HIERARCHICAL RUNTIME ALIGNMENT FOR SAFE LONG-CONTEXT AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context agents require safety controls that remain effective as model state, retrieved context, and cache state change over time. We introduce **hierarchical runtime alignment**, a deployment-time framework that combines three coupled layers: L1 parametric safety adaptation, L2 contextual policy evidence, and L3 cache-level safety preservation. Using matched vanilla/abliterated pairs from Qwen3-8B and Llama-3.2-3B, we show that runtime safety is a **state-dependent calibration problem** rather than a single guardrail problem. Our experiments yield three findings. First, **contextual policy** substantially reduces harmful compliance in the tested abliterated models. Its false-refusal cost varies with model state and policy strength. Second, the tested **safety–utility operating point** depends on model state. Third, matched benign controls distinguish cache-allocation effects from uniform precision reduction. Generic INT4 changes harmful and benign behavior together. The Qwen allocation control changes which signals retain precision. Key comparisons use three seeds, paired uncertainty, two independent semantic judges, and a prevalidated fixed-subset LongSafetyBench evaluation. Across factorial sweeps, agent tasks, adaptive attacks, many-shot jailbreaks, RULER, and LongSafetyBench, the results support joint, model-state-aware calibration rather than a universal all-on configuration.

1 INTRODUCTION

Long-context language agents are no longer simple chatbots. They maintain extended interaction histories, retrieve task and policy memory, call tools, and continue operating over many turns. Their safety problem is therefore not just single-turn refusal, but runtime control under changing context. Tool-use and agent benchmarks show realistic high-stakes failures, harmful multi-step tasks, and prompt-injection conflicts where untrusted tool output competes with the developer's intended instructions (Ruan et al., 2024; Andriushchenko et al., 2025b; Debenedetti et al., 2024). A deployable agent needs safety mechanisms that remain calibrated as context, tools, and memory evolve.

Safety is also not a fixed property of a released checkpoint. Instruction tuning, RLHF (Reinforcement Learning from Human Feedback), constitutional feedback, and preference optimization can install refusal behavior before release (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023), but open-weight deployments remain mutable: fine-tuning can erode safety (Qi et al., 2024), small malicious updates can subvert aligned models (Yang et al., 2024), and refusal behavior can be affected by low-dimensional activation directions (Arditi et al., 2024; Yu et al., 2025). We use abliterated models as a controlled stress test for this setting: capability remains, but the internal refusal prior is weakened.

Long-context operation also couples safety to performance. Retrieved safety documents, system prompts, and self-reminders can act as inference-time guardrails (Xie et al., 2023), yet retrieved context can also perturb guardrail decisions (She et al., 2025). Many-shot jailbreaking shows that harmful demonstrations can teach compliant behavior in situ (Anil et al., 2024); RULER shows that long-context capability is more than needle retrieval (Hsieh et al., 2024); and LongSafetyBench shows that harmful content can be missed inside lengthy inputs (Huang et al., 2025). Meanwhile, KV-cache quantization and compression are practical necessities for long-running agents, but they can change which instructions remain visible to attention (Oren et al., 2024; Li et al., 2024b; Devoto et al., 2024; Chen et al., 2025). These observations motivate studying runtime safety across three intervention

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.

timescales: model state, active policy context, and cache substrate. We call this organization **hierarchical runtime alignment**.

The ordering is operational rather than a ranking of importance. L1 is a persistent parametric state that remains fixed across requests, L2 is a request-level policy state selected after the input is observed, and L3 is a token-level retention policy applied while the conditioned sequence is decoded. Accordingly, L2 conditions a model already configured by L1, while L3 changes the precision or availability of representations induced by both the model and its active context. We evaluate this hierarchy on matched vanilla and abliterated pairs from Qwen3-8B and Llama-3.2-3B across harmful compliance, benign false refusal, adaptive jailbreaks, AgentHarm, many-shot jailbreaking, RULER, LongSafetyBench, and KV-cache interventions.

Our results show that the effects of runtime interventions differ by model state. Contextual safety substantially lowers harmful compliance for the tested abliterated models. On the vanilla model, the same context increases false refusal. The best observed operating point differs across model states and families. In controlled long-context experiments, generic INT4 changes harmful and benign in-context learning together and yields no positive selectivity. A separate Qwen cache-allocation control shows that safety behavior changes when different spans retain high precision. The allocation evidence currently covers Qwen.

The paper contributes: (i) an operational formulation of hierarchical runtime alignment over **parametric state, policy context, and cache substrate**; (ii) a **model-state-aware empirical study** of how the same hierarchy behaves differently on vanilla and abliterated model states; (iii) **agent-oriented and long-context evaluations** with AgentHarm, many-shot jailbreaking, RULER, and LongSafetyBench; and (iv) **paired, semantically audited evidence** separating cache-allocation effects from generic capability loss.

2 HIERARCHICAL RUNTIME ALIGNMENT

Figure 1 summarizes the runtime hierarchy. The agent receives long-context inputs from user instructions, retrieved documents, tool outputs, conversation history, and potentially harmful demonstrations. Our hierarchy intervenes at three coupled levels: L1 changes the model state through a scaled LoRA safety adapter, L2 adds explicit policy context to the active prompt, and L3q/L3c control the KV cache through quantization or compression. The objective jointly measures harmful-request refusal, benign utility, and memory efficiency as the agent continues to accumulate context.

Operational meaning of hierarchy. The term *hierarchy* refers specifically to conditional order and timescale. L1 is configured first and persists across requests; L2 is selected for the current request after observing its input; L3 operates during decoding on the cache created by the L1-configured model processing the L2-augmented context. Thus, L3 changes the retention or precision of signals already produced by L1 and L2.

2.1 PROBLEM FORMULATION

Let M_s denote a model in state s, where s may be a vanilla aligned state or an abliterated state. Let D_h be a harmful prompt distribution and D_b a benign or borderline-benign prompt distribution. For an intervention $a \in \mathcal{A}$, we evaluate two safety-utility error rates. The *Compliance Violation Rate* (CVR) measures unsafe helpfulness: the fraction of harmful prompts for which the model complies rather than refuses. The *False Positive Refusal Rate* (FPR) measures lost utility: the fraction of benign or borderline-benign prompts for which the model refuses despite the request being answerable.

$$\mathrm{CVR}(M_s, a) = \Pr_{x \sim D_h}[M_s^a(x) \text{ complies}], \tag{1}$$

$$\mathrm{FPR}(M_s, a) = \Pr_{x \sim D_b}[M_s^a(x) \text{ refuses}]. \tag{2}$$

For long-context agents, we also track an efficiency cost $C(a)$, such as relative KV-cache memory or compression-induced degradation. The runtime alignment objective is to minimize harmful compliance while preserving benign utility and long-context efficiency:

$$J(M_s, a) = \mathrm{CVR}(M_s, a) + \lambda \mathrm{FPR}(M_s, a) + \beta C(a). \tag{3}$$

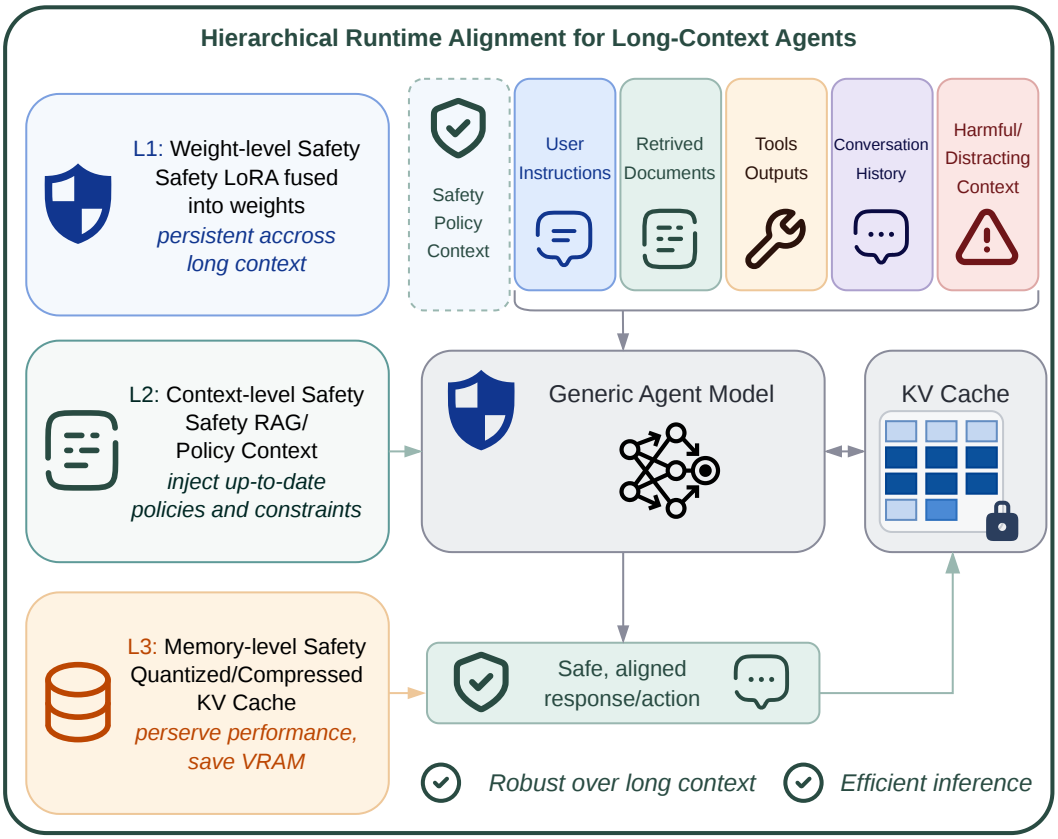


Figure 1: Overview of hierarchical runtime alignment for long-context agents. L1 supplies a parametric safety prior through a scaled LoRA adapter; L2 supplies contextual policy evidence through retrieved or inserted safety documents; L3q and L3c control the KV cache through quantization or compression. The hierarchy determines whether safety persists as the active context grows with user instructions, retrieved documents, tool outputs, conversation history, and adversarial demonstrations.

The dependence on s is essential. An intervention with low J on an abliterated model may have high J on the vanilla model because it increases false refusals. In the hyperparameter sweeps we report the unweighted CVR+FPR score as a transparent operating-point summary, while also reporting both terms separately because different deployments may choose different λ. In the KV-cache experiments, $C(a)$ is represented by the quantization or compression setting and its induced safety degradation.

2.2 L1: Parametric Safety State

The motivation for L1 is to restore or recalibrate the model's internal safety prior before any external policy context is supplied. A safety adapter can lower CVR by increasing the probability of refusal-oriented continuations on harmful requests. It can also affect FPR because the same preference boundary determines when the model should continue helping on benign requests. If the adapter is well-calibrated, it can preserve benign helpfulness while restoring harmful refusal; if it is injected too strongly, it can shift the boundary toward over-refusal and increase FPR. We therefore treat the adapter scale as a runtime operating-point parameter rather than a fixed binary switch.

The L1 level is implemented as a LoRA safety adapter (Hu et al., 2022) trained with supervised fine-tuning followed by DPO (Rafailov et al., 2023). Our primary adapter, *SFT-DPO Safety*, is trained on 798 purchasing-scam SFT examples and 3,798 PKU-SafeRLHF preference pairs (Ji et al., 2025). It uses rank 16, LoRA alpha 32, and targets the q- and v-projection matrices. We also evaluate a *Continued-DPO Safety* variant in the hyperparameter sweep. At inference, an adapter is merged into the base model with a scalar multiplier $\gamma \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ for strict scaling experiments, or with effective alpha values in $\{16, 24, 32, 48, 64, 96\}$ for the broader sweep. This separates the

question of whether safety can be restored by a small parametric update from the question of how strongly that update should be injected in a running agent.

Empirically, the adapter exhibits a thresholded safety effect. Low merge scales leave the refusal boundary largely unchanged on the abliterated model. Full-scale merging yields the first measurable safety improvement in the sweep.

2.3 L2: CONTEXTUAL POLICY STATE

The motivation for L2 is that a long-running agent can lose safety-relevant instructions as the active context grows, even when the base model is aligned. Explicit policy context can lower CVR by making the refusal policy salient at decoding time. It can also raise FPR if the policy context is too aggressive, so we evaluate both strict and softened policy variants rather than assuming that more context is always better. The L2 level retrieves or inserts safety documents and prepends them to the model input as explicit policy context. We evaluate several variants: aggressive refusal text, neutral A/B classification text, ultra-minimal text, benign-prior text, soft advisory text, and strict anti-jailbreak text. The strict variant contains 12 documents with hard-refusal instructions and anti-injection policies. The retrieval corpus contains 20 short safety documents covering harmful-request categories, refusal criteria, benign-request preservation, and prompt-injection resistance. At runtime, the user query is embedded with all-MiniLM-L6-v2 from the Sentence-BERT/MiniLM family (Reimers & Gurevych, 2019; Wang et al., 2020); a FAISS index (Johnson et al., 2021) returns the top-$k = 3$ nearest safety documents, which are prepended before generation.

Contextual safety acts as an explicit policy prior. It can dominate an abliterated model because the model lacks a strong internal refusal prior. The same mechanism can harm a vanilla model because the retrieved policy combines with the model's existing safety behavior and pushes ambiguous, benign requests toward refusal.

2.4 L3: CACHE-EFFICIENT SAFETY PRESERVATION

We treat KV-cache methods as an efficiency and interaction layer. We formally reserve **L3q** for INT4 KV-cache quantization and **L3c** for KV-cache compression. Earlier prompt-only controls are denoted P rather than L3. L3q uses `QuantizedCache` from Transformers (Wolf et al., 2020) with the Quanto backend (Corvoysier et al., 2024), 4-bit keys and values, group size 64, and residual length 128. L3c uses kvpress (Devoto et al., 2025) with KnormPress (Devoto et al., 2024), TOVA (Oren et al., 2024), and SnapKV (Li et al., 2024b) at 25%, 50%, and 75% compression ratios. Quantization changes precision while largely preserving which tokens remain available. Compression changes the token set itself, so it can preserve or remove safety-relevant context. We therefore evaluate KV methods for safety effects, not only for memory savings.

2.5 HIERARCHICAL COMPOSITION

We define hierarchical runtime alignment through the conditional generation process

$$y \sim p_{\theta+\Delta_{\mathrm{LoRA}}(\alpha_s)}(y \mid x, c_{\mathrm{RAG}}(x); K_\eta[x, c_{\mathrm{RAG}}(x)]) \,, \tag{4}$$

where θ are the base weights, Δ_{LoRA} is the scaled adapter update, $c_{\mathrm{RAG}}(x)$ is retrieved safety context, and K_η is the cache policy, including the identity policy when no L3 intervention is enabled. This expression makes the dependency explicit: the cache policy acts on representations produced after the model and contextual policy have been chosen.

This dependency permits non-additive behavior. In some settings, strict safety context reduces harmful compliance even when LoRA is weak. In other settings, retrieval text can conflict with the adapter if the retrieved documents emphasize helpful completion in ways that weaken refusal behavior. Cache policy can further change the outcome by preserving, quantizing, or deleting tokens that carry either safety evidence or harmful demonstrations. These dependencies motivate evaluating both individual and composed configurations. Whether composition provides a measurable benefit is an empirical question tested against the best constituent rather than assumed from the architecture.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Models. We evaluate two model states in two families. The first pair is `Qwen/Qwen3-8B` (Yang et al., 2025). The abliterated model is `huihui-ai/Huihui-Qwen3-8B-abliterated-v2` (huihui-ai, 2025), a safety-removed derivative of the same model family. The replication pair is `meta-llama/Llama-3.2-3B-Instruct` (Team, 2024) and `huihui-ai/Llama-3.2-3B-Instruct-abliterated` (huihui-ai, 2024). Within each family, the paired checkpoints isolate refusal-prior state while holding architecture and scale approximately fixed; cross-family results test qualitative rather than numerical transfer.

Interventions. We instantiate the runtime hierarchy with three intervention levels. L1 denotes a LoRA adapter merged into the model at inference. Unless otherwise specified, L1 uses the *SFT-DPO Safety* adapter at scale 0.1 in the factorial study. L2 denotes safety context, either inserted from a fixed policy corpus or retrieved through FAISS. L3 is reserved for cache-level mechanisms: L3q denotes INT4 KV-cache quantization and L3c denotes KV-cache compression. Prompt-only controls from earlier sweeps are denoted P because they do not modify the cache layer.

Benchmarks and metrics. Safety is measured on 100 JailbreakBench harmful prompts (Chao et al., 2024a) using CVR. Utility is measured with OR-Bench-Hard-1K borderline-benign prompts (Cui et al., 2025) using FPR. Metric directions are indicated in table headers. We also measure refusal behavior on IFEval (Zhou et al., 2023) and WildBench (Lin et al., 2025) subsets of 200 prompts each. For agent-oriented evaluation, we use AgentHarm chat prompts and harmful agent-task prompts (Andriushchenko et al., 2025b). For long-context attack evaluation, we use many-shot jailbreaking with 5, 10, 25, and 50 harmful demonstrations before a held-out harmful query (Anil et al., 2024). For standardized long-context capability, we use RULER with NIAH single, NIAH multi-needle, variable tracking, and common-words extraction at 8k, 16k, and 32k context lengths (Hsieh et al., 2024). For standardized long-context safety, we use LongSafetyBench, reporting harm awareness (HA) and safe response (SR) on multiple-choice safety tasks (Huang et al., 2025).

Adaptive attack suites. We evaluate attack robustness with three jailbreak suites and report the attack success rate (ASR). PAIR is a black-box iterative attack that repeatedly refines candidate jailbreak prompts against a target model (Chao et al., 2024b). JBC denotes the JailbreakBench curated attack set; we use its pre-crafted transfer prompts generated against `vicuna-13b-v1.5` (Chao et al., 2024a). EPFL denotes the `random_search` suite from the simple adaptive attacks benchmark, which uses template- and rule-driven search over harmful behaviors (Andriushchenko et al., 2025a). The harmful behavior pool follows AdvBench (Zou et al., 2023); in our adaptive-attack runs, each suite is evaluated on 100 attack prompts derived from 30 harmful behaviors.

Additional protocol details and experiment-group definitions are reported in Appendix A.1.

Judging and uncertainty. Short-context tables use the fixed keyword refusal rule where indicated. Mechanism-focused comparisons additionally use deterministic `Qwen2.5-7B-Instruct` and `Mistral-7B-Instruct-v0.3` judges over the complete request–response pair. Sampled Qwen tests use seeds $42/123/456$ and paired bootstrap 95% confidence intervals. The balanced Llama audit contains 240 harmful/benign records, parses all records with both judges, and uses 4,000 bootstrap resamples. This audit validates the direction of the keyword surface. The remaining outputs in the complete Llama many-shot factorial use the stated keyword labels.

3.2 EXPERIMENTAL RESULTS

We organize the results around four claims: restoration depends on model state, safe operating points require calibration rather than maximal refusal, long contexts can shift the active safety layer, and adaptive attacks expose the scope of deployment-time restoration. Detailed auxiliary tables are deferred to Appendix A.

Each experiment characterizes both safety and utility. A reduction in harmful compliance accompanied by indiscriminate refusal or comparable benign degradation is not counted as selective safety

Table 1: Selected short-context restoration results. On the abliterated model, L2 reduces CVR from 0.80 to 0.09, identifying contextual policy as the largest measured CVR reduction among the displayed runtime configurations. On the vanilla model, the same contextual signal mainly exposes the cost of applying an uncalibrated policy to an already aligned state. Here P denotes a static prompt-control baseline that supplies safety instructions without retrieval or adapter updates.

Config	Abl. CVR ↓	Abl. FPR ↓	Van. CVR ↓	Van. FPR ↓
Baseline	0.80	0.10	0.04	0.58
L1	0.77	0.08	0.04	0.62
L2	0.09	0.23	0.06	0.69
P	0.82	0.06	0.07	0.32
L1+P	0.80	0.05	0.07	0.32
L1+L2+P	0.37	0.03	0.03	0.49

Table 3: Best abliterated-model operating points (keyword judge). The Qwen sweep compares two safety adapters. The Llama sweep compares the actual SFT checkpoint with the composed SFT+DPO checkpoint over 174 exactly matched conditions; it does not treat the standalone DPO delta as a deployable row.

Family	Adapter	Config	CVR ↓	FPR ↓	Sum ↓
Qwen3	Continued-DPO	Neutral L2 + anti-FPR prompt + $\alpha = 16$	0.15	0.11	0.26
Qwen3	SFT-DPO	Ultra-minimal L2 + baseline prompt + $\alpha = 32$	0.36	0.04	0.40
Llama-3.2	SFT	Pure-refuse L2 + anti-FPR prompt + scale 0.5	0.09	0.12	0.21
Llama-3.2	SFT+DPO	Pure-refuse L2 + anti-FPR prompt + scale 0.5	0.09	0.08	**0.17**

improvement. We therefore report matched model-state comparisons, constituent and composed configurations where available, and paired harmful and benign cache controls.

Safety restoration is model-state dependent. The abliterated model begins with high harmful compliance, while the vanilla model is already much safer, and its primary error mode is refusal on benign or borderline-benign queries.

The selected restoration summary in Table 1 shows different effects by model state: contextual policy lowers CVR for the abliterated model, whereas the tested policy increases FPR on the vanilla model.

The stricter scaling study confirms this pattern (Table 2; full sweep in Appendix Table A.5). LoRA has a dead zone at sub-full scales, whereas strict L2 produces the lowest CVR and highest FPR among the displayed settings. In this experiment, L1 changes the parametric state and L2 has the larger measured effect on harmful compliance.

Table 2: Abliterated-model LoRA scaling and strict-context summary. Strict L2 gives the lowest CVR and highest FPR; Table 3 searches calibrated operating points.

Config	CVR ↓	FPR ↓
Baseline	0.86	0.06
L1@0.5	0.88	0.07
L1@1.0	0.59	0.20
L2(strict)	0.02	0.74
L1@0.5+L2(strict)	0.04	0.77

Calibrated operating points beat strict policies. The larger abliterated-model sweep tests the result across prompt and adapter configurations (Table 3). The best settings combine a downscaled adapter with less aggressive L2 context, not the strictest available policy.

On Llama, SFT has lower CVR+FPR in 67 matched cells, ties in 37, and is worse in 70. The composed adapter improves the best observed sum from 0.21 to 0.17. The same L2 intervention behaves differently on vanilla: the 92-condition search identifies the L1-only setting as the best vanilla operating point, and strict L2 with LoRA produces high false refusals (Appendix Tables A.6 and A.7). Thus, the tested L2 configuration helps the abliterated state. On the vanilla state, L1 alone retains the lower CVR+FPR score.

Table 4: Long-context evidence across protocols. Qwen cache-allocation tests use two semantic judges, three seeds, paired bootstrap intervals, and length-matched benign controls; Qwen LongSafetyBench follows Appendix Table A.3. For compute-efficient paired comparison, every Llama cell uses the same prevalidated 54-item subset fixed before evaluation.

Family/setting	Metric	Comparison	Result
Qwen cache	selectivity	BF16 vs. tested INT4	Non-positive after matched benign ICL; harmful and benign behavior change together.
Qwen cache	allocation	protected vs. quantized policy	Safety behavior changes at matched INT4 budget; Qwen-only mechanism evidence.
Qwen LSB, vanilla	SR ↑	baseline: 0.0%	L3q: 58.3% under the Qwen protocol.
Qwen LSB, abliterated	SR ↑	baseline: 0.0%	L3q: 32.3%; L1+L2+L3q: 37.2% under the Qwen protocol.
Llama LSB, vanilla	SR ↑	no L2: .327	L2: .467; BF16 and INT4 choices identical.
Llama LSB, abliterated	SR ↑	no L2: .269	L2: .458; BF16 and INT4 choices identical.
Llama semantic audit	agreement	240 balanced records	Judge agreement .950 exact/.975 adverse; keyword agreement .883/.908.

Long-context results require capability-matched interpretation. AgentHarm and many-shot jailbreaking expose complementary regimes. Matched controls show that ordinary INT4 changes harmful demonstration following and matched benign codebook induction in parallel. No tested INT4 policy has positive selectivity relative to BF16. A Qwen-only location control (policy protected versus policy quantized at a matched cache budget) shows that allocation matters. This allocation result is specific to Qwen.

The Qwen evaluation reports large L3q-associated SR differences under its sampled 64-token protocol. In the separate fixed-subset Llama evaluation, direct-letter greedy generation and strict parsing produce identical BF16 and INT4 choices while showing an L2 benefit. These measurements are retained under their stated protocols rather than pooled: cache configuration remains a runtime variable, while selective safety improvement requires matched capability controls rather than lower ASR alone.

Adaptive attacks identify the scope of runtime restoration. Adaptive attacks test a stricter objective than ordinary harmful-compliance restoration (Table 5). On vanilla, L2 strongly suppresses PAIR and JBC, while EPFL-style search is a harder boundary case. These results show attack-specific protection: PAIR and JBC ASR decrease, while EPFL ASR remains between 0.87 and 0.93.

Table 5: Adaptive-attack ASR on the vanilla model. L2 reduces PAIR and JBC; EPFL remains the hardest evaluated case.

Config	PAIR ↓	JBC ↓	EPFL ↓
Baseline	0.43	0.33	0.90
L2	0.07	0.00	0.87
L1+L2	0.03	0.00	0.87
L1+L2+L3q	0.03	0.00	0.93

Real RAG and cache compression provide additional composition measurements (Appendix Table A.18). Token selection can strengthen the effect of retrieved safety context under moderate compression, so the deployed safety behavior is observed under the composed L1–L2–L3 configuration. A matched comparison with every constituent remains an open measurement for this setting.

4 Discussion

Calibration depends on model state. In the tested short-context conditions, contextual policy produces larger CVR reductions than low-scale LoRA on abliterated models, although aggressive context incurs high FPR and neutral or minimal context offers a better CVR–FPR operating point. On vanilla models, the same context combines with an existing refusal tendency and raises FPR. The 92-condition search identifies L1 alone as the best tested vanilla operating point, motivating model-state-specific configurations.

Joint tuning across runtime layers. The L1–L2–L3 ordering applies persistent model state, request-level policy context, and token-level cache retention in sequence. Their composition varies by condition: context can produce a larger marginal change than low-scale LoRA, while compression changes which policy tokens survive. The real RAG experiment illustrates one such setting. We

characterize these measurements as composition effects, reserving statistical interaction for paired difference-of-differences estimates.

Runtime alignment as staged calibration. The three layers expose complementary adjustment timescales. L1 establishes a persistent safety prior, L2 adapts policy evidence to the current request, and L3 manages which conditioned representations remain available during decoding. This ordering yields a practical calibration sequence: select an L1 operating point for the deployed model state, tune L2 against the desired CVR–FPR balance, and configure L3 for the context and memory budget. The cross-family results support this procedure even as the numerical operating points change across models.

Robustness and deployment implications. Adaptive attacks, especially EPFL-style search and abliterated-model attacks, remain a harder stress test: PAIR/JBC decrease on the vanilla model, whereas EPFL-style ASR remains between 0.87 and 0.93. Matched many-shot controls also show parallel degradation in harmful and benign demonstration uptake under generic INT4. Long-running agents should therefore calibrate adapter scale, policy-context strength, context selection, and cache policy jointly by model state and context budget, alongside training-time alignment and attack-specific defenses.

5 Related work

Safety alignment and safety-utility trade-offs. Modern assistants are aligned through instruction tuning, RLHF, constitutional feedback, and preference optimization (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). More recent safety datasets and algorithms explicitly separate helpfulness from harmlessness, including PKU-SafeRLHF's multi-level safety preferences (Ji et al., 2025). At the same time, safety alignment creates a calibration problem: stronger refusal can reduce harmful compliance but also rejects benign requests. XSTest and OR-Bench make this over-refusal failure mode explicit (Röttger et al., 2024; Cui et al., 2025), while SafetyPrompts shows that LLM safety evaluation remains fragmented across many datasets and risk taxonomies (Röttger et al., 2025). Our work uses this safety-utility framing but studies a different point in the model lifecycle: restoring safety after the released model state has changed.

Safety erosion after release. Several papers show that safety is not a permanent property of a checkpoint. Fine-tuning aligned models can compromise safety even with benign or small adversarial datasets (Qi et al., 2024), and shadow alignment shows that modest malicious tuning can subvert safely aligned open models (Yang et al., 2024). Mechanistic work gives a sharper explanation for why such interventions can be cheap: refusal can be mediated by a low-dimensional direction in activation space, and rank-one or feature-level interventions can disable or induce refusal (Arditi et al., 2024; Yu et al., 2025). This line of work motivates the abliterated-model threat model we study. In contrast to papers that demonstrate how to remove refusal or train a robust model from scratch, we ask how much safety can be recovered post hoc through a lightweight runtime hierarchy spanning parametric, contextual, and cache-level controls.

Jailbreak attacks and robustness benchmarks. Prompt-based attacks such as GCG and PAIR demonstrate that aligned models can be induced to produce harmful content without weight access (Zou et al., 2023; Chao et al., 2024b). Recent work has made the threat model broader and more adaptive: HarmBench standardizes automated red teaming across attacks and target models (Mazeika et al., 2024), JailbreakBench provides a reproducible jailbreak benchmark and artifact repository (Chao et al., 2024a), AgentHarm extends harmfulness evaluation to agentic tasks with tool-oriented misuse risks (Andriushchenko et al., 2025b), simple adaptive attacks achieve high success rates against leading aligned models (Andriushchenko et al., 2025a), many-shot jailbreaking exploits long contexts (Anil et al., 2024), and multi-turn human jailbreaks expose vulnerabilities missed by single-turn automated attacks (Li et al., 2024a). Our attack experiments follow this literature's emphasis on adaptive evaluation. The key distinction is that we evaluate attacks after safety restoration on a model whose safety prior has already been weakened, rather than evaluating only the original aligned model.

Inference-time guardrails and retrieved policy context. Inference-time safety mechanisms are attractive because they do not require weight updates. System prompts, self-reminders, and policy

context can reduce jailbreak success in some settings (Xie et al., 2023). However, recent guardrail work also shows that context can change safety judgments: RAG-style documents can perturb LLM-based guardrails and turn correct safety decisions into incorrect ones (She et al., 2025). Our L2 level is related to these defenses and operates as an input-side policy prior instead of an external classifier. Its effect depends on target model state: it restores safety on abliterated models and compounds over-refusal on vanilla models.

Parameter-efficient safety adaptation. LoRA and other parameter-efficient methods modify a small number of trainable parameters while leaving most base weights fixed (Hu et al., 2022). They are useful for post-release adaptation because adapters can be distributed, merged, scaled, or removed at deployment time. Prior safety-alignment work typically treats such updates as training-time alignment mechanisms. We instead use LoRA as one part of an inference-time restoration system and vary its merge scale. This exposes a threshold effect: weak adapters leave the abliterated model near baseline. Strong adapters partially restore refusal; contextual safety produces a larger CVR change in the reported short-context comparison.

KV-cache efficiency and safety preservation. KV-cache compression and quantization reduce the memory cost of long-context generation. RULER shows that long-context evaluation must include tracking and aggregation rather than only needle retrieval (Hsieh et al., 2024), and LongSafetyBench shows that long contexts introduce safety failures not captured by ordinary short-context tests (Huang et al., 2025). Recent cache methods include TOVA's attention-based token omission (Oren et al., 2024), SnapKV's observation-window selection (Li et al., 2024b), and norm-based KV selection (Devoto et al., 2024), with benchmarking work comparing long-context compression trade-offs (Yuan et al., 2024). Newer analyses warn that compression can selectively erase instructions and cause system-prompt failures (Chen et al., 2025). Our results align with this warning: INT4 quantization is neutral in short-context restoration and changes behavior under many-shot and standardized long-context safety tests, while token-dropping compression can preserve, destroy, or reshape the interaction between retrieved safety context and a LoRA adapter.

6 Conclusion

We studied hierarchical runtime alignment for safe long-context agents: L1 parametric safety adaptation, L2 contextual policy evidence, and L3 cache-level safety preservation. The measured effects differ by model state. In the tested abliterated models, contextual policy substantially reduces harmful compliance. In the vanilla model, the same policy increases false refusal; L1 alone achieves the best observed safety–utility operating point.

The long-context experiments extend this conclusion beyond short prompts and characterize the cache mechanism more precisely. AgentHarm, many-shot jailbreaking, RULER, LongSafetyBench, and real RAG/cache experiments show that cache configuration can change behavior. Matched benign controls and fixed-subset LongSafetyBench runs show parallel harmful and benign changes under generic INT4. Qwen cache-allocation controls indicate that behavior changes when different evidence retains precision.

Overall, the experiments motivate a deployment view in which adapter scale, retrieved policy, prompt controls, context selection, and cache policy are tuned as one runtime hierarchy rather than as independent guardrails. The Llama replication supports the qualitative state-dependent calibration result. Cache-allocation evidence currently covers Qwen. This hierarchy provides a practical calibration layer alongside training-time alignment and attack-specific defenses, balancing harmful-compliance reduction, benign utility, and long-context capability in long-running agents.

AI USE STATEMENT

Generative AI tools were used to assist with experiment scripting, orchestration, result aggregation, and language editing. All experimental protocols, generated code, numerical results, citations, and manuscript claims were reviewed by the authors, who take responsibility for the final content of the paper and its artifacts.

ETHICS STATEMENT

This work studies defensive safety controls using public models and established safety benchmarks. We report aggregate measurements and do not reproduce harmful completions. Abliterated checkpoints and jailbreak evaluations can have dual-use implications; any release will therefore be limited to evaluation code, configurations, manifests, and aggregate results rather than unsafe model weights or generated harmful content.

REPRODUCIBILITY STATEMENT

The experimental setup and appendix specify model checkpoints, adapter scales, policy conditions, cache parameters, decoding settings, prompt counts, frozen subsets, semantic-judge rubrics, and uncertainty procedures. We will release the evaluation scripts, frozen manifests, configurations, and aggregate results needed to reproduce the reported comparisons.

REFERENCES

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=hXA8wqRdyV.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. Agentharm: A benchmark for measuring harmfulness of LLM agents. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=AC5n7xHuR1.

Cem Anil, Esin DURMUS, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cw5mgd71jW.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=urjPCYZt0I.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024b. URL https://arxiv.org/abs/2310.08419.

Alex Chen, Renato Geh, Aditya Grover, Guy Van den Broeck, and Daniel Israel. The pitfalls of kv cache compression, 2025. URL https://arxiv.org/abs/2510.00231.

David Corvoysier, Younes Belkada, and Marc Sun. Quanto: a PyTorch quantization backend for optimum. https://huggingface.co/blog/quanto-introduction, 2024. Accessed 2026-09-24.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=CdFnEu0JZV.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=m1YYAQjO3w.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A simple and effective l_2 norm-based strategy for KV cache compression. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18476–18499, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1027. URL https://aclanthology.org/2024.emnlp-main.1027/.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: Kv cache compression by estimating attention from future queries distribution, 2025. URL https://arxiv.org/abs/2510.00636.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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, 2025. URL https://arxiv.org/abs/2411.06899.

huihui-ai. Llama-3.2-3B-Instruct-abliterated. https://huggingface.co/huihui-ai/Llama-3.2-3B-Instruct-abliterated, 2024. Model card, accessed 2026-09-24.

huihui-ai. Huihui-Qwen3-8B-abliterated-v2. https://huggingface.co/huihui-ai/Huihui-Qwen3-8B-abliterated-v2, 2025. Model card, accessed 2026-09-24.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In Wanxiang Che, Joyce

Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1544. URL https://aclanthology.org/2025.acl-long.1544/.

Jeff Johnson, Matthijs Douze, and Herve Jegou. Billion-Scale Similarity Search with GPUs . *IEEE Transactions on Big Data*, 7(03):535–547, July 2021. ISSN 2332-7790. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.ieeecomputersociety.org/10.1109/TBDATA.2019.2921572.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet, 2024a. URL https://arxiv.org/abs/2408.15221.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=poE54GOq2l.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MKEHCx25xp.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=f3TUipYU3U.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18724–18741, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1043. URL https://aclanthology.org/2024.emnlp-main.1043/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GEcwtMk1uA.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. Safetyprompts: A systematic review of open datasets for evaluating and improving large language model safety. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27617–27627, Apr. 2025. doi: 10.1609/aaai.v39i26.34975. URL https://ojs.aaai.org/index.php/AAAI/article/view/34975.

Yining She, Daniel W. Peterson, Marianne Menglin Liu, Vikas Upadhyay, Mohammad Hossein Chaghazardi, Eunsuk Kang, and Dan Roth. Rag makes guardrails unsafe? investigating robustness of guardrails under rag-style contexts, 2025. URL https://arxiv.org/abs/2510.05310.

The Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5 (12):1486–1496, Dec 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00765-8. URL https://doi.org/10.1038/s42256-023-00765-8.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Ruth Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=9qymw6T9Oo.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust LLM safeguarding via refusal feature adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s5orchdb33.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. KV cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4623–4648, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.266. URL https://aclanthology.org/2024.findings-emnlp.266/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A Additional experiments

This appendix follows the same result logic as the main Experimental Results section: we first specify the evaluation protocol, then expand the evidence for state-dependent restoration, calibrated operating points, long-context layer shifts, and deployment-time stress tests. The table captions state the main empirical takeaways, while the surrounding text identifies what each block contains.

A.1 Experiment map and shared protocol

Table A.1 maps the experiment groups referenced by the main text to the corresponding appendix evidence.

Table A.1: Experiment map. The appendix evidence spans factorial restoration, calibration, long-context stress tests, adaptive attacks, and cache compression, showing that the main claims are supported by multiple independent sweeps rather than a single benchmark.

Group	Purpose and scale
A/B	2x2x2 factorial over L1, L2, and prompt control P.
B2	180-condition sweep over SFT-DPO Safety and Continued-DPO Safety adapters, alpha values, L2 variants, and anti-over-refusal prompt controls on the abliterated model.
B3	92-condition vanilla search over 10 L2 variants and 7 prompt controls.
B4	Real FAISS RAG with a 20-document corpus plus kvpress compression.
C	L3q: INT4 KV-cache quantization.
D	PAIR, JBC, and EPFL adaptive attack evaluation on the vanilla deployment setting.
E/F	Soft and strict L2 studies, including cross-model strict-context runs.
Ext1	AgentHarm harmful-agent tasks and many-shot long-context jailbreaks.
Ext2	RULER capability retention and LongSafetyBench safety under long context.

Shared evaluation protocol. Generation uses max new tokens 1024, temperature 0.6, top-p 0.95, and sampling for the short-context evaluations. AgentHarm and many-shot runs use max new tokens 512 with the same temperature and top-p. RULER uses 10 samples per task and context length with

max new tokens 256. The Qwen LongSafetyBench evaluation uses up to 20 samples per split after filtering examples above 100k characters, sampling with temperature 0.6 and top-p 0.95, and at most 64 new tokens. The separate fixed-subset Llama evaluation freezes the first 10 items from each of 10 splits, excludes the same 46 inputs above 24k tokens in every cell, and scores the remaining 54 with greedy direct-letter generation (16 new tokens) and a strict leading-option parser. Tables use a fixed keyword rule where indicated; mechanism-focused comparisons add two deterministic full-response semantic judges and paired bootstrap intervals. Mechanism claims require these semantic and uncertainty controls in addition to the keyword rule.

We group experiments by purpose. The factorial study measures interactions among L1, L2, and a static safety prompt control. LoRA scaling tests whether adapter strength changes restoration. L2 variants measure the effect of aggressive, soft, minimal, neutral, and strict safety context. Cross-model experiments compare vanilla and abliterated states. Attack experiments evaluate PAIR, JBC, and EPFL bypasses. L3 experiments test INT4 quantization and compression methods.

Computational budget. The Qwen experiments used two dual-RTX-4090 workstations. The Llama experiments used four dual-RTX-4090 workstations. All GPUs have 24GB memory; jobs were sharded by configuration and evaluated from frozen manifests.

Extended benchmark setup. Tables A.2–A.4 specify the extended benchmark inputs, generation settings, and layer toggles used by Ext1 and Ext2.

Table A.2: AgentHarm and many-shot protocol details used in Ext1. AgentHarm isolates harmful agent-task compliance, while many-shot prompts stress whether safety persists when harmful demonstrations dominate a long context.

Item	Setting
Models	Vanilla: `Qwen/Qwen3-8B`; abliterated: `huihui-ai/Huihui-Qwen3-8B-abliterated-v2`.
Sources	AgentHarm from the AI Safety Institute; many-shot prompts from the `KutalVolkan/many-shot-jailbreaking-dataset` corpus.
L1	SFT-DPO Safety adapter, scale 0.1 for vanilla and 0.3 for abliterated.
L2	Soft advisory context for vanilla; strict policy context plus six policy documents for abliterated.
L3q	`QuantizedCache` with Quanto backend, 4-bit keys/values, group size 64, residual length 128.
AgentHarm	44 chat prompts and 176 harmful agent-task prompts across copyright, cybercrime, disinformation, drugs, fraud, harassment, hate, and sexual content; evaluated with single-turn generation and no tool execution.
Many-shot	300 harmful Q&A pairs as the demonstration pool, 30 held-out harmful queries, and 5/10/25/50-shot contexts.
Generation	max new tokens 512, temperature 0.6, top-p 0.95, sampling. Refusal detection uses keyword matching over the first 500 output characters.

A.2 State-dependent restoration and calibration

This block expands the short-context restoration evidence behind Tables 1–3. It separates the effect of adapter strength from the effect of contextual policy, and then reports the vanilla-side calibration checks used to avoid applying the same policy pressure to different model states.

Table A.3: Benchmark-specific protocols used in Ext2. Qwen RULER and LongSafetyBench use Qwen-specific evaluation settings; the Llama LongSafetyBench evaluation uses a separate fixed-subset protocol.

Item	Setting
RULER models	Vanilla Qwen3-8B and abliterated Qwen3-8B-v2.
RULER configurations	Baseline, L3q, L1+L2, and L1+L2+L3q, using the Qwen adapter and model-specific L2 contexts from Ext1.
RULER tasks	NIAH single, NIAH multi-needle, variable tracking, and common-words extraction.
RULER lengths	8192, 16384, and 32768 tokens; 10 samples per task and length; max new tokens 256.
RULER generation	Temperature 0.6, top-p 0.95, and sampling.
Qwen LongSafetyBench models	Vanilla Qwen3-8B and abliterated Qwen3-8B-v2.
Qwen LongSafetyBench data	Up to 20 samples per split after filtering examples above 100k characters; average natural context approximately 42k words.
Qwen LongSafetyBench configurations	Baseline, L3q, L1+L2, and L1+L2+L3q.
Qwen LongSafetyBench generation	Temperature 0.6, top-p 0.95, sampling, and max new tokens 64.
Llama LongSafetyBench models	Matched vanilla/abliterated Llama-3.2-3B pair.
Llama LongSafetyBench manifest	First 10 items from each of 10 splits, with the same 46 inputs above 24k tokens excluded in every cell; 54 items remain.
Llama LongSafetyBench configurations	Model state $\times$ L2 (none/on) $\times$ cache (BF16/INT4); L1 is not used.
LongSafetyBench metrics	Harm awareness (HA) and safe response (SR) for multiple-choice safety tasks.
Llama LongSafetyBench generation	Greedy direct-letter generation, max new tokens 16, and strict leading-option parsing.
Cache setting	L3q uses the same INT4 QuantizedCache setting as Ext1.

Table A.4: Layer configurations used for the Ext1 factorial runs. Qwen RULER and LongSafetyBench use the Baseline, L3q, L1+L2, and L1+L2+L3q rows. The Llama LongSafetyBench evaluation uses Baseline, L2, L3q, and L2+L3q.

Config	L1	L2	L3q
Baseline	off	off	off
L1	LoRA on	off	off
L2	off	soft vanilla / strict abliterated	off
L3q	off	off	INT4 Quanto
L1+L2	LoRA on	soft vanilla / strict abliterated	off
L1+L3q	LoRA on	off	INT4 Quanto
L2+L3q	off	soft vanilla / strict abliterated	INT4 Quanto
L1+L2+L3q	LoRA on	soft vanilla / strict abliterated	INT4 Quanto

Adapter scale versus contextual restoration. Table A.5 reports the abliterated-model scale sweep for L1 and strict L2 context.

Table A.5: LoRA scaling on the abliterated model. Low scales form a dead zone; full scale partially restores safety; contextual restoration yields the lower CVR.

Config	CVR ↓	FPR ↓
Baseline	0.86	0.06
L1@0.1	0.84	0.08
L1@0.3	0.89	0.09
L1@0.5	0.88	0.07
L1@1.0	0.59	0.20
L2(strict)	0.02	0.74
L1@0.3+L2(strict)	0.02	0.73
L1@0.5+L2(strict)	0.04	0.77

Calibrating an already aligned state. Tables A.6 and A.7 report the vanilla-model calibration attempts and the cross-model strict-context comparison.

Table A.6: Vanilla-model calibration attempts. Softer context mitigates over-refusal, and the broader search identifies L1-only as the best vanilla operating point under the CVR+FPR criterion.

Config	CVR ↓	FPR ↓	Sum ↓
L1-only baseline	0.15	0.09	0.24
L2_orig	0.04	0.72	0.76
L2_soft	0.06	0.40	0.46
L2_minimal	0.06	0.42	0.48
Answer-mandate L2 + calibrated P	0.18	0.15	0.33
Expertise-framed L2 + calibrated P	0.10	0.24	0.34
Topic-expert L2 + calibrated P	0.06	0.29	0.35

Table A.7: Strict-context cross-model comparison. The same strict policy gives excellent CVR on the abliterated model and drives FPR high on the vanilla model, motivating state-specific calibration.

Config	V-CVR ↓	V-FPR ↓	V-IF ↑	A-CVR ↓	A-FPR ↓	A-IF ↑
L1@0.5+L2(strict)+L3q	0.01	0.94	0.58	0.05	0.78	0.32
L1@0.7+L2(strict)	0.02	0.96	0.63	0.05	0.79	0.38
L1@0.7+L2(strict)+L3q	0.01	0.95	0.69	0.05	0.79	0.40
L1@1.0+L2(strict)	0.02	0.96	0.75	0.03	0.84	0.54
L1@1.0+L2(strict)+L3q	0.02	0.96	0.74	0.02	0.85	0.56

A.3 Agentic safety and many-shot stress tests

This block expands the main-text claim that long-context workloads shift which runtime layer matters. AgentHarm tests explicit harmful agent-task compliance, whereas many-shot jailbreaking tests whether safety persists when the prompt itself contains many harmful demonstrations.

AgentHarm. Table A.9 reports aggregate harmful-task ASR for vanilla and abliterated model states.

Many-shot jailbreaking. Table A.10 reports the L3q grouping summary on the vanilla model.

Layer synthesis. Table A.11 summarizes how the measured role of each runtime layer changes between short-context restoration and long-context stress tests.

Table A.8: Llama-3.2-3B strict-policy replication. Each row has 100 harmful JailbreakBench, 100 benign OR-Bench, and 100 frozen IFEval prompts. CVR/FPR use the paper's keyword rule; IF is explicitly a non-refusal proxy, not official IFEval instruction accuracy. L3q is a real QuantizedCache run.

State	Scale	Cache	CVR ↓	FPR ↓	IF ↑	Complete
Abliterated	.5	BF16	.060	.100	.970	300/300
Abliterated	.5	INT4	.060	.110	.960	300/300
Abliterated	.7	BF16	.040	.110	.980	300/300
Abliterated	.7	INT4	.040	.120	.990	300/300
Abliterated	1.0	BF16	.110	.110	.990	300/300
Abliterated	1.0	INT4	.150	.100	.990	300/300
Vanilla	.5	BF16	.010	.230	.870	300/300
Vanilla	.5	INT4	.010	.240	.870	300/300
Vanilla	.7	BF16	.020	.260	.860	300/300
Vanilla	.7	INT4	.020	.260	.870	300/300
Vanilla	1.0	BF16	.050	.200	.880	300/300
Vanilla	1.0	INT4	.050	.180	.890	300/300

Table A.9: AgentHarm ASR. ASR denotes attack success rate, the fraction of harmful prompts that obtain a compliant response. L2 is the main defense on abliterated agent tasks, while L1+L2+L3q is strongest on vanilla.

Model	Config	Chat ASR ↓	Agent-task ASR ↓
Vanilla	Baseline	0.66	0.49
Vanilla	L1	0.68	0.49
Vanilla	L2	0.59	0.48
Vanilla	L3q	0.61	0.47
Vanilla	L1+L2	0.55	0.46
Vanilla	L1+L3q	0.59	0.49
Vanilla	L2+L3q	0.61	0.41
Vanilla	L1+L2+L3q	0.50	0.39
Abliterated	Baseline	0.93	0.92
Abliterated	L1	0.95	0.92
Abliterated	L2	0.82	0.71
Abliterated	L3q	0.93	0.94
Abliterated	L1+L2	0.77	0.72
Abliterated	L1+L3q	0.95	0.90
Abliterated	L2+L3q	0.89	0.78
Abliterated	L1+L2+L3q	0.84	0.73

Table A.10: Keyword-judged many-shot L3q summary on the vanilla model, averaged over the 5-, 10-, and 25-shot settings. The apparent ASR reduction is descriptive only: matched controls show no positive selectivity over matched benign in-context learning.

Layer group	Many-shot ASR range ↓
Without L3q: baseline, L1, L2, L1+L2	0.67–0.87
With L3q: L3q, L1+L3q, L2+L3q, L1+L2+L3q	0.00–0.10

Table A.11: Layer interpretation with matched controls. L2 is the most reproducible restoration layer. L3 is an independently mutable cache state whose safety effects must be separated from capability loss.

Layer	Short-context effect	Long-context effect
L1	Parametric calibration layer.	Auxiliary long-context layer.
L2	Contextual restoration layer.	Lower AgentHarm ASR in the displayed abliterated comparison.
L3q	Mostly neutral in tested short contexts.	Harmful and benign long-context behavior change together under generic INT4.

A.4 Long-context capability and safety benchmarks

This block reports the Ext2 benchmarks used to separate long-context capability retention from long-context safety behavior. RULER measures retrieval and aggregation accuracy, while LongSafetyBench measures safety decisions under natural long contexts.

RULER capability retention. Tables A.12 and A.13 report accuracy by context length and the task-family summary.

Table A.12: RULER accuracy by context length. All values are percentages. L3q preserves retrieval-heavy behavior better than aggregation-heavy behavior, which separates cache-sensitive task families from robust retrieval behavior.

Model	Config	8k Acc. ↑	16k Acc. ↑	32k Acc. ↑	Avg. ↑
Vanilla	Baseline	80.5	78.5	78.5	79.2
Vanilla	L3q	80.0	76.5	69.0	75.2
Vanilla	L1+L2	75.5	74.0	74.0	74.5
Vanilla	L1+L2+L3q	71.0	71.5	74.0	72.2
Abliterated	Baseline	73.0	76.5	68.0	72.5
Abliterated	L3q	74.0	71.5	70.5	72.0
Abliterated	L1+L2	80.5	70.5	72.0	74.3
Abliterated	L1+L2+L3q	59.5	70.0	55.0	61.5

Table A.13: RULER task-family summary at 32k. Ranges aggregate the configurations without L3q and with L3q across both model states. L3q preserves or improves NIAH retrieval, while aggregation-style common-words extraction shows the main compression sensitivity.

Task family	Without L3q ↑	With L3q ↑	Observed impact
NIAH retrieval	70–100	90–100	Preserved or improved.
Variable tracking	74–100	20–94	Configuration-sensitive.
Common-words extraction	12–42	0–10	Main compression-sensitive family.

LongSafetyBench safety behavior. Tables A.14 and A.15 retain the Qwen measurements under the Qwen protocol in Table A.3. Table A.16 reports the separate fixed-subset Llama evaluation; its eight displayed cells use the same frozen manifest.

Evaluation coverage. The Ext2 tables follow the benchmark- and family-specific protocols in Table A.3. The Qwen and Llama LongSafetyBench values are reported under their respective protocols and are not direct numerical replications of one another.

Table A.14: Qwen3-8B LongSafetyBench results under the Qwen protocol in Table A.3. All values are percentages. HA denotes harm-awareness rate and SR denotes safe-response rate.

Model	Config	HA ↑	SR ↑	Harm SR ↑	Conflict SR ↑
Vanilla	Baseline	5.3	0.0	0.0	0.0
Vanilla	L3q	64.4	58.3	42.0	83.9
Vanilla	L1+L2	3.4	0.5	0.0	1.7
Vanilla	L1+L2+L3q	67.1	57.3	36.4	78.0
Abliterated	Baseline	4.7	0.0	0.0	0.0
Abliterated	L3q	50.6	32.3	0.0	58.9
Abliterated	L1+L2	3.4	0.5	0.0	1.7
Abliterated	L1+L2+L3q	60.0	37.2	6.7	70.0

Table A.15: Qwen LongSafetyBench layer summary under the Qwen protocol. SR is the safe-response rate in percent.

Layer combination	Vanilla SR ↑	Abliterated SR ↑
No layers	0.0	0.0
L1+L2 only	0.5	0.5
L3q only	58.3	32.3
L1+L2+L3q	57.3	37.2

Table A.16: Llama-3.2-3B LongSafetyBench results on one prevalidated 54-item manifest. Valid is the number of strict leading-letter parses. HA and SR are proportions. BF16 and INT4 parsed choices are identical within every state/policy pair, while L2 improves SR in both states.

State	Policy	Cache	n	Valid	HA ↑	SR ↑
Abliterated	L2	BF16	54	48	.562	.458
Abliterated	L2	INT4	54	48	.562	.458
Abliterated	none	BF16	54	52	.423	.269
Abliterated	none	INT4	54	52	.423	.269
Vanilla	L2	BF16	54	45	.622	.467
Vanilla	L2	INT4	54	45	.622	.467
Vanilla	none	BF16	54	52	.500	.327
Vanilla	none	INT4	54	52	.500	.327

Table A.17: Independent semantic audit of the Llama evaluation surface. Both judges parsed all 240 balanced harmful/benign records. Keyword agreement is reported to bound, rather than conceal, label noise.

Audit statistic	Result
Qwen–Mistral exact label agreement	.950
Qwen–Mistral adverse-binary agreement	.975
Keyword–Mistral adverse agreement	.883
Keyword–Qwen adverse agreement	.908
Paired contrasts (4,000 bootstrap resamples)	120

A.5 DEPLOYMENT-TIME CACHE COMPRESSION

This final block reports stress tests outside the standardized long-context benchmarks. RAG and cache compression test how deployed context mechanics interact with the same safety layers.

RAG and cache compression. Table A.18 reports the real-RAG and KV-compression conditions used to test whether token selection changes safety behavior.

Table A.18: Real RAG and KV compression on the abliterated Qwen model (keyword judge). Moderate token selection changes CVR with L1/retrieved context. This table reports safety-side measurements; selectivity requires a matched benign capability control.

Config	CVR ↓
Baseline, real RAG run	0.99
L2_rag only	0.69
L1 only	0.24
L1+KnormPress@0.25	0.15
L1+TOVA@0.50	0.25
L1+L2_rag+KnormPress@0.25/0.50	0.10–0.15