Rethinking Deep Alignment Through the Lens of Incomplete Safety Learning

Thong Bach[1], **Dung Nguyen**[1], **Thao Minh Le**[2], **Truyen Tran**[1]

[1]Applied Artificial Intelligence Intiative (A2I2), Deakin University
[2]Pennsylvania State University
{t.bach,dung.nguyen,truyen.tran}@deakin.edu.au, mxl6224@psu.edu

Abstract

Large language models exhibit systematic vulnerabilities to adversarial attacks despite extensive safety alignment through supervised fine-tuning and reinforcement learning from human feedback. These vulnerabilities manifest as differential safety behavior across token positions, with safety modifications concentrating in early positions while later positions show minimal distributional changes from base models. We provide a mechanistic analysis of safety alignment training dynamics, revealing that gradient concentration during autoregressive training creates signal decay across token positions. This leads to incomplete distributional learning where safety training fails to fully transform model preferences in later response regions. We introduce base-favored tokens as computational indicators of incomplete safety learning. Analysis reveals that while early positions undergo substantial distributional changes, later positions retain concerning base model preferences in safety-critical contexts, indicating systematic incomplete learning due to insufficient training signals. We develop a targeted completion method that addresses these undertrained regions through adaptive penalties and hybrid teacher distillation. Experimental evaluation across Llama and Qwen model families demonstrates remarkable improvements in adversarial robustness, with dramatic reductions in attack success rates across multiple attack types while fully preserving general capabilities.

Introduction

Large language models undergo multi-stage safety alignment through supervised fine-tuning and reinforcement learning from human feedback to reduce harmful outputs (Bai et al. 2022; Rafailov et al. 2023; Ethayarajh et al. 2024; Azar et al. 2024). Despite extensive alignment efforts, these models exhibit systematic vulnerabilities to adversarial attacks (Andriushchenko, Croce, and Flammarion 2024; Chao et al. 2024; Zou et al. 2024; Xie et al. 2024; Huang et al. 2024b) , fine-tuning degradation (Che et al. 2025; Arditi et al. 2024; Lyu et al. 2024; Li et al. 2024), and context manipulation (Wei et al. 2024; Huang et al. 2024c; Qi et al. 2023; Chen et al. 2025). Recent work (Qi et al. 2024) identifies a key pattern underlying these vulnerabilities: safety alignment concentrates primarily in early token positions while later positions show minimal distributional changes from base models, termed "shallow alignment."

While this empirical characterization explains *what* happens during safety alignment, the fundamental question of *why* this pattern emerges during training remains unexplored. Understanding the mechanistic origins of shallow alignment is crucial for developing principled solutions that address root causes rather than symptoms. Moreover, if safety training systematically affects some token positions more than others (first few tokens), while the later ones do not change much between safe and base models, this suggests that alignment may be incomplete in certain regions of the response sequence. However, current approaches lack computational methods to detect where such incomplete learning occurs.

We address these gaps through mechanistic analysis of safety alignment training dynamics. Our investigation reveals that shallow alignment arises from gradient concentration and signal decay inherent to autoregressive training: early positions receive strong gradient signals due to shorter dependency chains, while later positions experience signal decay. Consequently, safety training incompletely transforms distributional preferences, where early positions undergo substantial changes while later positions retain base model patterns for formatting, punctuation, and linguistic preferences.

To detect incomplete distributional transformation, we introduce base-favored tokens: vocabulary positions where base model preferences exceed aligned model preferences. Unlike aggregate measures such as KL divergence, base-favored tokens provide fine-grained indicators of incomplete alignment. These tokens, predominantly formatting and structural elements, exhibit distinct patterns: early positions show many base-favored tokens due to training contention, while later positions retain base preferences due to insufficient gradient signal.

Building on this understanding, we develop a targeted completion method that addresses incomplete learning through adaptive penalties on base-favored tokens and hybrid teacher distillation. This approach completes the distributional transformation that safety alignment began but could not finish, achieving comprehensive safety alignment throughout response sequences.

Our contributions are: **(1)** The mechanistic analysis ex-

plaining why shallow alignment occurs during safety alignment training, establishing gradient concentration and signal decay as fundamental causes. **(2)** Base-favored tokens as computational indicators of incomplete distributional alignment, enabling fine-grained detection of undertrained regions. **(3)** A targeted completion framework that surgically addresses incomplete learning without expensive retraining. **(4)** Comprehensive experimental validation demonstrating substantial improvements in adversarial robustness (96-98% attack reduction across model families), safety recovery capabilities, and deliberative reasoning under adversarial conditions.

Limitations of Autoregressive Safety Alignment

Safety alignment vulnerabilities arise from limitations inherent to autoregressive training, independent of the specific alignment methodology employed. The sequential loss structure underlying all language model training—supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO)—creates systematic gradient concentration and error accumulation that compromises alignment effectiveness in later token positions.

Gradient Concentration in Sequential Loss Functions

Token Position Notation: We use $t \in \{1, 2, \ldots, T\}$ to denote the position index within a response sequence, where $t = 1$ represents the first response token after the instruction, and T represents the maximum response length. For a given context $(x, y_{<t})$, x denotes the input instruction and $y_{<t} = (y_1, y_2, \ldots, y_{t-1})$ represents the response prefix preceding position t.

All autoregressive training optimizes the standard language modeling objective:

$$\mathcal{L} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\sum_{t=1}^{T} \log P(y_t|x, y_{<t}) \right] \tag{1}$$

where the sequential factorization follows from the chain rule: $P(y|x) = \prod_{t=1}^{T} P(y_t|x, y_{<t})$.

We can measure gradient concentration by examining how gradient magnitudes vary across token positions:

$$\text{GradMag}(t) = \left\| \frac{\partial}{\partial \theta} \log P(y_t|x, y_{<t}) \right\|_2 \tag{2}$$

This measures the gradient magnitude from the loss term at position t specifically.

This fundamental structure creates position-dependent gradient magnitudes through two mechanisms: **Computational Path Length:** Early tokens experience shorter dependency chains in the loss function, receiving stronger gradient signals. Later tokens depend on longer context sequences, leading to gradient dilution through complex attention computations. **Context Dependency Asymmetry:** Early tokens influence all subsequent predictions through the autoregressive context, causing parameters affecting early positions to receive gradient contributions from multiple loss terms. Later tokens contribute only to their own prediction terms, resulting in weaker parameter updates.

Error Accumulation in Autoregressive Generation

The autoregressive nature of LLMs leads to **error accumulation**, where alignment deviations in early tokens compound through dependency chains. When early positions fail to establish a proper safety context, this misalignment propagates to later positions that lack a sufficient gradient signal to correct course.

Formally, let $\epsilon_i = \text{KL}(\pi_{\text{aligned}}(\cdot|x, y_{<i}) \| \pi_{\text{base}}(\cdot|x, y_{<i}))$ represent the distributional alignment deviation at position i. The influence of position i on position t can be quantified through gradient-based analysis:

$$\text{Influence}(i \to t) = \left\| \frac{\partial \log P(y_t|x, y_{<t})}{\partial h_i} \right\|_2 \tag{3}$$

where h_i is the hidden state at position i and $\| \cdot \|_2$ denotes the L2 norm. The cumulative alignment error becomes:

$$\text{Error}(t) = \sum_{i=1}^{t-1} \epsilon_i \cdot \text{Influence}(i \to t) \tag{4}$$

While computing influence scores for all position pairs has $O(t^2)$ complexity, this framework captures how early alignment deviations propagate through transformer attention mechanisms. In safety-critical contexts, adversarial inputs exploit this accumulation: high early deviations (ϵ_i for small i) compound through strong influence weights, overwhelming later positions that received insufficient training signal.

Adversarial Safety Contexts

These autoregressive limitations are most evident in adversarial safety contexts where attackers exploit incomplete alignment in later token positions.

Definition 1 (Adversarial Safety Contexts)**.** *An adversarial safety context is a tuple $(x, y_{<t})$ where attackers exploit incomplete alignment by combining:*

- *$x \in \mathcal{X}_{harmful}$: a harmful instruction from established benchmarks such as AdvBench (Zou et al. 2023)*
- *$y_{<t} \in \mathcal{Y}_{bypass}$: a response prefix designed to bypass initial refusal mechanisms, forcing generation past safety guardrails*

The set of all adversarial safety contexts is:

$$\mathcal{C}_{adversarial} = \{(x, y_{<t}) \mid x \in \mathcal{X}_{harmful} \wedge y_{<t} \in \mathcal{Y}_{bypass}\} \tag{5}$$

Base-Favored Tokens: Indicators of Incomplete Distributional Alignment

Our gradient dynamics analysis explains *why* shallow alignment occurs, but **how can we detect where incomplete learning happens?** Standard metrics like KL divergence average over entire distributions, masking fine-grained patterns where specific vocabulary positions remain undertrained.

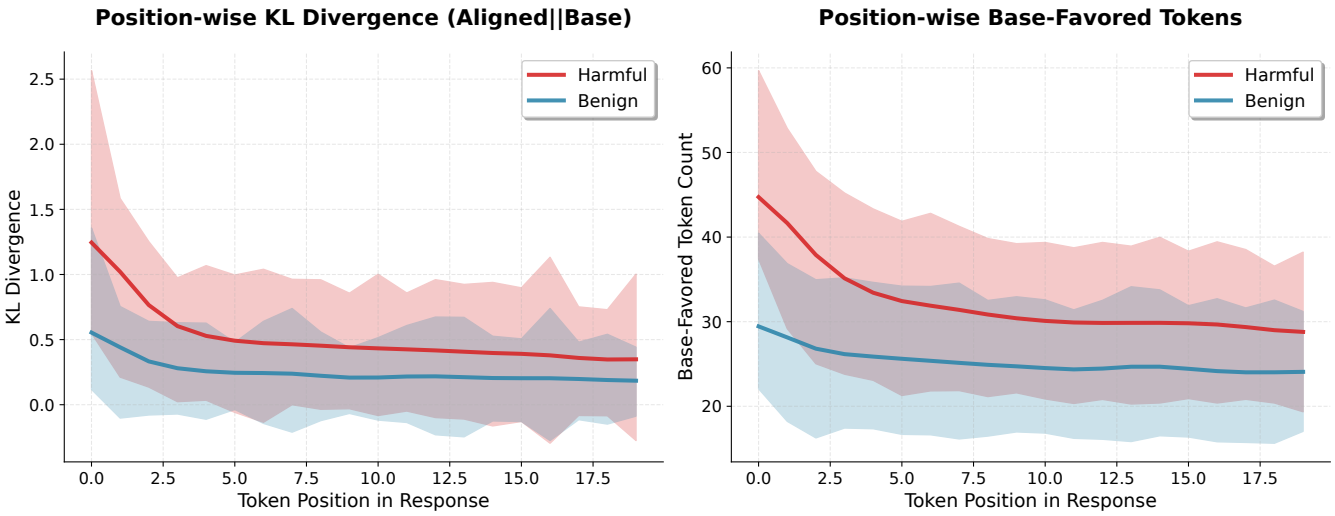

Figure 1: Position-wise analysis of shallow alignment detection. Left: KL divergence between aligned and base models shows declining distributional differences across token positions. Right: Base-favored tokens exhibit the same shallow alignment pattern, with adversarial safety contexts showing systematically higher counts than benign contexts (45 vs 30 tokens at early positions, 35 vs 25 at later positions). Base-favored tokens validate shallow alignment detection while providing vocabulary-level identification of specific undertrained tokens that aggregate measures cannot localize.

We introduce **base-favored tokens** as computational indicators of incomplete distributional alignment. These tokens reveal how safety training creates broad distributional shifts affecting formatting, punctuation, and linguistic preferences, but this transformation remains incomplete due to gradient decay across token positions.

Base-Favored Tokens: Definition and Detection

Definition 2 (Base-Favored Tokens)**.** *For a given context $(x, y_{<t})$ at response position t (defined in the previous section), base-favored tokens are vocabulary elements where the base model assigns a higher probability than the aligned model:*

$$\mathcal{B}_t(x, y_{<t}) = \{v \in \mathcal{V} : \pi_{base}(v|x, y_{<t}) > \pi_{aligned}(v|x, y_{<t})\} \tag{6}$$

where $\mathcal{V}$ denotes the vocabulary and the subscript t indicates the position-dependent context (Algorithm 1).

Algorithm 1: Base-Favored Token Detection

Require: Adversarial context $(x, y_{<t}) \in \mathcal{C}_{\text{adversarial}}$, models $\pi_{\text{base}}, \pi_{\text{aligned}}$, threshold k

1: Compute $L_{\text{base}} \leftarrow \text{logits}_{\text{base}}(x, y_{<t})$
2: Compute $L_{\text{aligned}} \leftarrow \text{logits}_{\text{aligned}}(x, y_{<t})$
3: Calculate preference difference: $\Delta L \leftarrow L_{\text{base}} - L_{\text{aligned}}$
4: Extract top-k base-favored tokens: $\mathcal{B}_t \leftarrow \text{TopK}(\Delta L, k)$
5: **return** Base-favored token set $\mathcal{B}_t$

Empirical Analysis: Incomplete Learning Patterns

We analyze base-favored token patterns during step-by-step generation, comparing harmful contexts (HEx-PHI (Qi et al. 2024) prompts with adversarial prefixes) versus benign contexts (Databricks Dolly instructions test set) using Llama-3.1-8B (base) and Llama-3-8B-Instruct (aligned). At each generation step, we measure Jensen-Shannon divergence, top-100 token overlap, and base-favored token counts where $\pi_{\text{base}}(v|x, y_{<t}) > \pi_{\text{aligned}}(v|x, y_{<t})$. This dynamic analysis reveals how incomplete distributional alignment manifests during actual response generation (detailed methodology in the Appendix). Our analysis reveals three key patterns:

Validation of Shallow Alignment Detection. Adversarial safety contexts exhibit systematically higher base-favored token counts across all positions compared to benign contexts, demonstrating that incomplete alignment is context-dependent and more pronounced when processing harmful content. Early positions (0-2) show 45 vs 30 tokens (50% increase), while later positions (15+) maintain 35 vs 25 tokens (40% increase) (Figure 1). This declining pattern confirms shallow alignment detection consistent with KL divergence trends.

Vocabulary-Specific Detection Tool. Base-favored tokens identify precise vocabulary positions where base model preferences persist, enabling targeted intervention on specific undertrained tokens that aggregate distributional measures cannot localize.

Vocabulary Distribution Analysis. The analysis of the most frequent base-favored tokens in Figure 2 reveals they are predominantly formatting elements, punctuation, and structural tokens rather than explicitly harmful content. This indicates that safety alignment modifies the model's probability distribution over the entire vocabulary, affecting stylistic and structural preferences, rather than only suppressing specific harmful words. The presence of these non-harmful tokens as base-favored suggests incomplete distributional

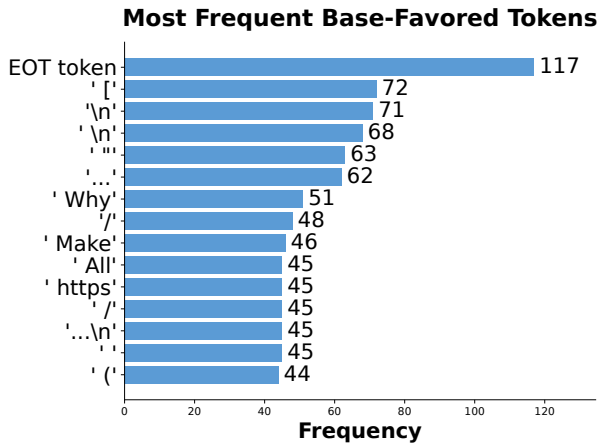

Figure 2: Base-Favored Tokens Reveal Distributional Differences. Most frequent base-favored tokens for Llama-3-8B using harmful instruction-response pairs. Tokens are predominantly formatting elements (punctuation, special tokens), common words, and structural elements rather than explicitly harmful content, supporting the distributional alignment hypothesis.

transformation where safety training partially altered general linguistic patterns but failed to complete this broader distributional shift.

Functional Validation via Inference-Time Contrastive Decoding

To validate that base-favored tokens represent *functional vulnerability mechanisms* (computational patterns that directly enable adversarial exploitation rather than statistical artifacts), we implement inference-time contrastive decoding.

Contrastive Decoding Methodology: The intervention operates through token detection and penalty application. At each generation step, we identify base-favored tokens where $\pi_{\text{base}}(v|x, y_{<t}) > \pi_{\text{aligned}}(v|x, y_{<t})$ and apply contrastive penalties proportional to preference differences:

$$\text{penalty}(v) = \alpha \cdot (\text{logit}_{\text{base}}(v) - \text{logit}_{\text{aligned}}(v)) \quad (7)$$

where α controls penalty strength. These penalties are subtracted from original logits before softmax normalization, reducing the selection probability of base-favored tokens. The sustained KL divergence in Figure 3 results from continuous suppression of base-favored tokens throughout generation, maintaining distributional separation from the base model across all positions.

Experimental Results Table 1 presents the performance of inference-time contrastive decoding on Llama-3.1-8B-Instruct. The intervention reduces prefill attack [1] success rates from 47.5% to 0.2% while maintaining performance across utility benchmarks. Utility metrics show minimal deviation from baseline performance, with most benchmarks exhibiting changes within 1-2 percentage points.

Position-Wise Safety Enhancement Figure 3 illustrates how contrastive decoding modifies the KL-divergence profile across token positions. The baseline aligned model exhibits high early-position KL-divergence (approximately 1.8) that decays to low late-position values (approximately 0.5), consistent with shallow alignment patterns. Contrastive decoding maintains elevated KL-divergence throughout the sequence (10.0 to 6.0), indicating sustained distributional differences from the base model across all positions. This KL-divergence profile suggests that targeted intervention on base-favored tokens extends safety-aligned behavior to later token positions where original training signals were insufficient. The sustained distributional separation provides evidence that base-favored token penalties address the underlying mechanisms of shallow alignment vulnerability.

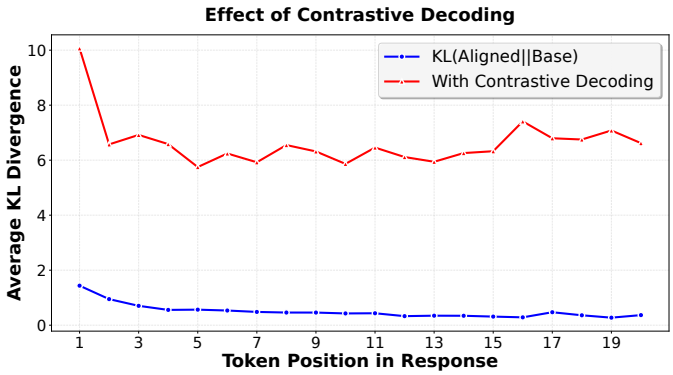

Figure 3: KL-divergence across token positions on LLama 3.1 8B between aligned and base model using normal decoding (blue) and contrastive decoding intervention (orange) in safety-critical contexts (Hex-PHI dataset). Contrastive decoding maintains higher KL-divergence throughout the sequence, indicating sustained safety alignment in later positions.

Targeted Learning Completion

The inference-time contrastive decoding results provide compelling validation that base-favored tokens represent the actual mechanisms underlying shallow alignment vulnerabilities. However, the computational constraints of real-time intervention, which require concurrent model loading, additional forward passes, and penalty calculations at each generation step, limit practical deployment scalability. This motivates a fundamental question: can we achieve equivalent safety benefits by incorporating these insights directly into model parameters during training, eliminating inference-time overhead while maintaining comparable protection?

[1] In this validation experiment, we use 4-token prefill attacks to demonstrate the base-favored token mechanism under controlled conditions. In main experiments (Section), we employ full targeted sequence prefills to maximize attack strength and evaluate complete robustness, as these represent the strongest adversarial conditions for comprehensive safety evaluation.

Method	Prefill Attack	ARC-E	ARC-C	BoolQ	HellaSwag	Winogrande	MMLU	ToxiGen	TriviaQA	TruthfulQA
Baseline	47.5	52.0	81.8	84.1	59.1	73.8	68.0	53.3	51.6	45.6
Contrastive	0.2	52.7	82.1	84.4	58.8	74.0	66.8	48.6	52.4	46.3

Table 1: Inference-Time Contrastive Decoding Validation. Contrastive decoding dramatically reduces prefill attack success rates from 47.5% to 0.2% while maintaining utility performance across benchmarks, providing functional validation that base-favored tokens represent exploitable vulnerability mechanisms rather than distributional artifacts.

Our targeted completion framework addresses this challenge by translating the successful inference-time intervention into a training-time approach. Rather than detecting and penalizing base-favored tokens during every generation, we identify harmful instruction-response pairs where incomplete learning occurs and apply focused training interventions to complete the suppression process that the alignment phase began but could not finish due to gradient decay.

Safety Learning Completion Framework

Given harmful training contexts $(x, y_{<t})$ where base-favored tokens $\mathcal{B}_t(x, y_{<t})$ indicate incomplete learning, we seek to minimize:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{distillation}} + \alpha_{\text{adaptive}} \cdot \mathcal{L}_{\text{completion}} \tag{8}$$

where $\mathcal{L}_{\text{completion}}$ provides targeted intervention on identified incomplete learning locations, analogous to the penalty mechanism validated in our inference-time experiments.

Adaptive Penalty-Based Completion

We introduce a focused training method that applies adaptive L_2 penalties specifically to base-favored tokens in harmful contexts, directly inspired by the successful contrastive penalty mechanism.

Definition 3 (Targeted L_2 Completion Loss)**.** *For harmful training context $(x, y_{<t})$ with detected base-favored tokens $\mathcal{M} = \textit{TopK}(\textit{logits}_{\textit{base}} - \textit{logits}_{\textit{aligned}}, k)$:*

$$\mathcal{L}_{\textit{completion}} = \lambda_{\textit{reg}} \sum_{v \in \mathcal{V}} \mathcal{I}_{\mathcal{M}}[v] \cdot (\textit{logits}_{\textit{aligned}}[v])^2 \tag{9}$$

where $\mathcal{I}_{\mathcal{M}}[v] = 1$ if $v \in \mathcal{M}$ and 0 otherwise, k is the number of top base-favored tokens, and $\lambda_{\textit{reg}} > 0$ is the regularization strength.

This approach directly suppresses the logit magnitudes of tokens where base model preferences exceed aligned model preferences. To adapt intervention strength based on incomplete learning severity, we scale penalties using base-favored token density:

$$\text{risk}_{\text{level}} = \frac{|\mathcal{B}_t(x, y_{<t})|}{|\mathcal{V}|} \tag{10}$$

$$\alpha_{\text{adaptive}} = \alpha_{\text{base}} \cdot (1 + \gamma \cdot \text{risk}_{\text{level}}) \tag{11}$$

where $\alpha_{\text{base}} > 0$ is the base penalty strength and $\gamma \geq 0$ controls adaptive scaling.

Hybrid Teacher Construction

Definition 4 (Hybrid Teacher Model)**.** *The teacher model combines aligned and base model knowledge through weighted interpolation:*

$$\textit{logits}_{\textit{teacher}} = \lambda \cdot \textit{logits}_{\textit{aligned}} + (1 - \lambda) \cdot \textit{logits}_{\textit{base}} \tag{12}$$

where $\lambda > 1$ (1.2 in our experiments) amplifies aligned model preferences while retaining base model information for utility preservation (Huang et al. 2024a).

The complete training objective integrates knowledge distillation with targeted completion:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KL}}(\text{student}, \text{teacher}) + \alpha_{\text{adaptive}} \cdot \mathcal{L}_{\text{completion}}(\text{student}, \mathcal{M}) \tag{13}$$

where the student model is trained using temperature-scaled KL divergence:

$$\mathcal{L}_{\text{KL}} = \text{KL}\left(\text{softmax}\left(\frac{\text{logits}_{\text{student}}}{\tau}\right) \| \text{softmax}\left(\frac{\text{logits}_{\text{teacher}}}{\tau}\right)\right) \cdot \tau^2 \tag{14}$$

with temperature parameter $\tau = 2.0$.

Experimental Evaluation

Deep Safety Alignment Capacity

Our first evaluation demonstrates that targeted completion achieves **deep safety alignment**, a robust safety behavior that appears across token positions and resists sophisticated adversarial manipulation.

Method	Prefill (%)	GCG (%)	Fine-tuning HRR
Baseline	23.0	51.0 ± 42.9	21.4
Safety Augmentation	0.8	1.6 ± 3.6	4.4
Ours	**0.5**	**0.4 ± 0.9**	**4.4**

Table 2: Comprehensive Deep Alignment Validation on Llama-2-7B-Chat. Our method demonstrates exceptional resistance across multiple attack vectors, achieving near-zero GCG attack success rates (0.4% vs 51.0% baseline), outperforming both baseline and safety augmentation (Qi et al. 2024) methods.

Experimental Setup. We evaluate across four model families: Llama-2-7B-Chat (Touvron et al. 2023), Llama-3.1-8B-Instruct (Dubey et al. 2024), Qwen-2.5-7B-Instruct (Bai et al. 2023), Qwen-3-8B-Instruct (Yang et al. 2025) using three complementary attack protocols. *Prefilling attacks* force models to begin responses with harmful continuations, exploiting incomplete learning in later positions. *Fine-tuning robustness* tests safety persistence after benign adaptation on Dolly instruction-following data using LoRA (Hu

et al. 2022) (rank 32, learning rate 2×10^{-4}, 1 epoch), measuring Harmfulness Rejection Rate (HRR) on AdvBench. *GCG optimization attacks* (Zou et al. 2023) employ adversarial suffix optimization to exploit undertrained regions. Our method uses HEx-PHI data (330 harmful pairs) for completion loss computation with top-100 base-favored token selection and adaptive L2 penalties, while the GSM8K training set provides distillation supervision. Training employs hybrid teacher construction ($\lambda = 1.2$, 20 epochs).

Attack Resistance. In Table 3, our method reduces attack success rates by 48–96% across four model families, with prefilling attacks dropping from 23–96% to 0.5–44% and consistent fine-tuning robustness improvements.

Besides, on LLama-2-7B-Chat, GCG optimization attacks achieve only 0.4%±0.9% success rate versus 51.0%±42.9% baseline (99.2% reduction), acquiring comparable performance with shallow alignment safety augmentation (1.6%±3.6%), with cross-dataset generalization (HEx-PHI training, AdvBench evaluation) confirming our method addresses fundamental incomplete learning (Table 2).

Enhanced Deliberative Reasoning Under Adversarial Conditions

Beyond attack resistance, our method reveals an emergent capability: **enhanced deliberative reasoning** under adversarial conditions, suggesting deep alignment unlocks advanced cognitive processes rather than simple refusal mechanisms.

Experimental Setup. We evaluate Qwen-3-8B-Instruct under prefill attacks using 384 AdvBench prompts, classifying responses by (1) safety outcome (harmful/safe) and (2) reasoning engagement (explicit deliberation about safety). This yields four categories: harmful/safe with/without reasoning.

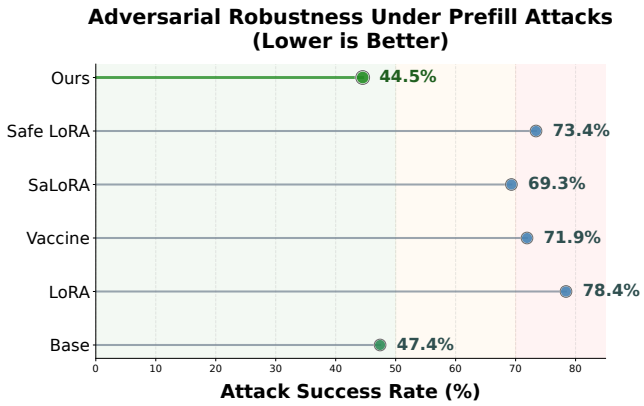

Figure 4: Deep Alignment Recovery Under Adversarial Attack. Prefill attack success rates demonstrate that our method (44.5% ASR) significantly outperforms existing safety preservation methods (69.3-73.4% ASR) and approaches the robustness of uncompromised base models (47.4% ASR), validating comprehensive deep alignment restoration.

Key Findings. In Table 5, our method transforms both safety and reasoning: harmful responses drop from 96.1% to 16.4%, while reasoning engagement increases from 37.8% to 60.2%. Most significantly, safe responses with reasoning increase 22-fold ($9 \rightarrow 196$), indicating a shift from *reactive* safety (recovery after harmful content) to *proactive* reasoning (prevention through deliberation).

The baseline model shows predominantly reactive patterns: reasoning typically occurs after harmful content generation (136/145 reasoning cases are harmful). Our method enables proactive reasoning: 85% of reasoning responses (196/231) result in safe outcomes, with deliberation occurring to prevent harmful content generation.

Implications. These results suggest deep alignment enhances cognitive sophistication beyond attack resistance, enabling complex deliberative processes that evaluate and respond appropriately to adversarial inputs through enhanced reasoning rather than simple refusal mechanisms.

Alignment Recovery After Fine-tuning

In this setting, we demonstrate that targeted completion can **restore deep safety alignment** in models that have experienced safety degradation through task-specific fine-tuning, a critical capability for production deployment where models must be adapted to specific use cases.

Experimental Setup. We simulate real-world deployment scenarios where safety-aligned models undergo task-specific fine-tuning that degrades their safety properties. Starting with aligned models (Llama-2-7B-Chat, Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct), we fine-tune on the Dolly (Conover et al. 2023) instruction-following dataset using LoRA adaptation (rank 32, learning rate 2×10^{-4}, batch size 128, 1 epoch), which introduces substantial safety degradation as documented in prior work (Qi et al. 2023). We then apply various recovery methods and evaluate safety restoration via Harmfulness Rejection Rate (HRR) on AdvBench and utility preservation across ARC-Challenge, GSM8K, ToxiGen, and TruthfulQA benchmarks.

Baselines. We compare against state-of-the-art safety preservation methods: Vaccine (Huang et al. 2024d) (perturbation-based robustness), SaLoRA (Liu et al. 2024) (safety-orthogonal projections), and SafeLoRA (Rando et al. 2022) (alignment-preserving subspaces), alongside standard LoRA fine-tuning.

Safety Restoration. In Table 4, our targeted completion method achieves near-complete safety recovery across all model families, with HRR values of 1.0%, 0.5%, and 0.3% for Llama-3.1-8B, Qwen-2.5-7B, and Llama-2-7B, respectively. This result dramatically improves over standard LoRA fine-tuning and achieves comparable performance with existing safety preservation methods. These results approach or match the safety levels of original base models while preserving utility, indicating comprehensive alignment restoration.

Adversarial Robustness Validation. Figure 4 presents our deep alignment recovery approach achieves a 44.5% attack success rate under prefill attacks, substantially outperforming existing safety preservation methods (Vaccine:

Model	Method	Deep Alignment Metrics		Utility Preservation				
		Prefill ASR ↓	Fine-tuning HRR ↓	MMLU	ARC-C	BoolQ	HellaSwag	Winogrande
Llama-2-7B-Chat	Baseline	23.0	21.4	45.0	43.4	80.6	57.8	67.0
	Ours	**0.5**	**4.4**	44.3	42.2	81.0	57.6	67.6
Llama-3.1-8B-Instruct	Baseline	90.1	25.3	68.0	51.6	84.1	59.1	73.6
	Ours	**14.8**	**12.3**	68.9	51.9	84.1	59.0	73.7
Qwen-2.5-7B-Instruct	Baseline	85.9	24.7	71.8	52.9	86.4	62.0	70.1
	Ours	**44.3**	**13.8**	71.8	52.0	86.3	62.4	69.8
Qwen-3-8B-Instruct	Baseline	96.1	10.7	73.0	55.5	86.6	57.1	68.1
	Ours	**16.4**	**4.7**	72.7	54.9	86.5	56.9	67.6

Table 3: Deep Alignment Achievement: Adversarial Robustness and Utility Preservation. Our targeted completion method achieves dramatic attack resistance across model families while preserving general capabilities, demonstrating successful completion of safety alignment throughout response sequences.

Models	Methods	Deep Alignment Recovery		Utility Preservation			
		Eval Loss ↓	HRR ↓	ARC-C	GSM8K	ToxiGen	TruthfulQA
Llama-3.1-8B	Base	1.9	1.4	52.0	75.2	53.3	45.5
	LoRA	1.2	25.5	51.2	72.4	44.9	39.0
	Vaccine	1.3	21.3	44.3	39.5	43.4	34.1
	SaLoRA	1.2	8.1	52.3	75.7	49.3	41.8
	Safe LoRA	1.3	11.0	51.1	75.6	48.7	42.0
	Ours	1.3	**1.0**	52.0	78.1	53.5	46.6
Qwen-2.5-7B	Base	3.6	0.0	53.0	76.4	57.2	56.3
	LoRA	1.2	24.7	55.0	60.2	57.2	44.5
	Vaccine	1.2	19.3	54.6	74.3	57.9	44.5
	SaLoRA	1.2	3.4	55.0	69.5	57.2	49.2
	Ours	1.3	**0.5**	52.7	73.1	57.6	54.7
Llama-2-7B	Base	2.5	0.0	43.3	20.1	52.9	37.2
	LoRA	1.1	21.4	44.4	19.6	44.7	32.3
	Vaccine	1.1	16.7	42.6	11.6	41.1	31.7
	SaLoRA	1.1	**0.0**	45.9	23.6	49.5	34.7
	Safe LoRA	1.2	**0.0**	45.6	21.5	43.8	33.1
	Ours	1.2	0.3	45.7	21.5	47.1	35.6

Table 4: Deep Alignment Recovery: Post-Training Safety Restoration. Our targeted completion method achieves superior safety recovery while enhancing utility performance, demonstrating effective restoration of deep safety alignment after fine-tuning degradation.

Response Category	Baseline	Deep Alignment
Harmful + Reasoning	136	35
Harmful + No Reasoning	233	28
Total Harmful	*369 (96.1%)*	*63 (16.4%)*
Safe + Reasoning	9	196
Safe + No Reasoning	6	125
Total Safe	*15 (3.9%)*	*321 (83.6%)*
Total with Reasoning	**145 (37.8%)**	**231 (60.2%)**

Table 5: Enhanced Deliberative Reasoning Under Adversarial Attack. Our method achieves dramatic safety improvement (96.1% → 16.4% harmful) while enhancing reasoning engagement (37.8% → 60.2%). The 22-fold increase in safe reasoning responses (9 → 196) demonstrates proactive safety reasoning.

71.9%, SaLoRA: 69.3%, Safe LoRA: 73.4%) and approaching the robustness of uncompromised base models (47.4%). This superior adversarial resistance demonstrates that our base-favored token completion approach addresses root causes of alignment vulnerabilities rather than merely preserving existing safety features during fine-tuning.

Conclusion

This work provides a mechanistic account of safety alignment failures, showing that gradient concentration leads to systematic undertraining. We introduce a framework to detect and restore these regions, offering a principled alternative to broad retraining. Our results also suggest that completing distributional alignment improves deliberative reasoning, hinting at a deeper link between alignment completeness and model cognition. Future work should scale these methods to larger models and examine their relationship to other safety dimensions.

References

Andriushchenko, M.; Croce, F.; and Flammarion, N. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*.

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37: 136037–136083.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Chao, P.; Debenedetti, E.; Robey, A.; Andriushchenko, M.; Croce, F.; Sehwag, V.; Dobriban, E.; Flammarion, N.; Pappas, G. J.; Tramer, F.; et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37: 55005–55029.

Che, Z.; Casper, S.; Kirk, R.; Satheesh, A.; Slocum, S.; McKinney, L. E.; Gandikota, R.; Ewart, A.; Rosati, D.; Wu, Z.; et al. 2025. Model tampering attacks enable more rigorous evaluations of llm capabilities. *arXiv preprint arXiv:2502.05209*.

Chen, R.; Perin, G. J.; Chen, X.; Chen, X.; Han, Y.; Hirata, N. S.; Hong, J.; and Kailkhura, B. 2025. Extracting and understanding the superficial knowledge in alignment. *arXiv preprint arXiv:2502.04602*.

Conover, M.; Hayes, M.; Mathur, A.; Xie, J.; Wan, J.; Shah, S.; Ghodsi, A.; Wendell, P.; Zaharia, M.; and Xin, R. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, J. Y.; Sengupta, S.; Bonadiman, D.; Lai, Y.-a.; Gupta, A.; Pappas, N.; Mansour, S.; Kirchhoff, K.; and Roth, D. 2024a. Deal: Decoding-time alignment for large language models. *arXiv preprint arXiv:2402.06147*.

Huang, T.; Hu, S.; Ilhan, F.; Tekin, S.; and Liu, L. 2024b. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37: 104521–104555.

Huang, T.; Hu, S.; Ilhan, F.; Tekin, S. F.; and Liu, L. 2024c. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.

Huang, T.; Hu, S.; Ilhan, F.; Tekin, S. F.; and Liu, L. 2024d. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Liu, B.; Tan, J.; Mu, Z.; Ding, W.; Wu, R.; Wang, J.; et al. 2024. Salora: Safety-aware low-rank adaptation for large language models. *arXiv preprint arXiv:2405.09859*.

Lyu, K.; Zhao, H.; Gu, X.; Yu, D.; Goyal, A.; and Arora, S. 2024. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *Advances in Neural Information Processing Systems*, 37: 118603–118631.

Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2024. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramer, F. 2022. Safe lora: Safer low-rank adaptation via subspace alignment. *arXiv preprint arXiv:2405.16833*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei, B.; Huang, K.; Huang, Y.; Xie, T.; Qi, X.; Xia, M.; Mittal, P.; Wang, M.; and Henderson, P. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*.

Xie, T.; Qi, X.; Zeng, Y.; Huang, Y.; Sehwag, U. M.; Huang, K.; He, L.; Wei, B.; Li, D.; Sheng, Y.; et al. 2024. Sorry-bench: Systematically evaluating large language model safety refusal. *arXiv preprint arXiv:2406.14598*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Kolter, J. Z.; Fredrikson, M.; and Hendrycks, D. 2024. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37: 83345–83373.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.