WHY DOES LONG-CONTEXT LLM SAFETY BREAK? A POSITIONAL ENCODING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have recently demonstrated strong long-context capabilities, enabling them to process and reason over increasingly extensive inputs. However, this capability also exposes them to many-shot jailbreaking (MSJ), in which long sequences of adversarial demonstrations can weaken refusal behavior on harmful requests. We find that the RoPE base is closely associated with this vulnerability: increasing it substantially lowers attack success, but directly doing so also reduces long-context utility. An analysis of RoPE interactions shows how input length and the base jointly shape rotary phases across token distances. This suggests a way to use positional variation during training without changing the model's inference setting. We introduce `MFA`, a multi-frequency alignment framework that samples RoPE bases during GRPO training and strengthens safety updates for longer prompts. At inference, the model retains its native RoPE base. On Llama-3.1-8B-Instruct under 256-shot MSJ, `MFA` reduces attack success rate from 65.5% to 8.8%, compared with 28.0% for GRPO trained at a single base. Across the evaluated models and attack lengths, `MFA` improves many-shot safety while maintaining performance on general and long-context tasks. Code is available at https://anonymous.4open.science/r/mfa-EBA4.

1 INTRODUCTION

Long-context capability has become central to modern large language models (LLMs) (Jaech et al., 2024; Yang et al., 2025). A single prompt may now contain extensive documents, code, prior dialogues, and retrieved information before the model responds to a user request. However, this extended context window simultaneously broadens the attack surface. In many-shot jailbreaking (MSJ), numerous harmful query–response demonstrations precede a malicious request, steering the model to follow the demonstrated behavior instead of its safety guardrails (Anil et al., 2024; Kim et al., 2025). Consequently, a request that is robustly refused in isolation may still elicit an unsafe answer when embedded at the end of a long adversarial context.

While safety alignment via human feedback and preference optimization (Ouyang et al., 2022; Rafailov et al., 2023; Meng et al., 2024) effectively instills refusal behaviors for standard interactions, these guardrails frequently degrade in long-context scenarios. Recent studies on MSJ and related attacks (Shah et al., 2025) suggest that this vulnerability stems directly from the positional structure of the prompt rather than its semantic content alone. Since modern LLMs process sequence positions primarily through Rotary Positional Embeddings (RoPE), this positional sensitivity is fundamentally tied to the RoPE frequency base. By determining how token distances modulate query–key interactions, the frequency base emerges as a natural variable for dissecting long-context vulnerabilities.

This mechanistic link motivates the core research questions of our paper: *Does the RoPE base dictate an aligned model's vulnerability to long-context jailbreaks, and if so, can we leverage its variation during training to fortify long-context safety without sacrificing utility?*

Empirical study (Section 3). To answer the first question, we probe positional sensitivity by varying the RoPE base θ at inference time under fixed attack settings. Examining two aligned models with different native bases, we observe a stark trade-off. Under 32-shot PANDAS (an MSJ variant) (Ma et al., 2025), artificially increasing the base of Llama-3.1-8B-Instruct from $500{,}000$ to $5{,}000{,}000$ drops the attack success rate (ASR) from 77.50% to 21.75%. However, its long-context utility (LOONG score) concurrently falls from 28.4 to 21.5. Qwen2.5-7B-Instruct exhibits a similar pattern:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ASR decreases from 48.50% to 23.75%, while LOONG decreases from 31.0 to 25.0. This establishes that simply tuning the base at inference is an insufficient defense. Instead, this forced safety–utility trade-off highlights the frequency base as a powerful lever for safety training.

Positional analysis (Section 4). Before integrating this variable into training, we formally analyze what changes under the base intervention. In RoPE, the base sets the frequency of each rotary pair, while input length determines the range of token separations to which these frequencies apply. We introduce *phase pressure*, $\Phi = L/\theta$, to characterize the phase span at the low-frequency limit of a continuous RoPE spectrum for a sequence of length L. This analysis cleanly decouples the distinct roles of the RoPE base and context length in positional processing, providing the theoretical foundation for corresponding design choices in our training methodology.

Multi-frequency alignment (Section 5). Building on these insights, we propose `MFA` (Multi-Frequency Alignment) within the GRPO framework (Shao et al., 2024) to fortify robustness while strictly retaining the model's native base θ_0 at inference. During training, each response group shares a dynamically sampled RoPE base, ensuring responses are compared under a consistent positional configuration while exposing the policy to diverse frequency scales across groups. Furthermore, a length-dependent factor scales the safety advantage *after* group normalization, granting longer adversarial prompts a proportionately greater influence on the policy update. A KL penalty regularizes the policy against a frozen reference model evaluated at the native base θ_0. Crucially, at inference, the trained model operates solely at θ_0, requiring no base search or additional computational overhead.

We evaluate `MFA` on three models against standard MSJ and PANDAS across multiple shot counts. A matched fixed-base GRPO control isolates the benefits of our two MFA components, while ablations assess base sampling and length weighting separately. On Llama-3.1-8B-Instruct under 256-shot MSJ, ASR decreases from 65.5% for the original model to 28.0% for fixed-base GRPO, and drastically down to 8.8% for `MFA`. The full method also outperforms either single-component variant. Across all evaluated models, these substantial safety gains are achieved without compromising general and long-context benchmarks.

Contributions. To address the critical degradation of LLM safety in long-context scenarios, our contributions are threefold. First, we establish through inference-time RoPE sweeps on fixed checkpoints that many-shot jailbreak vulnerability is strongly associated with a model's positional configuration. Because directly adjusting the base at inference imposes a strict safety–utility trade-off, we formalize a phase pressure analysis that decouples the effects of base and input length to guide our training interventions. Second, we develop `MFA`, a methodology that resolves this trade-off by integrating group-shared base sampling, length-weighted safety advantages, and a frozen native-base reference within GRPO. Third, comprehensive evaluations across models and attack settings demonstrate that `MFA` substantially reduces long-context attack success rates while seamlessly preserving native-base inference and maintaining general utility.

2 PRELIMINARIES

Rotary Positional Embeddings (RoPE). In self-attention, a token's query is compared with keys from other positions in the prompt. RoPE introduces positional information by rotating paired components of these queries and keys. For d rotary dimensions and RoPE base θ, pair $j \in \{0, \ldots, d/2-1\}$ has angular frequency $\omega_j(\theta) = \theta^{-2j/d}$. Consequently, a query at position n and a key at position m interact through a relative phase $(n-m)\omega_j(\theta)$. The base controls the range of frequencies, while prompt length affects the token distances over which they operate. We use θ_0 to denote a model's native base. These phases are part of learned, content-dependent attention interactions; they provide a way to examine how the positional structure of a longer prompt may affect model behavior.

Many-Shot Jailbreaking (MSJ). MSJ places examples of harmful requests and unsafe answers before a harmful target request, encouraging the model to follow the pattern of the examples rather than refuse the target (Anil et al., 2024; Kim et al., 2025). Let x_{adv} denote the target request and $e_i = (q_i, a_i)$ the i-th query–response demonstration. An n-shot prompt can be written as $x_L = [e_1, \ldots, e_n, x_{\text{adv}}]$, where L is its length in tokens. Variants such as PANDAS differ in how demonstrations are constructed and sampled (Ma et al., 2025). As shots are added, the model encounters more adversarial examples, and the distances between parts of the prompt can also grow.

Shot count n and token length L therefore describe related but distinct aspects of the attack. This coupling of content and position motivates examining positional sensitivity alongside attack success.

3 Empirical Study

We begin by examining how the RoPE configuration affects long-context safety in fixed model checkpoints. We vary the RoPE base at evaluation time while keeping model weights constant, and jointly measure jailbreak success and model utility. This controlled intervention tests whether positional configuration provides a useful training dimension and reveals the limitations of changing the base directly at inference.

3.1 Experimental Setup

Metrics. We assess model performance along three dimensions: (1) Safety: We report attack success rate (ASR) on HarmBench (Mazeika et al., 2024) under the 32-shot PANDAS setting, measured as the proportion of evaluated queries for which the model generates a response classified as unsafe. (2) General Utility: We use MMLU-Pro (Wang et al., 2024b) to evaluate general reasoning and knowledge capabilities. (3) Long-Context Utility: We use LOONG (Wang et al., 2024a) to evaluate the model's ability to process and reason over extended inputs. Together, these metrics assess safety gains alongside changes in general and long-context utility.

Models and intervention. We evaluate Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, whose native RoPE bases are $\theta_0 = 500{,}000$ and $\theta_0 = 1{,}000{,}000$, respectively. For each checkpoint, we vary the evaluation base θ while keeping model weights fixed. We use θ for an intervened configuration and reserve θ_0 for the model's native base throughout the paper.

3.2 Results and Implications

Figure 1 shows that both safety and utility respond to the base intervention. For Llama-3.1-8B-Instruct, increasing θ from $500k$ to $1M$ reduces ASR from 77.50% to 37.50%, while LOONG decreases from 28.4 to 27.5 and MMLU-Pro changes from 0.4425 to 0.4401. At $\theta = 5M$, ASR falls to 21.75%, accompanied by LOONG and MMLU-Pro scores of 21.5 and 0.4161. For Qwen2.5-7B-Instruct, increasing the base from its native $1M$ to $5M$ lowers ASR from 48.50% to 23.75%, while LOONG decreases from 31.0 to 25.0 and MMLU-Pro from 0.5764 to 0.5570. Appendix B.8 reports the full sweep.

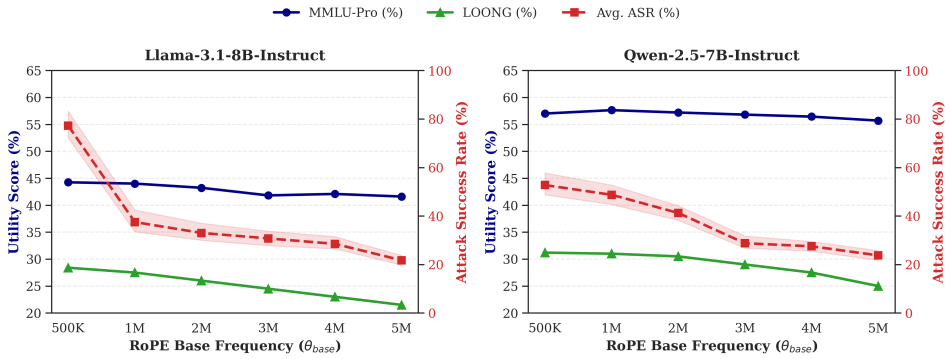

Figure 1: Effect of the RoPE base θ on 32-shot PANDAS ASR, MMLU-Pro, and LOONG. Model weights remain fixed throughout the sweep.

The decrease in ASR shows that many-shot safety is sensitive to the positional configuration. The accompanying decrease in LOONG, however, makes the base change difficult to use directly as a defense: lower attack success coincides with weaker long-context utility. Benign-context controls in Appendix B.9 further characterize how the same intervention affects ordinary in-context behavior. These observations motivate training across positional configurations and evaluating whether the resulting policy retains utility at its native base.

The sweep changes the base without adapting the model weights. Context-extension methods may combine a larger base with additional training (Liu et al., 2024b); our results characterize an intervention on existing checkpoints rather than an intrinsic limit on their ability to use a larger base.

Empirical Finding I (Safety–utility trade-off). With model weights fixed, increasing the RoPE base reduces MSJ attack success across both models, but also degrades long-context utility.

4 Understanding Positional Sensitivity

The empirical study shows that changing the positional configuration affects both many-shot safety and long-context utility. We next examine how input length and the RoPE base enter the long-range query–key interactions through which early instructions can affect later generation. This analysis connects the observed sensitivity to the two training choices in `MFA` while accounting for learned, content-dependent attention behavior.

4.1 Positional Sensitivity in RoPE Interactions

Consider a query at position n and a key at position m in one attention head, separated by $\Delta = n - m$ tokens. For rotary pair $j \in \{0, \ldots, d/2 - 1\}$, where d is the rotary dimension, RoPE assigns the angular frequency and relative phase

$$\omega_j(\theta) = \theta^{-2j/d}, \qquad \varphi_j(\Delta; \theta) = \Delta\, \omega_j(\theta). \tag{1}$$

Writing the pre-rotation queries and keys in complex form, the identity $e^{in\omega_j}\overline{e^{im\omega_j}} = e^{i\Delta\omega_j}$ gives the rotary contribution to their scaled pre-softmax interaction:

$$\begin{aligned} S(n, m; \theta) &= \frac{1}{\sqrt{d_h}} \operatorname{Re} \sum_{j=0}^{d/2-1} c_j(n, m; \theta) e^{i\varphi_j(\Delta;\theta)}, \\ c_j(n, m; \theta) &= q_n^{(j)}(\theta)\overline{k_m^{(j)}(\theta)}, \end{aligned} \tag{2}$$

where d_h is the attention-head dimension. The query–key coefficients may themselves depend on θ through preceding layers. Increasing input length exposes the model to larger token separations, while changing θ reshapes the frequencies across rotary pairs. Equation equation 2 shows why the frequency schedule alone cannot determine a learned attention score or the weight assigned to a token after softmax.

We use *phase pressure* as a compact coordinate for a length–base configuration:

$$\Phi(L, \theta) = \frac{L}{\theta}, \tag{3}$$

where L is the input token count. In a continuous parameterization of the standard RoPE spectrum, $1/\theta$ is the low-frequency endpoint; L/θ describes its phase span at distance scale L. For finite rotary dimension d, the lowest actual frequency is $\theta^{-1+2/d}$. Thus, Φ summarizes one end of the spectrum but does not describe all rotary phases or predict the learned attention weights. At fixed L, increasing θ reduces this endpoint phase span; the accompanying change in ASR is an empirical observation, not a consequence of Φ alone. Appendix B.10 gives the finite-dimensional expression.

Figure 2 shows how the plotted signed interaction varies with token distance and the RoPE base in the evaluated models. Its negative values identify it as a pre-softmax quantity. The attention mass assigned to an early safety prefix additionally depends on the logits of all competing tokens, so the figure alone does not determine that mass. Appendix B.11 defines the measurements and describes how the plotted statistic is extracted. Together with the fixed-checkpoint sweep, the figure motivates testing whether safety behavior can be maintained across positional configurations.

Insight II (Positional sensitivity). Long-context safety is sensitive to the positional configuration under which the input is processed. Context length and the RoPE base jointly determine the rotary phases encountered across token distances, although neither $\Phi = L/\theta$ nor the frequency schedule alone fully determines the model's learned attention behavior.

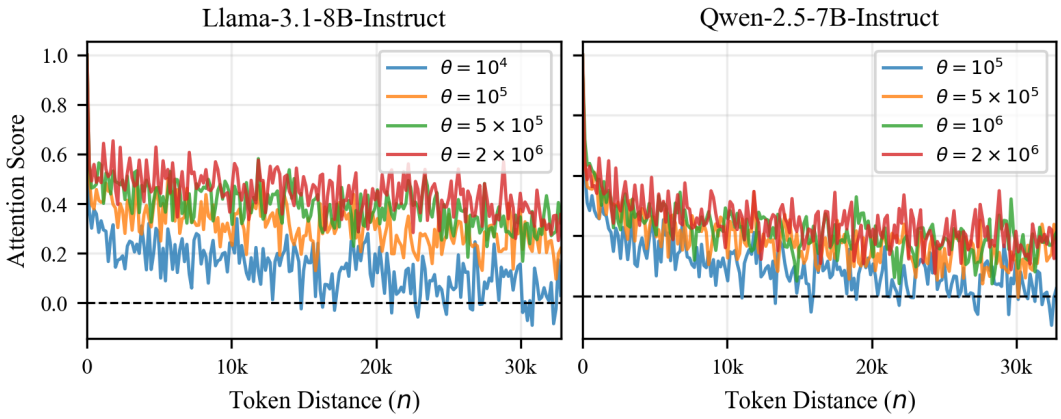

Figure 2: Signed pre-softmax RoPE interaction across token distance under different base settings. Negative values are possible; the plotted quantity is not a post-softmax attention probability.

4.2 From Positional Sensitivity to Safety Training

At inference, the model uses its native base θ_0, while longer inputs expose it to a wider range of relative positions. Covering every target length during training is expensive because rollout memory and computation grow with sequence length. Varying θ at a fixed input length perturbs the phase spectrum through which the same safety example is processed. Sampling several bases therefore exposes the policy to positional variation without requiring every rollout to use the longest target context.

Changing L and changing θ affect Eq. equation 1 differently. A longer input introduces larger token separations; a different base changes frequencies by different amounts across rotary pairs. We use base sampling as positional augmentation and retain input length as a separate training variable. Giving longer prompts greater weight in the safety update addresses the regime in which many-shot attacks are most consequential. These two design choices lead to `MFA`.

Insight III (Training implication). The observed positional sensitivity motivates safety optimization across multiple RoPE configurations. RoPE-base sampling broadens the positional conditions encountered during training, while length-aware advantage weighting gives longer adversarial inputs greater influence on the policy update.

5 Multi-Frequency Alignment (`MFA`)

Multi-Frequency Alignment (`MFA`) integrates these two design choices into GRPO (Shao et al., 2024). For a many-shot prompt x_L with L input tokens, let $\pi_W(\cdot \mid x_L; \theta)$ denote the trainable response policy with parameters W, evaluated under RoPE base θ. We initialize W from the frozen reference parameters W_{ref}. During training, the policy generates responses under randomly sampled bases and is updated with length-weighted, group-relative safety advantages.

Multi-frequency rollouts. At each rollout iteration, we freeze a snapshot of the behavior policy $\pi_{W_{\text{old}}}$. For each prompt x_L, we sample a single base

$$\theta \sim \text{Uniform}(\Theta), \qquad \Theta = [\theta_{\min}, \theta_{\max}], \tag{4}$$

and draw G independent responses $y_i \sim \pi_{W_{\text{old}}}(\cdot \mid x_L; \theta)$, $i = 1, \ldots, G$. Since responses in a group share both the prompt and the sampled base, their rewards are compared under the same positional configuration. Across prompts and iterations, base sampling exposes the policy to a range of positional configurations beyond the single one seen by fixed-base GRPO. The sampled θ is retained so that the same base is used to evaluate the current and behavior policies in the subsequent update.

Length-aware advantages. A safety judge assigns each response y_i a binary reward $r_i \in \{+1, -1\}$, where $+1$ indicates a safe response and -1 an unsafe one. Let μ_r and σ_r denote the mean and standard

deviation of the rewards in the group of y_i. If $\sigma_r > 0$, we define the length-aware advantage

$$\widehat{A}_i(L) = \gamma(L)\,\frac{r_i - \mu_r}{\sigma_r}, \tag{5}$$

where $\gamma(L) > 0$ is a weight that increases with the input length; we adopt a logarithmic schedule specified in Appendix B.2. If $\sigma_r = 0$, all advantages in the group are set to zero. We apply $\gamma(L)$ after group normalization, since scaling all rewards in a group beforehand would be canceled by the normalization. Consequently, among groups that carry a relative safety signal, longer inputs contribute more strongly to the policy update, while the ordering of responses within each group remains unchanged. Appendix B.3 provides further normalization details.

Policy optimization. The response-level advantage in Eq. equation 5 is assigned to every token of y_i; for brevity, we write $\widehat{A}_i = \widehat{A}_i(L)$. Let $T_i = |y_i|$ denote the response length and $h_{i,t} = (x_L, y_{i,<t})$ the prefix preceding token $y_{i,t}$. We minimize the clipped GRPO objective

$$\mathcal{L}_{\mathrm{MFA}}(W) = \mathbb{E}_{x_L,\theta,y_{1:G}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\Big(-\min\big\{\rho_{i,t}\widehat{A}_i,\ \mathrm{clip}_\epsilon(\rho_{i,t})\widehat{A}_i\big\} + \alpha\,\widehat{D}_{\mathrm{KL},i,t}\Big)\right], \tag{6}$$

where the expectation is taken over the prompt, base, and response sampling described above, $\mathrm{clip}_\epsilon(u) = \min\{\max\{u, 1-\epsilon\}, 1+\epsilon\}$, and α controls the strength of reference-policy regularization. The importance ratio and the reference divergence are

$$\begin{aligned}\rho_{i,t} &= \frac{\pi_W(y_{i,t} \mid h_{i,t};\theta)}{\pi_{W_{\mathrm{old}}}(y_{i,t} \mid h_{i,t};\theta)},\\ D_{\mathrm{KL},i,t} &= D_{\mathrm{KL}}\big(\pi_W(\cdot \mid h_{i,t};\theta)\,\big\|\,\pi_{W_{\mathrm{ref}}}(\cdot \mid h_{i,t};\theta_0)\big),\end{aligned} \tag{7}$$

and $\widehat{D}_{\mathrm{KL},i,t}$ is a per-token sampled estimator of $D_{\mathrm{KL},i,t}$. Both policies in the importance ratio use the sampled base θ, whereas the frozen reference policy uses the model's native base θ_0. The KL term thus anchors the policy under sampled positional configurations to the reference behavior at the native base. Moreover, since $\gamma(L)$ scales only the clipped policy term while α is fixed, longer inputs receive greater weight in the update relative to reference regularization. Appendix B.4 specifies the KL estimator and the remaining optimization details.

Inference. At inference time, we use the trained policy at the native base θ_0, i.e., $\pi_W(\cdot \mid x_L;\theta_0)$. Since base sampling and length weighting are applied only during training, `MFA` requires neither a frequency search nor any additional model calls at deployment. Algorithm 1 in Appendix C summarizes the full training procedure.

6 Experiment

We evaluate `MFA` under MSJ attacks and measure its general and long-context utility. A matched fixed-base GRPO control separates the benefit of ordinary safety tuning from the additional effect of RoPE-base sampling and length-aware advantages.

6.1 Experimental Setup

Training. We optimize on 256 target harmful queries (drawn from ManyHarm (Kim et al., 2025)) for three epochs. Each GRPO rollout generates $G = 16$ responses per prompt, with binary safety rewards assigned by Llama-Guard-3-8B. For `MFA`, a group shares one base sampled uniformly from $[\theta_{\min}, 5\theta_{\min}]$, and the length factor $\gamma(L)$ is applied after group advantage normalization. Policy updates use clipping threshold $\epsilon = 0.2$ and reference KL coefficient $\alpha = 1.0$; the frozen reference uses each model's native base θ_0. At inference, we use θ_0 except in the positional extrapolation comparison. Optimizer, sampling, and length-schedule settings are given in Appendix B.2.

Models. We evaluate Llama-3.1-8B-Instruct ($\theta_0 = 500k$) (Dubey et al., 2024), Qwen2.5-7B-Instruct ($\theta_0 = 1M$), and Qwen2.5-72B-Instruct ($\theta_0 = 1M$) (Yang et al., 2024a). The two Qwen models allow us to assess the role of model scale within one family and native-base setting.

Baselines. We compare MFA with inference-time defenses SmoothLLM (Robey et al., 2023) and Self-Reminder (Xie et al., 2023), and training-time defenses MTSA (Guo et al., 2025) and MUSE (Yan et al., 2025). We also train a matched fixed-base GRPO control on the same prompts and rewards; its results and the two-component ablation appear in Table 3.

Evaluation Metrics and Benchmarks. Following prior work (Chen et al., 2025a), we use Llama-Guard-3-8B (Dubey et al., 2024) as the main safety judge and report ASR. On HarmBench (Mazeika et al., 2024), we evaluate standard MSJ (Anil et al., 2024), which prepends many demonstrations before a harmful query, and PANDAS (Ma et al., 2025), an MSJ variant that combines positive affirmations, negative demonstrations, and adaptive sampling. We report 0, 32, 64, 128, and 256 shots, and we report rule-based refusal matching (ASR-R) on AdvBench in Appendix B.12.

We assess general utility with GSM8K (Cobbe et al., 2021), MMLU-Pro (Wang et al., 2024b), GPQA (Rein et al., 2024), and MT-Bench (Zheng et al., 2023), and long-context utility with LongBench V2 (Bai et al., 2025) and LOONG (Wang et al., 2024a). Benchmark and judge definitions are given in Appendix B.1. The pilot in Section 3 uses one RoPE-base sweep per fixed checkpoint. The ASRs in Table 1 average repeated evaluations with resampled demonstrations because results can vary with the choice of shots in long prompts.

6.2 Main Results

Resilience Against Many-Shot Attacks. Table 1 reports ASR across shot counts and attack types. MFA attains the lowest ASR in every nonzero-shot setting shown. On Llama-3.1-8B-Instruct under MSJ, its ASR ranges from 6.5% to 8.8% across 32–256 shots, compared with 63.7%–74.6% for the base model and 22.0%–26.0% for MUSE. The improvement also holds for PANDAS: on Qwen2.5-72B-Instruct at 256 shots, MFA achieves 7.0% ASR versus MUSE's 19.0%, a reduction of 12.0 percentage points.

The matched control in Table 3 clarifies the source of the gain. At 256-shot MSJ on Llama-3.1-8B-Instruct, fixed-base GRPO already lowers ASR from 65.5% to 28.0%; adding both MFA components reduces it further to 8.8%. The corresponding additional reduction is 19.2 percentage points. These results decouple the intrinsic benefit of safety training from the structural improvements introduced by RoPE-base sampling and length-aware advantages.

Table 1: Safety performance under MSJ and PANDAS across various shot counts. Lower values indicate stronger safety resilience. Numbers in green parentheses indicate the absolute ASR reduction compared to the Base model.

Model	Method	MSJ					PANDAS				
		0	32	64	128	256	0	32	64	128	256
Llama-3.1-8B-In	Base	0.0	63.7	74.6	70.2	65.5	0.0	77.5	84.7	82.2	77.5
	SmoothLLM	0.0	54.0 (↓9.7)	63.5 (↓11.1)	59.5 (↓10.7)	55.0 (↓10.5)	0.0	66.0 (↓11.5)	72.0 (↓12.7)	69.5 (↓12.7)	66.0 (↓11.5)
	Self-Reminder	0.0	56.5 (↓7.2)	66.5 (↓8.1)	62.5 (↓7.7)	58.5 (↓7.0)	0.0	69.0 (↓8.5)	76.0 (↓8.7)	73.5 (↓8.7)	69.0 (↓8.5)
	MTSA	0.0	25.5 (↓38.2)	29.5 (↓45.1)	28.0 (↓42.2)	26.0 (↓39.5)	0.0	31.0 (↓46.5)	34.0 (↓50.7)	32.5 (↓49.7)	31.0 (↓46.5)
	MUSE	0.0	22.0 (↓41.7)	26.0 (↓48.6)	24.5 (↓45.7)	23.0 (↓42.5)	0.0	28.0 (↓49.5)	31.0 (↓53.7)	29.5 (↓52.7)	28.0 (↓49.5)
	MFA	0.0	**6.5** (↓57.2)	**8.0** (↓66.6)	**8.5** (↓61.7)	**8.8** (↓56.7)	0.0	**7.5** (↓70.0)	**9.5** (↓75.2)	**10.0** (↓72.2)	**10.2** (↓67.3)
Qwen2.5-7B-In	Base	11.2	39.5	41.0	41.4	36.0	11.2	48.5	50.0	51.5	49.8
	SmoothLLM	9.5 (↓1.7)	33.5 (↓6.0)	34.8 (↓6.2)	33.5 (↓7.9)	30.5 (↓5.5)	9.5 (↓1.7)	41.0 (↓7.5)	42.5 (↓7.5)	43.8 (↓7.7)	42.3 (↓7.5)
	Self-Reminder	10.0 (↓1.2)	35.5 (↓4.0)	36.8 (↓4.2)	35.5 (↓5.9)	32.0 (↓4.0)	10.2 (↓1.0)	43.5 (↓5.0)	45.0 (↓5.0)	46.5 (↓5.0)	44.8 (↓5.0)
	MTSA	4.0 (↓7.2)	14.5 (↓25.0)	15.5 (↓25.5)	15.0 (↓26.4)	13.5 (↓22.5)	4.5 (↓6.7)	18.0 (↓30.5)	19.0 (↓31.0)	19.8 (↓31.7)	19.0 (↓30.8)
	MUSE	3.5 (↓7.7)	13.0 (↓26.5)	14.0 (↓27.0)	13.5 (↓27.9)	12.5 (↓23.5)	4.0 (↓7.2)	16.5 (↓32.0)	17.5 (↓32.5)	18.5 (↓33.0)	17.8 (↓32.0)
	MFA	**2.0** (↓9.2)	**6.0** (↓33.5)	**7.0** (↓34.0)	**7.2** (↓34.2)	**7.5** (↓28.5)	**2.5** (↓8.7)	**7.5** (↓41.0)	**9.0** (↓41.0)	**9.5** (↓42.0)	**9.8** (↓40.0)
Qwen2.5-72B-In	Base	18.3	42.5	45.0	44.3	43.8	18.3	51.5	52.8	52.0	51.3
	SmoothLLM	15.5 (↓2.8)	36.0 (↓6.5)	38.5 (↓6.5)	37.5 (↓6.8)	37.0 (↓6.8)	15.5 (↓2.8)	44.0 (↓7.5)	45.0 (↓7.8)	44.2 (↓7.8)	43.6 (↓7.7)
	Self-Reminder	16.5 (↓1.8)	38.0 (↓4.5)	40.5 (↓4.5)	39.5 (↓4.8)	39.0 (↓4.8)	16.5 (↓1.8)	46.0 (↓5.5)	47.5 (↓5.3)	46.8 (↓5.2)	46.0 (↓5.3)
	MTSA	6.5 (↓11.8)	16.5 (↓26.0)	17.5 (↓27.5)	17.0 (↓27.3)	16.5 (↓27.3)	6.8 (↓11.5)	20.5 (↓31.0)	21.0 (↓31.8)	20.8 (↓31.2)	20.5 (↓30.8)
	MUSE	5.5 (↓12.8)	14.5 (↓28.0)	15.5 (↓29.5)	15.0 (↓29.3)	14.8 (↓29.0)	6.0 (↓12.3)	18.5 (↓33.0)	19.5 (↓33.3)	19.2 (↓32.8)	19.0 (↓32.3)
	MFA	**1.5** (↓16.8)	**4.0** (↓38.5)	**5.0** (↓40.0)	**5.2** (↓39.1)	**5.5** (↓38.3)	**2.0** (↓16.3)	**5.5** (↓46.0)	**6.5** (↓46.3)	**6.8** (↓45.2)	**7.0** (↓44.3)

Maintaining General and Long-Context Capabilities. Table 2 shows that MFA avoids the common "alignment tax" associated with defensive tuning. Across all three model scales, general reasoning and knowledge scores (GSM8K, MMLU-Pro, GPQA, and MT-Bench) shift by a tightly bounded margin of $+0.0$ to $+0.2$ points relative to the unmodified base models. Furthermore, extended-context retrieval and multi-document reasoning remain fully intact. Rather than degrading under the modified RoPE sampling, Qwen2.5-72B-Instruct slightly improves from 35.9 to 36.8 on LOONG and 41.0 to 42.0 on LongBench-v2. Contrasting these stable utility metrics with the results in Table 1 confirms that MFA can reduce adversarial success rates by up to 60 percentage points without penalizing the model's core mathematical reasoning or sequence-processing capabilities.

Table 2: General capabilities and long-context reasoning performance across diverse benchmarks. Numbers in dark red parentheses indicate the performance shift (↑ / ↓) relative to the Base model.

Model	Method	General Capabilities				Long-Context	
		GSM8K	MMLU-PRO	GPQA	MT-Bench	LOONG	LB-v2
Llama-3.1-8B-In	Base	78.1	44.3	29.5	7.14	28.4	27.0
	SmoothLLM	73.4 (↓4.7)	40.1 (↓4.2)	25.3 (↓4.2)	6.62 (↓0.52)	24.6 (↓3.8)	23.4 (↓3.6)
	Self-Reminder	75.0 (↓3.1)	42.7 (↓1.6)	26.6 (↓2.9)	6.78 (↓0.36)	26.0 (↓2.4)	25.4 (↓1.6)
	MTSA	75.6 (↓2.5)	41.8 (↓2.5)	27.1 (↓2.4)	6.84 (↓0.30)	26.1 (↓2.3)	25.1 (↓1.9)
	MUSE	76.9 (↓1.2)	43.1 (↓1.2)	28.3 (↓1.2)	6.98 (↓0.16)	27.2 (↓1.2)	26.2 (↓0.8)
	MFA	**78.2** (↑0.1)	**44.3** (0.0)	**29.5** (0.0)	**7.15** (↑0.01)	**28.9** (↑0.5)	**27.8** (↑0.8)
Qwen2.5-7B-In	Base	91.6	56.3	36.2	7.96	31.2	30.0
	SmoothLLM	87.1 (↓4.5)	52.4 (↓3.9)	32.1 (↓4.1)	7.35 (↓0.61)	27.3 (↓3.9)	26.1 (↓3.9)
	Self-Reminder	87.5 (↓4.1)	52.5 (↓3.8)	33.0 (↓3.2)	7.48 (↓0.48)	27.8 (↓3.4)	26.3 (↓3.7)
	MTSA	89.2 (↓2.4)	54.1 (↓2.2)	34.0 (↓2.2)	7.58 (↓0.38)	29.1 (↓2.1)	28.0 (↓2.0)
	MUSE	90.6 (↓1.0)	55.4 (↓0.9)	35.1 (↓1.1)	7.78 (↓0.18)	30.2 (↓1.0)	29.1 (↓0.9)
	MFA	**91.7** (↑0.1)	**56.4** (↑0.1)	**36.3** (↑0.1)	**7.98** (↑0.02)	**31.9** (↑0.7)	**30.8** (↑0.8)
Qwen2.5-72B-In	Base	95.8	71.1	49.0	9.10	35.9	41.0
	SmoothLLM	92.2 (↓3.6)	67.5 (↓3.6)	45.2 (↓3.8)	8.55 (↓0.55)	32.6 (↓3.3)	37.6 (↓3.4)
	Self-Reminder	93.0 (↓2.8)	68.3 (↓2.8)	46.5 (↓2.5)	8.70 (↓0.40)	34.0 (↓1.9)	38.6 (↓2.4)
	MTSA	93.8 (↓2.0)	69.1 (↓2.0)	47.0 (↓2.0)	8.82 (↓0.28)	34.1 (↓1.8)	39.2 (↓1.8)
	MUSE	94.9 (↓0.9)	70.2 (↓0.9)	48.1 (↓0.9)	9.00 (↓0.10)	35.1 (↓0.8)	40.3 (↓0.7)
	MFA	**95.9** (↑0.1)	**71.2** (↑0.1)	**49.1** (↑0.1)	**9.12** (↑0.02)	**36.8** (↑0.9)	**42.0** (↑1.0)

Scalability and Model Capacity. Evaluating defense mechanisms across different model scales is critical, as larger models possess more complex attention dynamics and are notoriously harder to regularize without degrading their advanced reasoning capabilities. To verify that our intervention does not merely exploit the limited capacity of 7B/8B models, we test MFA on Qwen2.5-72B-Instruct. We observe that scale alone offers no inherent protection: despite its 72B parameters, the base model remains highly susceptible to density-based attacks, exhibiting 43.8% (MSJ) and 51.3% (PANDAS) ASR at 256 shots. MFA effectively neutralizes this adversarial surface, suppressing ASR to 5.5% and 7.0% respectively. Crucially, it achieves this without penalizing the model's robust utility baselines (Table 2). This confirms that our mechanism addresses a fundamental architectural vulnerability and scales seamlessly to high-capacity models without imposing a prohibitive alignment tax.

6.3 Ablations

To isolate the structural contributions of MFA from the baseline benefits of ordinary safety training, we systematically ablate its core components. We then analyze the model's sensitivity to the reference KL coefficient, assess its orthogonality with positional extrapolation techniques, and evaluate its resilience against diverse adaptive attacks.

Table 3: Matched component ablation on Llama-3.1-8B-Instruct. Trained variants are reported as mean $\pm$ standard deviation over three runs; Base denotes the untrained checkpoint. ASR is measured under MSJ. Lower ASR and higher utility are better.

Method	Sample θ	$\gamma(L)$	ASR@64	ASR@256	LOONG	MMLU-Pro
Base	–	–	74.6	65.5	28.4	44.3
Fixed-base GRPO	No	No	26.0 ± 2.4	28.0 ± 2.6	27.6 ± 0.4	44.0 ± 0.3
Base sampling only	Yes	No	14.0 ± 2.0	17.5 ± 2.2	28.5 ± 0.4	44.2 ± 0.2
Length weighting only	No	Yes	16.0 ± 2.1	15.0 ± 1.9	27.9 ± 0.4	44.1 ± 0.3
MFA	Yes	Yes	$\mathbf{8.0 \pm 1.5}$	$\mathbf{8.8 \pm 1.7}$	$\mathbf{28.9 \pm 0.3}$	$\mathbf{44.3 \pm 0.2}$

Contributions of Base Sampling and Length Weighting. Table 3 compares MFA with fixed-base GRPO and two single-component variants under the same training budget. At 256 shots, fixed-base GRPO reduces ASR from 65.5% to 28.0%, demonstrating a substantial benefit from ordinary safety training. Adding RoPE-base sampling alone lowers ASR to 17.5%, while length weighting alone reaches 15.0%. Combining both components further reduces ASR to 8.8%: an additional reduction of 19.2 percentage points over fixed-base GRPO and 6.2 points over the stronger single-component variant. Both components also improve upon the control at 64 shots, where the full method achieves the lowest mean ASR of 8.0%. Meanwhile, its LOONG and MMLU-Pro scores remain close to those

of the original checkpoint. These results support the complementary benefits of positional variation and length-aware updates beyond fixed-base safety training.

Sensitivity to the KL Coefficient α. We vary α in Eq. 6, which controls the strength of regularization against the frozen reference policy. During training, the current policy uses the sampled RoPE base, whereas the reference uses the native base θ_0. Increasing α therefore places greater weight on matching the reference distribution relative to the length-weighted safety update. Figure 3 shows that small coefficients leave ASR high, even though LOONG remains close to its base-model score. Around $\alpha = 1.0$, ASR falls below 10% while LOONG remains near this baseline. For $\alpha > 1.2$, ASR stays low but LOONG drops sharply, indicating a less favorable utility outcome. We therefore use $\alpha = 1.0$ as the default to retain the observed safety gains without the pronounced utility loss at larger coefficients. The LOONG decline alone does not establish over-refusal.

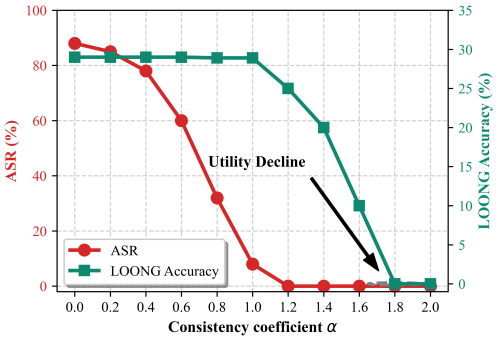

Figure 3: Effect of the reference KL coefficient α on ASR and LOONG.

Synergy with Positional Extrapolation. While RoPE is the ubiquitous standard in modern LLMs, many advanced models employ secondary extrapolation techniques (e.g., Position Interpolation (Chen et al., 2023), YaRN (Peng et al., 2023), and LongRoPE (Ding et al., 2024)) to further extend their context windows post-training. Since MFA manipulates the RoPE frequency base during safety alignment, a critical question is whether it conflicts with these context-extension mechanisms. Table 4 reports an exploratory comparison under 128-shot MSJ. Standard extrapolation methods without defense remain highly vulnerable, yielding ASRs between 45.1% and 56.1%. In contrast, seamlessly integrating MFA with LongRoPE drives ASR down to 7.4% while maintaining a competitive LongBench-v2 score of 27.8. This synergy demonstrates that MFA is structurally compatible with state-of-the-art extrapolation techniques, providing safety benefits without breaking extended-context functionality.

Table 4: Safety and utility performance of extending Llama-3.1-8B-Instruct via different secondary positional interpolation methods.

Extension Method	ASR (↓)	LongBench-v2 (↑)
Llama-3.1-8B-Instruct	70.2	27.0
+ PI (Chen et al., 2023)	53.5	25.3
+ YaRN (Peng et al., 2023)	56.1	26.4
+ LongRoPE Ding et al. (2024)	45.1	26.8
+ MFA	11.2	27.8
+ MFA + LongRoPE	**7.4**	**27.8**

Resilience to Adaptive Attacks. Beyond many-shot scenarios, MFA generalizes effectively to standard adaptive attacks. Against GCG, AutoDAN, and DRA, it consistently matches or outperforms strong baselines like MUSE (e.g., achieving 0% ASR on AutoDAN; Appendix B.5). However, we candidly establish the boundary of this resilience through extreme length-stress tests (Appendix B.6). Under a combined 512/1024-shot out-of-distribution evaluation, MFA's ASR rebounds to 28% (compared to 31% for MUSE). This indicates that while MFA provides state-of-the-art protection within its training distribution, uniform robustness at extreme, unseen context lengths remains an open challenge for future alignment research.

7 Conclusion

We investigate whether positional configurations can elucidate and mitigate the vulnerability of large language models to MSJ. While increasing the RoPE base at inference time substantially reduces attack success, it forces a strict trade-off by concurrently degrading long-context utility. We therefore analyze how input length and the frequency base jointly shape rotary phases, using the phase-pressure proxy $\Phi = L/\theta$ to motivate positional variation as an augmented training dimension. Building on this insight, we introduce MFA, which dynamically samples RoPE bases and applies length-weighted advantages during preference optimization. By anchoring these updates to a frozen native-base reference, MFA ensures robust safety across diverse positional conditions without altering standard inference. Empirically, this approach drastically reduces attack success rates compared to matched fixed-base GRPO controls while fully preserving general and long-context utility.

AI USE STATEMENT

Generative AI tools were used for basic copyediting and notation checking. All AI-assisted material was reviewed by the authors, who made all final research and writing decisions and take responsibility for the complete content of this work.

ETHICS STATEMENT

This work studies many-shot jailbreaking to improve defenses for long-context LLMs. The attacks and generated harmful responses are dual-use. We conduct evaluations on established safety benchmarks and do not target deployed systems. An additional safety-judge check includes three annotators who assessed a subset of 200 model responses; these annotations are reported as a limited evaluation rather than a complete human assessment. We do not collect private personal data. Released artifacts will follow the applicable dataset and model licenses and will omit credentials, private logs, and harmful outputs unnecessary for reproducing the reported evaluations.

REPRODUCIBILITY STATEMENT

Sections 3–6 specify the fixed-checkpoint RoPE intervention, the `MFA` objective, the evaluated attacks and models, and the reported metrics. Appendix B.1 describes the benchmarks and shot-generation and safety-refusal templates; Appendix B.2 reports rollout and optimization settings, and Appendix C gives the training procedure. The appendix also reports the full pilot sweep, additional judge checks, and complementary safety results. Implementation and evaluation materials are linked through the anonymous repository in the abstract.

REFERENCES

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. Secalign: Defending against prompt injection with preference optimization. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2833–2847, 2025a.

Yingfa Chen, Yutong Wu, Chenyang Song, Zhen Leng Thai, Xingyu Shen, Xu Han, Zhiyuan Liu, and Maosong Sun. Cost-optimal grouped-query attention for long-context modeling. *arXiv preprint arXiv:2503.09579*, 2025b.

Shuai Cheng, Shu Meng, Haitao Xu, Haoran Zhang, Shuai Hao, Chuan Yue, Wenrui Ma, Meng Han, Fan Zhang, and Zhao Li. Effective {PII} extraction from {LLMs} through augmented {Few-Shot} learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 8155–8173, 2025.

Yixin Cheng, Markos Georgopoulos, Volkan Cevher, and Grigorios G Chrysos. Leveraging the context through multi-round interactions for jailbreaking attacks. *arXiv preprint arXiv:2402.09177*, 2024.

Yuan Cheng, Fengzhuo Zhang, Yunlong Hou, Cunxiao Du, Chao Du, Tianyu Pang, Aixin Sun, and Zhuoran Yang. Demystifying the slash pattern in attention: The role of rope. *arXiv preprint arXiv:2601.08297*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Weiyang Guo, Jing Li, Wenya Wang, Yu Li, Daojing He, Jun Yu, and Min Zhang. Mtsa: Multi-turn safety alignment for llms through multi-round red-teaming. *arXiv preprint arXiv:2505.17147*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. *arXiv preprint arXiv:2502.01563*, 2025.

Sangyeop Kim, Yohan Lee, Yongwoo Song, and Kimin Lee. What really matters in many-shot attacks? an empirical study of long-context vulnerabilities in llms. *arXiv preprint arXiv:2505.19773*, 2025.

Nathalie Maria Kirch, Constantin Niko Weisser, Severin Field, Helen Yannakoudakis, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 480–520, 2025.

Pengcheng Li, Jie Zhang, Tianwei Zhang, Han Qiu, Zhang Kejun, Weiming Zhang, Neng H. Yu, and Wenbo Zhou. State-dependent safety failures in multi-turn language model interaction. 2026a. URL https://api.semanticscholar.org/CorpusID:286579446.

Songze Li, Ruishi He, Xiaojun Jia, Jun Wang, and Zhihui Fu. Knowledge-driven multi-turn jailbreaking on large language models. *ArXiv*, abs/2601.05445, 2026b. URL https://api.semanticscholar.org/CorpusID:284596923.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4711–4728, 2024a.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. In *International Conference on Learning Representations*, volume 2024, pp. 48160–48185, 2024b.

Avery Ma, Yangchen Pan, and Amir-massoud Farahmand. Pandas: Improving many-shot jailbreaking via positive affirmation, negative demonstration, and adaptive sampling. *arXiv preprint arXiv:2502.01925*, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 35181–35224, 2024.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

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

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. *arXiv preprint arXiv:2504.13203*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Derail yourself: Multi-turn llm jailbreak attack through self-discovered clues. 2024.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2421–2440, 2025.

Rishi Rajesh Shah, Chen Henry Wu, Shashwat Saxena, Ziqian Zhong, Alexander Robey, and Aditi Raghunathan. Jailbreaking in the haystack. *arXiv preprint arXiv:2511.04707*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5627–5646, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024b.

Mengjie Wu, Yihao Huang, Zhenjun Lin, Kangjie Chen, Yuhan Huang, Run Wang, Lina Wang, et al. Analogy-based multi-turn jailbreak against large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5 (12):1486–1496, 2023.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. Stress-testing long-context language models with lifelong icl and task haystack. *Advances in Neural Information Processing Systems*, 37:15801–15840, 2024.

Siyu Yan, Long Zeng, Xuecheng Wu, Chengcheng Han, Kongcheng Zhang, Chong Peng, Xuezhi Cao, Xunliang Cai, and Chenjuan Guo. Muse: Mcts-driven red teaming framework for enhanced multi-turn dialogue safety in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21293–21314, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2024a.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *arXiv preprint arXiv:2405.05610*, 2024b.

Zonghao Ying, Deyue Zhang, Zonglei Jing, Yisong Xiao, Quanchen Zou, Aishan Liu, Siyuan Liang, Xiangzheng Zhang, Xianglong Liu, and Dacheng Tao. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models. *arXiv preprint arXiv:2502.11054*, 2025.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, et al. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37: 60755–60775, 2024.

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9840–9855, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Weixiong Zheng, Peijian Zeng, Yiwei Li, Hongyan Wu, Nankai Lin, Junhao Chen, Aimin Yang, and Yongmei Zhou. Jailbreaking? one step is enough! In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11623–11642, 2025.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Improved few-shot jailbreaking can circumvent aligned language models and their defenses. *Advances in Neural Information Processing Systems*, 37:32856–32887, 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

LIMITATIONS

Our evaluation focuses on open-weight models where RoPE configurations can be dynamically optimized during training, providing a foundational framework for structural safety interventions. While our primary evaluations utilize Llama-Guard-3-8B for both training rewards and HarmBench ASR measurements, we proactively mitigate potential judge-specific correlations by incorporating independent cross-evaluations and complementary refusal-matching metrics (e.g., AdvBench ASR-R). Furthermore, while our pilot studies robustly associate RoPE-base interventions with joint shifts in attack success and long-context utility, we conservatively frame this as a macroscopic positional sensitivity rather than asserting a monolithic causal mechanism. To ensure a rigorous evaluation boundary, comprehensive Pilot and Benign Controls, alongside Defense-Aware ASR stress tests and evaluations such as the 256-Shot MSJ ASR, are extensively detailed in Appendices B.9 and B.6, which properly contextualize the robustness and generalizability of our approach.

Appendix roadmap. Appendix A reviews related work. Appendix B.1 provides comprehensive experimental details, including benchmark definitions (B.1) and training configurations (B.2–B.4). Crucially, it houses our extended defensive analyses and diagnostics: resilience to adaptive attacks (B.5), cross-judge and extreme-length stress tests (B.6), fixed-checkpoint pilot data (B.8), benign-context controls (B.9), mechanistic attention diagnostics (B.10–B.11), and a strict scope of causality (B.13). Appendix C outlines the algorithmic procedure, and Appendix B.14 offers a token-budget perspective on many-shot safety.

A RELATED WORK

Evolution of Long-Context Reasoning. LLMs have successfully expanded their context windows to million-token scales, primarily by employing positional encoding techniques such as Rotary Positional Embeddings (RoPE) (Su et al., 2024) and its various scaling derivatives (Xiong et al., 2024; Cheng et al., 2026). The performance of In-Context Learning (ICL) (Dong et al., 2024) scales predictably with the number of provided demonstrations, enabling models (Jaech et al., 2024; Yang et al., 2025) to navigate vast sequences for complex task execution. However, as sequence length L increases, the interaction between positional embeddings and the attention mechanism introduces significant architectural stress. Theoretical and empirical studies (Barbero et al., 2024; Zhang et al., 2024; Jin et al., 2025) indicate that while semantic capacity is enhanced, the efficiency of information retrieval and attention stability becomes a dominant bottleneck in extended context regimes.

Emerging Safety Risks in the Many-Shot Regime. The advancement of long-context capabilities has simultaneously created a new adversarial surface (Ren et al., 2024; Kirch et al., 2025; Russinovich et al., 2025; Li et al., 2026a; Rahman et al., 2025; Li et al., 2026b; Wu et al., 2025). Beyond traditional topic-based (Chao et al., 2025) or few-shot jailbreak (Zheng et al., 2024; Cheng et al., 2025; Zheng et al., 2025) attempts, Many-Shot Jailbreaking (MSJ) (Anil et al., 2024) reveals that attack effectiveness follows power-law patterns in extended context settings. By stacking hundreds or thousands of harmful query-answer pairs (Yang et al., 2024b; Ying et al., 2025), adversaries can systematically override model guardrails. Current defense mechanisms, while effective against standard interactions, often fail to provide invariant protection as the sequence length grows. This shift in the adversarial landscape indicates that safety is no longer just a matter of semantic intent but is increasingly tied to the model's behavior under extreme demonstration density.

Structural Vulnerabilities and Mechanistic Constraints. Prior work studies the effects of context length and attention behavior on model robustness (Cheng et al., 2024; Xu et al., 2024; Zhang et al., 2025; Chen et al., 2025b; Kim et al., 2025). These observations motivate a closer look at positional configuration, but they do not identify a universal cause of many-shot failures. Our fixed-checkpoint sweep and query–key diagnostic describe how safety outcomes and positional interactions vary in the evaluated models. The training objective in Section 5 then tests whether exposure to multiple RoPE bases improves robustness while retaining utility.

B EXPERIMENTAL DETAILS

B.1 DETAILS OF BENCHMARKS AND DATASETS

We describe the utility and safety benchmarks and the examples used to construct training prompts. The main MSJ/PANDAS results use HarmBench, whereas the positional-extrapolation comparison in Table 4 draws examples separately from ManyHarm.

B.1.1 UTILITY BENCHMARKS

To ensure MFA does not compromise the model's general capabilities while enhancing safety, we assess performance across reasoning, knowledge, and long-context understanding.

GSM8K (Cobbe et al., 2021): A grade-school mathematics reasoning dataset assessing step-by-step problem-solving abilities in arithmetic operations. The evaluation follows a chain-of-thought (CoT) reasoning protocol.

MMLU-Pro (Wang et al., 2024b): A multiple-choice benchmark extending MMLU (Hendrycks et al., 2020). We evaluate the selected subject areas used in our experiments, including physics, chemistry, computer science, engineering, biology, and economics; these scores need not represent the full benchmark.

GPQA (Rein et al., 2024): A rigorous diagnostic dataset evaluating expert-level comprehension of scientific concepts through difficult multiple-choice questions. We conduct evaluations under strict zero-shot conditions.

MT-Bench (Zheng et al., 2023): A multi-turn dialogue evaluation framework employing GPT-4o-0806 as the judge model to quantify conversational consistency across diverse dialogue paths and temperature settings.

LongBench V2 (Bai et al., 2025): We report its score as one measure of long-context utility, alongside LOONG.

LOONG (Wang et al., 2024a): A benchmark tailored for real-world multi-document question answering. It evaluates long-context processing through four specialized tasks: Spotlight Locating, Comparison, Clustering, and Chain of Reasoning. This dataset facilitates a granular assessment of model stability and reasoning integrity across extensive, multi-source sequences.

B.1.2 SAFETY BENCHMARKS

We evaluate MSJ and PANDAS on **HarmBench** (Mazeika et al., 2024), and use **AdvBench** (Zou et al., 2023) for a complementary refusal-matching analysis. For generated harmful query–answer shots, we screen the content with Llama-Guard-3-8B and apply a common question and bullet-point answer format. These constraints describe the harmful-shot pool, not every element of a PANDAS prompt. They reduce one source of formatting variation but do not by themselves establish that context length causes a change in safety behavior.

For HarmBench, we use Llama-Guard-3-8B (Dubey et al., 2024) as the automated safety judge and report the corresponding attack success rate (ASR). For AdvBench, we additionally report ASR-R, a rule-based attack success rate computed by checking whether the model response contains any refusal phrase from a predefined refusal list. A response is counted as a successful attack under ASR-R if it does not match any refusal phrase. Since ASR-R relies on surface-form matching, it provides a complementary view of refusal behavior but may overestimate jailbreak success when models produce implicit or non-template refusals.

B.1.3 ALIGNMENT DATASET

To implement the MFA framework, we construct a specialized training set $\mathcal{D}_{msj}$ focused on long context safety. We derive this collection from the ManyHarm framework, which includes 2,400 malicious demonstrations spanning 12 distinct categories based on the safety policies of Anthropic. These topics cover diverse risks such as child safety, misinformation, and infrastructure attacks. Following established many shot methodologies, we generate concise harmful shots to maximize the

density of adversarial examples within the limited attention budget. The malicious demonstrations are generated using the following prompt structure:

Shot Generation Template

malicious question Please provide a concise answer in bullet point format, keeping the response within a 200 token limit.

To counter these adversarial inputs, we generate corresponding defensive responses to serve as training targets. For each malicious query x_L in the many-shot sequence, we produce a refusal response using a standardized safety template.

Safety Refusal Generation Template

I cannot fulfill this request. My safety guidelines prohibit the generation of content related to *topic*. I am programmed to be a helpful and harmless assistant.

These paired samples provide safety-oriented training targets. The template specifies the desired response format; it does not, on its own, establish how the trained model represents safety internally.

B.2 Training Configuration

We optimize on 256 safety-alignment prompts for three epochs. Each rollout samples $G = 16$ responses per prompt at temperature 1.0 and top-p 0.95, with an 8192-token maximum sequence length and at most 1024 generated tokens. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.1, and peak learning rate 2×10^{-5}. We use GRPO clipping threshold $\epsilon = 0.2$ and reference KL coefficient $\alpha = 1.0$. One RoPE base is sampled for each response group from $[\theta_{\min}, 5\theta_{\min}]$; all responses within that group share the sampled base. The frozen reference remains at the model's native base θ_0.

For input length L, the length schedule used in the policy update is $\gamma(L) = 1 + \log(L/L_0)$ with $L_0 = 1024$. This expression increases with L and is positive only when $L > L_0/e$. The assumption $\gamma(L) > 0$ in Eq. equation 5 applies on that domain.

B.3 Length-Weighting and Reference Details

For a response group with reward mean μ_r and standard deviation $\sigma_r > 0$, multiplying every reward by the same positive factor $\gamma(L)$ *before* standardization has no effect:

$$\frac{\gamma(L)r_i - \gamma(L)\mu_r}{\gamma(L)\sigma_r} = \frac{r_i - \mu_r}{\sigma_r}. \tag{8}$$

We therefore form $\widehat{A}_i(L) = \gamma(L)(r_i - \mu_r)/\sigma_r$ after standardization, as in Eq. equation 5. When $\sigma_r = 0$, we set the whole group's advantages to zero. Here L is the *input* length, not the number of generated tokens T_i.

B.4 Reference-Policy Configuration

The token ratio in Eq. equation 7 evaluates both the current policy and the behavior snapshot at the group's sampled base θ. The KL term in Eq. equation 7 instead compares the current policy at θ with the frozen reference at θ_0, using the same text prefix. The two terms have separate coefficients: $\gamma(L)$ multiplies the group-relative advantage in the clipped policy term, while α weights the reference penalty. The KL divergence specifies the target quantity; the sampled estimator should follow the implementation.

B.5 Adaptive Attacks

We evaluate three prompt-based attacks; the "None" category in Figure 4 is a no-attack control, not a fourth attack method.

GCG (Zou et al., 2023) searches over adversarial suffix tokens with a budget of 500 iterations.

AutoDAN (Liu et al., 2023) searches for readable jailbreak prompts with a budget of 100 iterations.

DRA (Liu et al., 2024a) disguises a harmful request and asks the model to reconstruct it. We use a budget of 20 iterations.

Resilience to Adaptive Attacks We compare MFA with MUSE on Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct using the three attacks above. These tests complement the many-shot evaluation; Figure 4 does not by itself measure attacks embedded in long many-shot prompts.

Figure 4 shows lower ASR for MFA on GCG and DRA in the plotted settings; AutoDAN has 0% ASR for both methods on both backbones. Under DRA, MFA records 21.0% and 28.7% ASR on Llama and Qwen, versus 31.5% and 71.5% for MUSE. These observations support robustness on the tested attacks without establishing a representation-level causal mechanism or guaranteeing robustness to stronger search.

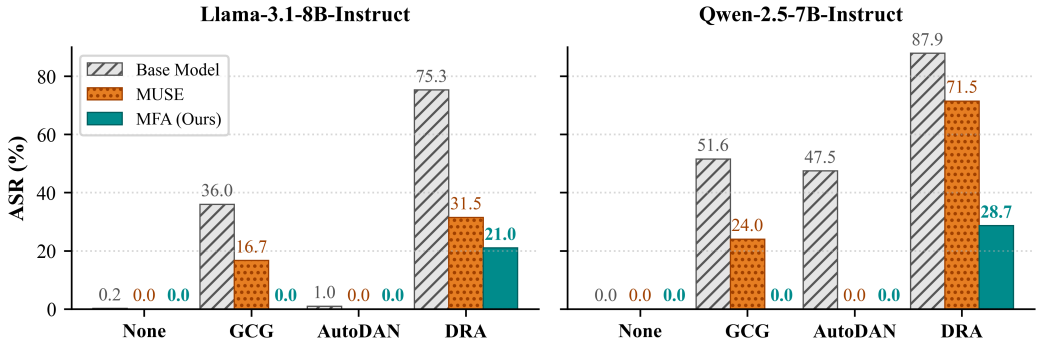

Figure 4: ASR under GCG, AutoDAN, and DRA on two backbones. "None" denotes the no-attack control. Lower is better.

B.6 Addressing Potential Reward Hacking

A critical concern in preference-based safety training is reward hacking, where the policy over-optimizes for the idiosyncrasies of the specific safety judge used during training. Since our primary optimization reward is provided by Llama-Guard-3-8B, evaluating solely with this judge (as in Table 1) risks overestimating the true safety gain. To rigorously defend against this, Table 5 cross-evaluates the Llama-3.1-8B-Instruct (256-shot MSJ) under independent evaluators: the HarmBench classifier, GPT-4o, and human annotators.

While the absolute ASR values vary due to different stringency levels across judges, the relative performance ordering remains strictly consistent. Crucially, MFA maintains a significant safety margin over both the Base model and MUSE across all independent evaluators. This confirms that the observed safety improvements reflect generalized robust refusal behavior rather than exploitation of the Llama-Guard-3-8B reward surface.

Table 5: MSJ ASR (%) at 256 shots on Llama-3.1-8B-Instruct under different safety judges.

Method	Llama-Guard-3	HarmBench cls.	GPT-4o	Human
Base	65.5	62	68	64
MUSE	23.0	25	29	27
MFA	8.8	13	16	15

Table 6 reports defense-aware and length-stress tests drawn from the same follow-up evaluation. In the out-of-range setting, ASR rises to 28% for MFA; this is a measurable limitation even though it remains below MUSE's 31% in that test. The reported out-of-range row combines a 512/1024-shot stress setting into one result and should not be read as a separate estimate for each shot count. The worst-base row changes the evaluation base within the training range and is a diagnostic rather than the native-base deployment setting.

Table 6: Defense-aware ASR (%) on Llama-3.1-8B-Instruct. Results are setting-specific and are not paired with the main HarmBench table.

Attack setting	Base	MUSE	MFA
GCG within a 256-shot context	78	38	22
Worst in-range θ over Θ	68	25	19
Out-of-range 512/1024-shot stress	70	31	28

B.7 Positional Extrapolation Techniques

Position Interpolation (PI) (Chen et al., 2023) rescales position indices to bring an extended context into the original positional range. It changes the effective phases across RoPE dimensions and may require adaptation to preserve utility.

YaRN (Peng et al., 2023) changes interpolation across RoPE frequency bands and uses attention scaling. Its position handling differs from the single scaling factor in PI.

LongRoPE (Ding et al., 2024) uses searched, nonuniform rescaling factors across dimensions. Table 4 reports its combination with MFA under 128-shot MSJ. That table uses randomly drawn ManyHarm examples that are not held fixed across methods; the reported 8.5% and 7.4% ASRs therefore do not estimate a paired incremental gain from LongRoPE.

B.8 Detailed Results of the Empirical Pilot Study

Table 7 reports the fixed-checkpoint RoPE-base sweep in Section 3. Each checkpoint is evaluated under multiple bases θ, with its native base θ_0 shown in bold. The 32-shot PANDAS pilot uses one sweep to show the within-checkpoint trend; the main safety table averages evaluations with resampled long-prompt demonstrations, so its ASRs need not coincide with this pilot.

Table 7: Fixed-checkpoint pilot under 32-shot PANDAS: ASR (%, lower is better), MMLU-Pro (fraction), and LOONG (score) versus RoPE base θ. Native bases θ_0 are in bold.

Model	θ	**ASR (↓)**	**MMLU-Pro (↑)**	**LOONG (↑)**
Llama-3.1-8B-Instruct	**500,000**	77.50	0.4425	28.4
	1,000,000	37.50	0.4401	27.5
	2,000,000	33.00	0.4323	26.0
	3,000,000	30.75	0.4182	24.5
	4,000,000	28.50	0.4209	23.0
	5,000,000	21.75	0.4161	21.5
Qwen2.5-7B-Instruct	500,000	48.50	0.5701	31.2
	1,000,000	48.50	0.5764	31.0
	2,000,000	41.25	0.5720	30.5
	3,000,000	28.75	0.5681	29.0
	4,000,000	27.50	0.5644	27.5
	5,000,000	23.75	0.5570	25.0

B.9 Benign-Context Controls

Table 8 compares the pilot ASR with two benign tasks at matched context lengths for Llama-3.1-8B-Instruct. Benign in-context learning weakens somewhat as θ increases, whereas adherence to an earlier benign instruction improves. The safety change therefore does not coincide with a uniform failure to use preceding text. These controls do not show that the RoPE-base effect is specific to safety; LOONG also declines, and the intervention changes multiple aspects of long-context processing.

Table 8: Pilot and benign controls on Llama-3.1-8B-Instruct with fixed model weights at matched 32-shot context lengths. ASR is measured with PANDAS; other columns report their respective benchmark scores.

θ	ASR (%)	ICL gain	Prefix adherence (%)	LOONG
500k	77.50	+10.2	82	28.4
1M	37.50	+9.4	88	27.5
5M	21.75	+7.1	91	21.5

B.10 Finite-Dimensional Phase Proxy

For rotary pair $j = 0, \ldots, d/2 - 1$, Eq. equation 1 gives $\varphi_j(\Delta; \theta) = \Delta\theta^{-2j/d}$. At the lowest actual frequency, $j = d/2 - 1$, the phase at distance scale L is

$$\varphi_{\min}(L; \theta) = L\theta^{-1+2/d} = \underbrace{(L/\theta)}_{\Phi(L,\theta)} \theta^{2/d}. \tag{9}$$

Thus $\Phi = L/\theta$ is the continuum low-frequency endpoint; at finite rotary dimension its difference from this phase is the factor $\theta^{2/d}$. It does not specify the higher-frequency phases or the learned query–key coefficients in Eq. equation 2. A lower value of Φ therefore does not, on its own, imply lower ASR or greater attention to a safety prefix.

Furthermore, Φ serves purely as an asymptotic proxy for the low-frequency behavior. It deliberately abstracts away the high-frequency oscillatory components (small j) that dominate local token-to-token syntax interactions. Our training methodology leverages the base variation to perturb these macroscopic phase spans without disrupting local semantic coherence, which aligns with the observed preservation of general utility.

B.11 Interpreting the Attention Diagnostic

The quantity $S(n, m; \theta)$ in Eq. equation 2 is a signed, scaled *pre-softmax* rotary query–key contribution for positions n and m. A negative plotted value is possible and must not be interpreted as a negative attention probability. By contrast, if $z_{n,k}$ is the complete pre-softmax logit for key position k and P denotes a set of prefix positions, the post-softmax mass allocated to that set is

$$A_n(P) = \sum_{m \in P} \frac{\exp(z_{n,m})}{\sum_{k \le n} \exp(z_{n,k})}. \tag{10}$$

This nonnegative quantity depends on all competing logits and cannot be determined from one pairwise $S(n, m; \theta)$. Figure 2 should therefore be read as a diagnostic of learned positional interactions, not as a measurement of prefix attention mass or proof of a causal safety mechanism.

B.12 Detailed Results of AdvBench under Many-Shot Settings

Table 9 reports refusal-matching ASR-R on AdvBench across MSJ and PANDAS shot counts. On Llama-3.1-8B-Instruct, `MFA` has lower ASR-R than the displayed baselines at every nonzero shot count. On Qwen2.5-7B-Instruct, its 32-shot MSJ ASR-R (2.9%) is slightly above MUSE (2.3%); most other nonzero-shot cells favor `MFA`. Shot-count trends are not strictly monotonic, and ASR-R should not be compared numerically with the judge-based HarmBench ASR in Table 1.

B.13 B.13 SCOPE OF MECHANISTIC CLAIMS AND CAUSALITY

Throughout this work, we demonstrate that the RoPE base significantly modulates an LLM's vulnerability to MSJ. However, we explicitly avoid claiming a strict, monolithic causal law mapping the phase pressure Φ directly to attention weights or attack success. Modern LLM attention is profoundly non-linear; the post-softmax attention mass allocated to safety prefixes depends not only on positional encoding but fundamentally on the semantic interaction of query-key vectors across all competing tokens (as detailed in Appendix B.11).

Table 9: Safety performance on AdvBench under MSJ and PANDAS across various shot counts, measured by ASR-R. ASR-R denotes a rule-based attack success rate computed by matching model responses against a predefined list of refusal phrases. Lower values indicate stronger safety resilience.

MODEL	METHOD	MSJ (Anil et al., 2024)					PANDAS (Ma et al., 2025)				
		0	32	64	128	256	0	32	64	128	256
Llama-3.1-8B-Instruct	Base	0.0	75.4	86.5	84.2	78.1	0.0	89.6	96.1	95.0	89.4
	SmoothLLM	0.0	69.3	75.4	81.5	85.6	0.0	73.5	81.4	86.7	91.5
	Self-Reminder	0.0	68.8	78.0	84.3	85.1	0.0	83.8	91.3	96.0	94.4
	MTSA	0.0	21.8	23.6	25.1	25.8	0.0	22.4	24.1	25.5	26.0
	MUSE	0.0	21.2	22.9	24.3	25.0	0.0	21.6	23.2	24.8	25.4
	MFA	0.0	8.8	11.5	13.7	11.8	0.0	11.1	13.0	14.8	12.9
Qwen2.5-7B-Instruct	Base	2.0	6.1	8.0	12.3	12.0	2.0	14.5	18.3	18.8	20.7
	SmoothLLM	1.7	5.2	6.6	10.5	10.2	1.8	13.6	14.1	17.2	16.5
	Self-Reminder	1.8	4.1	8.6	9.8	12.3	1.9	13.1	16.7	17.2	19.2
	MTSA	1.5	2.6	4.8	7.1	8.5	1.6	6.9	8.2	9.6	10.8
	MUSE	1.4	2.3	4.3	6.5	7.8	1.5	6.4	7.6	8.9	10.2
	MFA	1.0	2.9	2.4	4.5	3.3	1.1	5.1	7.4	6.8	7.9
Qwen2.5-72B-Instruct	Base	7.0	8.7	13.3	14.6	17.5	7.0	16.9	18.4	23.1	23.0
	SmoothLLM	5.8	8.6	9.9	13.7	13.4	6.1	13.1	17.2	18.4	21.1
	Self-Reminder	6.4	7.9	12.2	13.4	16.1	6.6	15.4	17.1	21.2	21.2
	MTSA	4.9	5.8	8.4	10.6	12.0	5.3	9.5	11.8	13.6	14.8
	MUSE	4.5	5.2	7.8	9.9	11.1	4.9	8.8	11.1	12.9	14.1
	MFA	1.8	4.1	6.6	5.9	7.5	2.4	8.2	7.7	9.9	8.3

The residual ASR observed in our fixed-checkpoint sweep (e.g., 21.75% ASR at $\theta = 5M$ for Llama-3.1-8B-Instruct) underscores that positional manipulation is a powerful lever, but not a complete cure for semantic vulnerabilities. MFA leverages positional sensitivity as an augmented training dimension. Its success demonstrates that varying the positional configuration during training strictly upper-bounds the robustness compared to fixed-base tuning, even if RoPE is only one component of the complex attention dynamics underlying MSJ.

B.14 TOKEN-BUDGET VIEW OF MANY-SHOT SAFETY

Figure 5 replots the reported many-shot ASR against an approximate prompt token budget. In this prompt construction, larger shot counts add roughly 91 tokens per demonstration on average. Individual prompt lengths still vary with sampled demonstrations, so this axis summarizes shot-count settings rather than matching every response to its exact token length.

For Llama-3.1-8B-Instruct, the base model and several baselines have high ASR over the tested many-shot budgets. The Qwen results also vary across shot counts without a uniformly increasing trend. In the plotted settings, `MFA` keeps ASR comparatively low. Because shot count changes both length and demonstration content, this plot alone cannot attribute differences to token count.

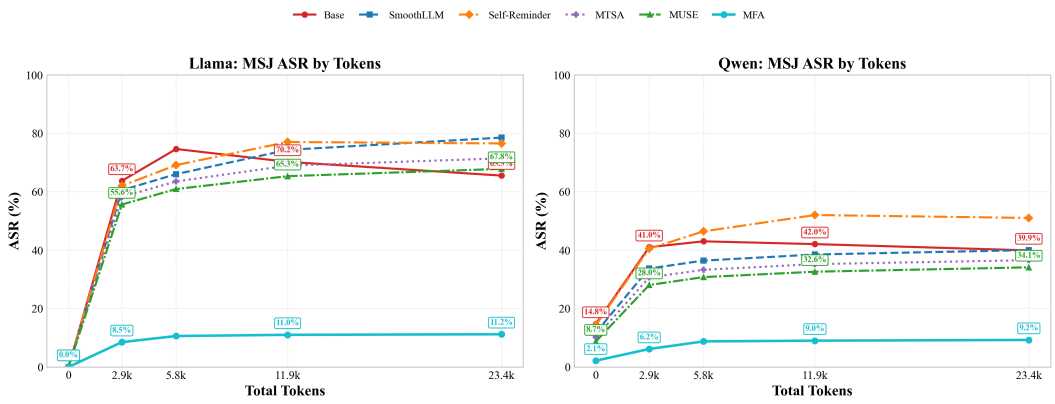

Figure 5: MSJ and PANDAS ASR replotted against approximate prompt token budgets for the tested shot counts.

C TRAINING ALGORITHM FOR MFA

The detailed procedural flow is presented in Algorithm 1.

Algorithm 1 MFA: multi-frequency rollouts with length-aware advantages

Input: Many-shot prompts $\mathcal{D}_{msj}$, base range $\Theta = [\theta_{\min}, 5\theta_{\min}]$, frozen native-base reference $\pi_{W_{\text{ref}}}(\cdot;\theta_0)$, KL coefficient α, group size G, iterations $\mathcal{T}$.
Initialize: $W \leftarrow W_{\text{ref}}$.
for iteration $\tau = 1$ **to** $\mathcal{T}$ **do**
  Freeze the behavior snapshot $W_{\text{old}} \leftarrow W$.
  Sample a batch of prompts from $\mathcal{D}_{msj}$.
  for each prompt x_L in batch **do**
    Set $L = |x_L|$ and evaluate $\gamma(L)$.
    Sample one base $\theta \sim \text{Uniform}(\Theta)$ for the group.
    Generate $y_i \sim \pi_{W_{\text{old}}}(\cdot \mid x_L;\theta)$ for $i = 1, \ldots, G$.
    Judge each response with Llama-Guard-3-8B to obtain $r_i \in \{+1, -1\}$.
    if $\sigma_r = 0$ **then**
      Set $\widehat{A}_i(L) = 0$ for all i.
    else
      Set $\widehat{A}_i(L) = \gamma(L)(r_i - \mu_r)/\sigma_r$ for all i.
    end if
    Evaluate importance ratios at θ and the reference penalty against $\pi_{W_{\text{ref}}}(\cdot;\theta_0)$.
  end for
  Update W by minimizing $\mathcal{L}_{\text{MFA}}$ in Eq. equation 6.
end for
Return: Trained policy $\pi_W(\cdot;\theta_0)$ for native-base inference.