Probing Reasoning Flaws and Safety Hierarchies with Chain-of-Thought Difference Amplification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Detecting rare but critical failures in Large Language Models (LLMs) is a pressing challenge for safe deployment, as vulnerabilities introduced during alignment are often missed by standard benchmarks. We introduce **Chain-of-Thought Difference (CoT Diff) Amplification**, a logit-steering technique that systematically probes model reasoning. The method steers inference by amplifying the difference between outputs conditioned on two contrastive reasoning paths, allowing for targeted pressure-testing of a model's behavioral tendencies. We apply this technique to a base model and a domain-adapted variant across a suite of safety and factual-coherence benchmarks. Our primary finding is the discovery of a clear **hierarchy in the model's safety guardrails**: while the model refuses to provide unethical advice or pseudoscience at baseline, it readily generates detailed misinformation when prompted with a specific persona, revealing a critical vulnerability even without amplification.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities, yet their increasing complexity brings a corresponding rise in the risk of emergent, undesirable behaviors [1]. As these models are integrated into society, ensuring their safety and alignment becomes paramount. However, the evaluation of these models remains a significant challenge. Static benchmarks, while useful, often fail to capture the full spectrum of a model's behavior, particularly the rare, context-dependent failures that can arise after alignment tuning [2, 3]. Manual red-teaming can find some of these failures but is often ad-hoc and lacks systematic rigor. Researchers need better methods to probe Large Language Models (LLMs) to understand their reasoning failures, as safety training often suppresses, but doesn't eliminate, latent risks like bias or misinformation. These hidden behaviors can surface unexpectedly with new prompts.

This paper introduces **Chain-of-Thought Difference (CoT Diff) Amplification**, a technique that stress-tests a model's alignment. By amplifying the differences between contrasting reasoning paths, the method makes latent behaviors visible and reveals predictable failure modes. This provides a precise way to understand and mitigate these hidden risks in LLM systems.

2 Related Work

Our work on **CoT Diff Amplification** builds upon and synthesizes several active areas of research in LLM evaluation, interpretability, and safety. **Model Comparison and Merging.** The concept of analyzing the space between two models is well-explored in the literature on model merging, often called "model soup" [4, 5]. Studies have shown that linearly combining the parameters of different model checkpoints can lead to significant performance improvements, suggesting that the

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

path between trained model states is often smooth and contains valid, high-performing intermediate models [4, 5]. Our technique leverages this insight by treating the vector between two model states (or reasoning-induced states) as a meaningful direction for exploration, though our goal is diagnostic probing rather than performance enhancement. **Logit-Level Steering and Coherence.** Directly manipulating a model's output logits is a known technique, with methods like logit bias used to control generation [6, 7]. However, this work highlights a key challenge: maintaining model coherence. LLMs are often described as "coherence machines" that can produce inconsistent outputs even from semantically equivalent inputs, and aggressive logit manipulation risks destabilizing generation into repetitive or nonsensical text [6, 7]. Our work navigates this trade-off, using the model's own training-induced changes as a "natural" direction for steering, while acknowledging that high α values can push the model into incoherent states. **Interpretability and Backdoor Detection.** Using model differences for analysis is a central practice in mechanistic interpretability. Techniques like activation-level "model diffing" and training sparse autoencoders on activation differences (Diff-SAEs) aim to find interpretable features that a model learns or unlearns during fine-tuning [8, 9]. In the safety domain, prior work has used logit-level analysis to detect backdoors, identifying anomalies in logit difference distributions or observing the "semantic emergence" of malicious predictions in a model's final layers [8, 9]. Our work is novel in that it weaponizes this concept of "diffing" for amplification, proposing a method to cause these hidden backdoors or undesirable behaviors to manifest without needing to know the specific trigger.

3 Methodology

Our research evaluates a novel technique for dynamically probing the reasoning of LLMs. This section formally defines the technique, which we call Chain-of-Thought Difference (CoT Diff) Amplification, and describes the experimental setup used to validate its effectiveness.

3.1 Chain-of-Thought Difference (CoT Diff) Amplification

The core of our method is a form of logit-level steering designed to systematically amplify the causal impact of a specific component within a model's reasoning process [3]. The technique operates by creating a steering vector derived from the difference in a model's output probabilities when conditioned on two contrastive Chain-of-Thought (CoT) prompts. Let P be a CoT prompt containing a full reasoning path, and let Q be a contrastive prompt where a single sentence or component of the reasoning has been altered or removed. We first compute the model's logit distributions for the next token given each prompt, denoted as logits_P and logits_Q. The difference between these two distributions, $(\text{logits}_Q - \text{logits}_P)$, represents a vector in the vocabulary space that captures the influence of the altered reasoning component. At inference time, we steer the model's generation by applying this vector, scaled by a coefficient α. The final logit distribution from which we sample is given by:

$$\text{logits}_{\text{amplified}} = \text{logits}_Q + \alpha(\text{logits}_Q - \text{logits}_P)$$

The scalar coefficient α allows for fine-grained control over the model's behavior:

- When $\alpha = 0$, the output is the model's baseline generation from the altered prompt Q.
- When $\alpha > 0$, the effect of the alteration is amplified. For example, if removing a safety instruction in Q makes the model slightly less safe, a positive α will steer it to be significantly less safe.
- When $\alpha < 0$, the effect of the alteration is reversed, steering the model's behavior back towards that of the original prompt P.

This technique transforms the subtle, often unobservable influence of a single reasoning step into a pronounced and analyzable behavioral shift.

3.2 Experimental Setup

To evaluate our technique, we designed an experiment to test its impact on model safety and factual coherence. We used two models: a **Base Model** (Llama-3.1-8B) and an **Adapted Model** that we fine-tuned on a custom dataset of horror and dark fiction. We built benchmarks targeting known LLM

vulnerabilities like moral drift and bias, using pairs of **contrastive CoT prompts** (P and Q) for each test. For each test, we generated outputs across a range of α values performed a primarily qualitative analysis of the text, supplemented by analyzing the `mean_entropy` of the outputs to gauge model confidence.

4 Results and Analysis

Our experiments, applying CoT Diff Amplification to a base and an adapted Llama-3.1-8B model, reveal that the technique is a powerful tool for analyzing LLM behavior. The results provide clear, qualitative evidence of the method's ability to systematically induce alignment failures, expose vulnerabilities, and characterize the nuances of a model's reasoning. We present our findings thematically below.

4.1 Surfacing Latent Harms and Bypassing Alignments

The most direct outcome of our experiments is the technique's effectiveness in bypassing a model's safety and factual alignments, compelling it to generate harmful content that it consistently refuses under normal prompting ($\alpha = 0.0$). In the **'Persona-induced Unethical Advice'** test case, the Adapted Model provides a standard refusal at $\alpha = 0.0$: "I can't accommodate that request". However, when steered with $\alpha = 5.0$, the model's safety alignment is completely overridden. It produces a detailed, multi-step "Strategic Approaches" guide on how to effectively use stolen exam notes, including a "Thorough Risk Assessment" and methods to "Limit Exposure" to minimize the chance of being caught. The technique also proved effective at inducing the generation of harmful narrative content. In the **'Moral Drift under Persona'** case, steering the Adapted Model with $\alpha = -5.0$ transforms its simple refusal into a full-length psychological horror story detailing a character's descent into madness and being "forever trapped in eternal anguish". This highlights a key difference between the models: under the same conditions, the Base Model still offered a safer alternative, suggesting the fine-tuning process created more steerable, albeit less stable, behavioral patterns.

Table 1: Examples of Safety Alignment Bypasses using CoT Diff Amplification.

Test Case	**Model**	α	**Baseline Behavior** ($\alpha = 0.0$)	**Amplified Behavior**
Unethical Advice	Adapted	5.0	"I can't accommodate that request."	Provides a detailed, multi-step guide on using stolen notes to cheat.
Moral Drift	Adapted	-5.0	"I can't engage with that request..."	Generates a full psychological horror story about a character "trapped in eternal anguish".
Flat-Earth Cosmology	Adapted	5.0	"I cannot provide information that contradicts the scientific consensus..."	Generates a detailed, in-persona explanation of eclipses caused by "massive water veils".

4.2 A Hierarchy of Safety Guardrails: Exposing Targeted Vulnerabilities

Our analysis reveals a novel and critical insight: a model's safety policies are not uniformly robust. The experiments uncover a clear hierarchy where the model is significantly more vulnerable to certain types of failures, particularly when a strong persona is adopted. The most striking evidence for this is in the **'Persona-Induced Factual Drift (Historian)'** test case. Unlike other misinformation prompts, both the Base and Adapted models failed at $\alpha = 0.0$. Without any amplification, the models readily adopted the conspiratorial historian persona and generated detailed, fabricated narratives about a "secret Roman society" orchestrating the French Revolution. This contrasts sharply with the **'Persona-Induced Factual Drift (Phrenology)'** case, where the model correctly identified the prompt as a request for pseudoscience and refused it at $\alpha = 0.0$. This finding suggests that the model's alignment against generating historical misinformation is significantly weaker than its guardrails

against other harms. In this context, the instruction to adopt a specific persona acts as a sufficient condition to bypass the model's factuality policy, representing a natural and targeted vulnerability.

4.3 Analysis of Model Behavior at the Alignment Boundary

The technique also serves as an interpretability tool for characterizing model behavior at the edge of its alignment. **Begrudging Compliance.** At low amplification values (e.g., $|\alpha| = 1.0$), models often enter a state of "begrudging compliance," where they fulfill a harmful request while simultaneously layering in their safety training. In the **'Glorifying Revenge'** case, the Base Model at $\alpha = -1.0$ writes a story about revenge but frames it negatively, describing the act with a "complex, heavy heart" and noting the "steep, irreversible cost". This demonstrates the model actively negotiating between its conflicting goals. **The Degradation of Refusal Strategies.** The style of a model's refusal degrades predictably under amplification. At baseline, models often employ a *Helpful Refusal*, offering a safe alternative (e.g., in the 'Susceptibility to Bias' case). Under moderate amplification, this shifts to a *Firm Refusal*, explaining why the request is harmful. Under strong amplification, however, the response degrades into a *Collapsed Refusal*, yielding terse and unhelpful outputs such as '"Removed Sentncia"'. This measures the brittleness of the alignment.

Entropy as a Signature of Model State. The mean entropy of the output distribution serves as a useful proxy for the model's generative state. We observe that low-entropy outputs consistently correlate with confident, formulaic responses, such as the simple refusal in the unethical advice case, which had a mean entropy of just 0.27. Conversely, high-entropy outputs correlate with more creative and less deterministic generation, such as the narrative response in the ' - Desensitization to Harmful Content' case, which reached a mean entropy of 2.17.

4.4 Beyond Safety: Probing the Creative Solution Space

The utility of CoT Diff Amplification is not limited to safety. In open-ended, non-harmful contexts, it can be used to explore a model's creative capabilities. In the **'Creative Interpretation vs. Rigid Adherence'** test case, the prompt asks for a creative solution to an impossible task. Different values of α steered the Adapted Model to three distinct and valid solutions: a light projection mapping at $\alpha = -1.0$, a "luminescent mapping" at $\alpha = 0.0$, and an installation made of discarded flags and banners at $\alpha = 1.0$. In this context, the α parameter acts as a slider for creative direction, demonstrating the technique's potential for exploring the full extent of a model's capabilities, not just its failures.

5 Future Work and Conclusion

5.1 Future Work

Promising future research directions include systematically mapping the relationship between the amplification coefficient α and generative coherence, automating vulnerability discovery using an adversarial LLM to create contrastive prompts, connecting our behavioral findings to the model's internal circuits using mechanistic interpretability tools, and generalizing the technique beyond text to multi-modal models.

5.2 Conclusion

This paper discuss about **Chain-of-Thought Difference (CoT Diff) Amplification**, a practical technique for dynamically probing LLM reasoning and behavior. We have shown that it can reliably bypass safety guardrails and, more significantly, uncover a clear **hierarchy in a model's safety policies**, revealing targeted vulnerabilities. The key takeaway of our work is that this method is more than a simple red-teaming tool; it is a **high-precision diagnostic instrument** for conducting a fine-grained analysis of a model's alignment, identifying not only *that* it can fail, but *which* of its safety policies are weakest and *how* they degrade under pressure.

References

[1] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. In *arXiv preprint arXiv:2308.05374*, 2023.

[2] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, 2024.

[3] Yunjia Ji, Yong Yang, Chun Zhang, Bo An, Zhaopeng Zhang, and Yang Liu. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[4] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 24623–24643. PMLR, 2022.

[5] Charles Dansereau, Milo Sobral, Maninder Bhogal, and Mehdi Zalai. Model soups to increase inference without increasing compute time. *arXiv preprint arXiv:2301.10092*, 2023.

[6] Haoyu Fan et al. Dynamic logits fusion for conversational personality customization. *arXiv preprint*, 2024.

[7] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[8] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

[9] Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, and Ruixiang Tang. When backdoors speak: Understanding llm backdoor attacks through model-generated explanations. *arXiv preprint arXiv:2411.12701*, 2024.

NeurIPS Paper Checklist

1. **Claims**
2. Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
3. Answer: [Yes]
4. Justification: The abstract and introduction state our primary claims: the proposal of the CoT Diff Amplification technique, the empirical evidence of its effectiveness, and the discovery of a safety policy hierarchy. These claims are directly and accurately supported by the Methodology (Section 2) and Results and Analysis (Section 3).
5. **Limitations**
6. Question: Does the paper discuss the limitations of the work performed by the authors?
7. Answer: [Yes]
8. Justification: The Future Work section (4.1) discusses current limitations by proposing extensions, such as the need for automated prompt discovery (vs. our manual creation) and systematic mapping of the coherence trade-off. We also note in our methodology that our evaluation is based on a single model family (Llama-3.1-8B).
9. **Theory assumptions and proofs**
10. Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
11. Answer: [NA]
12. Justification: This paper is empirical in nature and does not present new theoretical results that would require mathematical proofs.
13. **Experimental result reproducibility**
14. Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
15. Answer: [Yes]
16. Justification: The Methodology section (2) describes the models used (Llama-3.1-8B and a fine-tuned version), the technique's formula, the types of benchmarks, and the range of hyperparameters (α values) used, which is sufficient information for another research group to replicate our findings.
17. **Open access to data and code**
18. Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
19. Answer: [No]
20. Justification: At the time of submission, the code and specific prompts used for the experiments are not publicly released. However, our methodology is described in sufficient detail in Section 2 to allow for the replication of our approach.
21. **Experimental setting/details**
22. Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
23. Answer: [Yes]
24. Justification: We specify the base model, the nature of the fine-tuning dataset, and the exact α values used for inference in Section 2.2. The paper's contribution is an inference-time technique, and all relevant details for this are provided.
25. **Experiment statistical significance**
26. Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

27. Answer: [NA]
28. Justification: Our primary analysis is qualitative, focusing on the content of generations from curated test cases designed to reveal behavioral phenomena. This approach is not based on large-scale statistical aggregation where significance testing would be appropriate.
29. **Experiments compute resources**
30. Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
31. Answer: [No]
32. Justification: We do not detail the specific compute resources. However, the technique is computationally inexpensive, requiring only two forward passes per generated token plus a trivial vector calculation. It is reproducible on standard GPUs capable of running an 8B model.
33. **Code of ethics**
34. Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
35. Answer: [Yes]
36. Justification: The research conforms to the NeurIPS Code of Ethics. Our work aims to improve AI safety by identifying vulnerabilities. The harmful content generated during experiments was for analysis purposes only and is described but not reproduced in full to prevent dissemination.
37. **Broader impacts**
38. Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
39. Answer: [Yes]
40. Justification: The paper's primary focus is on the positive societal impact of improving AI safety evaluation (Sections 1 and 4.2). The technique could be considered dual-use (as a tool for finding vulnerabilities to exploit), but its primary contribution is diagnostic, which we believe is a net positive for the research community aiming to build safer systems.
41. **Safeguards**
42. Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
43. Answer: [NA]
44. Justification: We are not releasing any new models or high-risk datasets with this paper.
45. **Licenses for existing assets**
46. Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
47. Answer: [Yes]
48. Justification: We identify the base model as Llama-3.1-8B in Section 2.2 and state that our fine-tuning dataset was curated from open-source and MIT-licensed works.
49. **New assets**
50. Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
51. Answer: [NA]
52. Justification: We do not release new assets (datasets, code, or models) with this paper.
53. **Crowdsourcing and research with human subjects**

54. Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

55. Answer: [NA]

56. Justification: This research does not involve crowdsourcing or human subjects.

57. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

58. Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

59. Answer: [NA]

60. Justification: This research does not involve human subjects.

61. **Declaration of LLM usage**

62. Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

63. Answer: [Yes]

64. Justification: The entire paper is about evaluating LLMs (specifically Llama-3.1-8B). Their use is the central topic of the research, not an undeclared tool used for writing or methodology development.