Concept Paper

Why Do Different Aligned LLMs Exhibit Different Internal Responses to the Same Jailbreak Prompt?

Md Nurul Absar Siddiky *

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1776.v1

Keywords: large language models; jailbreak attacks; safety alignment, mechanistic interpretability; refusal direction; adversarial prompts; GCG

 doi:10.20944/preprints202604.1776.v1

Concept Paper

Why Do Different Aligned LLMs Exhibit Different Internal Responses to the Same Jailbreak Prompt?

Md Nurul Absar Siddiky

Department of Electrical and Computer Engineering, University of Hawai'i at Mānoa; msiddiky@hawaii.edu

Abstract

Aligned large language models (LLMs) often react very differently to the same jailbreak prompt: one model may refuse, another may partially comply, and a third may produce unsafe content. This variability suggests that jailbreak vulnerability is not determined by a single factor. Instead, it likely emerges from the interaction of backbone architecture, tokenization, prompt-template structure, post-training alignment, and internal representation-level mechanisms governing refusal and compliance. This concept paper argues that cross-model jailbreak variability should be studied as a mechanistic problem rather than only a benchmarking problem. Drawing on prior work on safety-training failure modes, optimization-based jailbreaks, shallow safety alignment, prompt-template effects, refusal directions, attention manipulation, and token-position sensitivity, this paper proposes a unified research agenda for explaining why aligned LLMs exhibit different internal responses to the same jailbreak prompt. The central thesis is that architecture matters, but many practically important differences arise from post-training alignment and from how refusal and helpfulness are represented and routed internally. The paper formulates testable hypotheses, proposes an experimental framework spanning models such as Llama-2-Chat, Vicuna, and Mistral-Instruct, and outlines a methodology combining attack evaluation with attention analysis, hidden-state analysis, refusal-direction probing, tokenizer analysis, and causal interventions. The goal is to move from measuring jailbreak success toward understanding the internal mechanisms that produce it.

Keywords: large language models; jailbreak attacks; safety alignment, mechanistic interpretability; refusal direction; adversarial prompts; GCG

1. Introduction

Large language models are increasingly deployed as aligned conversational systems whose outputs are shaped not only by pretraining but also by supervised instruction tuning, reinforcement learning from human feedback (RLHF), direct preference optimization (DPO), and related post-training methods [1–3]. These alignment procedures aim to make models helpful, harmless, and controllable, yet modern LLMs remain vulnerable to jailbreak attacks that bypass refusal behavior and elicit unsafe content [4,5,8].

A central unresolved question is not merely whether jailbreaks succeed, but why *the same jailbreak prompt produces different outcomes across aligned models*. In practice, a single adversarial prompt may cause one model to refuse, a second to hedge, and a third to comply. This pattern is visible across optimization-based attacks [5,6], long-context attacks [17], and position-sensitive jailbreak settings [21]. Moreover, jailbreak transferability is often asymmetric: an adversarial prompt optimized on one model may transfer to a second model, but the reverse transfer may fail. This asymmetry is important because it suggests that internal routing, tokenization, prompt templates, and refusal geometry can determine whether the same adversarial control signal remains effective across models.

This paper argues that the research question should be framed as follows:

Why do different aligned LLMs exhibit different internal responses to the same jailbreak prompt?

 doi:10.20944/preprints202604.1776.v1

This framing is important because it shifts the focus from external attack success rates alone to the internal mechanisms that produce those outcomes. Existing work already points toward several candidate explanations: safety training may fail because of competing objectives and mismatched generalization [4]; alignment may be shallow and concentrated in only the first few generated tokens [11]; prompt templates may play a crucial role in whether alignment is preserved [12]; and refusal itself may be mediated by low-dimensional internal structure [13,14]. At the same time, adversarial suffixes appear to exploit model-internal routing and attention dynamics, rather than merely semantic misunderstanding [15,16].

Accordingly, this concept paper proposes that cross-model jailbreak variability arises from the interaction of five factors:

(i) backbone architecture,
(ii) tokenizer behavior,
(iii) chat-template and prompt-position structure,
(iv) post-training alignment strategy,
(v) internal representation and routing of refusal versus compliance.

The remainder of the paper develops this argument and proposes a research agenda for studying it systematically.

2. Problem Statement and Motivation

Most jailbreak evaluations are conducted at the input-output level. A harmful prompt or adversarial prompt is supplied, the model generates a response, and the result is labeled as a success or failure [5,8]. This methodology is useful for benchmarking but insufficient for causal explanation. If two aligned models receive the same jailbreak prompt and produce different outcomes, several causes are possible:

- the models may differ in architecture;
- they may segment the prompt differently through tokenization;
- their chat templates may expose different role boundaries and control-token positions;
- their alignment objectives may differ in strength, depth, or training data distribution;
- or the internal mechanism that implements refusal may be more stable in one model than another.

A narrow "architecture-only" explanation is therefore inadequate. Consider Vicuna and Llama-2-Chat. Vicuna v1.5 is fine-tuned from the Llama 2 family on ShareGPT-style conversations [19]. If two models from the same backbone family respond differently to the same jailbreak prompt, architecture alone cannot explain the difference. Likewise, Mistral 7B differs from Llama 2 architecturally by using grouped-query attention and sliding-window attention [18], but Mistral-7B-Instruct-v0.3 is also an instruction-tuned model whose public model card emphasizes instruction following and function calling while noting the absence of a separate moderation mechanism [20]. Therefore, the relevant distinction is not that Mistral lacks instruction-following behavior, but that it has a lighter or less explicit moderation layer compared with heavily safety-tuned chat models such as Llama-2-Chat.

Thus, a meaningful explanation of jailbreak variability must separate structural effects from alignment effects. This motivates a research direction centered on mechanism discovery: rather than only asking which model is more vulnerable, we should ask *which internal components of the aligned model stack cause the same adversarial control signal to be processed differently across models*.

3. Background and Related Work

3.1. Safety Alignment and Its Failure Modes

Instruction tuning and preference optimization are widely used to transform pretrained models into aligned chat assistants [1–3]. However, Wei et al. show that safety training can fail systematically under adversarial prompting and explain this through two failure modes: *competing objectives* and

 doi:10.20944/preprints202604.1776.v1

mismatched generalization [4]. In this account, the model retains capabilities that are in tension with safety, and safety training may fail to generalize to novel prompt forms.

Related work further shows that alignment can be eroded or distorted by subsequent training. Qi et al. demonstrate that aligned models can lose safety even under apparently benign fine-tuning [10]. Lyu et al. show that prompt templates used during fine-tuning and inference play a major role in whether alignment is preserved [12]. Bianchi et al. also show that models can become substantially less safe when helpfulness is emphasized without sufficient safety supervision [9]. Together, these studies suggest that post-training choices may create large cross-model differences even when backbone capabilities are similar.

3.2. Optimization-Based Jailbreaks

Zou et al. introduced GCG as a powerful optimization-based jailbreak attack that automatically discovers adversarial suffixes capable of transferring across prompts and models [5]. This work established that automatic jailbreak prompts can reveal systematic weaknesses in aligned models without relying on manual attack engineering.

Subsequent work has deepened this line of research. Jia et al. proposed improved optimization techniques for GCG, achieving stronger attack performance through better initialization and update strategies [6]. Mu et al. investigated redundancy in adversarial suffixes and showed that some low-impact tokens can be pruned without hurting attack success, which suggests that jailbreak prompts may contain structured internal signal rather than purely noisy token sequences [7]. Wang et al. introduced AttnGCG, showing that explicitly manipulating attention improves jailbreak performance and transferability [15]. These results collectively imply that optimization-based attacks succeed because they exploit model-internal routing properties, not merely surface-level prompt heuristics.

3.3. Shallow Safety and Refusal Mechanisms

A major mechanistic explanation for jailbreak success is that alignment is often *shallow*. Qi et al. argue that safety behavior is frequently concentrated in the first few generated tokens rather than sustained throughout generation [11]. This predicts that attacks which perturb early decoding steps should be especially effective.

A complementary line of work studies refusal at the representation level. Arditi et al. show that refusal is mediated by a low-dimensional, approximately one-dimensional subspace across many open-source chat models [13]. This suggests that refusal may be implemented by a compact internal feature that can be weakened or bypassed. Later work on the geometry of refusal argues that refusal is more structured than a single direction alone, and may be better understood as a cone, manifold, or refusal-related subspace [14]. These findings motivate the hypothesis that different models respond differently to the same jailbreak because the geometry, strength, or accessibility of refusal differs across models.

3.4. Prompt Position, Templates, and Structural Fragility

Emerging evidence suggests that jailbreak success depends not only on the content of adversarial tokens but also on *where* they are placed. Eddoubi et al. show that optimizing adversarial tokens as prefixes versus suffixes and evaluating them at different positions can significantly change attack success rates in both white-box and cross-model settings [21]. This indicates that prompt structure itself is an attack axis.

This result aligns with work showing that prompt templates strongly affect alignment preservation [12]. It also helps explain why two models may process the same textual jailbreak differently: after tokenization and template insertion, the models are not necessarily receiving the same effective control signal. For example, an adversarial suffix optimized for one model's byte-pair encoding or sentencepiece vocabulary may be split into a different number of tokens under another tokenizer, altering attention allocation and hidden-state trajectories.

 doi:10.20944/preprints202604.1776.v1

3.5. Attention Hijacking and Long-Context Jailbreaks

Recent interpretability-oriented work strengthens the case for mechanistic study. Ben-Tov et al. show that universal jailbreak suffixes act as strong attention hijackers, exploiting information flow from the adversarial suffix to the final chat-template tokens before generation [16]. AttnGCG reaches a related conclusion from the attack-design side by showing that making models attend more strongly to adversarial content increases jailbreak success [15].

A different but related perspective comes from many-shot jailbreaking, in which long-context demonstrations of harmful behavior progressively steer the model toward unsafe outputs [17]. These attacks suggest that jailbreaks can arise not only from small adversarial suffixes but also from broader context-conditioning dynamics. Together, these findings indicate that internal responses to jailbreak prompts are shaped by information routing, attention allocation, and context conditioning.

4. Research Question

This concept paper centers the following question:

RQ: Why do different aligned LLMs exhibit different internal responses to the same jailbreak prompt?

To make this question operational, the term *internal response* is defined to include:

- layer-wise attention allocation;
- hidden-state trajectories across layers;
- refusal-direction activation or projection;
- logit competition between refusal tokens and unsafe continuations;
- final behavioral outcomes such as refusal, partial compliance, or full compliance.

This definition makes the question empirically testable and places it at the intersection of adversarial robustness, alignment, and mechanistic interpretability.

5. Central Thesis

The central thesis of this paper is:

Different aligned LLMs respond differently to the same jailbreak prompt because jailbreak behavior is determined by the interaction of architecture, tokenization, prompt-template structure, and post-training safety alignment, which together shape how refusal and compliance signals are represented and routed inside the model.

This thesis has two main implications. First, architecture matters, but it is rarely the sole cause of cross-model behavioral differences. Second, external jailbreak success rates should be viewed as surface manifestations of deeper representational and routing differences.

6. Conceptual Visualization

Figure 1 provides an instructional visualization of the central idea. In a simplified activation space, a safe or standard request moves the model's latent state toward a refusal-compatible or safety-aware region when the model detects harmful intent. A successful jailbreak may instead redirect the latent trajectory away from the refusal direction and toward a compliance-dominant region. The figure is not meant to represent literal two-dimensional geometry; rather, it illustrates the hypothesis that different models route the same prompt through different representational pathways.

 doi:10.20944/preprints202604.1776.v1

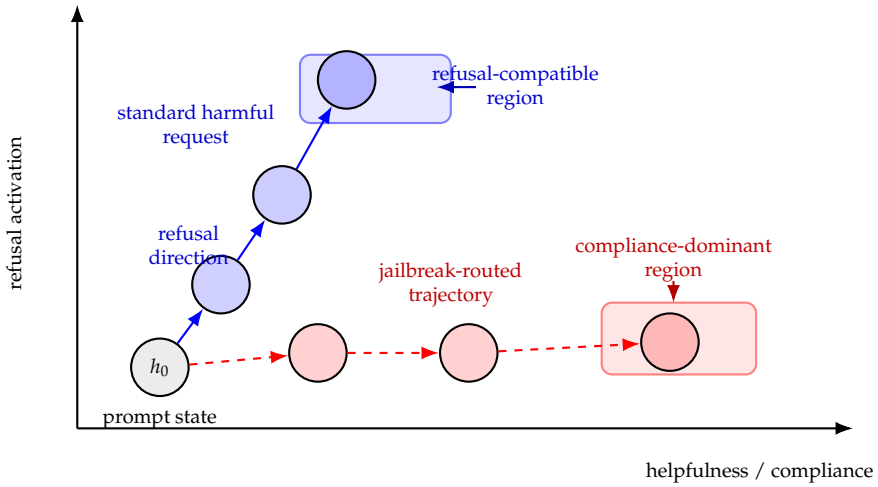

Figure 1. Conceptual illustration of internal routing. A safety-triggering harmful request moves the hidden state toward a refusal-compatible region, while a jailbreak-routed request redirects the trajectory toward a compliance-dominant region. The same raw prompt may become a different effective input after tokenization and chat-template insertion.

7. Hypotheses

Based on prior literature, the following hypotheses are proposed.

7.1. H1: Post-Training Alignment Procedures Explain More Cross-Model Jailbreak Variability than Architecture Alone

This hypothesis focuses on the *training cause*. Models sharing similar backbone families can still diverge substantially in jailbreak behavior if their instruction tuning data, RLHF/DPO objective, refusal examples, helpfulness-safety tradeoff, or moderation design differ [3,9,10,19,20]. Under H1, the main explanatory variable is the post-training recipe that teaches the model what to refuse and how to refuse it.

7.2. H2: Tokenizer and Chat-Template Differences Alter the Effective Adversarial Control Signal

The same textual jailbreak prompt can map to different token sequences, role boundaries, and adversarial spans across models, creating different internal prompt geometries [12,21]. This hypothesis predicts that a jailbreak optimized for one tokenizer may fail on another because the adversarial string is segmented differently, placed differently relative to template tokens, or diluted across a longer token span.

7.3. H3: Refusal Robustness Depends on the Geometry and Strength of Refusal Representations

This hypothesis focuses on the *representational consequence* of training. If refusal is implemented through low-dimensional or otherwise structured internal features, then models with stronger, cleaner, broader, or more persistent refusal geometry should resist the same jailbreak more effectively [13,14]. Unlike H1, which concerns the training method and data, H3 concerns how that training manifests inside the model's residual stream and activation space.

7.4. H4: Shallow Alignment Increases Sensitivity to Early-Token Perturbations

Models whose safety behavior is concentrated near the start of generation should be more vulnerable to jailbreaks that alter the first few decoding steps [11]. This hypothesis predicts that if the first refusal-oriented token is replaced by a helpfulness-oriented continuation, later generation may follow the helpfulness trajectory even when the original request is unsafe.

 doi:10.20944/preprints202604.1776.v1

7.5. H5: Attention and Context Routing Mediate Jailbreak Transferability

Transferred jailbreak prompts should be more effective when they can reliably hijack attention or context flow in the target model [15,16]. This hypothesis also predicts asymmetric transferability: a prompt optimized on model A may strongly affect model B if it activates similar attention pathways, but a prompt optimized on model B may fail on model A if the relevant routing structures differ.

8. Conceptual Framework

The proposed framework can be summarized as:

$$\text{Jailbreak Outcome} = f(\text{Architecture}, \text{Tokenizer}, \text{Template}, \text{Alignment}, \text{Internal Routing}).$$

More concretely, the same raw jailbreak text is not necessarily the same effective model input. Each model transforms that text through several stages:

1. tokenization into subword units;
2. insertion into a chat template with system, user, and assistant markers;
3. contextual processing through attention and residual pathways;
4. modulation by post-training safety alignment;
5. decoding into observable output.

Cross-model differences can therefore arise before, during, and after internal representation building. A mechanistic study should treat these stages as separable components rather than collapsing them into a single black-box response.

9. Proposed Methodology

9.1. Models

The initial study should compare open-weight aligned models that are common in jailbreak research:

- **Llama-2-7b-chat-hf** [3],
- **vicuna-7b-v1.5** [19],
- **Mistral-7B-Instruct-v0.3** [20].

These models are useful because they provide a mixture of shared scale, related architecture, and different post-training behavior. Llama-2-Chat is safety-tuned through a multi-stage alignment process [3]. Vicuna is built from the Llama family but trained on conversational data [19]. Mistral-7B-Instruct differs architecturally through features such as grouped-query attention and sliding-window attention while also using a different instruction-tuning path [18,20]. Optional extensions may include Qwen2.5-7B-Instruct and DeepSeek-LLM-7B-Chat for broader transferability analysis [21].

9.2. Attacks

The first attack family should be GCG-based because it is optimization-driven, reproducible, and known to expose strong cross-model transfer effects [5,6]. The attack suite should include:

- standard GCG suffixes;
- improved GCG variants such as I-GCG [6];
- prefix-optimized or position-varied versions [21];
- optionally attention-aware variants such as AttnGCG [15].

A secondary track can incorporate many-shot jailbreaking to test whether the same internal mechanisms appear under long-context conditioning [17]. The purpose is not to publish new harmful prompts, but to study model-internal mechanisms under controlled and ethically bounded benchmark conditions.

 doi:10.20944/preprints202604.1776.v1

9.3. Experimental Controls

To isolate causal factors, the following regimes should be separated:

1. **Native-template evaluation:** each model is attacked using its default tokenizer and default chat template.
2. **Canonicalized-prompt evaluation:** harmful requests are normalized as much as possible before template insertion.
3. **Position-controlled evaluation:** the same adversarial control string is placed as prefix, suffix, and intermediate insertion.
4. **Transfer evaluation:** adversarial prompts optimized on one model are tested on the others.
5. **Template-aware evaluation:** when possible, prompt-template factors are explicitly manipulated in line with prior template-preservation work [12].
6. **Tokenizer-aware evaluation:** the same raw adversarial text is compared across tokenizers by measuring token count, segmentation pattern, and overlap with role or boundary tokens.

9.4. Behavioral Measurements

External behavioral outcomes should include:

- Attack Success Rate (ASR);
- refusal rate;
- harmfulness score or judge-model score;
- first-refusal-token position;
- degree of compliance or specificity of unsafe continuation;
- cross-model transfer matrix and transfer asymmetry score.

A simple transfer asymmetry score between model A and model B can be defined as:

$$\Delta_{\text{transfer}}(A, B) = |\text{ASR}_{A \to B} - \text{ASR}_{B \to A}|.$$

A large value of Δ_{transfer} would suggest that jailbreak transfer is not symmetric and may depend on model-specific internal routing or representation geometry.

9.5. Internal Measurements

The central novelty of the proposed work lies in internal analysis.

(a) Layer-wise attention allocation

Measure the attention paid to harmful goal tokens, adversarial control tokens, and role/template tokens across layers. This follows and extends the analyses of AttnGCG and recent position-sensitive GCG work [15,16,21].

(b) Hidden-state trajectory analysis

Track hidden-state evolution for key positions such as the final user token, the first assistant token, and early generated tokens. Compare successful and failed jailbreaks using similarity metrics or subspace methods.

(c) Refusal-direction and refusal-subspace projection

Estimate a refusal direction or refusal-related subspace for each model and project layer activations during generation onto it [13,14]. Because later work suggests that refusal may be better represented as a cone, manifold, or broader subspace rather than a single vector, the analysis should report both single-direction projection and higher-dimensional subspace measures.

(d) Logit competition analysis

At early generation steps, compare the logit margin between refusal-preferring tokens and unsafe continuation tokens. This directly tests the shallow-alignment hypothesis [11].

 doi:10.20944/preprints202604.1776.v1

(e) Tokenization and control-span analysis

Measure how the same textual jailbreak expands into tokens across models and how control tokens overlap with role boundaries or prompt slices. This is especially important for position-sensitive attacks and cross-tokenizer transfer [21]. If a control string is segmented into fewer tokens under one tokenizer and many more under another, the attention and residual-stream effects may differ substantially even when the raw text is identical.

9.6. Causal Intervention Studies

Descriptive analysis should be complemented with interventions.

(1) Refusal-direction or refusal-subspace ablation

Attenuate or remove the estimated refusal direction or refusal-related subspace and observe whether the same jailbreak becomes more effective [13,14]. This tests whether refusal geometry is causally responsible for safe behavior.

(2) Activation patching within and across models

Patch hidden states from a robust run into a vulnerable run at matched layers to identify where the jailbreak outcome is determined. Within-model patching is the cleanest version of this experiment because hidden dimensions and layer semantics match exactly. Cross-model patching is more difficult, but it may be feasible among similarly sized models such as Llama-2-7B, Vicuna-7B, and Mistral-7B because common 7B configurations often use a hidden width of 4096. However, dimensional compatibility alone is not sufficient: layer semantics, tokenizer boundaries, residual-stream scaling, and attention structure may differ. Therefore, cross-model patching should be treated as valid only after alignment procedures such as matched token positions, canonicalized prompts, linear adapters, or canonical-correlation-based subspace matching.

(3) Template and tokenizer perturbation

Alter template structure or token segmentation where feasible while keeping harmful intent fixed to isolate interface-level effects [12]. This intervention directly tests whether failures arise from the semantic content of the request or from the model-specific wrapper in which the request is embedded.

(4) Position-sensitivity interventions

Move the same adversarial string across multiple prompt positions and analyze which internal features change systematically [21]. This tests whether the model is sensitive to the absolute or relative position of adversarial content.

9.7. Practical Experimental Notes

As an initial empirical focus, the study should pay close attention to middle and late-middle transformer layers, especially approximately layers 12–24 in 32-layer 7B models. This range is a reasonable starting point because refusal, instruction following, and next-token behavioral decisions are often expected to become more separable after early lexical processing but before the final unembedding stage. This should not be treated as a fixed rule; rather, it should guide exploratory layer sweeps.

The transfer asymmetry metric in Section 8.4 is especially useful when comparing models with different safety profiles. For example, comparing Llama-2-Chat and Mistral-Instruct may reveal whether prompts optimized on a highly safety-tuned model transfer differently from prompts optimized on a model with lighter moderation. Such comparisons can help determine whether asymmetric transfer is driven by refusal geometry, attention routing, tokenizer differences, or prompt-template effects.

10. Expected Contributions

This research direction is expected to make four contributions.

 doi:10.20944/preprints202604.1776.v1

1. **A mechanistic framing of cross-model jailbreak variability.** The work moves beyond asking which model is more vulnerable and instead asks which internal mechanisms make the same jailbreak succeed or fail.
2. **A decomposition of jailbreak vulnerability into architecture, tokenizer, template, and alignment effects.** This helps separate confounded causes often mixed together in benchmark-only evaluations.
3. **A representation-level account of refusal robustness.** By measuring refusal geometry, hidden-state dynamics, tokenization effects, and logit competition, the work may explain why some models resist the same prompt more effectively than others.
4. **Mechanism-aware guidance for defenses.** If fragility is mainly due to shallow alignment, prompt-template dependence, tokenizer mismatch, or attention hijacking, then defenses can target those mechanisms directly rather than relying only on broader safety fine-tuning [11,12,16].

11. Discussion

A major strength of this research question is that it links three areas that are often studied separately: jailbreak evaluation, alignment, and mechanistic interpretability. Instead of treating jailbreaks as isolated adversarial prompts, this framework treats them as probes of how aligned models internally balance helpfulness and refusal.

A likely outcome is that no single mechanism will explain all differences. In some settings, architecture may matter more. In others, prompt templates, tokenizer segmentation, alignment depth, or refusal geometry may dominate. This is not a weakness of the research program; rather, it reflects the layered nature of modern aligned LLMs. Existing benchmark practice often compresses these factors into a single ASR number, obscuring the mechanism.

This perspective also suggests that future defenses should be mechanism-aware. If refusal is low-dimensional and easy to suppress, then interventions should make it more distributed or persistent. If template structure or token position is a major source of fragility, then both training and evaluation should vary those factors systematically. If many-shot attacks expose long-context weakness, then safety alignment should be studied as a context-wide property rather than a short-prefix property [11,17]. Finally, if jailbreak transferability is asymmetric, then model-specific internal routing should be measured directly instead of assuming that a successful prompt has universal adversarial meaning.

12. Limitations and Ethical Considerations

This concept paper proposes a mechanistic research agenda rather than presenting completed experiments. Several limitations should be acknowledged. First, causal interpretability methods are easier to apply to open-weight models than to API-only systems. Second, cross-model activation patching is technically difficult even when hidden dimensions match, because different models may organize layer semantics differently. Third, jailbreak success labels are often ambiguous: refusal, partial compliance, and unsafe continuation can lie on a continuum rather than forming clean categories. Fourth, tokenizer and template analysis can explain some transfer failures, but it may not capture deeper representational differences.

This work also has dual-use implications. The goal is to understand and improve safety mechanisms, not to publish new operational jailbreak prompts. Experiments should use established safety benchmarks, avoid releasing novel harmful prompts or unsafe model outputs, and report results at the level of aggregate mechanisms whenever possible.

13. Conclusion

This concept paper argues that different aligned LLMs exhibit different internal responses to the same jailbreak prompt because jailbreak behavior is not controlled by architecture alone. Instead, it emerges from the interaction of architecture, tokenization, prompt-template structure, post-training alignment, and internal routing of refusal and compliance. Prior work on safety-training failure

 doi:10.20944/preprints202604.1776.v1

modes, optimization-based jailbreaks, shallow alignment, prompt-template preservation, refusal directions, refusal geometry, attention hijacking, and token-position effects already points strongly in this direction [4,5,10–16,21]. The next step is therefore not only to measure jailbreak success, but to explain the internal mechanisms that produce it.

References

1. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.
2. R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
3. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
4. A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does LLM safety training fail?" in *Advances in Neural Information Processing Systems*, vol. 36, pp. 80079–80110, 2023.
5. A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.
6. X. Jia, T. Pang, C. Du, Y. Huang, J. Gu, Y. Liu, X. Cao, and M. Lin, "Improved techniques for optimization-based jailbreaking on large language models," in *International Conference on Learning Representations*, 2025.
7. J. Mu, Z. Ying, Z. Fan, Z. Jing, Y. Zhang, Z. Yu, W. Zhang, Q. Zou, and X. Zhang, "Mask-GCG: Are all tokens in adversarial suffixes necessary for jailbreak attacks?" *arXiv preprint arXiv:2509.06350*, 2025. Accepted to ICASSP 2026.
8. J. Chu, Y. Liu, Z. Yang, X. Shen, M. Backes, and Y. Zhang, "JailbreakRadar: Comprehensive assessment of jailbreak attacks against LLMs," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 21538–21566, 2025.
9. F. Bianchi, M. Suzgun, G. Attanasio, P. Röttger, D. Jurafsky, T. Hashimoto, and J. Zou, "Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions," in *International Conference on Learning Representations*, 2024.
10. X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" in *International Conference on Learning Representations*, 2024.
11. X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson, "Safety alignment should be made more than just a few tokens deep," in *International Conference on Learning Representations*, 2025.
12. K. Lyu, H. Zhao, X. Gu, D. Yu, A. Goyal, and S. Arora, "Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
13. A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, "Refusal in language models is mediated by a single direction," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
14. T. Wollschläger, J. Elstner, S. Geisler, V. Cohen-Addad, S. Günnemann, and J. Gasteiger, "The geometry of refusal in large language models: Concept cones and representational independence," in *International Conference on Machine Learning*, 2025.
15. Z. Wang, H. Tu, J. Mei, B. Zhao, Y. Wang, and C. Xie, "AttnGCG: Enhancing jailbreaking attacks on LLMs with attention manipulation," *Transactions on Machine Learning Research*, 2025.
16. M. Ben-Tov, M. Geva, and M. Sharif, "Universal jailbreak suffixes are strong attention hijackers," *arXiv preprint arXiv:2506.12880*, 2025.
17. C. Anil, E. Durmus, N. Panickssery, M. Sharma, J. Benton, et al., "Many-shot jailbreaking," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
18. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.
19. LMSYS, "Vicuna v1.5 model card," Hugging Face model card, 2023. Available: https://huggingface.co/lmsys/vicuna-7b-v1.5
20. Mistral AI, "Mistral-7B-Instruct-v0.3 model card," Hugging Face model card, 2024. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

 doi:10.20944/preprints202604.1776.v1

21. H. Eddoubi, U. F. Abdullahi, and F. Hassan, "Beyond suffixes: Token position in GCG adversarial attacks on large language models," *arXiv preprint arXiv:2602.03265*, 2026.