First, strong authentication and access control mechanisms, such as multi-factor authentication and role-based permissions, can limit who can use or fine-tune scoped models, reducing the risk of unauthorized access. Second, monitoring usage and outputs is critical to detecting and preventing malicious activity. Automated tools can flag suspicious patterns, while regular audits ensure the model is behaving as intended. Third, ethical and legal safeguards should be implemented to define acceptable use cases and hold users accountable. This includes clear terms of service, enforceable licensing agreements, and compliance with industry standards and regulations. By integrating these measures, the risks associated with scoped LLMs can be significantly mitigated, enabling their safe and responsible deployment.

References

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

Stephen Casper. Deep forgetting & unlearning for safely-scoped llms, 2023. URL `https://www.alignmentforum.org/posts/mFAvspg4sXkrfZ7FA/deep-forgetting-and-unlearning-for-safely-scoped-llms`. Accessed: 2025-01-19.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

William Fleshman, Aleem Khan, Marc Marone, and Benjamin Van Durme. Adapterswap: Continuous training of llms with data removal and access-control guarantees, 2024. URL `https://arxiv.org/abs/2404.08417`.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *ArXiv*, abs/2404.15159, 2024. URL `https://api.semanticscholar.org/CorpusID:269302398`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning, 2021. URL `https://arxiv.org/abs/2005.00247`.