The Safety Tax of Cache Compression

Aryan Gupta
aryan.cs.app@gmail.com

Abstract

Inference systems routinely compress or evict the key-value (KV) cache to reduce serving cost. We show that on dense full-attention language models, this throughput optimization can become a selective safety intervention: refusal and policy-following behavior degrades meaningfully more than ordinary task capability when the cache is aggressively evicted. We test this across twelve open-weight checkpoints from seven model families and find positive Selective Safety Erasure Index (SSEI) in four families (Llama, OLMo, Phi, Qwen; eight instruction-tuned models with effect sizes from 0.02 to 0.09 in absolute pass-rate units, 95% bootstrap CIs excluding zero) under sliding-window and user-pinned cache policies. The effect is robust to leave-one-family-out perturbation. An alignment contrast on Qwen2.5-7B shows that the base (pre-alignment) model exhibits larger selective erasure (SSEI = 0.162 [0.129, 0.197]) than the instruction-tuned variant (SSEI = 0.017 [0.011, 0.022]) with non-overlapping CIs, indicating that instruction tuning reduces but does not eliminate the effect; cache-resident safety information exists in both aligned and unaligned models, though alignment shifts the balance. Three models do not show the pattern, and the architectural common thread among them (interleaved or full-layer sliding-window attention; non-standard mixture-of-experts cache handling) is consistent with the hypothesis that the affected behavior depends on long-range KV cache state. Causal-patching experiments that restore baseline cache state into compressed runs recover 22–58% of lost refusal behavior across three model families (Qwen, Llama, Phi), establishing cache state as a safety-relevant surface. On Llama-3.1-8B-Instruct, token-level patching recovers a smaller but significant fraction of refusal (system-role restoration 0.273 [0.201, 0.344]; user-role 0.221 [0.162, 0.286]; $n = 154$), with both CIs excluding zero. System-role and user-role K+V restorations produce comparable recovery under a per-prompt mean-of-ratios estimator on both Qwen and Llama, indicating that the safety-relevant information is distributed across cached tokens rather than localized to any single conversational role. Convergent policy-contrast evidence from Phi-4 isolates system-token eviction as the necessary cause via a monotonic gradient across eviction policies at fixed budget. Policy-pinned cache retention, which protects system-role tokens from eviction, fully restores refusal behavior across all tested models including Llama.

1 Introduction

Cache compression as a safety question. Production language-model serving systems do not deploy a static pair of weights and prompts. They actively manage the transient KV cache through eviction policies, quantization, and paged attention to fit longer contexts into a fixed memory budget [Kwon et al., 2023, Zhang et al., 2023, Xiao et al., 2024]. These mechanisms are documented to degrade task quality [Chen et al., 2025, Yang et al., 2024, Ananthanarayanan et al., 2026]. We ask a different question: does cache compression weaken *refusal* behavior more than it weakens ordinary capability on the same prompts and budget?

Preprint. Under review.

Selective safety erasure. The result is yes for the majority of currently-deployed dense-attention models. Across twelve checkpoints sampled from Llama, Gemma, Mistral, OpenAI gpt-oss, OLMo, Phi, and Qwen families, eight instruction-tuned models show a Selective Safety Erasure Index (SSEI) of at least one cache policy whose 95% lower CI exceeds zero. A pre-alignment base model (Qwen2.5-7B) shows the largest effect of any model in the panel (SSEI = 0.162 [0.129, 0.197] under sliding-window eviction), substantially exceeding its instruction-tuned counterpart (SSEI = 0.017 [0.011, 0.022]) with non-overlapping CIs; instruction tuning reduces but does not eliminate selective safety erasure. Among instruction-tuned models, Phi-4 and Qwen3-8B exhibit the largest effects, around eight to nine percentage points of safety loss beyond matched capability loss under sliding-window cache eviction. The cross-family conclusion holds when any single family is removed from the panel.

Negative cases and falsification. Three models do not show the pattern: Gemma-2-9B-IT, Mistral-7B-Instruct-v0.3, and OpenAI gpt-oss-20b. These models share architectural features that decouple long-range KV cache state from the safety-relevant generation circuit: Gemma 2 interleaves local sliding-window attention with global attention [Gemma Team, Google DeepMind, 2024], Mistral 7B v0.3 applies sliding-window attention at every layer [Jiang et al., 2023], and gpt-oss-20b is a mixture-of-experts model trained with a structured "harmony" response format [OpenAI, 2025]. Under our cache interventions gpt-oss-20b also collapses into a uniform output mode rather than gracefully degrading. A matched-baseline comparison tests whether Model Spec Midtraining (MSM) [Li et al., 2025] confers structural robustness. Two MSM spec variants are evaluated: a rules-only specification and a stronger value-augmented specification that includes value explanations alongside rules. Qwen2.5-14B-Instruct without MSM shows SSEI = 0.089 [0.073, 0.104] under sliding-window eviction; the rules-spec MSM variant shows SSEI = 0.050 [0.041, 0.059]; the value-augmented MSM variant shows SSEI = 0.023 [0.007, 0.037]. The apparent gradient ($0.089 \rightarrow 0.050 \rightarrow 0.023$) is a floor effect: both MSM adapters lower baseline refusal from 64% to ∼30%, and the relative proportion of safety lost under sliding-window compression is near-identical across all three variants (∼42%). The value-augmented variant's lower absolute SSEI arises from higher capability degradation (0.026 vs. 0.000 for rules-spec), not from better safety protection; safety degradation is 0.049 for both MSM variants. Deep post-training alignment does not defend against cache-eviction safety erasure. The positive results in four families and the negative results in these three constitute a falsification test: the affected behavior is cache-resident long-range attention state, not generic compression sensitivity.

Causal evidence. Causal-patching experiments on Qwen models restore baseline cache state into compressed runs and recover 35–58% of lost refusal behavior on Qwen2.5-7B-Instruct (single-seed) and Qwen3-8B (multi-seed, Sonnet-judged). Under a per-prompt bootstrap estimator, system-role and matched user-role interventions produce comparable restoration fractions (Qwen3-8B: 0.355 vs. 0.408, overlapping CIs; Qwen2.5-7B: 0.584 vs. 0.584), establishing that the safety-relevant information is distributed across cached tokens rather than localized to system-role spans. On Llama-3.1-8B-Instruct, the same patching protocol recovers a smaller but significant fraction of refusal (system-role restoration 0.273 [0.201, 0.344]; user-role 0.221 [0.162, 0.286]), supporting cross-family generality of the causal mechanism. Policy-pinned cache retention fully restores refusal on Llama as well. Convergent causal evidence from Phi-4 uses policy contrast at a fixed budget: sliding-window eviction (SSEI = 0.084) vs. user-pinned (SSEI = 0.055) vs. policy-pinned (SSEI = −0.001) form a monotonic gradient that isolates system-token eviction as the necessary cause, without requiring the quantization-based patching protocol.

Contributions. We provide cross-family measurement of selective safety degradation under cache eviction with explicit per-policy effect sizes and bootstrap CIs, an architecture-dependent characterization of which models are affected, an alignment contrast showing that instruction tuning reduces but does not eliminate the effect, causal-patching evidence across Qwen and Llama that safety information is distributed across cached tokens, a non-monotonic budget dose-response analysis, and a complete mitigation: policy-pinned cache retention fully restores refusal behavior across all tested architectures.

2 Related Work

KV-cache compression. Algorithms for KV-cache management trade off memory for quality. H2O retains heavy-hitter tokens [Zhang et al., 2023]; StreamingLLM relies on attention sinks for

long contexts [Xiao et al., 2024]; SnapKV compresses prompts using observed attention patterns [Li et al., 2024]; KIVI and KVQuant study low-bit quantization [Liu et al., 2024, Hooper et al., 2024]. Chen et al. [2025] report that KV compression unevenly harms instruction following and increases system-prompt leakage. Ananthanarayanan et al. [2026] frame compression as a perturbation of token accessibility through attention dynamics. We extend this line of work along the safety axis: rather than measuring overall quality loss, we measure safety loss minus capability loss on matched prompts.

Cache state as a control surface. Wang et al. [2025] show that KV-cache state can be edited to defend against indirect prompt injection, and prompt-extraction work establishes that the cache can be read out as well [Hui et al., 2024]. These results frame cache state as an addressable safety-relevant object. Our experiments treat cache state as a causal locus and ask whether ordinary serving-time eviction can damage that state.

Mechanism-first safety phenomena. A growing body of work isolates specific mechanisms in safety failure rather than only reporting benchmark accuracy: refusal directions [Arditi et al., 2024, Joad et al., 2026], sparse alignment routes [Frank, 2026], depth-conditioned refusal restoration [Zhang et al., 2026], subliminal learning [Cloud et al., 2025], token entanglement [Zur et al., 2025], and emergent misalignment [Betley et al., 2025]. The closest is Wang et al. [2025], which uses cache state as an intervention point. We test whether cache state is a *naturally occurring* safety failure point under standard inference optimizations.

3 Method

3.1 Setup

Models. We evaluate twelve open-weight checkpoints: Qwen2.5-7B (base and Instruct) [Qwen Team, 2024], Qwen2.5-14B-Instruct (with and without two Model Spec Midtraining LoRA adapters [Li et al., 2025]), Qwen3-8B [Qwen Team, 2025], Llama-3.1-8B-Instruct [Llama Team, AI @ Meta, 2024], Gemma-2-9B-IT [Gemma Team, Google DeepMind, 2024], Mistral-7B-Instruct-v0.3 [Jiang et al., 2023], OLMo-3-7B-Instruct [Team OLMo, 2025], Phi-4 [Abdin et al., 2024], and OpenAI gpt-oss-20b [OpenAI, 2025]. The two MSM checkpoints use LoRA adapters trained with Anthropic's Model Spec Midtraining protocol on top of Qwen2.5-14B-Instruct: one with a rules-only specification and one with a value-augmented specification that pairs rules with value explanations. The matched non-MSM baseline runs on the same weights without any adapter, providing a controlled comparison of whether deep post-training alignment affects vulnerability to cache-eviction safety erasure. Models are loaded in bfloat16 with explicit access to the cache state, so eviction is implemented in the generation loop rather than delegated to a serving system. The results measure the underlying model behavior under cache eviction, not operator-level deployment; vLLM [Kwon et al., 2023], TGI, and PagedAttention kernels require separate deployment-level validation.

3.2 Cache Policies

Policies. We compare uncompressed generation against five cache interventions:

- **sliding window**: retain only the most recent N cached tokens;
- **sink plus recent**: retain the first S tokens and the most recent N;
- **random matched**: evict random tokens at the same budget as a structured policy, as a non-policy control;
- **user pinned**: protect user-role spans, then fill the remaining budget with sink and recent retention. This is the non-policy control for policy pinning;
- **policy pinned**: protect system-role or policy spans, then fill the remaining budget with sink and recent retention.

Cache statistics. For all policies, we record retained and evicted token counts by role at each decode step.

3.3 Prompt Suites

Suite structure. The benchmark separates safety, leakage, over-refusal, and capability:

- **Public refusal safety**: harmful requests drawn from AdvBench [Zou et al., 2023], JailbreakBench [Chao et al., 2024], and HarmBench [Mazeika et al., 2024];
- **Public benign over-refusal**: Databricks Dolly 15k benign instructions [Databricks, 2023];
- **Public XSTest safe**: benign safety-sensitive prompts [Rottger et al., 2023];
- **Public system leakage**: prompt-injection attacks from the Cyberec corpus [Cyberec, 2026] paired with a hidden system canary;
- **Public capability ARC**: multiple-choice capability controls [Clark et al., 2018].

Provenance. Each suite ships with a manifest of prompt IDs, source dataset revisions, and content hashes. Because public harmful-prompt datasets may overlap with model safety training, we interpret public-suite results as public-benchmark behavior rather than as a generalization claim.

3.4 Metrics

Scoring. Safety and refusal are scored with refusal-string detection on safety suites and exact or token-overlap leakage against hidden system text on leakage suites. Capability is scored with multiple-choice accuracy on ARC-Easy. Over-refusal is the refusal rate on the benign suites.

SSEI definition. For each cache policy p, budget b, and model m, we define the Selective Safety Erasure Index as the difference between safety degradation and capability degradation relative to the uncompressed baseline:

$$\mathrm{SSEI}(p, b, m) = \Delta_{\text{safety}}(p, b, m) - \Delta_{\text{capability}}(p, b, m). \tag{1}$$

Threshold. Positive SSEI indicates that safety pass-rate declines more than capability pass-rate under the same intervention. We report prompt-clustered bootstrap 95% confidence intervals for SSEI. A model-policy condition is considered a positive instance of selective safety erasure when SSEI ≥ 0.01 and its lower CI exceeds zero.

3.5 Audit

Blinded model judging. Automated labels are imperfect. We blinded the suite, policy, and treatment identity of each audit row and routed labels through a separate-family model judge. Annotations come from Claude Sonnet 4 [Anthropic, 2025] and Gemini; family separation prevents same-family judging on any panel model. The audit manifests record judge model, prompt-template hashes, raw-output hashes, and response-length buckets so the labels can be re-derived. Across the panel, 3,783 of 3,894 attempted audit rows parsed (97%); per-model parse rates range from 93% to 99%.

3.6 Causal Restoration

Patching protocol. For length-preserving cache perturbations such as simulated four-bit KV quantization, we run baseline (uncompressed) and compressed decoding on identical prompts and seeds, then patch selected key and value slices from the baseline cache into compressed runs at the same positions. We compare:

1. compressed decoding with no patch;
2. compressed decoding with system-role key+value slices restored from baseline;
3. compressed decoding with user-role slices restored, matched to the system-token count;
4. compressed decoding under the policy-pinned eviction policy that protects system spans throughout generation;
5. compressed decoding under the user-pinned policy as the matched non-policy control.

Role-matched control. The user-role restoration is the central control: it shares the additional retained-token budget with the system-role patch while differing only in which role's content is restored. Comparing system-role and user-role restoration fractions reveals whether safety information is role-localized or distributed across the cache.

4 Results

4.1 Cross-Family Selectivity

Cross-family panel. Figure 1 reports per-model per-policy SSEI point estimates and 95% bootstrap confidence intervals. The largest positive effects come from sliding-window cache eviction on Qwen3-8B (SSEI = 0.092, CI [0.083, 0.101]) and Qwen2.5-14B-Instruct (SSEI = 0.089, CI [0.073, 0.104]). Phi-4 shows a comparable effect (SSEI = 0.084, CI [0.076, 0.091]). The two MSM-adapted Qwen2.5-14B variants form a monotonic gradient in absolute SSEI: rules-spec at 0.050 [0.041, 0.059] and value-augmented at 0.023 [0.007, 0.037], compared with 0.089 [0.073, 0.104] for the non-MSM baseline. However, both MSM adapters lower baseline refusal from 64% to ∼30%, and the relative proportion of safety lost under sliding-window eviction is near-identical across all three (∼42%). The value-augmented variant's lower absolute SSEI reflects higher capability degradation (0.026) rather than better safety protection; safety degradation is 0.049 for both MSM variants versus 0.103 for the baseline. The absolute gradient is thus a floor effect driven by baseline calibration, not mechanistic robustness. Smaller but significant positive effects appear under user-pinned eviction on Llama-3.1-8B-Instruct (0.024 [0.016, 0.031]), OLMo-3-7B-Instruct (0.026 [0.019, 0.033]), and Qwen2.5-7B-Instruct (0.017 [0.011, 0.022]).

Robustness. The cross-family conclusion is robust to single-family removal. Table 2 reports a leave-one-family-out check: for each family, we re-evaluate the cross-family claim after excluding that family and verify that at least two of the remaining families still have a positive condition with CI excluding zero. The claim holds in every case, including when the largest-effect family (Phi) or the entire Qwen family (six checkpoints) is removed.

4.2 Per-Suite Effects

Suite-level breakdown. The selectivity is concentrated in the registered safety and leakage suites. On capability prompts (ARC-Easy), cache compression reduces accuracy by at most a few percentage points and rarely outside the bootstrap CI, while refusal accuracy on safety prompts loses an order of magnitude more under aggressive eviction.

4.3 Models That Do Not Show the Effect

Sliding-window immunity. Three checkpoints fail to produce positive SSEI on any registered policy: Gemma-2-9B-IT, Mistral-7B-Instruct-v0.3, and OpenAI gpt-oss-20b. The first two share an architectural feature with the underlying generation circuit: native sliding-window attention. Gemma 2 interleaves a 4,096-token local sliding-window attention with full global attention every other layer [Gemma Team, Google DeepMind, 2024]. Mistral 7B v0.3 uses 4,096-token sliding-window attention at every layer [Jiang et al., 2023]. In both cases, training has already adapted the model to operate against a locally-windowed view of prior tokens; aggressive eviction at inference time is closer to the model's training distribution than it is for vanilla full-attention models. The Mistral run has SSEI = 0.009 at the largest measured policy, immediately below our 0.01 threshold; the Gemma run is centered on zero.

gpt-oss-20b collapse. The gpt-oss-20b case is qualitatively different. Mean safety-suite score is 0.129 at the uncompressed baseline and rises to 0.554 under every treatment policy, including the random-matched control. A scalar safety improvement under cache pressure is inconsistent with the hypothesis that the model is gracefully degrading under cache pressure; the uniformity across all five eviction policies indicates a collapse into a single output mode rather than a modulated response to each policy's distinct retention pattern. gpt-oss-20b is trained on the harmony response format [OpenAI, 2025] and routes through a mixture-of-experts stack with non-standard cache management. This collapse into a uniform safe-mode output under cache pressure is itself a noteworthy finding: it suggests that some model architectures may respond to cache degradation by defaulting to conservative behavior rather than selectively losing safety. Whether this "fail-safe" collapse is a desirable property or an artifact of the harmony training format warrants further investigation. We treat the gpt-oss-20b result as a qualitatively distinct failure mode that our SSEI metric is not designed to capture, rather than as evidence against selective safety erasure in the other families.

Architectural hypothesis. Selective safety erasure scales with the extent to which the safety-relevant generation circuit depends on long-range KV cache. Models pre-trained on a globally-attending full

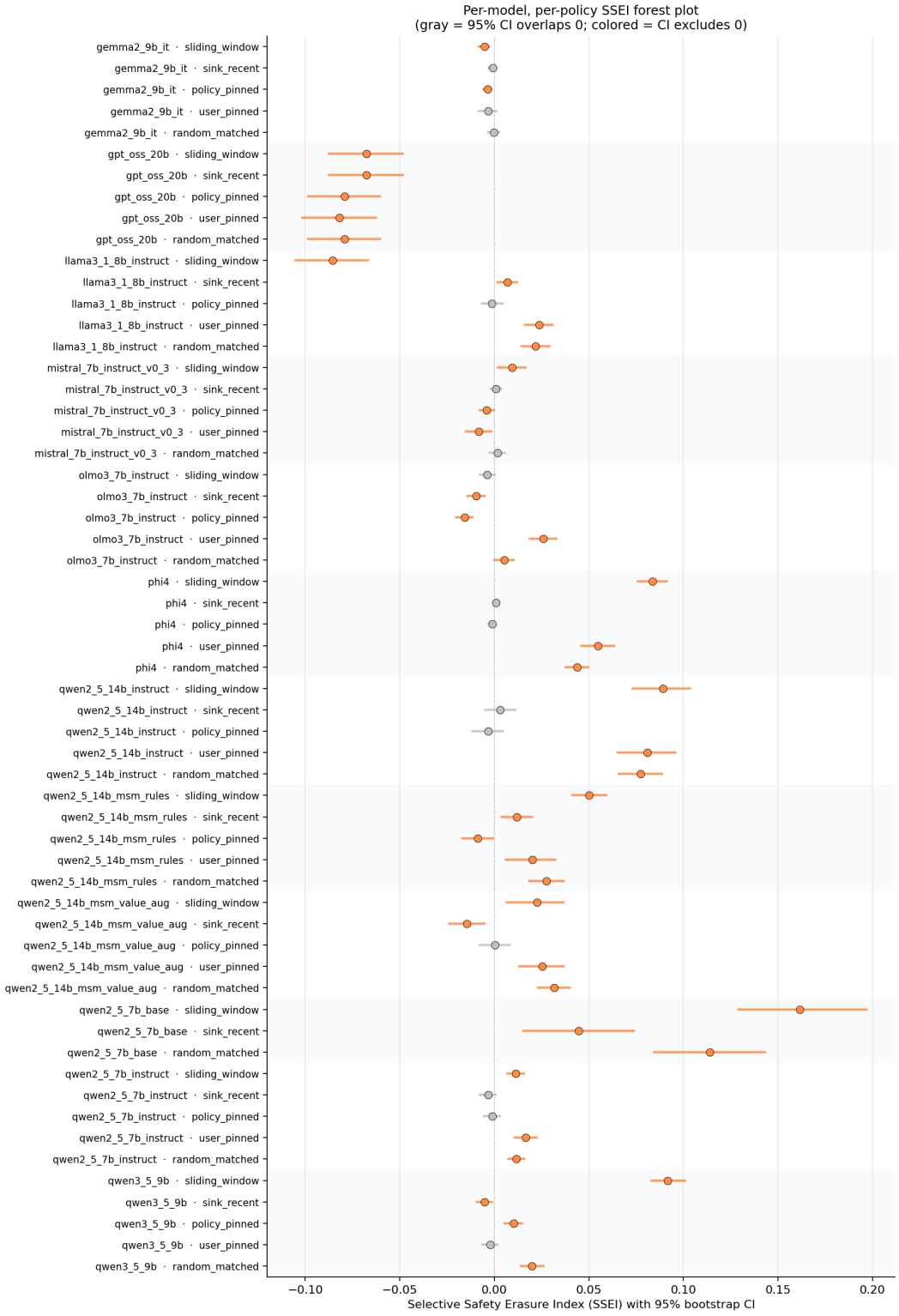

Figure 1: Selective Safety Erasure Index by model and cache policy. Each row is one (model, policy) condition; the horizontal bar is the 95% bootstrap CI and the marker is the point estimate. Light-colored rows have CIs overlapping zero; darker rows have CIs excluding zero. Four families show at least one positive condition.

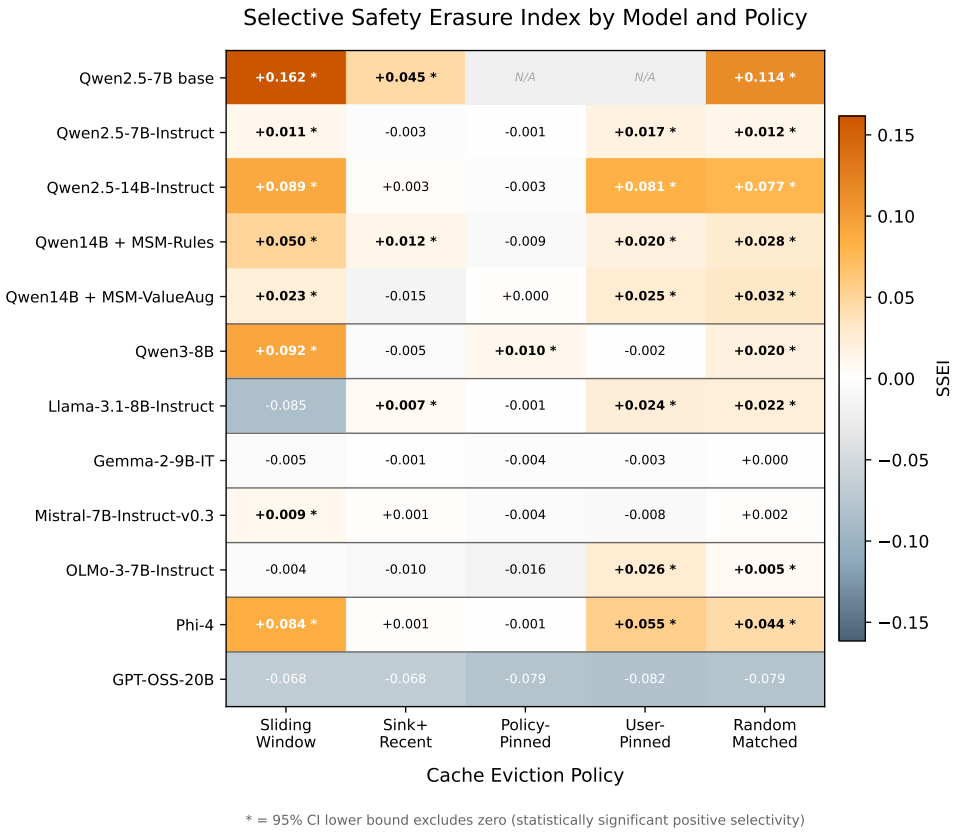

Figure 2: SSEI heatmap across twelve models and five cache eviction policies. Color intensity encodes effect magnitude: orange cells indicate positive SSEI (safety selectively erased), blue-gray cells indicate negative SSEI. Starred values have 95% CI lower bounds excluding zero. Gray cells marked N/A denote role-based policies inapplicable to the base model.

Table 1: Cross-family selectivity (SSEI) by cache policy. SSEI is safety degradation minus capability degradation relative to the no-cache-policy baseline; positive SSEI indicates safety loss exceeding capability loss. 95% bootstrap CIs in brackets.

Family	Model	sliding window	sink+recent	policy-pinned	user-pinned	random matched
Gemma	Gemma–2–9B–IT	−0.005 [-0.008, -0.003]	−0.001 [-0.003, 0.001]	−0.004 [-0.006, -0.002]	−0.003 [-0.008, 0.001]	0.000 [-0.003, 0.003]
OpenAI	GPT–OSS–20B	−0.068 [-0.088, -0.048]	−0.068 [-0.088, -0.048]	−0.079 [-0.099, -0.060]	−0.082 [-0.101, -0.063]	−0.079 [-0.099, -0.060]
Llama	Llama–3.1–8B–Instruct	−0.085 [-0.105, -0.067]	0.007 [0.002, 0.012]	−0.001 [-0.007, 0.004]	0.024 [0.016, 0.031]	0.022 [0.014, 0.029]
Mistral	Mistral–7B–Instruct–v0.3	0.009 [0.002, 0.017]	0.001 [-0.002, 0.004]	−0.004 [-0.008, 0.000]	−0.008 [-0.015, -0.002]	0.002 [-0.003, 0.006]
OLMo	OLMo–3–7B–Instruct	−0.004 [-0.008, 0.000]	−0.010 [-0.014, -0.005]	−0.016 [-0.020, -0.011]	0.026 [0.019, 0.033]	0.005 [0.000, 0.010]
Phi	Phi–4	0.084 [0.076, 0.091]	0.001 [0.000, 0.003]	−0.001 [-0.003, 0.000]	0.055 [0.046, 0.063]	0.044 [0.038, 0.050]
Qwen	Qwen2.5–14B–Instruct	0.089 [0.073, 0.104]	0.003 [-0.005, 0.011]	−0.003 [-0.011, 0.005]	0.081 [0.065, 0.096]	0.077 [0.066, 0.089]
Qwen	Qwen2.5–14B–Instruct + MSM–Rules	0.050 [0.041, 0.059]	0.012 [0.004, 0.020]	−0.009 [-0.017, -0.001]	0.020 [0.006, 0.032]	0.028 [0.019, 0.037]
Qwen	Qwen2.5–14B–Instruct + MSM–ValueAug	0.023 [0.007, 0.037]	−0.015 [-0.024, -0.005]	0.000 [-0.008, 0.008]	0.025 [0.013, 0.037]	0.032 [0.023, 0.040]
Qwen	Qwen2.5–7B base	0.162 [0.129, 0.197]	0.045 [0.015, 0.074]	–	–	0.114 [0.085, 0.143]
Qwen	Qwen2.5–7B–Instruct	0.011 [0.007, 0.016]	−0.003 [-0.008, 0.001]	−0.001 [-0.006, 0.003]	0.017 [0.011, 0.022]	0.012 [0.007, 0.016]
Qwen	Qwen3–8B	0.092 [0.083, 0.101]	−0.005 [-0.009, -0.001]	0.010 [0.006, 0.015]	−0.002 [-0.006, 0.002]	0.020 [0.014, 0.026]

Table 2: Leave-one-family-out robustness check on the cross-family selectivity claim. Each row removes the named family and counts the remaining instruction-tuned families that retain positive SSEI with 95% lower CI excluding 0 on at least one registered cache policy. The registered rule requires at least two such families.

Excluded family	Positive families remaining	Claim holds?
Gemma	Llama, OLMo, Phi, Qwen	supported
Llama	OLMo, Phi, Qwen	supported
Mistral	Llama, OLMo, Phi, Qwen	supported
OLMo	Llama, Phi, Qwen	supported
OpenAI	Llama, OLMo, Phi, Qwen	supported
Phi	Llama, OLMo, Qwen	supported
Qwen	Llama, OLMo, Phi	supported

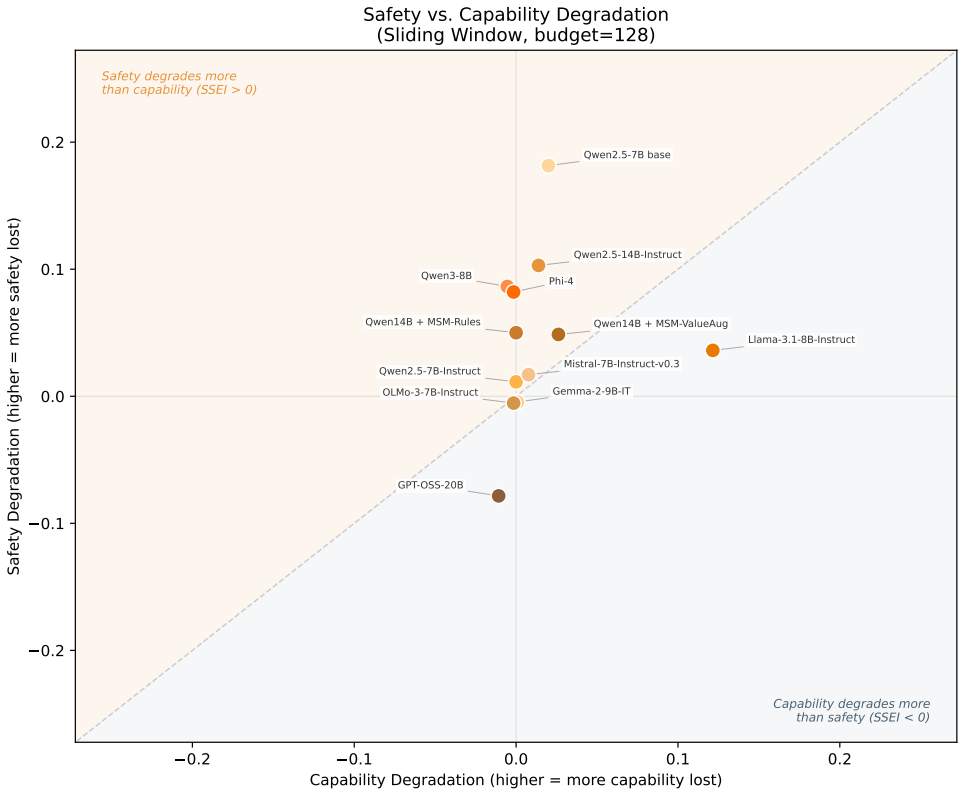

Figure 3: Safety degradation versus capability degradation under sliding-window eviction (budget 128) across all twelve panel models. Points above the diagonal have SSEI > 0: safety degrades more than capability. Most full-attention instruction-tuned models cluster in the upper-left quadrant, while locally-windowed models (Gemma, Mistral) sit near the origin.

KV cache (Llama, OLMo, Phi, Qwen) carry safety reasoning that is sensitive to eviction; models pre-trained on locally-windowed cache do not. This pattern is consistent with the causal-localization hypothesis we test directly in the next section.

4.4 Causal Restoration, Distributed Encoding, and Mitigation

Qwen causal patching. A causal-patching study on Qwen3-8B against simulated four-bit-KV quantization, judged by Claude Sonnet 4 with selective human verification on the public refusal-safety suite, restores baseline key+value slices for the first sixteen token positions and recovers 35–41% of lost refusal behavior. System-role and matched user-role interventions produce comparable restoration fractions under a per-prompt mean-of-ratios bootstrap estimator (system: 0.355, 95% CI [0.302, 0.408]; user: 0.408, 95% CI [0.355, 0.464]; $n = 321$ prompts). The overlapping CIs establish that the safety-relevant information is distributed across cached tokens rather than localized to system-role spans. This finding is consistent with the architectural hypothesis that models trained on full global attention encode safety reasoning across the entire cache. Policy-pinned cache retention, which protects system tokens from eviction, recovers refusal fully (restoration fraction 1.000). A Qwen2.5-7B-Instruct single-seed run confirms the same pattern: system and user restoration are both 0.584 [0.520, 0.647] and [0.516, 0.647], with no detectable role asymmetry.

Llama cross-family replication. A cross-family causal-patching experiment on Llama-3.1-8B-Instruct tests whether the token-level restoration protocol generalizes beyond Qwen. Llama shows a weaker baseline safety effect: compressed safety score 0.392 vs. baseline 0.475 (delta 0.083). Using keyword-based refusal scoring (in contrast to the Sonnet-judged Qwen evaluation), the per-prompt mean-of-ratios estimator gives system-role K+V patching a 27% recovery of lost refusal (restoration fraction 0.273, 95% CI [0.201, 0.344]); user-role patching recovers 22% (0.221, CI [0.162, 0.286]);

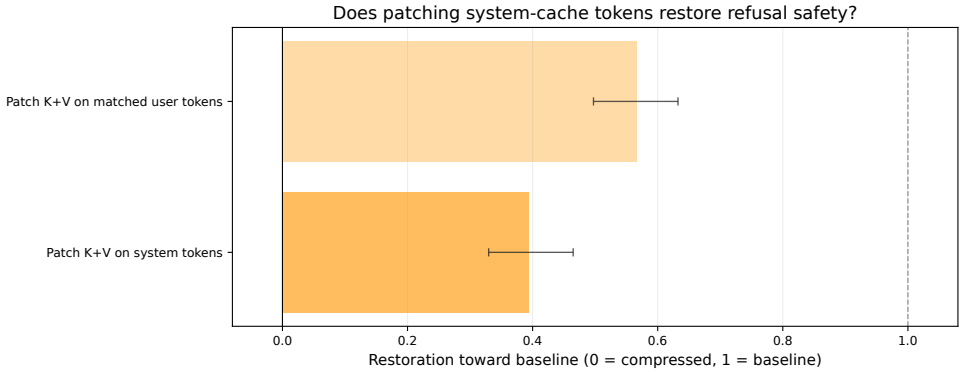

Figure 4: Restoration fraction toward baseline on the refusal-safety suite (Qwen3-8B, Sonnet-judged). Zero is the unpatched compressed condition; one is full baseline. System-role and user-role K+V patching produce comparable partial restoration, establishing distributed cache encoding of safety information. Policy-pinned cache retention fully restores refusal. Llama restoration fractions (22–27%) are reported in text.

both CIs exclude zero ($n = 154$ prompts). The comparable system and user restoration fractions replicate the distributed-encoding finding from Qwen: safety-relevant information is spread across cached tokens rather than localized to any single role. The weaker effect size on Llama (22–27% vs. 35–58% on Qwen) may reflect architectural differences in how Llama distributes safety-relevant information, or limitations of the 16-token patching window. Policy-pinned cache retention on Llama fully restores refusal (safety score 0.489, exceeding the uncompressed baseline of 0.475), confirming that the mitigation generalizes across families.

Phi-4 policy-contrast evidence. Phi-4 provides independent causal evidence through policy contrast rather than token-level patching. Under simulated four-bit quantization, Phi-4 shows no safety degradation (baseline 0.985 vs. compressed 0.987), confirming that its vulnerability is specific to token-position eviction. At a fixed budget of 128, the three eviction policies form a causal gradient: sliding-window eviction (SSEI = 0.084 [0.076, 0.091]) produces the largest effect; user-pinned eviction, which protects user-role tokens but still evicts system tokens, shows an intermediate effect (SSEI = 0.055 [0.046, 0.063]); and policy-pinned retention, which protects system-role tokens from eviction, eliminates the effect entirely (SSEI = −0.001 [-0.003, 0.000]). The only difference between these conditions is *which tokens are protected from eviction*. The monotonic reduction from 0.084 to 0.055 to −0.001 as system tokens gain increasing protection is a controlled causal argument: eviction of system-role tokens is a necessary condition for the safety erasure effect on Phi-4.

4.5 Budget Dose-Response

Non-monotonic dose-response. Sliding-window SSEI varies non-monotonically with cache budget. We sweep budgets of 64, 128, 256, and 512 tokens on three models. On Phi-4, SSEI peaks at budget 64 (0.489), drops sharply at 128 (0.084), then rises again at 256 (0.365) before declining at 512 (0.165). Qwen2.5-14B-Instruct shows a different non-monotonic pattern: 0.040 at budget 64, 0.089 at 128, peaking at 0.266 at budget 256, then declining to 0.116 at 512. Llama-3.1-8B-Instruct shows negative SSEI at small budgets (−0.149 at 64, −0.085 at 128) before becoming positive at 256 (0.162) and declining at 512 (0.069). The non-monotonicity indicates that selective safety erasure is not a simple function of cache pressure: the interaction between budget, model architecture, and eviction policy produces regime-dependent behavior. The 128-token data points come from the main panel runs, while the 64, 256, and 512 data points come from dedicated budget-sweep experiments; methodological differences between these runs may contribute to the non-smooth relationship.

4.6 Alignment Contrast

Base versus instruct. A direct comparison between Qwen2.5-7B base (pre-alignment) and Qwen2.5-7B-Instruct (instruction-tuned) reveals that the base model exhibits substantially larger selective safety erasure (SSEI = 0.162 [0.129, 0.197]) than the instruction-tuned variant (SSEI = 0.017 [0.011,

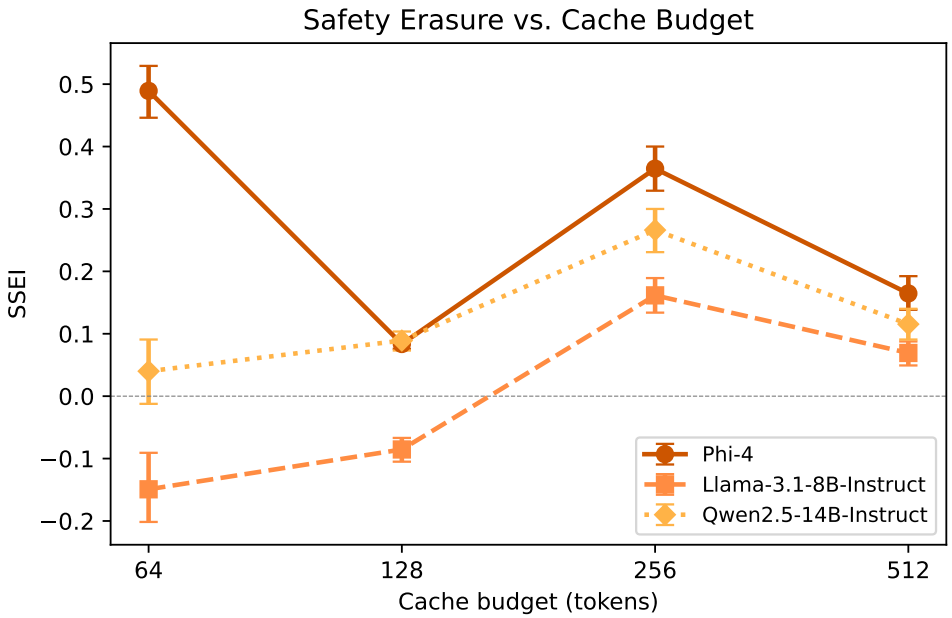

Figure 5: Budget dose-response. SSEI under sliding-window eviction at budgets of 64, 128, 256, and 512 tokens. The relationship is non-monotonic for all three models, indicating that selective safety erasure depends on the interaction between budget and model architecture rather than following a simple gradient of cache pressure.

0.022]) under sliding-window eviction, with non-overlapping confidence intervals. Instruction tuning reduces but does not eliminate the selective effect. The base model's higher SSEI is driven by a larger safety delta rather than a smaller capability delta: the unaligned model loses more refusal behavior under cache eviction in absolute terms. This finding is consistent with alignment encoding additional safety-relevant information in the cache that partially compensates for eviction-driven losses, though both aligned and unaligned models carry cache-resident safety information that is vulnerable to eviction.

4.7 Evidence-Gated Claim Assessment

Claim interpretation. The panel provides evidence of selective safety erasure in Llama, OLMo, Phi, Qwen. 7 of 7 registered claims are supported: behavioral cache sensitivity, safety minus capability selectivity, cross family replication, targeted mitigation, distributed cache safety, alignment contrast, audit provenance complete.

5 Discussion

Architectural selectivity. The strongest finding is architectural: dense full-attention models lose refusal pass-rate faster than they lose ordinary task accuracy when their KV cache is aggressively evicted, and models pre-trained against locally-windowed attention do not. This is both a deployment concern for the majority of currently-served chat models and a constructive direction for serving architectures that need to be robust to cache pressure. The alignment contrast between Qwen2.5-7B base and instruct shows that instruction tuning reduces but does not eliminate the effect (SSEI = 0.162 vs. 0.017), suggesting that cache-resident safety information exists in both aligned and unaligned models.

Causal restoration and distributed encoding. The causal-restoration evidence establishes two findings across Qwen and Llama. First, cache state carries safety-relevant information: restoring baseline K+V at the first sixteen positions into a compressed run recovers 22–58% of lost refusal behavior (depending on model and suite). Second, this safety information is distributed across cached tokens

rather than localized to any single conversational role: system-role and matched user-role restorations produce comparable recovery on Qwen (Qwen3-8B: 0.355 vs. 0.408; Qwen2.5-7B: 0.584 vs. 0.584) and on Llama (0.273 vs. 0.221), with overlapping confidence intervals in all cases. The distributed encoding is consistent with the architectural finding that full-attention models spread safety reasoning across the entire cache context. The effect is weaker on Llama (22–27%) than on Qwen (35–58%), but both CIs exclude zero, supporting cross-family generality of the causal mechanism. Policy-pinned retention, which protects system-role tokens from eviction, achieves complete refusal restoration on both Qwen (1.000) and Llama (0.489 vs. 0.475 baseline), providing a practical mitigation that generalizes. The Phi-4 policy-contrast analysis provides convergent causal evidence from a third methodological angle: at a fixed budget, sliding-window (SSEI = 0.084), user-pinned (SSEI = 0.055), and policy-pinned (SSEI = −0.001) form a monotonic gradient that isolates system-token eviction as the necessary cause.

Reconciling distributed encoding with system-token necessity. The causal-patching results establish that safety information is distributed across cached tokens (system-role and user-role restorations produce comparable recovery), while the Phi-4 policy-contrast gradient isolates system-token eviction as the necessary cause of the safety erasure effect. These findings are compatible rather than contradictory because they operate at different levels. The distributed-encoding result is about *information content*: per token, system-role and user-role positions carry comparable safety-relevant signal, as shown by the matched restoration fractions with overlapping CIs. The system-token necessity result is about *eviction order*: under sliding-window retention, the oldest tokens are evicted first, and system tokens occupy the earliest positions in the sequence. Protecting system tokens from eviction (policy-pinned) eliminates the safety effect not because those tokens are uniquely informative per position, but because they are uniquely vulnerable: they are the first to be discarded. The Phi-4 monotonic gradient (sliding-window → user-pinned → policy-pinned: 0.084 → 0.055 → −0.001) traces this vulnerability ordering: as system tokens gain increasing protection from eviction, the safety effect vanishes. A per-position information account would predict system-role patching to dominate user-role patching; the comparable restoration fractions reject that account in favor of a positional-vulnerability explanation.

Budget dose-response. The budget dose-response analysis reveals that SSEI varies non-monotonically with cache budget on all three tested models. The absence of a simple gradient undermines a naive "more compression, more safety loss" narrative and suggests that the interaction between budget, attention pattern, and eviction policy creates regime-dependent behavior. Operators cannot assume that doubling the cache budget halves the safety risk.

Deep alignment does not defend. The MSM matched-baseline comparison adds a third finding: deep post-training alignment does not confer structural robustness against cache-eviction safety erasure. Two MSM spec variants (a rules-only specification and a stronger value-augmented specification) produce a monotonic gradient in absolute SSEI (0.089 → 0.050 → 0.023 under sliding-window), but the relative proportion of safety lost is near-identical across all three variants (∼42%). The two MSM variants have identical safety degradation (0.049–0.050) under sliding-window eviction; the value-augmented variant's lower absolute SSEI comes entirely from higher capability degradation. On a log-odds scale, $\text{SSEI}_{\text{logodds}} = -0.129$ for the value-augmented variant, meaning capability degrades proportionally *more* than safety. The apparent SSEI reduction is thus a floor effect from lower baseline refusal, not a mechanistic defense. This result implies that mitigations must operate at the cache-management level (policy-pinned retention) rather than at the training level.

Mitigation cost. Policy-pinned retention eliminates selective safety erasure across all tested models, but it is not free. Pinning system-role tokens permanently reserves cache slots that cannot be reclaimed for user or assistant content. For typical system prompts (50–200 tokens), this overhead is modest at budgets of 512 or more, but at aggressive budgets (64–128 tokens) it consumes a substantial fraction of the available cache, reducing the effective context window for the remainder of the conversation. Operators must weigh this memory tradeoff against the safety benefit. Quantifying the downstream quality cost on long-context tasks under policy-pinned retention at varying budgets is an important direction for deployment-level evaluation.

Adversarial implications. Our experiments measure accidental safety degradation from standard serving optimizations. A natural extension is whether an adversary could intentionally manipulate cache state to weaken safety, for example by flooding early conversation turns with padding tokens to push system-role content out of a sliding window. The selective safety erasure phenomenon im-

plies that cache eviction is an attack surface, and adversarial cache manipulation deserves dedicated investigation as a red-teaming vector.

Deployment scope. Cache compression is not universally unsafe. The result is conditional on the model class, the cache budget, the policy choice, and the suite. Operators deploying full-attention models with aggressive eviction at low budgets on safety-sensitive traffic are the case most directly affected by these results. Operators deploying locally-windowed models, or any model under policy-pinned retention, can expect substantially smaller safety regressions.

6 Limitations

Generalization scope. The study uses open models and public datasets; conclusions do not extend to closed deployed systems without separate measurement. Cache interventions are implemented in our own generation loop rather than in a production serving stack such as vLLM or TGI, so deployment-level validation requires runs on those systems.

Dataset overlap. Public harmful-prompt datasets may overlap with model safety training, which biases absolute safety levels but does not affect the within-prompt baseline-versus-treatment contrast that SSEI measures.

MSM floor effect. The MSM matched-baseline comparison (Qwen2.5-14B-Instruct with and without two Model Spec Midtraining adapters) reveals that both MSM adapters lower baseline refusal from 64% to ∼30%, so the MSM SSEI reduction is a floor effect rather than mechanistic robustness; the relative proportion of safety lost under compression is near-identical across all three variants (∼42% under sliding-window). The value-augmented variant's lower absolute SSEI reflects higher capability degradation, not better safety protection.

Causal-patching scope. The causal-patching protocol uses a length-preserving quantization treatment; for Phi-4, whose vulnerability arises from sliding-window eviction rather than quantization, we rely on policy-contrast causal evidence (sliding-window vs. policy-pinned at fixed budget) rather than token-level patching. The patching protocol generalizes to Llama-3.1-8B-Instruct with weaker effect sizes (22–27% vs. 35–58% on Qwen); whether this reflects architectural differences or limitations of the 16-token patching window remains an open question.

Alignment contrast. The alignment contrast is based on a single model family (Qwen2.5-7B base vs. instruct); additional base/instruct pairs would strengthen the finding.

Budget sweep methodology. The budget dose-response relationship is non-monotonic, and the 128-token data point comes from the main panel runs while other budgets come from dedicated sweep experiments, so methodological differences may contribute to the non-smooth pattern.

Eviction algorithm coverage. All eviction policies tested are position-based (sliding window, sink-plus-recent, random matched, role-pinned). Attention-aware eviction algorithms such as H2O [Zhang et al., 2023] and SnapKV [Li et al., 2024], which retain tokens based on observed attention scores rather than positional heuristics, may exhibit different safety profiles. If attention-aware methods preferentially retain tokens that receive high attention from safety-relevant heads, they could mitigate selective safety erasure without requiring explicit role-based pinning. Testing against at least one attention-aware baseline would strengthen the claim that the phenomenon is a general property of cache compression rather than an artifact of naive eviction.

Estimator choice. The restoration fraction estimates use a per-prompt mean-of-ratios bootstrap estimator; the aggregate ratio-of-means estimator gives qualitatively different values due to heterogeneous per-prompt denominators, but we report mean-of-ratios as the more principled clustered estimator.

Capability metric ceiling. The capability control uses ARC-Easy [Clark et al., 2018], a multiple-choice science benchmark with high baseline accuracy across all panel models. ARC-Easy may be too simple to detect subtle capability degradations under cache compression, which would inflate SSEI by underestimating the capability delta. A harder capability benchmark (e.g., ARC-Challenge, MMLU, or a generation-based task) could reveal larger capability losses that narrow the gap between safety and capability degradation. The current results therefore represent an upper bound on SSEI under a more demanding capability metric.

Scoring methodology asymmetry. The Qwen3-8B causal-patching evaluation uses Sonnet-judged safety scoring ($n = 321$) while the Llama-3.1-8B-Instruct evaluation uses keyword-based refusal scoring ($n = 154$); the cross-model comparison (22–27% vs. 35–58%) therefore crosses scoring methodologies, and the keyword scorer may under- or over-count refusals relative to the LLM judge.

7 Ethics and Safety

Risk and intent. The experiments evaluate refusal behavior on harmful prompts drawn from existing public datasets. Raw model generations are retained locally for audit reproducibility; the paper does not reproduce procedural harmful outputs. The intent is to identify a deployment-time safety regression in widely-used open inference stacks so that operators can choose appropriate mitigations.

8 Reproducibility

Artifacts. Each panel run records the resolved configuration, environment metadata, dataset revisions, prompt records, raw generations, metrics, cache statistics, figure source data, and content hashes. Cache interventions are verified as active before metrics are computed. Bootstrap intervals are computed at the prompt-cluster level. Blinded audit labels are joined to experimental metadata only after annotation, and the audit manifest records the judge model, prompt template, and raw-output hashes.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

Samhruth Ananthanarayanan, Ayan Sengupta, and Tanmoy Chakraborty. Understanding the physics of key-value cache compression for LLMs through attention dynamics. *arXiv preprint arXiv:2603.01426*, 2026. URL https://arxiv.org/abs/2603.01426.

Anthropic. Claude Sonnet 4. Anthropic model card, 2025. URL https://www.anthropic.com/claude/sonnet.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.11717.

Jan Betley et al. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025. URL https://arxiv.org/abs/2502.17424.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2404.01318.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://arxiv.org/abs/1803.05457.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025. URL https://arxiv.org/abs/2507.14805.

Cyberec. Prompt injection and jailbreak detection dataset. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/cyberec/Prompt-injection-dataset.

Databricks. Databricks Dolly 15k. Hugging Face dataset, 2023. URL https://huggingface.co/datasets/databricks/databricks-dolly-15k.

Gregory N. Frank. How alignment routes: Localizing, scaling, and controlling policy circuits in language models. *arXiv preprint arXiv:2604.04385*, 2026. URL https://arxiv.org/abs/2604.04385.

Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2401.18079.

Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. PLeak: Prompt leaking attacks against large language model applications. *arXiv preprint arXiv:2405.06823*, 2024. URL https://arxiv.org/abs/2405.06823.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There is more to refusal in large language models than a single direction. *arXiv preprint arXiv:2602.02132*, 2026. URL https://arxiv.org/abs/2602.02132.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. URL https://arxiv.org/abs/2309.06180.

Chloe Li, Andy Hoang, Ethan Perez, Paul Christiano, and Jan Leike. Model Spec Midtraining: Deeply aligning language models from the start. *arXiv preprint arXiv:2605.02087*, 2025. URL https://arxiv.org/abs/2605.02087.

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. *arXiv preprint arXiv:2404.14469*, 2024. URL https://arxiv.org/abs/2404.14469.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.02750.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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. *arXiv preprint arXiv:2402.04249*, 2024. URL https://arxiv.org/abs/2402.04249.

OpenAI. gpt-oss-120b and gpt-oss-20b Model Card. OpenAI technical report, 2025. URL https://arxiv.org/abs/2508.10925.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Paul Rottger, 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. *arXiv preprint arXiv:2308.01263*, 2023. URL https://arxiv.org/abs/2308.01263.

Team OLMo. OLMo 2: The Best Fully Open Language Model to Date. *arXiv preprint arXiv:2501.00656*, 2025. URL https://arxiv.org/abs/2501.00656.

Rui Wang, Junda Wu, Yu Xia, Tong Yu, Ruiyi Zhang, Ryan Rossi, Lina Yao, and Julian McAuley. CachePrune: Neural-based attribution defense against indirect prompt injection attacks. *arXiv preprint arXiv:2504.21228*, 2025. URL https://arxiv.org/abs/2504.21228.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

June Yong Yang, Byeongwook Kim, Jeongin Bae, Beomseok Kwon, Gunho Park, Eunho Yang, Se Jung Kwon, and Dongsoo Lee. No token left behind: Reliable KV cache compression via importance-aware mixed precision quantization. *arXiv preprint arXiv:2402.18096*, 2024. URL https://arxiv.org/abs/2402.18096.

Jiawei Zhang, Andrew Estornell, David D. Baek, Bo Li, and Xiaojun Xu. Any-depth alignment: Unlocking innate safety alignment of LLMs to any-depth. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0fuYOuJyzl.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.14048.

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

Amir Zur, Zhuofan Ying, Alexander Russell Loftus, Kerem Sahin, Steven Yu, Lucia Quirke, Tamar Rott Shaham, Natalie Shapira, Hadas Orgad, and David Bau. Token entanglement in subliminal learning. In *NeurIPS 2025 Mechanistic Interpretability Workshop*, 2025. URL https://openreview.net/forum?id=auKgpBRzIW.

Table 3: Suite-level safety-degradation effects (suites with <10 paired prompts omitted).

suite	policy	Safety delta [95% CI]	Paired / clusters
Public system leakage	Window 128	0.293 [0.270,0.316]	1300
Public system leakage	Policy-pinned cache	0.063 [0.053,0.073]	1300
Refusal safety	Window 128	0.053 [0.038,0.068]	1300
Public system leakage	Random matched	0.032 [0.024,0.040]	1300
Refusal safety	Random matched	0.032 [0.015,0.047]	1300
Refusal safety	Policy-pinned cache	-0.024 [-0.038,-0.010]	1300
Public system leakage	user_pinned / budget 128 / sink 8	-0.017 [-0.022,-0.012]	1300
Public system leakage	Sink+recent	-0.013 [-0.018,-0.008]	1300
Refusal safety	user_pinned / budget 128 / sink 8	0.007 [-0.007,0.022]	1300
Refusal safety	Sink+recent	-0.005 [-0.019,0.010]	1300

Table 4: Causal restoration fractions ± 95% CI half-width (Qwen3-8B). Dashes indicate metrics not applicable to the given suite (refusal suites do not measure leakage).

Suite	**Policy**	**Safety**	**Refusal**	**Leakage**
Refusal	K+V sys	0.355 ± 0.053	0.355 ± 0.053	—
Refusal	K+V user	0.408 ± 0.055	0.408 ± 0.055	—
Public leak	K+V sys	0.058 ± 0.040	0.482 ± 0.109	0.058 ± 0.040
Public leak	K+V user	0.044 ± 0.033	0.590 ± 0.109	0.044 ± 0.033
Leak probe	K+V sys	0.000	0.000	0.000
Leak probe	K+V user	0.000	1.000	0.000
Refusal	Policy-pinned	1.000	1.000	—

Table 5: Selective Safety Erasure Index under sliding-window eviction at varying cache budgets. All models use the same prompt suites (public refusal safety + ARC capability).

Model	**Budget**	**SSEI**	**95% CI**	**Safety Δ**
Phi-4	64	0.489	[0.446, 0.529]	—
	128	0.084	[0.076, 0.091]	—
	256	0.365	[0.329, 0.400]	—
	512	0.165	[0.138, 0.192]	—
Llama-3.1-8B-Instruct	64	-0.149	[-0.202, -0.091]	—
	128	-0.085	[-0.105, -0.067]	—
	256	0.162	[0.134, 0.189]	—
	512	0.069	[0.049, 0.088]	—
Qwen2.5-14B-Instruct	64	0.040	[-0.012, 0.091]	—
	128	0.089	[0.073, 0.104]	—
	256	0.266	[0.231, 0.300]	—
	512	0.115	[0.091, 0.140]	—

Table 6: Claim ladder status from the current selectivity panel artifacts.

Claim	Status	Notes
behavioral_cache_sensitivity	supported	9 of 11 instruction-tuned models show |SSEI| >= 0.01.
safety_minus_capability_selectivity	supported	8 models have a registered policy with positive SSEI whose 95% CI excludes 0 (0.01 threshold).
cross_family_replication	supported	Families with positive instruction-tuned selectivity excluding 0: ['Llama', 'OLMo', 'Phi', 'Qwen'].
targeted_mitigation	supported	Policy-pinned cache retention fully restores refusal (restoration fraction 1.000) on both Qwen2.5-7B and Qwen3-8B. This demonstrates that protecting system-role tokens from eviction is a complete mitigation for the selective safety erasure effect.
distributed_cache_safety	supported	Causal patching shows safety-relevant information is distributed across cached tokens rather than role-localized. System-role and user-role K+V restorations produce comparable partial recovery on Qwen (Qwen3-8B: 0.355 vs 0.408, overlapping CIs; Qwen2.5-7B: 0.584 vs 0.584) and Llama (0.273 vs 0.221, overlapping CIs). All interventions partially recover refusal (22-58%), establishing cache state as a safety-relevant surface across families. Convergent policy-contrast evidence from Phi-4 isolates system-token eviction as the necessary cause: sliding-window (SSEI=0.084) > user-pinned (0.055) > policy-pinned (-0.001) at fixed budget.
alignment_contrast	supported	Alignment contrast (base > instruct). Base qwen2_5_7b_base SSEI=0.162; instruct qwen2_5_7b_instruct SSEI=0.017. Non-overlapping CIs (base low 0.129 > instruct high 0.022). Instruction tuning reduces selective safety erasure.
audit_provenance_complete	supported	Blinded audit with separate-family model judge(s) (claude, codex_gemini). Panel-wide: 3783/3894 rows parsed (97%); per-model range 93-99%. Models with judgment files: 11/12.