FROM DILUTION TO ALIGNMENT: UNDERSTANDING AND MITIGATING LONG-CONTEXT SAFETY DEGRADATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) become more capable, their safety risks become increasingly consequential. Existing safety research has focused primarily on short-context settings, leaving long-context safety alignment largely underexplored. Prior studies show that model safety degrades markedly as inputs grow longer, even when the surrounding context is entirely benign. We use safety-vector projection to trace this degradation through the attention and MLP pathways. Our analysis identifies two coupled mechanisms: input-side risk dilution, which reduces attention to risky queries, and downstream safety-perception decline in MLP layers. Guided by these findings, we propose SALSA (Safety-Aware Long-context Self-distillation Alignment). Its first component, Safety-Aware Group Sequence Policy Optimization (SA-GSPO), strengthens safety perception through response-level safety judgment. Its second component, Segment Summary On-Policy Self-Distillation (SS-OPSD), re-encodes diluted risk in segment summaries generated within reasoning and distills this behavior into a prompt-free policy. An exponential moving average (EMA) anchor on benign samples helps preserve response rates and general capabilities. Extensive experiments show that SALSA substantially improves long-context safety over Plain GSPO baselines and mitigates length-induced safety degradation, bringing long- and short-context safety rates into close agreement. General capability and response rates remain largely unaffected in both long- and short-context settings. Further mechanistic analysis also confirms that SALSA alleviates the attention- and MLP-level failures underlying long-context safety degradation.

1 INTRODUCTION

Large language models (LLMs) have advanced rapidly and are now widely deployed in everyday applications (Brown et al., 2020; Zhao et al., 2023). Their context windows have expanded from a few thousand tokens to hundreds of thousands. This expansion enables new capabilities but raises an underexplored question: does safety alignment remain effective as inputs grow longer? Recent works such as LongSafety (Lu et al., 2025) suggest otherwise: model safety degrades markedly with input length. Despite extensive research on safety alignment, existing methods have been developed almost exclusively for short contexts, leaving long-context safety largely underexplored. The causes of this degradation also remain poorly understood. Conventional short-context safety alignment methods optimize query–response pairs using feedback from a guard model. However, their effectiveness in long-context settings remains limited, leaving a substantial safety gap between short- and long-context inputs. Moreover, these methods can negatively affect general capability and response rates under long contexts.

To address these problems, we conduct a systematic empirical study and find that the degradation is remarkably general: safety declines sharply with input length regardless of whether the long context is benign or harmful, and regardless of whether it is relevant to the user instruction, pointing to length itself as the driving factor. To explain why long contexts induce risk, we adopt a safety-vector projection analysis (Zou et al., 2023) on LongSafety datasets (Lu et al., 2025) at both the attention and MLP levels. We find that as length grows, attention becomes increasingly diffuse and

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.

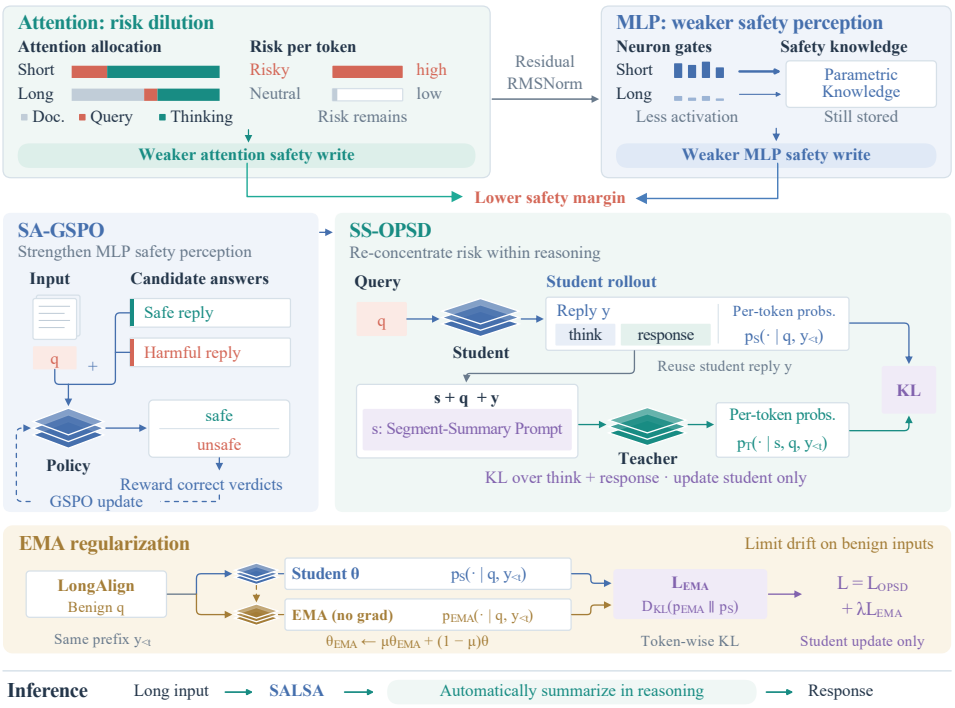


Figure 1: Overview of SALSA. The top strip illustrates a short-/long-context safety failure and the mechanism linking attention dilution to weaker MLP safety perception. SA-GSPO rewards correct safety judgments of candidate responses using balanced labels, rather than directly rewarding safe-answer generation. SS-OPSD transfers the prompted teacher's summarization behavior to a prompt-free student by matching the teacher's token-level distributions along student-sampled trajectories on the same long input. A benign-data EMA anchor regularizes behavior, and inference requires no summarization prompt.

the attention allocated to the risky query drops substantially—a safety dilution effect. This in turn weakens safety perception in the MLP layers and causes their safety-relevant writes into the residual stream to fail, ultimately giving rise to unsafe behavior.

This analysis identifies two coupled causes of long-context safety failures in LongSafety (Lu et al., 2025): input-side risk dilution and the downstream decline in safety perception. To address the latter, we propose Safety-Aware Group Sequence Policy Optimization (SA-GSPO). SA-GSPO pairs long-context risky queries with candidate responses collected from different models and trains the model to judge the safety of those responses, sharpening its ability to discriminate risk under long contexts. To address input-side dilution, we exploit a complementary observation: although attention to the risky query decreases with length, the model's generated thinking retains a substantial share of attention. We therefore relocate risk information from the long input into segment summaries generated within the model's own reasoning. Segment Summary On-Policy Self-Distillation (SS-OPSD) transfers this summarization behavior into a prompt-free policy. Together, SA-GSPO and SS-OPSD form SALSA (Safety-Aware Long-context Self-distillation Alignment), a unified framework that addresses both mechanistic sources of long-context safety degradation. To prevent safety-oriented training from degrading benign behavior, SS-OPSD also incorporates an exponential moving average (EMA) anchor. A slowly updated EMA copy of the model regularizes the policy on benign long-context inputs, improving safety without increasing over-refusal or sacrificing helpfulness.

Extensive experiments validate SALSA's effectiveness. Compared with RL baselines, SALSA substantially improves long-context safety without sacrificing general capability in either short- or long-context settings or inducing over-refusal. Unlike the Plain GSPO baseline, SALSA achieves long-context safety on par with short-context safety, with no pronounced length-induced decline, closing the safety gap that motivates this work. Mechanistic analysis further confirms that SALSA allevi-

ates safety failures in both the attention and MLP pathways. Framework of our method SALSA is demonstrated in Figure 1.

Our contributions are summarized as follows:

- We characterize long-context safety degradation and its underlying mechanism through safety-vector projection in LongSafety (Lu et al., 2025), identifying input-side risk dilution in the attention pathway and downstream safety-perception decline in the MLP pathway.
- We propose Safety-Aware Group Sequence Policy Optimization (SA-GSPO), which strengthens safety perception at the MLP level through response-level safety judgments and improves risk discrimination under long contexts.
- We introduce in-reasoning segment summaries to re-concentrate safety-relevant information and develop Segment Summary On-Policy Self-Distillation (SS-OPSD) to counter attention-side risk dilution.
- Across safety-rate and interpretability-based metrics, SALSA aligns safety behavior between short- and long-context settings, substantially improving long-context safety while largely preserving general capability and avoiding over-refusal.

2 Related Work

2.1 Long-Context Safety Risks

Existing long-context research has largely emphasized capability rather than safety. Benchmarks such as LongBenchv2 (Bai et al., 2025) and RULER (Hsieh et al., 2024) evaluate retrieval, reasoning, and long-range dependency handling without assessing safety. Many-shot Jailbreaking (Anil et al., 2024) exposes long-context safety risks by prepending tens or hundreds of demonstrations of undesirable behavior, with attack success increasing as the number of demonstrations grows. LongSafety (Lu et al., 2025) provides a systematic benchmark for open-ended long-context safety evaluation, comprising 1,543 cases across seven safety categories. A particularly notable finding is that simply extending an input with neutral or benign context can degrade model safety. We investigate the internal mechanisms underlying this phenomenon through safety-vector projection analysis in our paper.

2.2 Safety Alignment

Safety alignment has been studied extensively, but almost entirely in the short-context regime: reinforcement learning from human feedback (Ouyang et al., 2022), Constitutional AI (Bai et al., 2022), and preference-optimization variants such as DPO (Rafailov et al., 2023) establish safe behavior by training models to refuse harmful prompts. A smaller line of work extends alignment to long contexts, mainly by supervised fine-tuning on curated long-context safety data (Huang et al., 2025; Lu et al., 2025). Such data-driven approaches, however, are constrained by the scarcity of long-context safety examples and, more importantly, treat safety as a surface behavior rather than addressing its mechanistic origin. Consistent with this, our experiments show that reinforcement learning alone fails to halt the length-induced safety decline, motivating an alignment method that repairs the attention- and MLP-level causes of dilution rather than merely fitting safer outputs.

2.3 On-Policy Self-Distillation

Knowledge distillation (Hinton et al., 2015) transfers knowledge from teacher to student, while on-policy distillation (Agarwal et al., 2024) provides token-level teacher supervision on student-generated trajectories. On-Policy Self-Distillation (Zhao et al., 2026) uses the same model as both teacher and student to improve reasoning. Our SS-OPSD adapts this paradigm to long-context safety by internalizing segment summarization within reasoning, where attention remains substantial even in long contexts. These summaries re-concentrate diluted risk information, counteracting attention-side dilution.

3 INTERPRETABILITY ANALYSIS IN LONG-CONTEXT SAFETY PROBLEMS

Building on LongSafety (Lu et al., 2025), this section investigates how and why model safety varies with input length under two settings: one in which the long context is semantically relevant to the query, and one in which it is unrelated. We first report how the safety rate varies with context length under the LongSafety setting in Table 1. As shown, as the length grows, even when the long samples are mostly neutral or the context is irrelevant to the query, the model's safety rate drops significantly. We analyze safety at the *decision position* $t^\star$, the prediction step of the first response token after the model has produced its thinking segment. For the pre-normalized backbone, the last-layer residual stream $\mathbf{z}$ before the final RMSNorm decomposes exactly as $\mathbf{z} = \mathbf{e} + \sum_l \left(\mathbf{a}^{(l)} + \mathbf{m}^{(l)}\right)$, where $\mathbf{e}$ is the token embedding, and $\mathbf{a}^{(l)}$ and $\mathbf{m}^{(l)}$ are the attention and MLP outputs of layer l, respectively. All vectors are evaluated at $t^\star$, and the sum spans all layers.

Inspired by Zhao et al. (2025b), we define a unit safety direction $\mathbf{r}_{\text{ref}} = (\boldsymbol{\mu}_H - \boldsymbol{\mu}_B)/\|\boldsymbol{\mu}_H - \boldsymbol{\mu}_B\|_2$ from the mean harmful residual $\boldsymbol{\mu}_H$ and mean benign residual $\boldsymbol{\mu}_B$ at length $L{=}0$. Projecting the residual onto $\mathbf{r}_{\text{ref}}$ yields the pivot of our analysis, the *safety margin* S, which splits additively into a constant term, an attention-pathway write A, and an MLP-pathway write M:

$$S \;\equiv\; \mathbf{z}\cdot\mathbf{r}_{\text{ref}} \;=\; \underbrace{\mathbf{e}\cdot\mathbf{r}_{\text{ref}}}_{\text{const}} + \underbrace{\sum_l \mathbf{a}^{(l)}\cdot\mathbf{r}_{\text{ref}}}_{A} + \underbrace{\sum_l \mathbf{m}^{(l)}\cdot\mathbf{r}_{\text{ref}}}_{M}. \tag{1}$$

A larger S indicates a stronger tendency toward safe behavior, so the question of why long contexts erode safety reduces to how A and M evolve with input length.

Table 1: Safety rate, attention-pathway retention, and MLP-pathway retention of Qwen3-8B as the context length grows. Relevant contexts are drawn from LongSafety, whereas irrelevant contexts are drawn from RAGTruth (Niu et al., 2024). Retention is normalized to 1.000 at $L{=}0$, while safety rates are percentages. Δ denotes the change from $L{=}0$ to $L{=}8192$.

Metric	Context	L_0	L_{1024}	L_{4096}	L_{8192}	Δ
Safety Rate	Relevant	69.7%	51.1%	46.7%	42.7%	−27.0%
	Irrelevant	64.0%	60.0%	55.0%	50.5%	−13.5%
Attention Retention	Relevant	1.000	0.782	0.749	0.728	−0.272
	Irrelevant	1.000	0.878	0.808	0.812	−0.188
MLP Retention	Relevant	1.000	0.566	0.432	0.377	−0.623
	Irrelevant	1.000	0.863	0.702	0.667	−0.333

3.1 ATTENTION PATHWAY: INPUT-SIDE RISK DILUTION.

Expanding A over heads and keys, each key contributes its attention weight times the projection of its value onto the safety direction, $A = \sum_l \sum_h \sum_k \alpha_h^{(l)}(k)\left(\mathbf{v}_h^{(l)}(k)\cdot\mathbf{u}_h^{(l)}\right)$, where $\mathbf{u}_h^{(l)} = \mathbf{W}_{O,h}^{(l)\top}\mathbf{r}_{\text{ref}}$ is the OV projection of head h onto $\mathbf{r}_{\text{ref}}$. Here $\mathbf{u}_h^{(l)}$ acts as a per-head risk read-out axis in value space: the projection $\mathbf{v}_h^{(l)}(k)\cdot\mathbf{u}_h^{(l)}$ of each token's value onto it is precisely the risk concentration carried by token k, so that A **is the sum of these per-token risk concentrations weighted by the attention each token receives.** Appendix B derives this decomposition from the multi-head attention output. For each head, consider a fixed number k of risky query tokens and N neutral distractors, with shared logits z_q and z_d within the two groups. Writing $\Delta z = z_q - z_d$, softmax assigns the query a total attention mass $\text{mass}^{\text{query}} = \frac{ke^{z_q}}{ke^{z_q}+Ne^{z_d}} = \left(1+\frac{N}{k}e^{-\Delta z}\right)^{-1}$. If N grows proportionally to context length L and $|\Delta z|$ remains bounded, the denominator grows proportionally to L, yielding $\text{mass}^{\text{query}} = \Theta(1/L)$. With bounded positive query safety values and zero safety write from all other tokens, this implies $A \to 0$ in this simplified model. The experiments in Table 1 also verify the above formulation: as the context length grows from 0 to 8192, with relevant context, the attention-pathway write A decreases by 27.2%; even with irrelevant context, the attention-pathway write still decreases by 18.8%. We also analyze how the attention distribution across different segments changes with context length under a long-context safety scenario, as shown in Table 2. To avoid interference, we exclude the attention allocated to the template segment and re-normalize the

remaining attention masses. We find that as the context length increases, the attention share of the query decreases significantly. Although the share of generated thinking also decreases notably, it still accounts for a large proportion, which provides the intuition for our approach to addressing long-context safety; the detailed method is presented in Section 4.2.

Table 2: Fraction of attention mass at the decision position that Qwen3-8B allocates to each context segment as the input length grows. From L=0 to L=8192, attention to the risky query decreases ($0.151 \rightarrow 0.065$), while attention to neutral document tokens increases ($0.000 \rightarrow 0.467$).

Model	Segment	L_0	L_{1024}	L_{4096}	L_{8192}
Qwen3-8B	Neutral document	0.000	0.331	0.405	0.467
Qwen3-8B	Query	0.151	0.081	0.073	0.065
Qwen3-8B	Generated thinking	0.849	0.589	0.522	0.468

3.2 MLP PATHWAY: SAFETY PERCEPTION UNDER DILUTION.

Expanding the SwiGLU FFN (Shazeer, 2020) neuron by neuron, $\mathbf{m}^{(l)} \cdot \mathbf{r}_{\text{ref}} = \sum_n g_n^{(l)}(\mathbf{x})\, \kappa_n^{(l)}$, where $g_n^{(l)}(\mathbf{x})$ is the input-dependent gated activation of neuron n, combining its gate and up branches, and $\kappa_n^{(l)} = \mathbf{w}_{\text{down},n}^{(l)} \cdot \mathbf{r}_{\text{ref}}$ is its signed safety-writing strength per unit activation. Their product is the neuron's actual contribution along the safety direction. For a fixed model and reference direction, the coefficients $\kappa_n^{(l)}$ remain unchanged across input lengths, so changes in the MLP-pathway write M are determined by the weighted aggregate of neuronal activations. Appendix C derives this exact decomposition. The FFN input $\mathbf{x}^{(l)}$ is obtained by applying RMSNorm (Zhang & Sennrich, 2019) to a residual that already contains attention writes. Intuitively, attention-side risk dilution will weaken the effective risk cues reaching the MLP and reduce the net safety contribution of its gated activations, potentially amplifying the length-induced decline in safety information within the MLP pathway. Table 1 provides empirical evidence consistent with this intuition. As context length increases from 0 to 8192, the MLP safety write decreases by 62.3% with relevant context and 33.3% with irrelevant context, compared with attention-write decreases of 27.2% and 18.8%, respectively. The MLP pathway thus exhibits an amplified relative decline in safety write compared with the attention pathway.

4 METHOD

Our analysis of LongSafety (Lu et al., 2025) suggests that long-context safety degradation reflects two mechanistic failures rather than a loss of safety knowledge: (i) input-side *risk dilution* in attention and (ii) downstream *safety-perception decline* in the MLP pathway. These findings motivate targeted repair of the two pathways rather than uniform alignment. First, the conventional safety-alignment baselines we evaluated do not eliminate the length-induced decline, motivating SS-OPSD to re-concentrate risk signals and preserve safety-direction writes. Second, the MLP-side decline motivates SA-GSPO to restore internal safety perception in long contexts, rather than only shaping outputs. Together, SA-GSPO and SS-OPSD constitute our alignment framework, SALSA. We also incorporate Plain GSPO, which directly optimizes long-context generated responses using safety rewards assigned by Qwen3-Guard, into SALSA to further improve safety. Nevertheless, our extensive ablation study in Table 5 demonstrates that SALSA without Plain GSPO still significantly outperforms Plain GSPO alone, demonstrating the effectiveness of our paper.

4.1 SA-GSPO: RESTORING LONG-CONTEXT SAFETY PERCEPTION VIA RISK DISCRIMINATION

Our preceding analysis identifies impaired MLP safety perception as one source of long-context safety degradation. Accordingly, SA-GSPO directly targets long-context safety perception through response-level safety judgment rather than safe-answer generation. For a context c and query q, we collect candidate responses a from a model pool, label them with Qwen3Guard as $b \in \{0, 1\}$ (safe or unsafe), and filter to a balanced distribution $\mathcal{D}$ with $\Pr_{\mathcal{D}}(b{=}0) = \Pr_{\mathcal{D}}(b{=}1) = 1/2$. The judgment input $x = (c, q, a)$ places the candidate response inside the input, so the policy emits a verdict $\hat{b}(y)$ and is rewarded for agreement with the label, $r(x, b, y) = \mathbf{1}\{\hat{b}(y) = b\}$, with expected

reward $\mathcal{R}(\theta) = \frac{1}{2}\sum_{b=0}^{1}\mathbb{E}_{x\sim\mathcal{D}_b}\Big[\Pr_{y\sim\pi_\theta(\cdot|x)}\big(\hat{b}(y)=b\big)\Big]$. Because safe and unsafe examples are equally balanced, always predicting the same label yields a reward of only 0.5; achieving a higher reward requires the model to distinguish safe responses from unsafe ones. We keep the GSPO optimizer: with G rollouts per input, group-relative advantages $\widehat{A}_i$, and the length-normalized ratio $s_i(\theta) = \big(\pi_\theta(y_i \mid x)/\pi_{\theta_{\text{old}}}(y_i \mid x)\big)^{1/|y_i|}$, we maximize

$$\mathcal{J}_{\text{SA-GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\big\{s_i(\theta)\widehat{A}_i,\ \text{clip}\big(s_i(\theta),1-\epsilon,1+\epsilon\big)\widehat{A}_i\big\}\right]. \tag{2}$$

Appendix D.1 details the data construction and advantage estimator.

Risk discrimination, not blanket conservatism. Plain safety RL rewards safe generations, but on benign inputs it cannot separate an unnecessary refusal from a helpful answer, so safety can rise through indiscriminate conservatism. To make the contrast precise, Appendix D.2 analyzes a shared-feature model with a generation-side refusal bias u, a shared risk-discrimination strength ω, a judgment-side label bias v, and an effective risk-feature magnitude $\alpha > 0$ at a fixed length. Writing H and B for risky and benign inputs, the refusal and judgment probabilities are

$$f_H = \sigma(u+\alpha\omega), \quad f_B = \sigma(u-\alpha\omega), \qquad q_H = \sigma(v+\alpha\omega), \quad q_B = \sigma(v-\alpha\omega). \tag{3}$$

Here, f_H and f_B are the refusal probabilities on risky and benign inputs, respectively. Raising u increases refusal on both input types, whereas raising ω separates them; only the latter reflects improved risk perception. In the judgment task, q_H is the probability of predicting an unsafe response as unsafe, whereas q_B is the probability of incorrectly predicting a safe response as unsafe; hence, the correct-classification probability for a safe response is $1 - q_B$.

A safety–responsiveness trade-off. Proposition 1 (Appendix D.4) compares one expected GSPO update at a common rollout point, with balanced, symmetric judgment classes, $v = 0$, fixed α, common output length, and locally separable coordinates. Let $\eta_u, \eta_\omega > 0$ be the coordinate-wise update scales and $D, J > 0$ the task-specific gradient coefficients defined in the appendix. The generation-logit changes are

$$\begin{aligned} \text{Plain safety RL:} \quad & \Delta z_H = D(\eta_u + \alpha^2\eta_\omega), \quad \Delta z_B = D(\eta_u - \alpha^2\eta_\omega), \\ \text{Judgment RL:} \quad & \Delta z_H = J\,\alpha^2\eta_\omega, \qquad\qquad \Delta z_B = -J\,\alpha^2\eta_\omega. \end{aligned} \tag{4}$$

Both updates raise the risky-input refusal margin, but Plain safety RL also raises benign refusal whenever $\eta_u > \alpha^2\eta_\omega$, whereas judgment RL is less prone to increasing benign refusal under the stated assumptions. In this shared-feature model, judgment RL improves safety by sharpening the safe/unsafe distinction rather than by shifting an input-independent refusal bias. Complete coordinate separation is sufficient but not necessary: for $\Delta\omega > 0$, the favorable signs persist whenever a leaked bias change obeys $|\Delta u| < \alpha\Delta\omega$ (Appendix D.4).

Consequence for general ability. On a benign task where refusal scores zero, success factors as $U = (1 - f_B)Q$, with Q the conditional success probability given a substantive answer. Holding $Q > 0$ fixed, an increase in benign refusal lowers U. More generally, if $f_B^+ \le f_B$ and $|Q^+ - Q| \le \varepsilon_Q$, then $U^+ \ge U - \varepsilon_Q$, where $+$ denotes post-update quantities on the same input distribution (Appendix D.5). Table 4 reports safety and general performance, while Figure 2(b) provides complementary mechanistic evidence: after SA-GSPO, the MLP safety write decays less from length 0 to 8192 than in the base model.

4.2 SS-OPSD: Countering Attention Dilution through In-Reasoning Segment Summaries

The previous section improved long-context safety at the MLP level by strengthening safety perception; here we turn to the attention pathway and directly counter the input-side risk dilution identified in Section 3. As Table 2 shows, the attention mass allocated to both the risky query and the model's generated thinking drops markedly as the input length grows, yet the thinking segment decays far more slowly and still commands a substantial share of attention even at long lengths. This makes

Table 3: Effect of the segment-summary strategy on Qwen3-8B. Direct Summary applies the summarization system prompt at inference time on the SA-GSPO-trained models, whereas SALSA(including SS-OPSD) represents our OPSD-based method with the EMA anchor and internalized summarization behavior. Full and Query denote LongSafety Full and LongSafety Query. XS, LBv2, and OC-avg denote the XSTest response rate, LongBench v2, and the OpenCompass average, respectively.

	Long-context safety		Response rate	General capability	
Method (%)	Full↑	Query↑	XS↑	LBv2↑	OC-avg↑
Qwen3-8B	39.21	68.51	96.40	46.70	80.71
Direct Summary	88.59	96.50	68.20	43.96	78.03
SALSA(including SS-OPSD)	89.20	92.30	95.20	47.25	79.98

the generated thinking a natural vehicle for repairing dilution: if the model summarizes its input segment by segment, by semantic content, within its reasoning process, it both sharpens its semantic understanding of the context—so that risk is localized more reliably—and raises the effective concentration of the risky query, since a query of only one or two sentences can then occupy an entire summary segment inside the thinking rather than being diluted across the raw context.

However, we find that simply instructing the model to summarize through a system prompt, while it substantially improves long-context safety, also causes a pronounced drop in general capability, as shown in Table 3. To retain this behavior without the capability cost, we distill the system-prompted summarization pattern back into the model via On-Policy Self-Distillation (OPSD) (Zhao et al., 2026). Let c_{sum} denote the summarization system prompt; the prompted model is the teacher and the prompt-free model is the student, sharing weights θ. We sample trajectories from the prompt-free student and train it to match the prompted teacher's token distributions along these trajectories using $\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_x \, \mathbb{E}_{y\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)} \big[\sum_t D_{\mathrm{KL}}\big(\pi_{\theta_{\mathrm{old}}}(\cdot \mid c_{\mathrm{sum}}, x, y_{<t}) \,\big\|\, \pi_\theta(\cdot \mid x, y_{<t})\big)\big]$. At inference the model summarizes with no prompt. To protect general ability against this shift in reasoning style, we keep an exponential moving average (EMA) anchor $\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta$ with decay $\mu \in [0, 1)$ and, on benign long-context data (LongAlign (Bai et al., 2024)), regularize the policy toward it, giving the overall objective $\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{OPSD}}(\theta) + \lambda \, \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{benign}}} \big[\sum_t D_{\mathrm{KL}}\big(\pi_{\bar{\theta}}(\cdot \mid x, y_{<t}) \,\big\|\, \pi_\theta(\cdot \mid x, y_{<t})\big)\big]$.

Why summarization restores the attention write. We apply the token-level softmax separately to the baseline and summary-containing runs, then aggregate the normalized attention within each group. Let $m_q^{(0)}$ be the baseline query's total attention mass and m_s the summary group's total attention mass. The concentrations c_q and c_s are within-group attention-weighted averages of the token-level risk concentrations, not averages of logits. In the reduced model, all remaining keys form a background group whose write is set to zero. We assume that the query concentration is the same finite constant $c_q > 0$ in both runs and across lengths, and that summaries rescale the original query and background masses by $1 - m_s$ without changing their relative allocation. Under these explicit comparison assumptions,

$$\begin{aligned} A_{\mathrm{base}}(L) &= m_q^{(0)}(L) c_q, \\ A_{\mathrm{ss}}(L) &= \big(1 - m_s(L)\big) A_{\mathrm{base}}(L) + m_s(L) c_s. \end{aligned} \tag{5}$$

The new write is a weighted average of the baseline write and the summary concentration. Consequently, for $m_s(L) > 0$,

$$\begin{aligned} &A_{\mathrm{ss}}(L) - A_{\mathrm{base}}(L) = m_s(L)\big[c_s - A_{\mathrm{base}}(L)\big], \\ &A_{\mathrm{ss}}(L) > A_{\mathrm{base}}(L) \iff c_s > m_q^{(0)}(L) c_q. \end{aligned} \tag{6}$$

Under the baseline dilution assumption $m_q^{(0)}(L) = \Theta(1/L)$ as $L \to \infty$, any fixed $c_s > 0$ eventually exceeds the baseline write. If $c_s \geq c_q$, the write is no smaller at any length, and strictly larger whenever $c_s > c_q$ or the baseline assigns nonzero attention to the background. A non-vanishing write additionally requires the summaries to retain a positive attention share. Proposition 2 in Appendix E states these assumptions and separates the improvement condition from the positive-lower-bound result.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training Setup. Our experiments center on long-context safety. As such data are scarce, we pool samples from two complementary LongSafety corpora (Lu et al., 2025; Huang et al., 2025) to form a 3,000-example training set, and hold out 543 examples from Lu et al. (2025) as a disjoint test set. All safety labels come from Qwen3Guard (Zhao et al., 2025a). Plain GSPO is capped at two epochs—beyond which long-context general ability drops by roughly five points and both the judge training stage and the SS-OPSD stage are trained for one epoch. Details are shown in Appendix G.

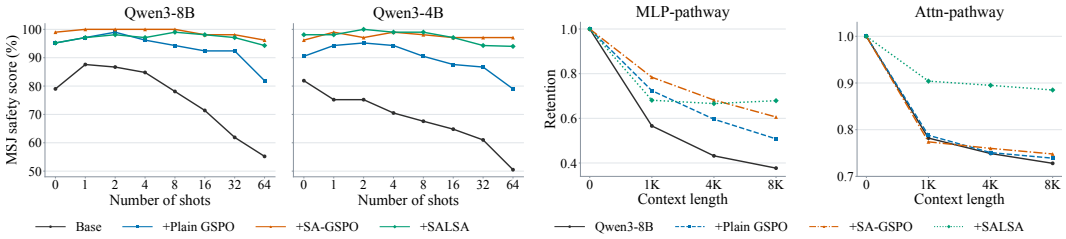
Evaluation. Long-context evaluation: For long-context safety, beyond the 543 held-out LongSafety cases (Lu et al., 2025), we evaluate robustness to Many-shot Jailbreaking (Anil et al., 2024) across different context lengths and measure long-context general ability on LongBench v2 (Bai et al., 2025). **Short-context evaluation:** We reuse the same 543 cases as single queries without long context to measure safety and the long–short safety gap, and additionally evaluate safety on JailbreakBench (JBB-Behaviors) (Chao et al., 2024) and StrongREJECT (Souly et al., 2024), as well as over-refusal on XSTest (Röttger et al., 2024) and the benign split of WildGuardTest (Han et al., 2024). **General capability:** We evaluate general capability with OpenCompass (OpenCompass Contributors, 2023) across Knowledge (MMLU-lite (Hendrycks et al., 2021), MMLU-Pro-lite (Wang et al., 2024), GPQA, and GPQA-Diamond (Rein et al., 2023)), Math (GSM8K (Cobbe et al., 2021)), Code (HumanEval and HumanEval+ (Chen et al., 2021; Liu et al., 2023), and MBPP and MBPP+ (Austin et al., 2021; Liu et al., 2023)), Instruction Following (IFEval (Zhou et al., 2023)), and Reasoning (BBH-lite (Suzgun et al., 2022) and AGIEval (Zhong et al., 2023)).

5.2 MAIN RESULTS

We conduct experiments on Qwen3-4B and Qwen3-8B (Yang et al., 2025), with the main results reported in Table 4. We evaluate general capabilities on 12 widely used benchmarks; Table 4 reports OC-Avg, the average over all 12 scores, and per-benchmark breakdowns are given in Tables 6 and 7. We also observe that SALSA and Plain GSPO achieve comparable safety on short-context inputs in the "Query" column. However, the gap between long-context and short-context safety for SALSA is only 3.10% and 4.89%, whereas Plain GSPO exhibits gaps of around 17% in both cases. We further evaluate how model safety changes under different numbers of demonstrations in the Many-shot Jailbreaking setting, as shown in Figure 2 (a). Both SA-GSPO and SALSA substantially improve robustness to MSJ, with only minor safety degradation as the number of demonstrations increases. In contrast, the Plain GSPO baseline exhibits a clear decline in safety at 64 shots. Notably, because safety dilution is less pronounced on MSJ, the benefit of SS-OPSD is limited in this setting. Nevertheless, SA-GSPO effectively strengthens long-context safety perception and safety-information writing, yielding substantial safety improvements over the baseline.

Table 4: Main evaluation results. Full and Query denote LongSafety Full and LongSafety Query, and Δ denotes the Long–Short gap (↓). JBB, S-R, XS, WG, LBv2, and OC-avg denote JailbreakBench Behaviors, StrongREJECT, XSTest, WildGuard Test, LongBench v2, and the OpenCompass average, respectively. XS and WG report response rates.

		Long-context safety			Safety benchmarks		Response rate		General capability	
Model	Method (%)	Full↑	Query↑	Δ ↓	JBB↑	S-R↑	XS↑	WG↑	LBv2↑	OC-avg↑
Qwen3-8B	Base	39.21	68.51	29.30	94.00	94.57	96.40	98.50	46.70	80.71
	+Plain GSPO	70.35	88.03	17.68	98.00	98.08	92.40	97.25	42.31	80.47
	+SA-GSPO	83.43	94.11	10.68	100.00	99.36	92.80	95.45	43.41	80.91
	+SALSA	89.20	92.30	3.10	99.00	99.68	95.20	96.02	47.25	79.98
Qwen3-4B	Base	38.67	68.12	29.45	96.00	93.29	95.20	97.88	40.11	77.55
	+Plain GSPO	71.09	88.60	17.51	98.00	98.72	92.80	96.83	35.71	77.57
	+SA-GSPO	85.09	94.29	9.20	99.00	99.36	92.80	96.30	35.07	77.92
	+SALSA	87.19	92.08	4.89	100.00	99.36	93.60	95.66	40.66	77.62



(a) Many-Shot Jailbreaking (b) Safety-Information Retention

Figure 2: (a) Many-shot Jailbreaking safety scores for Qwen3-8B and Qwen3-4B across different numbers of shots. (b) MLP-pathway and attention-pathway retention of safety-relevant information as the context length increases. All retention values are normalized to 1.0 at length 0.

Table 5: Ablation of Plain GSPO in SALSA for Qwen3-8B. Full and Query denote LongSafety Full and LongSafety Query, and Δ denotes the Long–Short gap (↓). JBB, S-R, XS, WG, LBv2, and OC-avg denote JailbreakBench Behaviors, StrongREJECT, XSTest, WildGuard Test, LongBenchv2, and OpenCompass average, respectively. XS and WG report response rates.

	Long-context safety			Safety benchmarks		Response rate		General capability	
Method (%)	Full↑	Query↑	Δ ↓	JBB↑	S-R↑	XS↑	WG↑	LBv2↑	OC-avg↑
Qwen3-8B	39.21	68.51	29.30	94.00	94.57	96.40	98.50	46.70	80.71
+Plain GSPO	70.35	88.03	17.68	98.00	98.08	92.40	97.25	42.31	80.47
+SALSA w/o Plain GSPO	85.82	90.98	5.16	100.00	99.04	98.00	97.04	45.60	80.51
+SALSA	89.20	92.30	3.10	99.00	99.68	95.20	96.02	47.25	79.98

5.3 Post-Alignment Interpretability Analysis and Ablation Study

The above results demonstrate SALSA's effectiveness through behavioral metrics; we next examine its internal mechanisms. The results are shown in Figure 2 (b). We find that SA-GSPO substantially improves the model's MLP retention, yielding a gain of over 0.2 at 8K context length. However, the base model, Plain GSPO, and SA-GSPO exhibit similar attention retention as context length increases, indicating that SA-GSPO cannot directly resolve attention dilution. In contrast, SALSA significantly improves attention retention. We also conduct an ablation study by removing Plain GSPO from SALSA, as shown in Table 5. Plain GSPO provides SALSA with additional safety gains, but its contribution is limited. Compared with Plain GSPO, SALSA without Plain GSPO performs substantially better on both safety and general-capability metrics, achieving a 15.47% safety gain in long-context settings and reducing the short–long safety gap from 17.68% to 5.16%. Furthermore, our method also substantially outperforms the baseline on long-context general-capability metrics such as LongBenchv2.

6 Conclusion

We systematically investigate the degradation of large language model safety in long-context settings. Through safety-vector projection analysis, we identify two interrelated internal mechanisms: the dilution of risk information in the attention pathway as the context grows, and the decline in safety perception in the MLP pathway. Based on these findings, we propose SALSA, a mechanism-driven alignment framework for long-context safety. SALSA employs SA-GSPO to strengthen risk discrimination under long contexts through response-level safety judgment training. It further uses SS-OPSD to re-concentrate risk information in segment summaries generated during reasoning and transfers this capability to a prompt-free model through OPSD, while introducing an EMA anchor to limit behavioral drift on benign data. Systematic evaluations show that SALSA substantially improves long-context safety, markedly narrows the safety gap between short- and long-context settings, and enhances robustness to Many-shot Jailbreaking. Meanwhile, long-context general capability, overall general capability, and over-refusal performance remain largely stable. Post-alignment interpretability analysis further shows that SA-GSPO mitigates the decline in safety perception in the MLP pathway, while SS-OPSD alleviates risk dilution in the attention pathway.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript, including improving grammar, clarity, concision, and LaTeX presentation. We have not used generative AI tools for research ideation or execution, drafting manuscript sections, or generating synthetic datasets. All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 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 (ACL)*, 2025.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

DeepSeek-AI et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Peter W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990. doi: 10.1145/84537.84552.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Mianqiu Huang, Xiaoran Liu, Shaojun Zhou, Mozhi Zhang, Qipeng Guo, Linyang Li, Chenkun Tan, Yang Gao, Pengyu Wang, Linlin Li, Qun Liu, Yaqian Zhou, Xipeng Qiu, and Xuanjing Huang. LongSafety: Enhance safety for long-context LLMs. *arXiv preprint arXiv:2411.06899*, 2025.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yida Lu, Jiale Cheng, Zhexin Zhang, Shiyao Cui, Cunxiang Wang, Xiaotao Gu, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. Longsafety: Evaluating long-context safety of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

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. In *Advances in Neural Information Processing Systems*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

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. *arXiv preprint arXiv:2311.12022*, 2023.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025a.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, 2025b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. URL https://arxiv.org/abs/2507.18071.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Detailed OpenCompass Results

We conduct a systematic OpenCompass (OpenCompass Contributors, 2023) evaluation spanning five categories: Knowledge (MMLU-lite (Hendrycks et al., 2021), MMLU-Pro-lite (Wang et al., 2024), GPQA, and GPQA-Diamond (Rein et al., 2023)), Math (GSM8K (Cobbe et al., 2021)), Code (HumanEval and HumanEval+ (Chen et al., 2021; Liu et al., 2023), MBPP and MBPP+ (Austin et al., 2021; Liu et al., 2023)), Instruction Following (IFEval (Zhou et al., 2023)), and Reasoning (BBH-lite (Suzgun et al., 2022) and AGIEval (Zhong et al., 2023)). The full results are reported in Tables 6 and 7. The OC-avg reported in the main text is the average over these values. Compared with the base model, our method has little impact on general capability.

Table 6: Detailed OpenCompass results for Qwen3-8B and its aligned variants. Higher is better.

Category	Dataset	Qwen3-8B	+Plain GSPO	+SA-GSPO	+SALSA
Knowledge	MMLU-lite	86.12	87.31	86.48	85.76
	MMLU-Pro-lite	72.37	72.14	73.96	73.02
	GPQA	53.72	54.28	53.85	52.21
	GPQA-Diamond	57.07	58.59	57.07	53.54
Math	GSM8K-0shot	95.38	95.15	95.38	95.15
Code	HumanEval	93.90	90.85	94.51	93.29
	HumanEval+	86.59	84.15	84.76	85.98
	MBPP	89.72	87.97	88.72	87.97
	MBPP+	72.93	72.93	73.93	73.18
IF	IFEval	82.62	84.66	84.29	82.02
Reasoning	BBH-lite	90.96	90.58	90.91	91.17
	AGIEval	87.18	87.03	87.06	86.45

Table 7: Detailed OpenCompass results for Qwen3-4B and its aligned variants. Higher is better.

Category	Dataset	Qwen3-4B	+Plain GSPO	+SA-GSPO	+SALSA
Knowledge	MMLU-lite	82.51	82.31	82.23	82.39
	MMLU-Pro-lite	69.49	68.66	69.27	67.30
	GPQA	45.07	45.09	46.14	45.52
	GPQA-Diamond	47.47	48.99	51.52	49.49
Math	GSM8K-0shot	94.16	94.16	94.31	94.47
Code	HumanEval	92.07	92.07	93.29	93.90
	HumanEval+	84.76	85.37	86.59	84.76
	MBPP	87.47	88.22	87.47	88.47
	MBPP+	72.43	72.68	72.18	73.68
IF	IFEval	82.81	80.96	80.41	79.26
Reasoning	BBH-lite	89.44	90.07	88.87	89.46
	AGIEval	82.95	82.22	82.75	82.77

B Detailed Derivation of the Attention-Pathway Decomposition

At the decision position $t^\star$, let l, h, and k index layers, heads, and visible key positions, respectively. Let $\alpha_h^{(l)}(k)$ be the softmax attention weight, $\mathbf{v}_h^{(l)}(k)$ the corresponding value vector, and $\mathbf{W}_{O,h}^{(l)}$ the block of the output matrix that maps head h into residual space. With attention dropout disabled and no output bias, the layer's attention output is

$$\mathbf{a}^{(l)} = \sum_h \mathbf{W}_{O,h}^{(l)} \left(\sum_k \alpha_h^{(l)}(k) \mathbf{v}_h^{(l)}(k) \right). \tag{7}$$

Projecting onto $\mathbf{r}_{\text{ref}}$ and using $(\mathbf{W}\mathbf{v}) \cdot \mathbf{r} = \mathbf{v} \cdot (\mathbf{W}^\top \mathbf{r})$ gives

$$\begin{aligned} A &= \sum_l \mathbf{a}^{(l)} \cdot \mathbf{r}_{\text{ref}} \\ &= \sum_{l,h,k} \alpha_h^{(l)}(k) \left(\mathbf{W}_{O,h}^{(l)} \mathbf{v}_h^{(l)}(k) \right) \cdot \mathbf{r}_{\text{ref}} \\ &= \sum_{l,h,k} \alpha_h^{(l)}(k) \left(\mathbf{v}_h^{(l)}(k) \cdot \mathbf{u}_h^{(l)} \right), \end{aligned} \tag{8}$$

where $\mathbf{u}_h^{(l)} = \mathbf{W}_{O,h}^{(l)\top} \mathbf{r}_{\text{ref}}$ is the read-out axis obtained by mapping the safety direction into the head's value space through the transpose of its output matrix. The signed read-out $c_h^{(l)}(k) = \mathbf{v}_h^{(l)}(k) \cdot \mathbf{u}_h^{(l)}$ is interpreted as the token's risk concentration for that layer and head; multiplying it by $\alpha_h^{(l)}(k)$ gives its actual contribution to A. Restricting the sum to one head yields $A = \sum_k \alpha(k) c(k)$, the single-head notation used in Appendix E.2.

C Detailed Derivation of the MLP-Pathway Decomposition

This appendix provides the step-by-step derivation of the MLP-pathway decomposition presented in Section 3.2, starting from the architectural primitives of the Qwen3 decoder layer (Yang et al., 2025).

SwiGLU feed-forward network. Modern Transformer models, including Qwen3, replace the standard two-matrix FFN with a three-matrix gated variant called SwiGLU (Shazeer, 2020). For layer l, the FFN output is

$$\mathrm{FFN}^{(l)}(\mathbf{x}) = \mathbf{W}^{(l)}_{\mathrm{down}}\Big(\mathrm{SiLU}\big(\mathbf{W}^{(l)}_{\mathrm{gate}}\mathbf{x}\big) \odot \mathbf{W}^{(l)}_{\mathrm{up}}\mathbf{x}\Big), \tag{9}$$

where $\mathbf{W}^{(l)}_{\mathrm{gate}}, \mathbf{W}^{(l)}_{\mathrm{up}} \in \mathbb{R}^{d_{ff}\times d}$ are the gate and value projection matrices, $\mathbf{W}^{(l)}_{\mathrm{down}} \in \mathbb{R}^{d\times d_{ff}}$ is the down-projection matrix, $\odot$ denotes element-wise multiplication, and $\mathrm{SiLU}(z) = z\cdot\sigma(z)$ with σ being the sigmoid function. The gate branch $\mathrm{SiLU}(\mathbf{W}^{(l)}_{\mathrm{gate}}\mathbf{x})$ acts as a data-dependent filter that controls which neurons are activated for a given input.

Neuron-by-neuron expansion. We write the n-th row of $\mathbf{W}^{(l)}_{\mathrm{gate}}$ as $\mathbf{w}^{(l)}_{\mathrm{gate},n}$ and similarly for $\mathbf{W}^{(l)}_{\mathrm{up}}$, and the n-th column of $\mathbf{W}^{(l)}_{\mathrm{down}}$ as $\mathbf{w}^{(l)}_{\mathrm{down},n}$. Equation 9 can then be rewritten as a sum over all d_{ff} neurons:

$$\mathrm{FFN}^{(l)}(\mathbf{x}) = \sum_{n=1}^{d_{ff}} \underbrace{\mathrm{SiLU}\big(\mathbf{w}^{(l)}_{\mathrm{gate},n}\cdot\mathbf{x}\big)\cdot\big(\mathbf{w}^{(l)}_{\mathrm{up},n}\cdot\mathbf{x}\big)}_{g^{(l)}_n(\mathbf{x})}\ \mathbf{w}^{(l)}_{\mathrm{down},n}, \tag{10}$$

where $g^{(l)}_n(\mathbf{x})$ is the *gated activation* of neuron n: a scalar that depends on the input $\mathbf{x}$ through both the gate and the value projections.

Projection onto the safety direction. The MLP-pathway write along $\mathbf{r}_{\mathrm{ref}}$ at layer l is $\mathbf{m}^{(l)}\cdot\mathbf{r}_{\mathrm{ref}} = \mathrm{FFN}^{(l)}(\mathbf{x})\cdot\mathbf{r}_{\mathrm{ref}}$. Substituting Equation 10:

$$\mathbf{m}^{(l)}\cdot\mathbf{r}_{\mathrm{ref}} = \sum_{n=1}^{d_{ff}} g^{(l)}_n(\mathbf{x})\ \underbrace{\mathbf{w}^{(l)}_{\mathrm{down},n}\cdot\mathbf{r}_{\mathrm{ref}}}_{\kappa^{(l)}_n}, \tag{11}$$

where $\kappa^{(l)}_n = \mathbf{w}^{(l)}_{\mathrm{down},n}\cdot\mathbf{r}_{\mathrm{ref}}$ is the neuron's *safety-writing strength per unit activation*: a signed coefficient measuring the alignment of its output column with the safety direction. The product $\kappa^{(l)}_n g^{(l)}_n(\mathbf{x})$ is its actual safety-direction contribution. For a positive threshold τ, neurons with $\kappa^{(l)}_n > \tau$ are termed *safety-encoding neurons*, referring to their positive output alignment. For fixed model parameters and a fixed reference direction, all $\kappa^{(l)}_n$ remain constant across input lengths. Consequently, changes in $M = \sum_l\sum_n \kappa^{(l)}_n g^{(l)}_n(\mathbf{x}^{(l)})$ are determined by changes in the weighted aggregate of gated activations. This decomposition does not require that any particular subset of neurons dominate M.

RMSNorm and the MLP input. In each Qwen3 decoder layer, the FFN input is obtained by applying RMSNorm (Zhang & Sennrich, 2019) to the post-attention residual:

$$\mathbf{x}^{(l)} = \mathrm{RMSNorm}\big(\mathbf{h}^{(l)}\big) = \boldsymbol{\gamma}^{(l)} \odot \frac{\mathbf{h}^{(l)}}{\sqrt{\mathrm{RMS}(\mathbf{h}^{(l)})^2 + \epsilon}}, \quad \mathrm{RMS}(\mathbf{h}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d} h_i^2}, \tag{12}$$

where $\mathbf{h}^{(l)}$ is the residual stream after the attention sub-layer, $\boldsymbol{\gamma}^{(l)}$ is the learnable element-wise scale, and $\epsilon > 0$ is added to the mean square inside the square root for numerical stability. This specifies the architectural connection between attention and the MLP; it does not by itself impose a monotonic relation between their safety writes.

Mechanistic intuition and empirical evidence. The exact decomposition ends at Equation 11; the following interpretation concerns how its input-dependent activations may change. Because the residual entering the MLP already contains attention writes, attention-side risk dilution may weaken the effective risk cues available to the gate and up branches. Intuitively, this can reduce the aggregate safety contribution of $g_n^{(l)}(\mathbf{x})$ and amplify the length-induced decline in safety information within the MLP pathway. Table 1 corroborates the predicted pattern of stronger relative attenuation: from length 0 to 8192, the MLP write decreases by 62.3% and 33.3% under relevant and irrelevant contexts, respectively, exceeding the corresponding attention-write decreases of 27.2% and 18.8%. This observed amplification supports declining safety perception, as measured by the MLP safety write, over the tested lengths.

D Analysis of SA-GSPO

This appendix explains how response-level judgment can strengthen risk discrimination without relying on a uniform increase in refusal. We first specify the training objective and a shared-feature model, derive the expected finite-group GSPO gradient, and then compare its effects on risky-input safety and benign-input responsiveness. Finally, we separate changes in responsiveness from changes in answer quality.

D.1 Data construction and training objective

Inputs, labels, and reward. Let $\mathcal{M} = \{\pi_1, \ldots, \pi_K\}$ be a pool of K response-generating models. Given a context c and query q, a model π_m, with $m \in \{1, \ldots, K\}$, generates a candidate response $a \sim \pi_m(\cdot \mid c, q)$. Qwen3Guard assigns a label $b \in \{0, 1\}$, where 0 means safe and 1 means unsafe. The judgment input is $x = (c, q, a)$: the candidate response a is part of the input, whereas the policy's output y is a newly generated judgment sequence. We filter the joint distribution $\mathcal{D}$ of (x, b) so that $\Pr_{\mathcal{D}}(b = 0) = \Pr_{\mathcal{D}}(b = 1) = 1/2$.

The trainable policy is π_θ, with parameters θ. Let $\hat{b}(y)$ be the verdict extracted from y, and let $\mathbf{1}\{\cdot\}$ denote the indicator, which equals one when its condition holds and zero otherwise. The judgment reward is

$$r(x, b, y) = \mathbf{1}\{\hat{b}(y) = b\}, \qquad y \sim \pi_\theta(\cdot \mid x). \tag{13}$$

An invalid verdict receives zero reward. Writing $\mathcal{D}_b$ for the distribution of x conditional on label b, the expected reward is

$$\mathcal{R}(\theta) = \frac{1}{2} \sum_{b=0}^{1} \mathbb{E}_{x \sim \mathcal{D}_b} \left[\Pr_{y \sim \pi_\theta(\cdot|x)} \left(\hat{b}(y) = b \right) \right]. \tag{14}$$

Thus the objective measures agreement with the automatic labels. Always predicting the same valid label achieves accuracy $1/2$; exceeding this value requires predictions that distinguish the two labeled classes.

Group-relative optimization. For each labeled input (x, b), independently sample $G \geq 2$ sequences $y_1, \ldots, y_G$ from the rollout policy $\pi_{\theta_{\text{old}}}$, where θ_{old} denotes the parameters used to collect the group. Let $r_i = r(x, b, y_i)$ and let $|y_i| > 0$ be the number of generated tokens. Define the group mean reward $\bar{r}$, reward standard deviation σ_r, normalized advantage $\widehat{A}_i$, and length-normalized sequence-probability ratio $s_i(\theta)$ by

$$\begin{aligned} \bar{r} &= \frac{1}{G} \sum_{j=1}^{G} r_j, \qquad \sigma_r = \sqrt{\frac{1}{G} \sum_{j=1}^{G} (r_j - \bar{r})^2}, \\ \widehat{A}_i &= \frac{r_i - \bar{r}}{\sigma_r + \delta}, \qquad \delta > 0, \\ s_i(\theta) &= \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)} \right)^{1/|y_i|}. \end{aligned} \tag{15}$$

Here δ stabilizes the advantage denominator. We maximize the GSPO surrogate

$$\mathcal{J}_{\text{SA-GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_i(\theta)\right], \qquad \mathcal{L}_i(\theta) = \min\Big\{s_i(\theta)\widehat{A}_i,\ \text{clip}\big(s_i(\theta), 1-\epsilon, 1+\epsilon\big)\widehat{A}_i\Big\}, \tag{16}$$

where $\mathcal{L}_i$ is the contribution of rollout i, $\epsilon > 0$ is the clipping radius, and $\text{clip}(s, 1-\epsilon, 1+\epsilon)$ restricts s to the indicated interval. The expectation averages over labeled inputs and sampled groups. During a surrogate update, rollout probabilities, rewards, and advantages are held fixed; only the current-policy ratios depend on θ. In a group containing both correct and incorrect judgments, correct judgments have positive advantages. If every judgment receives the same reward, every advantage is zero. SA-GSPO therefore changes the task and reward while retaining the GSPO update rule.

D.2 A SHARED-FEATURE MODEL OF SAFETY AND RESPONSIVENESS

Three coordinates and their roles. We use a reduced model with coordinates $\vartheta = (u, \omega, v)^\top$ to distinguish two ways of changing safety behavior. The scalar u is an input-independent refusal bias during answer generation; ω is the strength of a risk-discrimination feature shared by generation and judgment; and v is a judgment-specific bias toward predicting the unsafe label. These are explanatory coordinates rather than individual network weights. At a fixed context length, let $\alpha > 0$ be the finite magnitude of the risk feature, with values $+\alpha$ and $-\alpha$ for the two classes.

Generation: refusal versus answering. Let H denote risky inputs and B benign inputs. The model groups generated outputs into refusal and substantive answering. For each class, z_H or z_B is the log-odds of refusal versus answering, and f_H or f_B is the corresponding refusal probability:

$$z_H = u + \alpha\omega, \qquad z_B = u - \alpha\omega, \qquad f_H = \sigma(z_H), \qquad f_B = \sigma(z_B), \qquad \sigma(t) = \frac{1}{1+e^{-t}}. \tag{17}$$

Here σ is the sigmoid function, mapping a scalar log-odds t to a probability. Increasing u raises both refusal probabilities. Increasing ω instead raises f_H and lowers f_B: risky and benign inputs become more separated. We call $1 - f_B$ benign-input responsiveness, the probability of giving a substantive answer.

Judgment: unsafe prediction versus correct classification. In the judgment task, H and B refer instead to unsafe and safe candidate responses within their contexts. Using the same feature coordinate ω, define

$$q_H = \Pr(\hat{b} = 1 \mid H) = \sigma(v + \alpha\omega), \qquad q_B = \Pr(\hat{b} = 1 \mid B) = \sigma(v - \alpha\omega). \tag{18}$$

Both quantities are probabilities of predicting *unsafe*: q_H is a correct prediction, whereas q_B is a false positive. In the reduced model, judgment outputs always have one of the two valid verdicts. The classwise correctness probabilities are therefore

$$p_H^J(\vartheta) = q_H = \sigma(v + \alpha\omega), \qquad p_B^J(\vartheta) = 1 - q_B = \sigma(\alpha\omega - v). \tag{19}$$

The superscript J denotes judgment training. Increasing ω improves both correctness probabilities; increasing v favors the unsafe label on both classes. Sharing ω and the feature values across the two tasks is the model's explicit link between judgment learning and generation behavior.

Assumptions for the comparison. The following assumptions isolate that shared-discrimination mechanism:

- *Balanced, symmetric judgment.* The two judgment classes have equal probability and feature values $+\alpha$ and $-\alpha$. Both training alternatives start from the same finite coordinates $\vartheta_0 = (u_0, \omega_0, 0)^\top$. At this point, the two correctness probabilities equal $p_J = \sigma(\alpha\omega_0)$.

- *Fixed features and coordinate-wise updates.* The feature magnitude α is fixed during the update. The positive, finite scales η_u, η_ω, and η_v convert each coordinate's expected gradient into its parameter increment, without cross-coordinate mixing. Since judgment probabilities do not depend on u, judgment training leaves u unchanged in this model.
- *Matched sampling and output lengths.* The comparison uses the same group size $G \geq 2$ and advantage stabilizer $\delta > 0$. Every possible generated output in either task has the same token length $\ell > 0$. This common-length assumption will allow the length factor to be taken outside conditional expectations in the gradient derivation.

These conditions specify a local comparison model. The two-verdict and common-length assumptions simplify the analysis; the training objective above still permits invalid verdicts and variable-length outputs.

The two rewards being compared. Write D for Plain safety training and J for judgment training when identifying the two alternatives. To expose the refusal shortcut, assume that refusal is safe on both generation classes, while substantive answering is safe on benign inputs and unsafe on risky inputs. Let $\rho \in (0, 1]$ be the proportion of risky inputs in the *generation* task; it is distinct from the judgment task's fixed class weight $1/2$. The expected rewards are

$$\begin{aligned} \mathcal{R}_D &= \rho f_H + (1 - \rho), \\ \mathcal{R}_J &= \frac{1}{2}(p_H^J + p_B^J) \\ &= \frac{1}{2}\big[\sigma(v + \alpha\omega) + 1 - \sigma(v - \alpha\omega)\big]. \end{aligned} \tag{20}$$

The constant term $1 - \rho$ arises because both actions receive reward one on benign generation inputs. Thus this reward gives no preference to a useful safe answer over an unnecessary refusal. On risky inputs, reward one corresponds to refusal. This two-action restriction isolates a possible route to increased safety scores; actual policies can also answer risky queries safely, and our experiments measure safety with Qwen3Guard rather than equating it with refusal.

D.3 Finite-group GSPO gradient

The next step connects a binary reward to the *actual group-normalized GSPO gradient*. Throughout this subsection, fix an input x and its reward rule, which does not depend on the parameters. Assume a finite, parameter-independent output support with positive, differentiable probabilities near the rollout parameters $\theta_0 = \theta_{\text{old}}$. Define

$$p_\theta = \Pr_{y \sim \pi_\theta(\cdot|x)}(r = 1), \qquad p = p_{\theta_0} \in (0, 1).$$

Here p is the probability of receiving reward one: it is judgment correctness for the judgment task and risky-input refusal for the two-action generation task. All expectations below use the fixed rollout policy, and all derivatives are evaluated at θ_0.

Step 1: differentiate the sequence ratio. Let $\mathbf{s}(y) = \partial_\theta \log \pi_\theta(y \mid x)|_{\theta_0}$ be the score vector, where ∂_θ denotes the vector of parameter derivatives. It describes how the output's log-probability changes with the parameters and is distinct from the scalar length-normalized sequence ratio (Zheng et al., 2025),

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_0}(y_i \mid x)}\right)^{1/|y_i|}.$$

Holding the reference policy and the sampled output y_i fixed, the chain rule and $s_i(\theta_0) = 1$ give

$$\partial_\theta s_i(\theta)|_{\theta_0} = \frac{s_i(\theta)}{|y_i|}\partial_\theta \log \pi_\theta(y_i \mid x)\Big|_{\theta_0} = \frac{\mathbf{s}(y_i)}{|y_i|}.$$

Since $s_i(\theta_0) = 1$ lies strictly inside the clipping interval, clipping does not affect the gradient at this point. Differentiating the surrogate while holding sampled advantages fixed therefore gives the expected gradient for this input:

$$\mathbf{g}_x = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i \frac{\mathbf{s}(y_i)}{|y_i|}\right]. \tag{21}$$

Step 2: condition on the number of rewarded outputs. Let $N = \sum_{i=1}^{G} r_i$ count the reward-one outputs. Independent sampling gives $N \sim \mathrm{Binomial}(G, p)$. When $N = m$, define the rewarded fraction $t_m = m/G$ and the binary-reward variance $V_m = t_m(1 - t_m)$. The group's mean reward and variance are then exactly t_m and V_m. Reward-one outputs have advantage $(1-t_m)/(\sqrt{V_m}+\delta)$; reward-zero outputs have advantage $-t_m/(\sqrt{V_m} + \delta)$.

For $j \in \{0, 1\}$, define the conditional mean score per token by

$$\boldsymbol{\xi}_j = \mathbb{E}\left[\frac{\mathbf{s}(y)}{|y|} \,\middle|\, r = j\right].$$

In a mixed group, the two classes occupy fractions t_m and $1 - t_m$. Conditional on the reward class, their mean scores remain $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_0$ because the rollouts are independent. Consequently,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{G}\sum_i \widehat{A}_i \frac{\mathbf{s}(y_i)}{|y_i|} \,\middle|\, N = m\right] \\
&= t_m \frac{1 - t_m}{\sqrt{V_m} + \delta}\boldsymbol{\xi}_1 - (1 - t_m)\frac{t_m}{\sqrt{V_m} + \delta}\boldsymbol{\xi}_0 \\
&= \frac{V_m}{\sqrt{V_m} + \delta}(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_0).
\end{aligned} \tag{22}$$

When $m = 0$ or $m = G$, all advantages are zero, consistent with $V_m = 0$.

Step 3: average over all possible groups. Weighting each rewarded count m by its binomial probability yields

$$\begin{aligned}
\mathbf{g}_x &= H_G(p)(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_0), \\
H_G(p) &= \sum_{m=1}^{G-1} \binom{G}{m} p^m (1 - p)^{G-m} \frac{V_m}{\sqrt{V_m} + \delta} > 0.
\end{aligned} \tag{23}$$

The scalar $H_G(p)$ is the expected group weight contributed by mixed-reward groups. It is positive for $G \geq 2$ and $0 < p < 1$, since at least one mixed count has positive probability and positive variance. The mixed-group probability is $1 - p^G - (1 - p)^G$; even balanced training data can produce individual groups with no gradient.

Step 4: relate the mean scores to reward probability. The likelihood-ratio (score-function) identity (Glynn, 1990; Williams, 1992) gives

$$\begin{aligned}
\partial_\theta p_\theta|_{\theta_0} &= \sum_{y: r(y)=1} \pi_{\theta_0}(y \mid x)\mathbf{s}(y) = p\,\mathbb{E}[\mathbf{s}(y) \mid r = 1], \\
\mathbb{E}[\mathbf{s}(y)] &= p\,\mathbb{E}[\mathbf{s}(y) \mid r = 1] + (1 - p)\mathbb{E}[\mathbf{s}(y) \mid r = 0] = 0.
\end{aligned} \tag{24}$$

Here $r(y)$ abbreviates the fixed input's reward rule. The second identity follows by differentiating the normalization $\sum_y \pi_\theta(y \mid x) = 1$. Under the common-length assumption $|y| = \ell$, dividing these conditional score means by ℓ gives

$$\boldsymbol{\xi}_1 = \frac{\partial_\theta p_\theta|_{\theta_0}}{\ell p}, \qquad \boldsymbol{\xi}_0 = -\frac{\partial_\theta p_\theta|_{\theta_0}}{\ell(1 - p)}. \tag{25}$$

Subtracting the two expressions and substituting into Eq. (23) yields

$$\mathbf{g}_x = \Lambda_G(p)\,\partial_\theta p_\theta|_{\theta_0}, \qquad \Lambda_G(p) = \frac{H_G(p)}{\ell p(1 - p)} > 0. \tag{26}$$

Thus $\Lambda_G(p)$ is the positive coefficient relating the expected GSPO gradient to the gradient of reward-one probability. For a fixed input, group normalization changes the gradient magnitude but not its direction under the common-length assumption. For $G = 2$, direct substitution gives $H_2(p) = p(1 - p)/(1 + 2\delta)$ and $\Lambda_2(p) = 1/[\ell(1 + 2\delta)]$. For larger groups, the coefficient generally depends on p and must be retained when averaging different inputs. If rewards are constant ($p = 0$ or $p = 1$), all advantages are zero and the gradient is zero directly, without using the fractions above. With unequal output lengths, Eq. (23) remains valid, while the reduction to Eq. (26) requires additional conditions.

D.4 CONDITIONAL SAFETY–RESPONSIVENESS COMPARISON

Update notation. We now apply the gradient result to the reduced coordinates $\vartheta = (u, \omega, v)^\top$. For a training alternative $T \in \{D, J\}$, let $\mathbf{g}^T = (g_u^T, g_\omega^T, g_v^T)^\top$ be its expected GSPO gradient, averaged over that task's input classes at the common starting point $\vartheta_0 = (u_0, \omega_0, 0)^\top$. A superscript $+$ denotes the state after one update:

$$\vartheta_T^+ = \vartheta_0 + \mathrm{diag}(\eta_u, \eta_\omega, \eta_v)\mathbf{g}^T. \tag{27}$$

Here diag forms a diagonal matrix from the three update scales, and $\top$ denotes transpose, so the coordinates and gradients are column vectors. Thus $\Delta u^T = \eta_u g_u^T$, and similarly for ω and v. The notation $\Delta z_H^T = z_H(\vartheta_T^+) - z_H(\vartheta_0)$ denotes the resulting risky-input logit change; Δz_B^T is defined analogously. In what follows, unmarked probabilities and their derivatives are evaluated at ϑ_0.

Proposition 1 (discrimination versus blanket conservatism). Under the model and comparison assumptions of Appendix D.2, let $f_H = \sigma(u_0 + \alpha\omega_0)$ and $p_J = \sigma(\alpha\omega_0)$. Define

$$D = \frac{\rho H_G(f_H)}{\ell} > 0, \qquad J = \frac{H_G(p_J)}{\ell} > 0. \tag{28}$$

These scalars measure the expected gradient strengths for Plain safety RL and judgment RL, respectively; they are not probabilities. The update in Eq. (27) gives

$$\begin{aligned} \Delta z_H^D &= D(\eta_u + \alpha^2\eta_\omega), & \Delta z_B^D &= D(\eta_u - \alpha^2\eta_\omega), \\ \Delta z_H^J &= J\alpha^2\eta_\omega, & \Delta z_B^J &= -J\alpha^2\eta_\omega. \end{aligned} \tag{29}$$

Both updates increase risky-input refusal and hence safety in this two-action model. Plain safety RL increases benign refusal when $\eta_u > \alpha^2\eta_\omega$, leaves it unchanged at equality, and decreases it otherwise. Judgment RL decreases benign refusal under the stated symmetry and coordinate-separation assumptions. The logit increments are exact for this single expected-gradient update because the reduced logits are linear in the coordinates at fixed α; they provide a local explanatory comparison for the full network.

Proof, part 1: the Plain safety gradient. On benign inputs, both actions receive reward one, so every rollout group has zero advantages. Only risky inputs contribute a gradient, and their reward-one probability is f_H. Differentiating Eq. (17) gives

$$\partial_u f_H = f_H(1 - f_H), \qquad \partial_\omega f_H = \alpha f_H(1 - f_H), \qquad \partial_v f_H = 0. \tag{30}$$

By Eq. (26), the input-averaged gradient includes the risky-input probability ρ:

$$\begin{aligned} g_u^D &= \rho\Lambda_G(f_H) f_H(1 - f_H) = \frac{\rho H_G(f_H)}{\ell} = D, \\ g_\omega^D &= \alpha D, \qquad g_v^D = 0. \end{aligned} \tag{31}$$

Consequently, $\Delta u^D = \eta_u D$ and $\Delta\omega^D = \eta_\omega\alpha D$. Substituting these increments into $\Delta z_H^D = \Delta u^D + \alpha\Delta\omega^D$ and $\Delta z_B^D = \Delta u^D - \alpha\Delta\omega^D$ proves the first row of Eq. (29).

Proof, part 2: the judgment gradient. Start with the general correctness functions in Eq. (19), differentiate them, and then evaluate at $v = 0$. At this point $p_H^J = p_B^J = p_J$, but their derivatives with respect to the label bias have opposite signs:

$$\begin{aligned} \partial_\omega p_H^J &= \partial_\omega p_B^J = \alpha p_J(1 - p_J), \\ \partial_v p_H^J &= p_J(1 - p_J), \qquad \partial_v p_B^J = -p_J(1 - p_J). \end{aligned} \tag{32}$$

Equal correctness probabilities give the same GSPO coefficient $\Lambda_G(p_J)$ for both classes. Combining this coefficient with the class weights $1/2$ yields

$$\begin{aligned} g_v^J &= \tfrac{1}{2}\Lambda_G(p_J)\big[p_J(1 - p_J) - p_J(1 - p_J)\big] = 0, \\ g_\omega^J &= \tfrac{1}{2}\Lambda_G(p_J)\big[\alpha p_J(1 - p_J) + \alpha p_J(1 - p_J)\big] \\ &= \frac{\alpha H_G(p_J)}{\ell} = \alpha J, \qquad g_u^J = 0. \end{aligned} \tag{33}$$

Thus the label-bias gradients cancel, the discrimination gradients add, and the generation-bias gradient is zero because the judgment probabilities do not depend on u. The update is $\Delta\omega^J = \eta_\omega \alpha J$ and $\Delta u^J = \Delta v^J = 0$. Substituting into the generation logits proves the second row of Eq. (29). Finally, the sigmoid is strictly increasing, so each logit change has the same sign as its refusal-probability change. This completes the proof.

Interpretation: where the two factors of α arise. One factor of α appears when differentiating reward probability with respect to ω; the other appears when a change in ω is converted back into a generation-logit change. Plain safety RL also updates the input-independent bias u, whereas symmetric judgment training updates the shared discrimination coordinate without changing either bias at the comparison point. If dilution reduces α in this fixed coordinate system while the update scales remain unchanged, the condition $\eta_u > \alpha^2 \eta_\omega$ becomes easier to satisfy. Judgment's beneficial update also contains the factor α^2, so this comparison concerns the direction of the safety–responsiveness trade-off rather than a guarantee of efficient learning from arbitrarily weak signals.

Allowing imperfect task separation. The mechanism can tolerate a nonzero change in the generation bias. At fixed α, a judgment update with coordinate increments Δu and $\Delta\omega$ gives

$$\Delta z_H = \Delta u + \alpha \Delta\omega, \qquad \Delta z_B = \Delta u - \alpha \Delta\omega. \tag{34}$$

For $\Delta\omega > 0$, both favorable signs, $\Delta z_H > 0$ and $\Delta z_B < 0$, hold exactly when $|\Delta u| < \alpha \Delta\omega$. In words, the discrimination-induced change must exceed the magnitude of the shared refusal-bias shift. This provides a quantitative condition for extending the mechanism beyond complete coordinate separation.

D.5 General performance and the scope of the mechanism

Responsiveness and answer quality are separate factors. Consider a fixed distribution of benign evaluation inputs on which refusal receives score zero. Let f_B be the refusal probability on this distribution, $Q \in [0,1]$ the conditional success probability given a substantive answer, and $U \in [0,1]$ the overall success probability. Assume the probability of answering is positive so that Q is defined. Then

$$U = (1 - f_B)Q, \qquad dU = -Q\, df_B + (1 - f_B)\, dQ. \tag{35}$$

The differential separates two effects: changing how often the model answers and changing how successful its answers are. Holding $Q > 0$ fixed, increased benign refusal lowers performance, while reduced refusal raises it. This is the responsiveness channel identified by Proposition 1; changes in conditional answer quality require separate consideration.

A finite-update bound with quality drift. Let f_B^+, Q^+, and U^+ denote the corresponding post-update quantities on the same input distribution, again with positive answer probability. Subtracting the two success rates and adding and subtracting $(1 - f_B^+)Q$ gives the exact decomposition

$$\begin{aligned} U^+ - U &= (1 - f_B^+)Q^+ - (1 - f_B)Q \\ &= \underbrace{Q(f_B - f_B^+)}_{\text{responsiveness effect}} + \underbrace{(1 - f_B^+)(Q^+ - Q)}_{\text{answer-quality effect}}. \end{aligned} \tag{36}$$

Suppose $|Q^+ - Q| \le \varepsilon_Q$, where $\varepsilon_Q \ge 0$ bounds the change in conditional answer quality. If refusal does not increase, then $Q(f_B - f_B^+) \ge 0$, and hence

$$U^+ - U \ge Q(f_B - f_B^+) - (1 - f_B^+)\varepsilon_Q \ge -\varepsilon_Q.$$

Therefore,

$$f_B^+ \le f_B \quad \Longrightarrow \quad U^+ \ge U - \varepsilon_Q. \tag{37}$$

This is a bound on possible performance loss, rather than an unconditional increase in U. A sufficient condition for $U^+ \ge U$ is $Q(f_B - f_B^+) \ge (1 - f_B^+)\varepsilon_Q$: the responsiveness benefit then covers the largest allowed loss in answer quality. Evaluating these two factors on the same input distribution makes the comparison interpretable, even though the models may answer different subsets. The analysis thus isolates a mechanism for preserving responsiveness, while conditional answer quality and the associated MLP-write behavior are assessed empirically.

E Analysis of SS-OPSD

E.1 Segment-summary self-distillation and the EMA anchor

Let c_{sum} be a fixed system prompt that instructs segment-by-segment summarization of the input inside the reasoning trace. The teacher and student use the same model with and without this prompt, respectively. Trajectories y are drawn on-policy from the prompt-free student $\pi_{\theta_{\text{old}}}(\cdot \mid x)$. Along each student-generated prefix $y_{<t}$, the prompted teacher provides the distribution $\pi_{\theta_{\text{old}}}(\cdot \mid c_{\text{sum}}, x, y_{<t})$, which the student $\pi_\theta(\cdot \mid x, y_{<t})$ matches using the forward token-level KL defined in Section 4.2. The teacher snapshot is held fixed during each update, and only the student parameters are optimized. The EMA anchor is updated as $\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta$ with decay $\mu \in [0, 1)$ and is not itself optimized; on benign long-context inputs (LongAlign), the policy is regularized toward it using the objective defined in that section. The self-distillation term transfers the summarization behavior, and the EMA term limits drift in general behavior. Neither term guarantees capability preservation, which is assessed empirically.

E.2 Attention-write model and assumptions

Token-level normalization. We analyze one representative attention head at the safety-decision position $t^\star$. The superscript $r \in \{0, \text{ss}\}$ distinguishes the no-summary baseline from the summary-containing run. In each run, $z_i^{(r)}$ is the attention score of key token i before the softmax, and $\mathcal{K}^{(r)}$ is the set of keys visible at that position. We normalize separately in the two runs:

$$\alpha^{(r)}(i) = \frac{e^{z_i^{(r)}}}{\sum_{j \in \mathcal{K}^{(r)}} e^{z_j^{(r)}}}, \qquad \sum_{i \in \mathcal{K}^{(r)}} \alpha^{(r)}(i) = 1. \tag{38}$$

Thus $\alpha^{(0)}$ and $\alpha^{(\text{ss})}$ need not be the same distribution. For the following within-run definitions, we suppress r and the length L to simplify notation. Let $c(i) = \mathbf{v}(i) \cdot \mathbf{u}$ be the signed value-space read-out derived in Appendix B. We interpret $c(i)$ as the token's risk concentration for this head; the attention-weighted contribution is $\alpha(i)c(i)$. Summing these contributions gives the single-head write, denoted A in this appendix:

$$A = \sum_i \alpha(i)c(i). \tag{39}$$

Group aggregation after normalization. We partition the visible keys into risky-query tokens $\mathcal{H}$, summary tokens $\mathcal{S}$, and all remaining background keys $\mathcal{N}$. The background includes neutral context and any non-summary thinking or template tokens. For each group $\mathcal{I}$, define

$$\begin{aligned} m_{\mathcal{I}} &= \sum_{i \in \mathcal{I}} \alpha(i), \\ c_{\mathcal{I}} &= \begin{cases} \displaystyle\sum_{i \in \mathcal{I}} \frac{\alpha(i)}{m_{\mathcal{I}}} c(i), & m_{\mathcal{I}} > 0, \\ 0, & m_{\mathcal{I}} = 0. \end{cases} \end{aligned} \tag{40}$$

Here $m_{\mathcal{I}}$ is the group's *total* attention mass, whereas $c_{\mathcal{I}}$ is its attention-weighted *average* concentration: the weights $\alpha(i)/m_{\mathcal{I}}$ sum to one. If a per-token mean attention is desired, it is $m_{\mathcal{I}}/|\mathcal{I}|$ for a nonempty group, not the mass used in the write. The zero-mass convention covers empty groups, including the absent summary group in the baseline. In every case,

$$m_{\mathcal{I}} c_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \alpha(i)c(i). \tag{41}$$

This regrouping is exact and allows different logits and concentrations within a group. No averaging of logits before the softmax is required.

Assumptions for comparing the two runs. The regrouping identity alone does not specify how summarization changes attention or values. To isolate the summary contribution, we make the following reduced-model assumptions. First, the background-group concentration is set to exactly zero

in both runs. This is an idealization, not a claim that every non-summary token in a real model carries no risk information. Second, the query concentration is a finite positive constant shared by both runs and all lengths:

$$c_{\mathcal{H}}^{(0)}(L) = c_{\mathcal{H}}^{(\mathrm{ss})}(L) = c_q > 0. \tag{42}$$

This cross-length condition is needed for the baseline write to inherit the decay rate of its attention mass. Third, let $m_q^{(0)}(L) = m_{\mathcal{H}}^{(0)}(L)$, $m_q(L) = m_{\mathcal{H}}^{(\mathrm{ss})}(L)$, and $m_s(L) = m_{\mathcal{S}}^{(\mathrm{ss})}(L)$. We assume that the summary leaves the query/background allocation unchanged within the remaining attention budget:

$$m_q(L) = \big(1 - m_s(L)\big) m_q^{(0)}(L). \tag{43}$$

For example, if the summary receives 20% of the attention, the original groups retain 80% of their respective baseline masses. This rescaling is an explicit comparison assumption, not a consequence of the group definitions. Writing c_s for the summary group's actual attention-weighted concentration at the length under consideration gives

$$\begin{aligned} A_{\mathrm{base}} &= m_q^{(0)} c_q, \\ A_{\mathrm{ss}} &= m_q c_q + m_s c_s \\ &= (1 - m_s) A_{\mathrm{base}} + m_s c_s. \end{aligned} \tag{44}$$

The last line makes the comparison transparent: the new write is a weighted average of the old write and the summary concentration.

Baseline dilution as length grows. We assume $m_q^{(0)}(L) = \Theta(1/L)$ as $L \to \infty$. One sufficient softmax setting is that, relative to a common reference logit, the query has $\Theta(1)$ total unnormalized weight while the background has $\Theta(L)$ total unnormalized weight. Here an unnormalized weight is the exponential of an individual token's relative logit, not the exponential of an averaged logit. The background condition holds, for example, with $\Theta(L)$ background tokens whose average unnormalized weight is $\Theta(1)$. Since c_q is fixed and positive, the baseline write then satisfies $A_{\mathrm{base}}(L) = \Theta(1/L) \to 0$. These are asymptotic statements, not formulas to evaluate at $L = 0$; that notation denotes the finite reference input without additional context.

The following additional assumptions describe what the summaries must retain across lengths.

(B1) Summaries retain risk information. For simplicity, we model the actual summary-group concentration as a finite, length-independent value $c_s > 0$. Individual summary tokens need not have equal concentrations; only their attention-weighted group average is held fixed in this model.

(B2) Thinking retains attention. Let $\mathcal{T}$ be all generated thinking tokens in the summary-containing run, including $\mathcal{S} \subseteq \mathcal{T}$. Define $m_{\mathrm{think}}(L) = \sum_{i \in \mathcal{T}} \alpha^{(\mathrm{ss})}(i)$. Assume there is a length-independent $\beta > 0$ with $m_{\mathrm{think}}(L) \geq \beta$ for every $L \geq 0$. The constant β is a uniform lower bound, not a limiting value. Table 2 motivates this assumption over finite lengths but does not establish it at arbitrarily large lengths or for every summary-containing run.

(B3) Summaries retain a share of thinking attention. Assume $m_s(L) \geq \phi m_{\mathrm{think}}(L)$ for a length-independent $\phi \in (0, 1]$. This is a share of *attention*, not a share of token count. Together, (B2) and (B3) imply $m_s(L) \geq \phi\beta > 0$. This lower bound is an additional assumption; neither generating some summary tokens nor applying the softmax guarantees it.

E.3 Proposition 2 and proof

Proposition 2 (conditions for an improved and non-vanishing attention write). Under the comparison assumptions in Appendix E.2, the following statements hold.

(i) Improvement at a given length. At any length L with $m_s(L) > 0$,

$$A_{\mathrm{ss}}(L) > A_{\mathrm{base}}(L) \iff c_s > m_q^{(0)}(L) c_q. \tag{45}$$

If $c_s \geq c_q$, the write cannot decrease: it increases strictly unless $c_s = c_q$ and $m_q^{(0)}(L) = 1$. The latter equality means that the baseline assigns all attention to the query. Under the baseline

dilution condition and (B1), strict improvement holds at every sufficiently large length with positive summary mass. This conclusion does not require a uniform lower bound on m_s.

(ii) A non-vanishing write. Under (B1)–(B3),

$$\begin{aligned} A_{\mathrm{ss}}(L) &\geq \phi\beta c_s > 0 \qquad \text{for every } L \geq 0, \\ A_{\mathrm{base}}(L) &= \Theta(1/L) \to 0 \quad \text{as } L \to \infty. \end{aligned} \tag{46}$$

The first line follows from the summary concentration and attention bounds; the second additionally uses the baseline dilution condition. If the reference writes $A_{\mathrm{ss}}(0)$ and $A_{\mathrm{base}}(0)$ are finite and positive, the corresponding retention ratios $R_{\mathrm{ss}}(L) = A_{\mathrm{ss}}(L)/A_{\mathrm{ss}}(0)$ and $R_{\mathrm{base}}(L) = A_{\mathrm{base}}(L)/A_{\mathrm{base}}(0)$ have a positive lower bound and converge to zero, respectively.

Proof of (i): compare a new group with the old average. At a fixed length, suppress L and subtract the baseline from Eq. (44):

$$\begin{aligned} A_{\mathrm{ss}} - A_{\mathrm{base}} &= (1 - m_s)A_{\mathrm{base}} + m_s c_s - A_{\mathrm{base}} \\ &= m_s\big(c_s - A_{\mathrm{base}}\big) \\ &= m_s\big(c_s - m_q^{(0)} c_q\big). \end{aligned} \tag{47}$$

Since $m_s > 0$, the sign is determined by $c_s - A_{\mathrm{base}}$, proving Eq. (45). The summary need not have a larger concentration than the original query; it need only exceed the query's *diluted contribution* A_{base}. For example, $m_q^{(0)} = 0.1$, $c_q = 1$, $m_s = 0.2$, and $c_s = 0.4$ give $A_{\mathrm{base}} = 0.1$ and $A_{\mathrm{ss}} = 0.8 \times 0.1 + 0.2 \times 0.4 = 0.16$.

If $c_s \geq c_q$, then

$$c_s - m_q^{(0)} c_q = (c_s - c_q) + (1 - m_q^{(0)}) c_q \geq 0. \tag{48}$$

Both terms are nonnegative because $0 < m_q^{(0)} \leq 1$ and $c_q > 0$. They vanish together exactly when $c_s = c_q$ and $m_q^{(0)} = 1$, giving the stated equality case. Finally, the dilution condition and fixed c_q imply $A_{\mathrm{base}}(L) \to 0$. Under (B1), the fixed positive c_s therefore exceeds $A_{\mathrm{base}}(L)$ beyond some L_0. A positive but arbitrarily small m_s then gives a strict, possibly arbitrarily small, improvement; it does not by itself give a non-vanishing write.

Proof of (ii): retain both concentration and attention. The query contribution is nonnegative, so

$$A_{\mathrm{ss}}(L) = m_q(L) c_q + m_s(L) c_s \geq m_s(L) c_s. \tag{49}$$

Using $c_s > 0$ from (B1), followed by the attention bounds (B3) and (B2), gives

$$\begin{aligned} A_{\mathrm{ss}}(L) &\geq m_s(L) c_s \\ &\geq \phi\, m_{\mathrm{think}}(L) c_s \\ &\geq \phi\beta c_s > 0. \end{aligned} \tag{50}$$

All three constants are independent of L, so the lower bound cannot decay with input length. This step only needs a nonnegative query contribution and (B1)–(B3), not the baseline dilution rate or the rescaling relation. Separately, the baseline dilution condition gives $A_{\mathrm{base}}(L) = \Theta(1/L) \to 0$. Dividing by the assumed finite, positive reference writes yields

$$R_{\mathrm{ss}}(L) \geq \frac{\phi\beta c_s}{A_{\mathrm{ss}}(0)} > 0, \qquad R_{\mathrm{base}}(L) \to 0. \tag{51}$$

The reference $L = 0$ is handled by these finite values, not by substituting zero into $1/L$. This completes the proof.

Consequence for the safety margin. Equation (1) decomposes the safety margin as $S = \mathrm{const} + A + M$. The proposition shows that segment summaries can replace a vanishing attention contribution with a positive one under the stated assumptions. Holding the other terms fixed, this increases S and therefore favors safe behavior under the margin–behavior monotonicity used in Section 3. In the full network, however, other heads, layers, and the MLP term M may also change, so the proposition alone does not guarantee an increase in the total safety margin. Because the MLP receives a residual stream that already contains the attention write, preserving risk information in A can also alleviate downstream suppression of safety-encoding neurons; we treat this as a qualitative mechanism rather than a separate theorem.

E.4 SCOPE

The proposition provides a head-level account of how segment summaries affect the attention-pathway write. To isolate this mechanism, the comparison uses zero background write, preserved query concentration, and proportional rescaling of the original groups' attention masses. These modeling assumptions complement the exact aggregation of normalized token-level attention. The formulation also accommodates a shared nonzero background concentration c_n: under the same rescaling and concentration-preservation assumptions, the baseline becomes

$$A_{\text{base}} = m_q^{(0)} c_q + (1 - m_q^{(0)}) c_n,$$

while the difference remains $m_s(c_s - A_{\text{base}})$. The improvement criterion is therefore expressed relative to the full baseline write, whose long-context behavior depends on both query and background contributions. Changes in the original groups' concentrations or relative attention allocation can be incorporated as additional terms in this comparison.

The criterion $c_s > A_{\text{base}}$ highlights the role of the summary's risk concentration relative to the existing attention-weighted signal. Under the zero-background dilution model, even a small, fixed positive concentration eventually exceeds the declining baseline, with the gain scaled by the summary attention mass. The stronger non-vanishing-write result additionally uses (B1)–(B3) to retain both risk concentration and attention across lengths. The above analysis provides a mechanistic explanation of the attention-side benefit, while Figure 2 empirically validates that SS-OPSD mitigates risk-attention dilution and improves attention-pathway retention.

F SEGMENT-SUMMARY SYSTEM PROMPT

The following is an example of the system prompt used by the teacher during segment-summary on-policy self-distillation. It instructs the model to distinguish the user's actual request from the long document, construct segment-level summaries and a question–evidence checklist within the reasoning trace, and then answer the request without exposing these intermediate structures. During distillation, the teacher receives the long-context input together with this prompt, whereas the student receives the same long-context input without it.

```
You are an analytical assistant that works with very long
documents.

Before replying, think privately inside <think> ...
</think> and work through these three steps there, in
order:

Step 1 - Locate the request:  Quote the user's actual
request verbatim, and state whether you found it at the
beginning or at the end of the message.  Then say in one
sentence what the user is actually asking you to produce.

Step 2 - Segmented summary:  Read the document in order and
divide it into coherent segments at natural topic or
structural boundaries.  Use between 3 and 8 segments; never
output more than 8.  For every segment write one line:

[Segment <i>] Locator:  <section title, opening phrase, or
topic that identifies where this segment begins> | Summary:
<1-3 sentences on this segment's key facts, entities,
claims and data>

Step 3 - Questions to resolve:  As a numbered list, state
the specific questions you must settle in order to fulfil
the request you quoted in Step 1.  For each question, note
which segment(s) provide the evidence, and mark it "resolved
by context" or "not covered by context".

Then close the thinking block and write your reply.

Rules for the reply (everything after </think>):
```

```
- Reply directly to the request you quoted in Step 1, as an
ordinary answer addressed to the user.

- Never mention or reproduce the steps, the segment list or
the question list.  No "Step 1", no "Stage 3", no "Segmented
Summary" or "Questions To Resolve" headings, no
meta-commentary about your own thinking process.

- The reply must stand on its own:  a reader who never sees
the thinking block must still get a complete answer.

- Keep Steps 1-3 concise so that most of your effort goes
into the reply itself.
```

G TRAINING DETAILS

Data. Our experiments focus on long-context safety. Given the scarcity of such data, we pool samples from two complementary LongSafety corpora (Lu et al., 2025; Huang et al., 2025) to construct a training set of 3,000 examples. We hold out 543 examples from Lu et al. (2025) as a test set disjoint from the training data. All safety labels are assigned by Qwen3Guard (Zhao et al., 2025a) under a strict labeling protocol, where controversial cases are classified as unsafe.

Training configuration. We use ROLL for the GSPO-based experiments, including Plain GSPO and SA-GSPO, and Verl for SS-OPSD. Training uses a batch size of 96 and a learning rate of 10^{-6}, with eight rollouts per input for GSPO and four rollouts per input for SS-OPSD. All experiments are trained on eight H20 GPUs, each with 96 GB of memory. The EMA update rate for SS-OPSD is set to 0.01.

Training duration. Plain GSPO training is capped at two epochs, as training beyond this point reduces long-context general capability by roughly five points. The SA-GSPO judgment stage and the SS-OPSD stage are each trained for one epoch.

Model pool. Our model pool for SA-GSPO comprises three models, DeepSeek-V4 Flash (DeepSeek-AI et al., 2026), Llama-3.1-8B (Grattafiori et al., 2024), and Qwen3-14B (Yang et al., 2025).

Code availability. We will publicly release our training code and configurations to facilitate reproducibility.