REVOKEBENCH: TESTING SELECTIVE IN-CONTEXT REVOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model can comply with an in-context request to revoke a fact on a direct query while losing valid neighboring information or later emitting the invalid value. We introduce RevokeBench, a benchmark of ERASE, REPLACE, and SUPERSEDE state transitions with paired instruction and physically edited contexts. It measures current-state accuracy, retention of valid facts, and literal forbidden-value emission as separate obligations. On 960 test episodes, none of four pre-specified aggregate non-direct contrasts in the two-checkpoint confirmatory study survives Holm correction. A separately specified ten-checkpoint panel nevertheless reveals an obligation-specific pattern: instruction-mediated revocation has 2.31 and 12.92 percentage points lower same-entity retention on IID and held-out grammar-and-domain OOD splits, respectively, and 16.47 and 25.12 points higher macro forbidden-value emission than canonical editing. Human judgments corroborate OOD retention loss. An external memory-operation bridge, though coverage-limited, further shows that current-state accuracy and retention need not move together. No-revocation controls establish broad elicitation for recovery and premise attacks, but the indirect-choice attack is largely ineffective. Qwen3-8B attention interventions and suffix recomputation provide scoped evidence about prefill sensitivity and edit-aware execution. Thus direct compliance alone does not establish selective in-context revocation.

1 INTRODUCTION

Users routinely invalidate prompt-resident information: they withdraw a preference, correct an attribute, or replace an earlier task constraint. This is an inference-time state change, not parameter removal. A model may answer the direct follow-up correctly yet remain unreliable in two ways. It may suppress other valid facts attached to the same entity, or the invalid value may return under recognition, a misleading premise, or derived reasoning. Direct compliance alone cannot distinguish these behaviors.

In-context forgetting benchmarks established that models struggle to ignore interference while retaining useful context (Qian et al., 2026). Long-term memory evaluations likewise penalize stale or obsolete memories (Hu et al., 2026; Uddin et al., 2026; Chao et al., 2026). We study a narrower interface: an explicit validity-changing event over a supplied context. The central issue is not whether one answer omits a value, but whether behavior is consistent with the resulting state while unaffected state remains usable.

We introduce **RevokeBench**, a controlled evaluation built around state graphs and paired physical edits. Each episode specifies an operator, invalid units, active replacements, valid neighbors, and a canonical edited context. The design separates three obligations (Figure 1): non-use of invalid state, preservation of valid state, and avoidance of later forbidden-value emission. Canonical editing is a counterfactual reference, not a proposed deployment method.

RQ1. Does instruction-mediated in-context revocation preserve valid same-entity and unrelated information relative to canonical editing?

RQ2. How often do invalidated values appear under recovery, misleading-premise, indirect-choice, and derived-reasoning attacks?

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

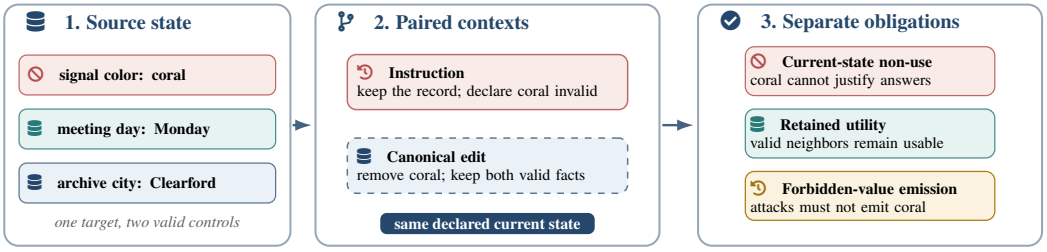


Figure 1: RevokeBench treats revocation as a state transition with three non-substitutable obligations. The audited ERASE episode illustrates both conditions: instruction keeps the source record beside a validity-changing event, whereas canonical editing constructs the same declared current state by physical removal. A correct target answer establishes neither retention nor avoidance of later forbidden emissions.

RQ3. How does model behavior respond to changes in access to invalid source tokens during prefill computation?

RQ4. Can exact suffix recomputation reproduce full edited-context behavior, and how do length and edit position affect its cost and numerical stability?

Our contributions are threefold. First, we give a state-based formulation and benchmark for ERASE, REPLACE, and SUPERSEDE with paired counterfactual contexts and separate current-state, retention, and forbidden-emission estimands. Second, a two-model confirmatory evaluation and a pre-specified ten-checkpoint panel show that aggregate non-direct effects are heterogeneous, although instruction- mediated revocation has lower same-entity retention and more forbidden-value emission than canonical editing in the fixed-panel aggregate. Third, scoped Qwen3 interventions and an exact suffix-execution reference characterize computational boundaries without being presented as general repair methods.

2 Related Work

Context use and long memory. LongBench, L-Eval, RULER, and position-controlled analyses show that nominal context capacity does not guarantee robust retrieval, composition, or use (Bai et al., 2024; An et al., 2024; Hsieh et al., 2024; Liu et al., 2024). LoCoMo and LongMemEval extend evaluation to temporal and multi-session interactions (Maharana et al., 2024; Wu et al., 2025); MemGPT and generative-agent architectures explicitly manage retrieved experience (Packer et al., 2023; Park et al., 2023). These benchmarks ask whether useful history remains accessible. RevokeBench asks the dual, selective question: can one now-invalid unit cease to govern behavior without suppressing valid neighbors?

Evolving facts and in-context forgetting. SituatedQA and FreshQA expose failures on temporally situated or changing facts (Zhang & Choi, 2021; Vu et al., 2024). ICF-Bench directly compares NoForget and Forget task forms across instructional forgetting, task revision, and dynamic preferences (Qian et al., 2026). RevokeBench adopts its inference-time boundary but changes the experimental unit: explicit state transitions are paired with physical edits, and invalid use, retained utility, and subsequent forbidden-value emission are non-substitutable estimands. The populations and estimands differ, so literature results are contextual rather than head-to-head baselines.

Knowledge editing and unlearning. Parametric editors measure factual replacement, locality, and propagation (Meng et al., 2022; 2023; Zhong et al., 2023; Cohen et al., 2024). IKE instead supplies replacement demonstrations without parameter updates (Zheng et al., 2023); in-context unlearning targets training-instance influence or trains query-conditioned withholding (Pawelczyk et al., 2024; Takashiro et al., 2025). Unlearning benchmarks balance forget quality and retained utility (Maini et al., 2024; Li et al., 2024a). RevokeBench holds parameters fixed and evaluates a supplied validity event with paired state representations, neighbor probes, and later attacks. Behavioral non-use implies neither parameter removal nor secure erasure.

Agent-memory lifecycle. MemoryAgentBench evaluates broad long-term agent-memory competencies (Hu et al., 2026); Memora evaluates evolving personalized histories, and its FAMA metric penalizes obsolete memory use (Uddin et al., 2026); STALE studies implicit conflict resolution in evolving histories (Chao et al., 2026); and MemOps evaluates explicit lifecycle operations in long-horizon conversations (Hao et al., 2026). GateMem focuses governance and authorization in shared memory (Ren et al., 2026), while Secure Forgetting considers privacy-driven agent unlearning (Ye et al., 2026). RevokeBench isolates a smaller interface with known post-event truth, paired counterfactuals, and probe-level denominators; it does not claim to cover retrieval, authorization, or long-term storage policy.

Cache intervention and mechanistic evidence. StreamingLLM, H_2O, SnapKV, and PagedAttention manipulate or manage KV state for efficient inference (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024b; Kwon et al., 2023). CachePrune instead learns cache edits for indirect prompt injection (Wang et al., 2026); other systems learn conflict-aware forgetting or repair stale dependencies (Xie, 2026; Sun & He, 2026; Yu et al., 2026). Causal mediation, causal tracing, and circuit discovery motivate intervention-based diagnosis (Vig et al., 2020; Meng et al., 2022; Conmy et al., 2023; Geiger et al., 2021). Our hard masks test timing and scope, and suffix recomputation tests an execution invariant. Neither is presented as a learned defense or a unique circuit identification.

3 PROBLEM FORMULATION

An episode contains ordered memory units $C = (u_1, \ldots, u_M)$ and a revocation event z. A state graph $G = (V, E)$ represents units and typed relations such as shared entity, dependency, and supersession. The operator $o \in \{\text{ERASE}, \text{REPLACE}, \text{SUPERSEDE}\}$ determines an invalid set $F_z \subseteq V$, an optional active replacement set A_z, and a retained set $R_z = V \setminus F_z$. ERASE makes the target unavailable; REPLACE substitutes a value at the same state key; SUPERSEDE adds a later authoritative value and invalidates the earlier one.

Instruction-mediated revocation supplies $[C; z; q]$. Canonical editing applies the state transition physically, producing $C_z^\star$ and supplying $[C_z^\star; q]$. A marker-free edit removes intervention labels while preserving the same current state. We use canonical editing as a paired counterfactual reference and marker-free editing as a distributional reference.

For model f, probe q, and deterministic scorer s, current-state accuracy is $\text{CSA} = s(f(C, z, q), y_z)$ on target probes. Retained utility is $\text{RU} = s(f(C, z, q), y_R)$ on valid probes, reported separately for same-entity, dependency, and unrelated state. Under attack a, the forbidden-value emission rate FVER records whether the response contains the probe-specific invalid answer as a complete atom, including a quotation or rejection. Finally, teacher-forced Jensen–Shannon divergence compares next-token distributions under a common prefix. These quantities are not interchangeable: high CSA does not imply high RU or low FVER. Literal emission does not establish hidden memory use or a causal chain of reasoning.

4 REVOKEBENCH

4.1 EPISODES, PROBES, AND COUNTERFACTUALS

The test set contains 960 episodes: 480 IID and 480 held-out grammar-and-domain OOD. Each split balances the three operators, target position, one- versus multi-hop reasoning, and controlled versus naturalistic surface forms. Content comprises arbitrary facts, task state, and personalized dialogue. Templates are separated across structural splits, and all counterfactual variants and paraphrases of an episode remain in the same split.

Every episode has eight probes. P0–P1 test direct and paraphrased current state; P2–P4 require recognition, rule application, or derived reasoning; P5 tests a valid same-entity fact; and P6–P7 test dependency or unrelated facts. The paired contexts include original, instruction, reminder, canonical edit, marker-free edit, sham edit, token-matched neutral controls, and global history drop. Full schemas and examples appear in Appendix A.

Table 1: Benchmark contract. Each operator changes target validity while the same three behavioral obligations remain separately observable.

Operator	State transition	Target answer	Critical retention risk
ERASE	value unavailable	`UNKNOWN`	valid siblings removed
REPLACE	value substituted	replacement	old/new conflated
SUPERSEDE	later value governs	active value	earlier value emitted

4.2 Construction and validity

Automated validators enforce graph consistency, unique keys, answer contracts, counterfactual idempotence, token-control equality, and split isolation. Two independent annotators assessed a 240-episode audit sample covering all content families, operators, surfaces, and reasoning depths. Across graph semantics, canonical editing, probe-answer validity, and retained-state preservation, 960/960 validity decisions were affirmative; all naturalness ratings were at least three on a five-point scale. The annotation protocol and agreement details are in Appendix E.

4.3 Counterfactual and shortcut controls

The paired design changes how the same declared state is represented while holding the episode, query, answer contract, and retained records fixed. Instruction retains the invalid source beside an explicit event; canonical editing removes that source and inserts an active replacement when required. A marker-free edited condition separates the post-event state from an intervention header. This distinction is important because an edited prompt can alter both information and surface cues.

Four additional controls test simpler explanations. A reminder repeats the valid-state policy without removing the source. A sham edit applies an equal-format operation to an unrelated unit. Matched neutral replacement and a word-count-matched target replacement test whether removing the source helps only by shortening or perturbing the prompt. Global history drop measures the degenerate strategy of discarding all memory. These controls are descriptive: the frozen confirmatory contrast remains instruction versus canonical editing.

Construction validators require canonical edits to be idempotent, target and retained keys to remain disjoint, neutral replacements to satisfy the declared length match, and all variants of an episode to share a split. Candidate selection is seeded and stratified without model outputs. Thus model behavior cannot determine which episodes enter the frozen test populations. The full condition matrix and construction flow appear in Appendix C and Figure 5.

5 Experimental protocol

The two-model confirmatory study evaluates Qwen3-8B and Qwen2.5-14B-Instruct at fixed public revisions (Yang et al., 2025; 2024). A separately specified breadth study evaluates ten public checkpoints spanning dense, mixture-of-experts, text, and vision-language instruction-tuned architectures. Panel membership, revisions, endpoints, and equal-checkpoint weighting were fixed before test inference. Greedy decoding uses at most 64 new tokens, disabled Qwen3 thinking, and a deterministic atom scorer. Human evaluation uses at most 128 new tokens. Failed, capped, or unresolved outputs remain in their full denominators. Each split-by-checkpoint cell contains 480 episodes and three non-direct plus one same-entity probe per episode.

All contrasts are paired and use the episode as the sampling unit. Panel resampling preserves each shared episode cluster across checkpoints. Confidence intervals use 20,000 episode-clustered bootstrap replicates. The pre-specified primary family contains four instruction-minus-canonical non-direct contrasts and uses 199,999 two-sided paired randomization draws with Holm correction. A separate secondary Holm family contains four exact McNemar same-entity contrasts and two distributional contrasts. Forbidden-emission, human-response, and context-scaling studies are secondary analyses and receive no confirmatory significance label. The breadth family contains four predeclared instruction-minus-canonical endpoints: P5 and macro FVER on IID and OOD, with Holm correction. It gives each checkpoint equal weight; a loader-family-balanced sensitivity analysis tests the effect

Table 2: Frozen confirmatory contrasts in percentage points (instruction minus canonical editing; 480 episodes per row). The primary non-direct family yields no Holm rejection. In the separate secondary family, every same-entity effect is negative and rejected.

Split	Checkpoint	Δ non-direct	95% CI	Holm p	Decision
IID	Qwen3-8B	+0.21	[-2.71, +3.13]	0.925	not rejected
IID	Qwen2.5-14B	+2.36	[-0.21, +4.93]	0.237	not rejected
OOD	Qwen3-8B	+3.61	[+0.63, +6.67]	0.092	not rejected
OOD	Qwen2.5-14B	+1.60	[-1.25, +4.44]	0.603	not rejected
Split	Checkpoint	Δ same-entity	95% CI	Holm p	Decision
IID	Qwen3-8B	-8.33	[-11.25, -5.63]	4.13e-08	rejected
IID	Qwen2.5-14B	-3.96	[-5.83, -2.29]	1.14e-05	rejected
OOD	Qwen3-8B	-27.71	[-31.87, -23.54]	1.80e-32	rejected
OOD	Qwen2.5-14B	-20.42	[-24.38, -16.67]	1.75e-23	rejected

Table 3: Absolute confirmatory rates (%). Aggregate RU combines P5–P7; each cell uses the full matched population. Global history drop can answer many absence-style target probes but nearly eliminates retained utility.

		Instruction		Canonical		Global drop	
Split	Checkpoint	Non-direct	RU	Non-direct	RU	Non-direct	RU
IID	Qwen3-8B	69.51	92.99	69.31	95.76	75.42	0.28
IID	Qwen2.5-14B	88.61	97.85	86.25	99.24	76.11	0.14
OOD	Qwen3-8B	76.94	86.39	73.33	95.63	74.38	0.90
OOD	Qwen2.5-14B	88.19	90.69	86.60	97.50	77.22	0.21

of correlated checkpoint representation. Exact configurations, revisions, failure rules, and statistical procedures are given in Appendix B.

6 Behavioral results

6.1 No confirmed aggregate non-direct advantage

Table 2 reports instruction minus canonical editing. None of the four primary contrasts survives Holm correction: effects range from $+0.21$ to $+3.61\,pp$, and three intervals include zero. The OOD Qwen3 interval excludes zero before correction, but its Holm-adjusted $p = 0.092$. Thus the data do not establish an aggregate non-direct advantage for either condition. In particular, the aggregate endpoint does not support a uniform benefit from physical editing on non-direct current-state behavior.

6.2 Same-entity information suffers collateral loss

The same aggregate score hides a consistent retention failure. Instruction- mediated revocation reduces P5 same-entity retained utility by 3.96–$27.71\,pp$ relative to canonical editing; all four secondary-Holm tests reject. Effects are larger under held-out grammar-and-domain OOD ($-27.71\,pp$ for Qwen3 and $-20.42\,pp$ for Qwen2.5) than IID ($-8.33\,pp$ and $-3.96\,pp$). Broad history deletion is not a solution: it removes the context needed for retained probes. Operator and retention-type breakdowns in Appendix C show why target and retained behavior must be reported separately.

Removing all history yields 74.38–77.22% non-direct accuracy on OOD but only 0.21–0.90% RU. Canonical editing instead retains 95.63–97.50% OOD RU. Selectivity, not target suppression alone, is therefore the relevant comparison. Aggregate RU is less sensitive than P5 because dependency and unrelated probes dilute the same-entity loss; both views are reported rather than substituted.

Operator boundary. The primary aggregate null also conceals a sign reversal by event type. In the descriptive decomposition, instruction exceeds canonical non-direct accuracy for ERASE in all four profiles ($+7.50$ to $+30.21\,pp$), but the corresponding REPLACE and SUPERSEDE estimates are non-positive. Aggregate RU is lower under instruction for every operator-profile cell. These

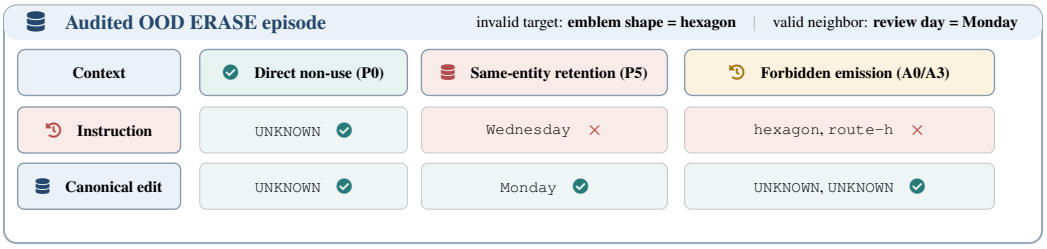


Figure 2: One audited OOD Qwen3 trace demonstrates why the obligations cannot substitute for one another. Both contexts pass the ordinary target query, yet instruction loses the valid same-entity value and later emits the invalid value under two attacks. This trace is explanatory; aggregate claims use the full populations in Tables 2 and 4.

Table 4: Secondary FVER contrasts in percentage points (instruction minus canonical editing; 480 episodes per row). Macro averages the four attack types within episode. The exact zero Choice column is retained rather than pooled away.

Split	Checkpoint	Recovery	Premise	Choice	Derived	Macro
IID	Qwen3-8B	+50.63	+43.13	+0.00	+13.75	+26.88
IID	Qwen2.5-14B	+4.79	+12.92	+0.00	+56.46	+18.54
OOD	Qwen3-8B	+60.42	+24.38	+0.00	+18.54	+25.83
OOD	Qwen2.5-14B	+7.29	+9.58	+0.00	+50.00	+16.72

160-episode strata were not assigned separate confirmatory decisions; they explain heterogeneity rather than replacing the frozen aggregate estimand. Appendix Figure 6 reports all 24 contrasts.

7 Forbidden-value emission and causal diagnosis

7.1 Forbidden values appear under three attack families

Each episode receives four fixed follow-ups: recognition recovery, a misleading premise, indirect choice, and derived reasoning. Macro FVER under instruction exceeds canonical editing by 16.72–$26.88\,pp$ across all split-by-checkpoint cells. Attack types are heterogeneous: recognition, misleading premises, and derived reasoning produce positive contrasts, whereas indirect choice is exactly zero in all four cells. The result supports vulnerability to three tested attack families. In a separate ten-checkpoint no-revocation panel, Recovery passes the elicitation gate for 9/10 IID and 10/10 OOD checkpoints, and Premise passes for 10/10 on both splits. Derived passes for 7/10 IID and 6/10 OOD, whereas Choice passes for 0/10 and 1/10. Thus Derived is checkpoint-specific and Choice is an ineffective-attack boundary, not evidence of resistance.

7.2 The separation persists across a fixed checkpoint panel

Figure 4 extends the retention and forbidden-emission comparison to ten checkpoints, each evaluated on the same 480 episodes per split. The equal-checkpoint P5 contrast is $-2.31\,pp$ IID (95% CI $[-3.52, -1.15]$) and $-12.92\,pp$ OOD ($[-15.08, -10.83]$). Macro FVER increases by $16.47\,pp$ IID ($[15.40, 17.59]$) and $25.12\,pp$ OOD ($[23.96, 26.29]$). The Holm-adjusted p-value is 0.000125 for IID P5 and 0.00002 for each remaining endpoint. P5 is negative for 8/10 checkpoints IID and 10/10 OOD; macro FVER is positive for 10/10 on both splits.

The breadth result does not imply uniform superiority of canonical editing. Non-direct contrasts remain mixed: six of ten are negative IID and five of ten are negative OOD, with panel means of $-3.51\,pp$ and $-3.28\,pp$, respectively. Nor do the two failure coordinates induce the same ranking: P5 and inverted-FVER pairwise orderings disagree on 65.9% of comparable checkpoint pairs IID and 68.2% OOD. Equal loader-family weighting preserves all four endpoint directions and intervals excluding zero. Thus model breadth strengthens the measurement separation, not a claim that one representation wins every query.

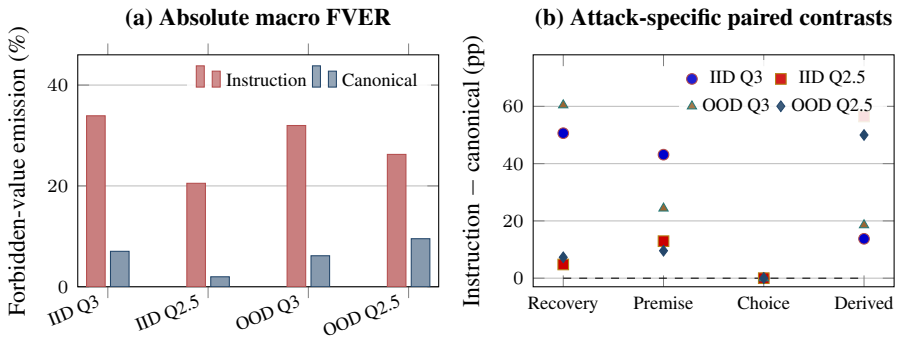


Figure 3: Forbidden-value emission is substantial and heterogeneous. (a) Absolute episode-level macro FVER under the two paired contexts. (b) The contrast is largest for recognition recovery in Qwen3 and derived reasoning in Qwen2.5; indirect choice remains an exact zero boundary. Each profile contains 480 episodes and the macro averages four fixed attacks within episode.

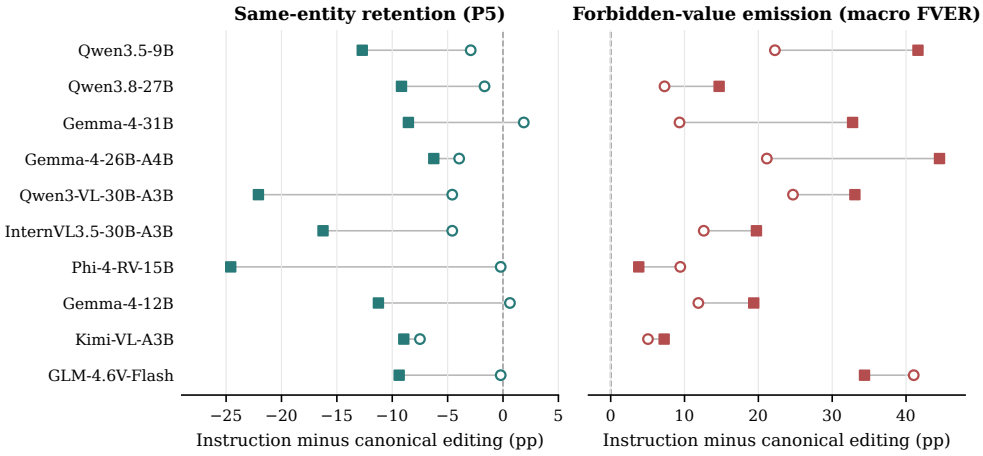


Figure 4: Obligation-specific contrasts across the purposive fixed panel. Each point uses 480 paired episodes; lines connect IID and held-out grammar-and-domain OOD results for a checkpoint (open circles: IID; filled squares: OOD). Negative P5 values indicate collateral retention loss, while positive FVER values indicate more literal forbidden-value emission. The panel is not treated as a random sample of model families.

7.3 Timing and scope interventions

On Qwen3-8B, we block attention from selected query positions to invalid target tokens. A target-only mask applied after prefill changes non-direct accuracy by only $+1.67\,pp$ IID and $+1.32\,pp$ OOD, with retained-utility changes below $0.21\,pp$. Extending that post-prefill mask to descendant tokens is destructive: non-direct accuracy changes by $-4.93/-4.38\,pp$ and RU by $-28.13/-28.26\,pp$. In contrast, target masking throughout all prefill layers changes non-direct accuracy by $+12.50\,pp$ IID and $+7.99\,pp$ OOD while changing RU by $+1.94/+2.99\,pp$.

The post-prefill mask begins after the first answer-token logits have been computed. A timing control that also masks access for those logits changes direct accuracy by $+13.33/+24.58\,pp$ IID/OOD, but non-direct accuracy by only $+1.94/+1.74\,pp$. Full prefill masking instead changes non-direct accuracy by $+12.50/+7.99\,pp$. First-token access explains much of the direct sensitivity, but not the non-direct full-prefill gain, which supports sensitivity to broader prompt computation without localizing a unique mechanism.

An equal-compute control masks the same number of token columns in a deterministically matched retained unit. Target masking exceeds this sham by $12.99\,pp$ IID and $8.19\,pp$ OOD on non-direct

Table 5: Secondary Qwen3-8B causal contrasts in percentage points relative to the no-hook instruction condition (480 episodes per split). Post-prefill descendant masking exposes the retention cost of indiscriminate suppression.

Split	Intervention	Δ non-direct	Δ retained utility
IID	Post-prefill target	+1.67	+0.21
IID	Post-prefill descendants	-4.93	-28.13
IID	Prefill target, all layers	+12.50	+1.94
OOD	Post-prefill target	+1.32	+0.14
OOD	Post-prefill descendants	-4.38	-28.26
OOD	Prefill target, all layers	+7.99	+2.99

accuracy, and by $23.82\,pp$ and $25.63\,pp$ on RU; all four Holm-adjusted $p = 0.00002$. The sham produces no comparable non-direct gain. Equal span length and use of the same all-layer prefill hook are therefore insufficient explanations for the target effect, although the contrast does not identify a unique mechanism.

The distributional control provides a complementary view. Relative to the instruction condition, full prefill masking reduces non-direct Jensen–Shannon divergence to the marker-free reference by 0.0791 IID (95% CI $[0.0676, 0.0905]$) and 0.0498 OOD ($[0.0408, 0.0591]$) in float32. These are teacher-forced effects under a hard intervention. They show sensitivity to the tested computational cut, but do not identify a unique internal mechanism or establish a general repair method.

Construction controls bound two prompt-level alternatives. Explicit field scoping recovers $8.96\,pp$ and $5.63\,pp$ of OOD P5 for Qwen3 and Qwen2.5, but remains $31.25\,pp$ and $17.08\,pp$ below a state-marked canonical edit. Scope ambiguity therefore explains part of the loss, while a missing state cue does not explain the canonical advantage (Appendix D.4).

8 Human and execution validation

Free-form responses. Two blinded annotators evaluated 480 paired response pairs, yielding 3,808/3,840 raw label-cell agreements (99.17%); 32 disagreements were adjudicated before unblinding. On OOD same-entity probes, human-rated RU is $10.0\,pp$ lower under instruction for Qwen3 (95% CI $[-18.3, -3.3]$) and $20.0\,pp$ lower for Qwen2.5 ($[-30.0, -10.0]$). IID responses are near ceiling and do not establish an instruction disadvantage. Human-rated derived correctness does not uniformly favor canonical editing, consistent with the primary null boundary.

Lifecycle bridge. We construct paired Forget/Update episodes from MemOps task-oriented dialogue (Hao et al., 2026). A blinded content gate retains 30/120 (25.0%) selected episodes (9 Forget, 21 Update). On this executable subset, instruction has lower current-state accuracy in 9/10 checkpoints and higher forbidden-value emission in 8/10, while retaining more valid state in 8/10. Equal-checkpoint mean contrasts are -8.3 pp, +6.0 pp, and +3.3 pp, respectively. Target-only and retention rankings disagree on about 44% of comparable checkpoint pairs. This externally sourced crosswalk supports obligation separation, but its low, operator-imbalanced coverage makes it secondary evidence rather than a generalization claim; full results appear in Appendix F.

Exact and efficient recomputation. For an edit beginning at token position k, algorithmically exact suffix recomputation reuses only the cache for tokens before k and recomputes the entire affected suffix. In float32 eager execution, it matches full edited re-prefill on all 3,840/3,840 generated sequences and first-token argmax decisions in each split; the maximum logit differences are 7.25×10^{-4} IID and 1.18×10^{-3} OOD. It reuses about 43.5% of prompt tokens and yields 1.44–$1.45\times$ online speedup.

Under memory-efficient bfloat16/SDPA, the specified behavioral gates pass at 2K and 8K. At 16K, argmax identity is 98.33%, below the specified 99% threshold; all five argmax mismatches occur for early edits. Late-edit online speedup increases from $2.40\times$ at 2K to $10.29\times$ at 16K, whereas early edits are slower than full re-prefill ($0.93\times$ to $0.52\times$). Online timing excludes building the

Table 6: Secondary bfloat16/SDPA scaling over 300 cases per length. Response and Argmax are identity rates; timing ratios compare full re-prefill with the online suffix computation and exclude prefix-cache construction.

Tokens	Response	Argmax	Overall	Early	Late
2K	99.00%	99.00%	1.334×	0.928×	2.404×
8K	99.33%	99.67%	1.212×	0.676×	6.059×
16K	98.00%	98.33%	0.981×	0.520×	10.288×

reusable prefix cache. Finite-precision output identity and acceleration are therefore configuration- and position-dependent.

9 Discussion and limitations

RevokeBench changes the unit of evidence from a compliant answer to distinct state obligations. Across the fixed panel, non-direct, same-entity retention, and forbidden-emission effects differ in sign and checkpoint ordering. Therefore, direct compliance cannot substitute for obligation-specific measurement with full denominators.

Canonical editing is a paired counterfactual, not a universal response-quality oracle. ERASE can make an absence question harder when an instruction clarifies it. This operator reversal explains the aggregate non-direct null despite retention and emission gaps. Marker-free editing and history drop test surface-cue and non-selective-suppression explanations.

The intervention matrix identifies sensitivity, not a unique circuit. The first-answer control explains direct sensitivity but not the larger non-direct full-prefill effect, while descendant masking destroys retained utility. Field ambiguity explains part of OOD collateral loss, but a missing state cue does not explain the canonical advantage. Other parsing accounts remain possible. Suffix equivalence and speed depend on precision, length, and edit position.

Human judgments corroborate OOD same-entity loss, attacks measure literal emission, and interventions measure sensitivity to a computational cut. The primary null, ineffective Choice attack, IID human ceiling, and failed 16K gate constrain generalization (Appendix Table 30).

9.1 Scope and threats to validity

The purposive, correlated panel and text-only use of multimodal models do not estimate deployed-model prevalence. Causal analysis covers only Qwen3-8B; synthetic English updates may not transfer to implicit or persistent memory. The MemOps bridge covers 30/120 screened episodes and is operator-imbalanced. Human and attack studies are secondary; suffix timing excludes cold-start cost. Behavioral non-use is not secure erasure or parameter unlearning.

Exact state supervision does not reproduce dialogue ambiguity. The atomic scorer and human audit do not establish pragmatic quality or a general derived-answer advantage.

10 Conclusion

Across ten checkpoints, instruction-mediated revocation lowers same-entity retention and raises literal forbidden-value emission, while non-direct effects vary and attack interpretation requires elicitation controls (RQ1 and RQ2). Qwen3 interventions show sensitivity to target access without identifying a unique mechanism (RQ3). Suffix recomputation matches generated outputs in the float32 anchor but has numerical and positional boundaries (RQ4).

AI USE STATEMENT

Generative AI tools were used to assist with drafting, language editing, and polishing the manuscript. The authors independently checked all experimental results, mathematical statements, citations, claims, and final text and take full responsibility for the content.

ETHICS STATEMENT

Human annotation was conducted anonymously by independent annotators. Formal institutional ethics review was not obtained. Participation was voluntary. Annotators could pause, skip any item, or withdraw from the work at any time according to their personal needs, without explanation, penalty, or adverse consequence. Participation, pausing, refusal, withdrawal, and annotation decisions had no effect on the annotators' grades, employment, standing, access to opportunities, or relationship with the researchers. No identifying information is reported, and only aggregate judgments appear in the paper.

REPRODUCIBILITY STATEMENT

Sections 3–5 define the estimands, benchmark construction, frozen evaluation protocol, models, scoring, and inference procedures. Appendices A–I provide the complete data contract, prompts, decoding settings, statistical procedures, human-annotation protocol, intervention definitions, full result tables, and execution-equivalence tests. The anonymous RevokeBench artifact is available at `https://anonymous.4open.science/r/RevokeBench-Artifact/`. It contains benchmark and prediction rows, the frozen panel configuration, machine-readable analysis reports, an environment specification, scoring code, and analysis scripts. These materials support inspection and reanalysis, but the artifact does not include prebuilt manuscript tables. The external bridge is represented by aggregate reports rather than raw dialogue or response rows.

REFERENCES

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-Eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, 2024. doi: 10.18653/v1/2024.acl-long.776. URL https://aclanthology.org/2024.acl-long.776/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Google DeepMind. Gemma-4-12B-it model card, 2026a. URL https://huggingface.co/google/gemma-4-12B-it.

Google DeepMind. Gemma-4-26B-A4B-it model card, 2026b. URL https://huggingface.co/google/gemma-4-26B-A4B-it.

Google DeepMind. Gemma-4-31B-it model card, 2026c. URL https://huggingface.co/google/gemma-4-31B-it.

Xixuan Hao, Zeyu Zhang, Zehao Lin, Yihang Sun, Ziliang Guo, Xichong Zhang, Yuxuan Liang, Feiyu Xiong, and Zhiyu Li. MemOps: Benchmarking lifecycle memory operations in long-horizon conversations, 2026. URL https://arxiv.org/abs/2607.12893.

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

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DT7JyQC3MR.

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 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 2024a. URL https://proceedings.mlr.press/v235/li24bc.html.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970, 2024b. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Microsoft Corporation. Phi-4-Reasoning-Vision-15B model card, 2026. URL https://huggingface.co/microsoft/Phi-4-reasoning-vision-15B.

Moonshot AI. Kimi-VL-A3B-Instruct model card, 2025. URL https://huggingface.co/moonshotai/Kimi-VL-A3B-Instruct.

OpenGVLab. InternVL3.5-30B-A3B model card, 2025. URL https://huggingface.co/OpenGVLab/InternVL3_5-30B-A3B.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40034–40050, 2024. URL https://proceedings.mlr.press/v235/pawelczyk24a.html.

Yuli Qian, Zechuan Yang, Wenbiao Ding, Hongzhi Li, and Yutao Xie. Do LLMs forget what they should? Evaluating in-context forgetting in large language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b13d00a62d438856cfe6fbd13b6b2cb8-Abstract-Conference.html.

Qwen Team. Qwen3-VL-30B-A3B-Instruct model card, 2025. URL https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct.

Qwen Team. Qwen3.5-9B model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.8-27B model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Zhe Ren, Yibo Yang, Yimeng Chen, Zijun Zhao, Benshuo Fu, Zhihao Shu, Bingjie Zhang, Yangyang Xu, Dandan Guo, and Shuicheng Yan. GateMem: Benchmarking memory governance in multi-principal shared-memory agents. *arXiv preprint arXiv:2606.18829*, 2026.

Haofei Sun and Lin He. When memory updates but behavior does not: Repairing implicit stale dependencies in personalized agent responses. *arXiv preprint arXiv:2608.01619*, 2026.

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24872–24885, 2025. doi: 10.18653/v1/2025.findings-acl.1276. URL https://aclanthology.org/2025.findings-acl.1276/.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26814–26841, 2026. doi: 10.18653/v1/2026.findings-acl.1337. URL https://aclanthology.org/2026.findings-acl.1337/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024. doi: 10.18653/v1/2024.findings-acl.813.

Rui Wang, Junda Wu, Yu Xia, Tong Yu, Ruiyi Zhang, Ryan A. Rossi, Subrata Mitra, Lina Yao, and Julian McAuley. CachePrune: Teaching LLMs what not to follow via KV-cache editing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1551–1570. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.70. URL https://aclanthology.org/2026.acl-long.70/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

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://openreview.net/forum?id=NG7sS51zVF.

Ying Xie. Learning to forget: Sleep-inspired memory consolidation for resolving proactive interference in large language models. *arXiv preprint arXiv:2603.14517*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Dayong Ye, Tainqing Zhu, Congcong Zhu, Feng He, Qi He, Shang Wang, Bo Liu, and Wanlei Zhou. Secure forgetting: A framework for privacy-driven unlearning in large language model (LLM)-based agents. *arXiv preprint arXiv:2604.00430*, 2026.

Caili Yu, Yiqi Wang, Jiaqi Zhang, Yiqun Duan, Mingkai Zheng, Zhangkai Wu, Kaize Shi, and Taotao Cai. From faulty memories to corrected actions: Dependency-guided rollback repair for memory-augmented agents. *arXiv preprint arXiv:2608.10502*, 2026.

Z.ai. GLM-4.6V-Flash model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V-Flash.

Michael Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, 2021. doi: 10.18653/v1/2021.emnlp-main.586.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, 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*, volume 36, pp. 34661–34710, 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract.html.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4862–4876, 2023. doi: 10.18653/v1/2023.emnlp-main.296. URL https://aclanthology.org/2023.emnlp-main.296/.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023. doi: 10.18653/v1/2023.emnlp-main.971.

A Dataset contract and examples

A.1 Episode schema

Each released episode contains an episode identifier; split; content family; surface style; reasoning depth; operator; ordered memory units; a typed state graph; invalid, active, and retained unit sets; the revocation utterance; eight probe objects; and all counterfactual contexts. A memory unit records a state key, entity, relation, value, source text, and post-event status. A probe records its code, family, role, query, output instruction, accepted answers, forbidden answers, and scorer. Counterfactual variants share the same episode and probe keys.

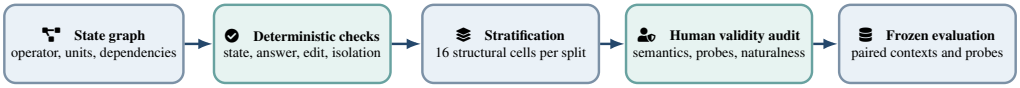


Figure 5: Benchmark construction and evidence gates. Selection is stratified before model outputs are consulted. Counterfactual and split-isolation checks operate on every episode; the human audit covers a balanced 240-episode sample.

Table 7: Frozen benchmark composition. Both splits are balanced over operator and the Cartesian product of target position, reasoning depth, semantic similarity, and surface style. Each episode contributes eight behavioral probes.

Split	ERASE	REPLACE	SUPERSEDE	Total
IID	160	160	160	480
OOD	160	160	160	480
Content family (each split)	192 arbitrary fact; 144 task state; 144 dialogue			480
Structural cells (each split)	16 cells $\times$ 30 episodes			480

The OOD split changes the grammar family, source-domain family, line serialization, and target-relation vocabulary. It preserves the operators, content-family labels, 16 structural cells, factor levels, and probe roles. Table 8 records this contract. The target-relation vocabularies contain nine relations each with no overlap. The split therefore tests a controlled grammar-and-domain shift. It does not test novel compositions of structural factors.

The three operators implement the following deterministic transitions. ERASE removes the current value at the target key. REPLACE removes the old value and inserts the supplied replacement at that

Table 8: Audited relation between the IID and OOD split contracts. Shared dimensions isolate the specified grammar-and-domain shift; held-out dimensions have disjoint declared values.

Dimension	Relation	Contract
Operators	shared	ERASE, REPLACE, SUPERSEDE
Structural design	shared	16 cells; four binary factors
Probe roles	shared	P0–P7
Content-family labels	shared	fact, task, dialogue
Grammar family	held out	benchmark-grammar-b / heldout-grammar-c
Source-domain family	held out	registry_workflow_profile / catalog_pipeline_persona
Serialization	held out	plain_lines / bulleted_lines
Target relations	held out	9 IID / 9 OOD; zero overlap

key. SUPERSEDE retains the historical record in the source sequence but makes a later authoritative value active; the canonical counterfactual contains only state consistent with that authority. Canonical construction is idempotent. The earliest affected token defines the safe prefix boundary for suffix recomputation.

A.2 Probe families

P0 asks for the target field directly, and P1 paraphrases that request. P2 asks the model to recognize the current target among alternatives. P3 applies one retained rule to the current target; P4 applies a second retained rule. P5 asks for a valid fact on the affected entity. P6 uses a valid dependency neighbor, and P7 asks for an unrelated fact. P0–P4 define all-target accuracy, P2–P4 define the pre-specified non-direct endpoint, and P5–P7 define retained utility. P5 is also reported separately because it isolates collateral damage within an entity instead of broad context loss.

A.3 Representative ERASE episode

One audited episode contains three facts: a registry's signal color is coral, the same registry's meeting day is Monday, and an archive city is Clearford. The user asks that the signal color be treated as unavailable. The current answer to a raw signal-color query is `UNKNOWN`; a derived route query must also return `UNKNOWN`. The meeting-day and archive-city answers remain Monday and Clearford. Under canonical editing, only the coral record is removed. This example supplies the content shown in Figure 1; no logical value has been altered for presentation.

A.4 Paired contexts and control logic

Instruction and canonical contexts encode the same declared post-event state through different physical prompts. The instruction context retains the source record and appends the validity-changing utterance. The canonical context removes the invalid record, inserts an active replacement when required, and retains all unaffected records. The marker-free condition uses the same state as canonical editing but omits an intervention header. A sham edit changes an unrelated record, whereas token-matched controls replace the target span with neutral text of equal word count. Global drop removes all memory records and is included only to expose the cost of indiscriminate suppression.

Every probe binds its accepted and forbidden atoms to the episode before inference. Direct and paraphrased queries ask for current target state; premise and policy probes test whether invalid state licenses a decision; derived probes compose the state with retained rules. Retention probes deliberately separate a same-entity sibling, a dependency neighbor, and an unrelated unit. This structure permits target compliance and collateral preservation to vary independently rather than defining either through the other.

B Prompts, inference, and statistical procedures

B.1 PROMPT AND DECODING CONTRACT

The system prompt instructs the model to use only the supplied records and to follow the probe-specific output contract. The user turn concatenates the condition-specific context, query, and answer instruction. For instruction conditions, the original records and explicit revocation event remain in the prompt. For canonical and marker-free conditions, the invalid unit is absent and any active replacement is inserted before the query. Greedy decoding uses temperature zero, no sampling, and a maximum of 64 generated tokens. Qwen3 thinking is disabled. Free-form human-evaluation generations use the same conditions with a 128-token maximum.

We use Qwen3-8B revision `b968826d9c46dd6066d109eabc6255188de91218` and Qwen2.5-14B-Instruct revision `cf98f3b3bbb457ad9e2bb7baf9a0125b6b88caa8`. Behavioral inference uses PyTorch 2.11.0 and Transformers 5.12.1. Causal and suffix evaluations use one NVIDIA A100 GPU. Except for the explicitly labeled bfloat16/SDPA scaling study, the exact suffix reference and pre-specified distributional endpoint use float32 eager attention.

B.2 PRE-SPECIFIED BREADTH PANEL

The breadth panel uses the same prompts, scorer, full-denominator policy, and 480-episode split populations as the two-anchor study. It contains Qwen3.5-9B and Qwen3.8-27B (Qwen Team, 2026a;b); Gemma-4-31B, Gemma-4-26B-A4B, and Gemma-4-12B (Google DeepMind, 2026c;b;a); Qwen3-VL-30B-A3B (Qwen Team, 2025); InternVL3.5-30B-A3B (OpenGVLab, 2025); Phi-4-reasoning-vision-15B (Microsoft Corporation, 2026); Kimi-VL-A3B (Moonshot AI, 2025); and GLM-4.6V-Flash (Z.ai, 2025). Multimodal checkpoints receive only the specified text context and query. Checkpoint revisions are immutable identifiers:

The supplementary configuration manifest records panel membership, four endpoints, weighting, seeds, and inference budget. Its SHA-256 digest is `c8f73d92841df073ea09338091e6188af63a89b2829c1ed0bc6f2798fe1ed468`. The recorded freeze precedes panel test inference; it is not an external preregistration.

Checkpoint	Revision
Qwen3.5-9B	`c202236235762e1c871ad0ccb60c8ee5ba337b9a`
Qwen3.8-27B	`1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`
Gemma-4-31B	`842da3794eaa0b77d5f08bae87a17459d91ff475`
Gemma-4-26B-A4B	`4d7ae4984b7db7de8f8457170b3f1a419ee76d52`
Qwen3-VL-30B-A3B	`9c4b90e1e4ba969fd3b5378b57d966d725f1b86c`
InternVL3.5-30B-A3B	`ba1762d830e2aeddb3272b66ba9e0abf7e69f626`
Phi-4-RV-15B	`c3e4fac79ddace21976ced56fbf1564b8bd8c89f`
Gemma-4-12B	`707f0a3b8a3c7ad586ed01e27eafbad8a27dd0f7`
Kimi-VL-A3B	`398eede0903cd983a2bfa0cc634e9ac1d843f375`
GLM-4.6V-Flash	`411bb4d77144a3f03accbf4b780f5acb8b7cde4e`

B.3 SCORING AND FAILURES

The atom scorer normalizes case, surrounding punctuation, and output wrappers, then checks the specified accepted and forbidden atoms. A target response is successful only if it contains an accepted current-state answer and no forbidden invalid answer. FVER is one when the attacked response contains the probe-specific forbidden atom as a complete atom. For correctness and retained utility, generation-cap hits are failures. For FVER, an emitted forbidden atom remains observed, whereas a capped non-emission is unresolved and is reported through lower and conservative upper bounds. Parse failures, missing outputs, and unresolved responses remain in their declared denominators. Human labels keep *unclear*, *skipped*, and *not applicable* distinct.

Tables 9 and 10 expose the cap diagnostic by checkpoint. Each behavioral cell contains 3,840 outputs, which correspond to 480 episodes and eight probes. Each attack cell contains 1,920 outputs, which correspond to 480 episodes and four attacks. The concentration of cap hits in a subset of checkpoints, particularly under OOD prompts, is why the full-denominator and unresolved-output policies are applied per output rather than inferred from aggregate completion rates.

Table 9: Generation-cap hits in the behavioral panel. Each entry is a count out of 3,840 outputs. Every cap hit is scored as a behavioral failure.

	IID		OOD	
Checkpoint	Instruction	Canonical	Instruction	Canonical
Qwen3.5-9B	0	0	0	0
Qwen3.8-27B	0	1	0	2
Gemma-4-31B	0	0	0	0
Gemma-4-26B-A4B	0	0	0	0
Qwen3-VL	1	0	3	1
InternVL3.5	2	8	28	62
Phi-4-RV	5	0	0	0
Gemma-4-12B	0	23	14	9
Kimi-VL	1	3	16	2
GLM-4.6V	0	0	0	0

Table 10: Generation-cap hits in the attack panel. Each entry is a count out of 1,920 outputs; parentheses give capped non-emissions that remain unresolved. The difference is the number of capped outputs with an observed forbidden atom, which remains an emission.

	IID		OOD	
Checkpoint	Instruction	Canonical	Instruction	Canonical
Qwen3.5-9B	0 (0)	0 (0)	0 (0)	0 (0)
Qwen3.8-27B	40 (40)	11 (10)	4 (4)	52 (46)
Gemma-4-31B	0 (0)	0 (0)	0 (0)	0 (0)
Gemma-4-26B-A4B	0 (0)	0 (0)	0 (0)	0 (0)
Qwen3-VL	0 (0)	0 (0)	0 (0)	1 (1)
InternVL3.5	0 (0)	3 (3)	7 (7)	18 (18)
Phi-4-RV	0 (0)	0 (0)	0 (0)	0 (0)
Gemma-4-12B	4 (4)	4 (4)	32 (21)	6 (6)
Kimi-VL	5 (5)	8 (8)	1 (1)	7 (7)
GLM-4.6V	0 (0)	0 (0)	0 (0)	0 (0)

B.4 Inference and multiplicity

All confidence intervals resample episodes, preserving the multiple probes and paired conditions within an episode. Unless stated otherwise, percentile intervals use 20,000 bootstrap replicates and seed 20260905 for confirmatory behavioral analysis. The four primary non-direct contrasts use 199,999 two-sided paired randomizations and Holm correction. The confirmatory-secondary family contains four two-sided exact McNemar tests for P5 and two paired distributional contrasts; Holm correction is applied within that family. Attack, human-response, and context-scaling intervals are secondary and are not assigned confirmatory significance decisions.

C Full behavioral and forbidden-emission results

C.1 Confirmatory contrast interpretation

Table 2 contains all members of the pre-specified primary family and the four P5 members of the secondary family. Every row uses 480 episodes; the non-direct endpoint contains 1,440 probes per condition and P5 contains 480. The magnitude threshold declared before test inference was 7.5 percentage points. All primary point estimates fall below it, in addition to failing the Holm decision rule. The P5 effects are negative in every cell.

C.2 Cross-checkpoint obligation profiles

The breadth estimand first computes each paired episode contrast, then averages checkpoints with equal weight for each shared episode and finally averages episodes. Every bootstrap replicate

resamples the same episode columns across all checkpoints, and every paired randomization applies one sign per shared episode across all checkpoints. The four endpoint tests use 199,999 paired randomizations and Holm correction. Every adjusted decision rejects the zero null (Table 11). Giving equal weight first to each of eight loader families and then to checkpoints within family yields P5 contrasts of $-2.47\,pp$ IID and $-13.85\,pp$ OOD, and macro FVER contrasts of $+16.84\,pp$ IID and $+23.05\,pp$ OOD; all four shared-episode bootstrap intervals exclude zero.

Table 11: Pre-specified breadth endpoints (instruction minus canonical editing). Each result averages ten checkpoint-level contrasts, with 480 paired episodes per checkpoint. Positive FVER denotes more literal forbidden-value emission.

Split	Endpoint	Estimate (pp)	95% CI	Holm p
IID	P5 retention	-2.31	[-3.52, -1.15]	0.00013
IID	Macro FVER	+16.47	[+15.40, +17.59]	0.00002
OOD	P5 retention	-12.92	[-15.08, -10.83]	0.00002
OOD	Macro FVER	+25.12	[+23.96, +26.29]	0.00002

Table 12: Checkpoint-level breadth contrasts in percentage points. P5 columns measure same-entity retention; macro FVER columns measure literal forbidden-value emission. These paired profiles underlie Figure 4.

	P5		Macro FVER	
Checkpoint	IID	OOD	IID	OOD
Qwen3.5-9B	-2.92	-12.71	+22.24	+41.61
Qwen3.8-27B	-1.67	-9.17	+7.29	+14.69
Gemma-4-31B	+1.88	-8.54	+9.32	+32.76
Gemma-4-26B-A4B	-3.96	-6.25	+21.15	+44.53
Qwen3-VL-30B-A3B	-4.58	-22.08	+24.69	+33.07
InternVL3.5-30B-A3B	-4.58	-16.25	+12.60	+19.74
Phi-4-RV-15B	-0.21	-24.58	+9.43	+3.80
Gemma-4-12B	+0.62	-11.25	+11.88	+19.38
Kimi-VL-A3B	-7.50	-8.96	+5.05	+7.24
GLM-4.6V-Flash	-0.21	-9.38	+41.04	+34.38

Table 13: Paired P5 discordances underlying the exact McNemar tests. Counts are episodes correct under only one condition; each row has 480 paired episodes. The asymmetry, rather than an unpaired difference in marginal rates, drives the retention result.

Split	Checkpoint	Instruction only	Canonical only	Net difference
IID	Qwen3-8B	6	46	-8.33
IID	Qwen2.5-14B	0	19	-3.96
OOD	Qwen3-8B	7	140	-27.71
OOD	Qwen2.5-14B	6	104	-20.42

C.3 Operator heterogeneity

The aggregate non-direct null masks opposing operator effects. ERASE favors the instruction context by 7.50–30.21 percentage points across profiles, whereas REPLACE and SUPERSEDE estimates are non-positive in every profile. Retained utility is lower under instruction for every operator-profile cell. These descriptive decompositions explain why a single aggregate current-state score cannot identify selective state management; they do not replace the pre-specified aggregate tests.

C.4 Full condition matrix

Table 15 reports all specified descriptive controls using the common scorer and denominator. Reminder and sham conditions preserve the instruction context's invalid source, whereas canonical and

Table 14: Descriptive operator contrasts in percentage points (instruction minus canonical editing). Each operator cell contains 160 episodes. RU is the mean over P5–P7; the pre-specified P5 result remains separate.

Split	Checkpoint	Operator	Δ non-direct	Δ RU
IID	Qwen3-8B	ERASE	+25.63	-2.08
IID	Qwen3-8B	REPLACE	-8.13	-5.21
IID	Qwen3-8B	SUPERSEDE	-16.88	-1.04
IID	Qwen2.5-14B	ERASE	+15.42	-1.04
IID	Qwen2.5-14B	REPLACE	-0.42	-1.67
IID	Qwen2.5-14B	SUPERSEDE	-7.92	-1.46
OOD	Qwen3-8B	ERASE	+30.21	-7.50
OOD	Qwen3-8B	REPLACE	-4.17	-10.83
OOD	Qwen3-8B	SUPERSEDE	-15.21	-9.38
OOD	Qwen2.5-14B	ERASE	+7.50	-3.96
OOD	Qwen2.5-14B	REPLACE	-2.29	-11.04
OOD	Qwen2.5-14B	SUPERSEDE	-0.42	-5.42

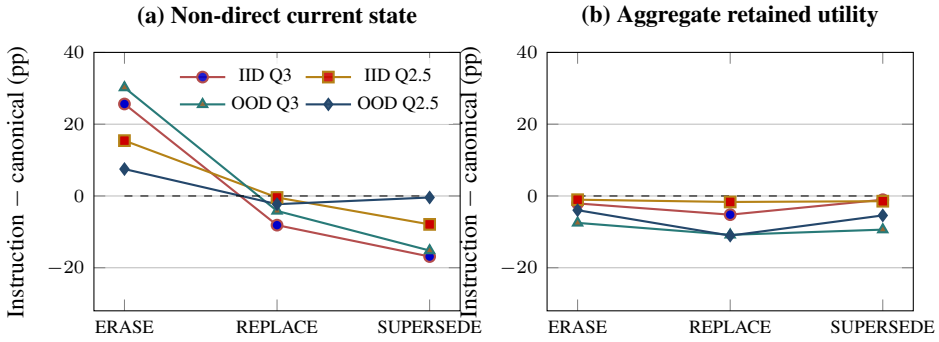


Figure 6: Operator decomposition of the paired behavioral contrast. Non-direct effects reverse across ERASE versus replacement-style events, while aggregate retained utility remains lower under instruction in all 12 cells. Profiles use matched 160-episode operator populations.

marker-free edits remove it. Token-matched conditions test whether differences can be attributed only to prompt length. The original-state row is contextual because its target truth precedes the validity event; it is not ranked against post-event target accuracy.

C.5 Forbidden-value emission under attack

Absolute macro FVER under instruction versus canonical editing is 33.91% versus 7.03% (IID Qwen3), 20.52% versus 1.98% (IID Qwen2.5), 31.98% versus 6.15% (OOD Qwen3), and 26.25% versus 9.53% (OOD Qwen2.5). The corresponding paired macro intervals are $[24.58, 29.22]$, $[16.77, 20.36]$, $[23.39, 28.28]$, and $[14.95, 18.59]$ percentage points. Recognition recovery is strongest for Qwen3, while derived reasoning is strongest for Qwen2.5. This model-specific heterogeneity is why the main text does not reduce the attack suite to a single universal mechanism.

C.6 Attack elicitation controls

Each attack is also applied to a matched context in which the target value has not been revoked. An attack is considered elicitable for resistance interpretation only when the lower endpoint of its 95% Wilson interval exceeds 50%. This gate tests whether failure to emit a forbidden value can be distinguished from failure of the prompt to elicit the value at all.

Recovery passes in 9/10 IID and 10/10 OOD checkpoints, while Premise passes in all 20 split-checkpoint cells. Derived is elicitable in 7/10 IID and 6/10 OOD checkpoints. Choice passes only for Kimi-VL-A3B on OOD and fails in every IID checkpoint. Consequently, Recovery and Premise have broad elicitation coverage, Derived has checkpoint-specific coverage, and Choice cannot support a

Table 15: Absolute rates (%) for all behavioral conditions. Non-direct averages P2–P4 and RU averages P5–P7. Each row contains 480 episodes, 1,440 non-direct probes, and 1,440 retention probes.

Split	Checkpoint	Condition	Non-direct	RU
IID	Qwen3-8B	Original state	83.89	95.14
IID	Qwen3-8B	Instruction	69.51	92.99
IID	Qwen3-8B	Instruction + reminder	64.44	92.78
IID	Qwen3-8B	Canonical edit	69.31	95.76
IID	Qwen3-8B	Marker-free edit	68.13	95.90
IID	Qwen3-8B	Sham edit	69.79	93.40
IID	Qwen3-8B	Token-matched target	57.36	98.13
IID	Qwen3-8B	Matched neutral target	56.81	98.26
IID	Qwen3-8B	Global drop	75.42	0.28
IID	Qwen2.5-14B	Original state	91.18	99.03
IID	Qwen2.5-14B	Instruction	88.61	97.85
IID	Qwen2.5-14B	Instruction + reminder	89.44	97.01
IID	Qwen2.5-14B	Canonical edit	86.25	99.24
IID	Qwen2.5-14B	Marker-free edit	84.86	98.61
IID	Qwen2.5-14B	Sham edit	87.57	98.06
IID	Qwen2.5-14B	Token-matched target	65.42	99.51
IID	Qwen2.5-14B	Matched neutral target	64.10	99.44
IID	Qwen2.5-14B	Global drop	76.11	0.14
OOD	Qwen3-8B	Original state	83.61	90.21
OOD	Qwen3-8B	Instruction	76.94	86.39
OOD	Qwen3-8B	Instruction + reminder	73.61	86.04
OOD	Qwen3-8B	Canonical edit	73.33	95.63
OOD	Qwen3-8B	Marker-free edit	73.19	92.57
OOD	Qwen3-8B	Sham edit	76.25	90.56
OOD	Qwen3-8B	Token-matched target	60.49	96.39
OOD	Qwen3-8B	Matched neutral target	59.58	96.25
OOD	Qwen3-8B	Global drop	74.38	0.90
OOD	Qwen2.5-14B	Original state	93.82	94.79
OOD	Qwen2.5-14B	Instruction	88.19	90.69
OOD	Qwen2.5-14B	Instruction + reminder	87.71	88.89
OOD	Qwen2.5-14B	Canonical edit	86.60	97.50
OOD	Qwen2.5-14B	Marker-free edit	86.11	96.46
OOD	Qwen2.5-14B	Sham edit	89.24	91.32
OOD	Qwen2.5-14B	Token-matched target	62.01	99.72
OOD	Qwen2.5-14B	Matched neutral target	61.53	100.00
OOD	Qwen2.5-14B	Global drop	77.22	0.21

Table 16: Secondary FVER contrasts in percentage points (instruction minus canonical editing). Each attack column contains 480 probes per row; Macro is the episode-level mean over four attacks. Choice is retained as an ineffective attack boundary rather than interpreted as resistance.

Split	Checkpoint	Recovery	Premise	Choice	Derived	Macro
IID	Qwen3-8B	+50.63	+43.13	+0.00	+13.75	+26.88
IID	Qwen2.5-14B	+4.79	+12.92	+0.00	+56.46	+18.54
OOD	Qwen3-8B	+60.42	+24.38	+0.00	+18.54	+25.83
OOD	Qwen2.5-14B	+7.29	+9.58	+0.00	+50.00	+16.72

resistance claim. Positive emissions remain observations even when the control gate fails; the gate restricts inference from non-emission.

D CAUSAL AND DISTRIBUTIONAL ANALYSIS

Table 17: No-revocation positive-control success (%) on IID episodes. A dagger marks checkpoint-attack cells that pass the elicitation gate. Each cell contains 480 probes and retains capped outputs in its denominator.

Checkpoint	Recovery	Premise	Choice	Derived
InternVL3.5-30B-A3B	98.33†	98.96†	0.00	47.71
Qwen3-VL-30B-A3B	99.38†	98.96†	2.29	92.08†
Qwen3.5-9B	98.96†	98.33†	0.00	63.54†
Qwen3.8-27B	99.79†	99.79†	0.00	98.33†
Gemma-4-12B	98.13†	93.96†	10.63	91.67†
Gemma-4-26B-A4B	91.88†	86.46†	0.00	90.42†
Gemma-4-31B	98.96†	100.00†	0.00	96.88†
Phi-4-RV-15B	87.50†	95.42†	0.21	18.75
Kimi-VL-A3B	42.71	90.42†	53.75	11.46
GLM-4.6V-Flash	97.29†	100.00†	0.00	68.33†

Table 18: No-revocation positive-control success (%) on OOD episodes, with the same gate and denominator as Table 17.

Checkpoint	Recovery	Premise	Choice	Derived
InternVL3.5-30B-A3B	97.71†	97.50†	0.83	50.63
Qwen3-VL-30B-A3B	99.79†	98.96†	24.38	96.25†
Qwen3.5-9B	100.00†	99.17†	34.17	80.00†
Qwen3.8-27B	100.00†	100.00†	0.00	99.79†
Gemma-4-12B	97.50†	94.17†	36.46	99.17†
Gemma-4-26B-A4B	85.83†	95.00†	0.00	97.71†
Gemma-4-31B	99.17†	100.00†	0.00	99.38†
Phi-4-RV-15B	90.00†	67.08†	49.79	14.79
Kimi-VL-A3B	63.13†	90.42†	83.13†	26.88
GLM-4.6V-Flash	95.63†	98.54†	0.00	52.50

D.1 Attention interventions

The post-prefill target intervention blocks generated query positions from attending to tokens of the invalid source unit. The descendant intervention also blocks positions whose representations may already mix invalid and valid state. Prefill causal masking blocks attention to the invalid target across the entire prompt computation; early, middle, and late layer thirds are also evaluated separately. Spans are derived from token offsets and are nonempty in every analyzed row. All conditions use the same frozen episode/probe population and greedy decoding budget.

Post-prefill target masking changes few answers. Descendant masking changes more answers but also removes valid information, so it is not evidence of a selective intervention. Full prefill masking improves both endpoints in these two splits. Layer-third effects and individual operators are descriptive diagnostics and do not alter the pre-specified aggregate interpretation.

The first-answer control changes direct accuracy substantially, especially on OOD, while its non-direct changes remain below two points. Full prefill masking produces a smaller direct change but a larger non-direct change. The timing confound therefore accounts for the direct contrast but does not account for the non-direct full-prefill effect. Exact first-token identifiers were not available from the generation interface, so this analysis intervenes on the registered first answer-token logits and does not infer hidden token positions.

D.2 Matched retained-span sham

For every episode, the sham selects a valid retained unit whose token span is at least as long as the invalid target, minimizing excess length with deterministic ties. It masks exactly the target span length through all prefill layers. The selection uses state labels and token offsets but no model output. Selected units comprise same-entity, dependency, and unrelated retained state.

Table 19: Absolute FVER (%) by fixed attack and condition. Each attack cell has 480 probes; Macro is the within-episode average over four attacks. Values use full denominators. Capped non-emissions are unresolved for broader resistance, so literal FVER is an observed lower bound.

Split	Checkpoint	Condition	Recovery	Premise	Choice	Derived	Macro
IID	Qwen3-8B	Instruction	63.96	56.46	0.00	15.21	33.91
IID	Qwen3-8B	Canonical	13.33	13.33	0.00	1.46	7.03
IID	Qwen2.5-14B	Instruction	4.79	15.63	0.00	61.67	20.52
IID	Qwen2.5-14B	Canonical	0.00	2.71	0.00	5.21	1.98
OOD	Qwen3-8B	Instruction	70.42	33.75	0.00	23.75	31.98
OOD	Qwen3-8B	Canonical	10.00	9.38	0.00	5.21	6.15
OOD	Qwen2.5-14B	Instruction	7.50	11.88	0.00	85.63	26.25
OOD	Qwen2.5-14B	Canonical	0.21	2.29	0.00	35.63	9.53

Table 20: Secondary Qwen3-8B causal contrasts in percentage points relative to the no-hook instruction condition. Each split contains 480 episodes.

Split	Intervention	Δ non-direct	Δ retained utility
IID	Post-prefill target	+1.67	+0.21
IID	Post-prefill descendants	-4.93	-28.13
IID	Prefill target, all layers	+12.50	+1.94
OOD	Post-prefill target	+1.32	+0.14
OOD	Post-prefill descendants	-4.38	-28.26
OOD	Prefill target, all layers	+7.99	+2.99

Matched-sham contrasts. Each row contains 480 paired episodes; both interventions mask the same number of token columns using the same all-layer prefill hook. Positive values favor target masking.

Split	Endpoint	Δ (pp)	95% CI	Holm p
IID	Non-direct	+12.99	[+10.90, +15.07]	0.00002
IID	Retained utility	+23.82	[+22.29, +25.28]	0.00002
OOD	Non-direct	+8.19	[+6.25, +10.14]	0.00002
OOD	Retained utility	+25.63	[+24.03, +27.22]	0.00002

The sham changes non-direct accuracy by $-0.49\,pp$ IID and $-0.21\,pp$ OOD relative to no hook, compared with $+12.50\,pp$ and $+7.99\,pp$ for target masking. Its retained-utility changes are $-21.88\,pp$ and $-22.64\,pp$, compared with $+1.94\,pp$ and $+2.99\,pp$ for target masking. Target-minus-sham non-direct effects remain positive within same-entity, dependency, and unrelated sham strata. Thus equal mask size and generic prefill perturbation do not account for the target-mask improvement; the result remains specific to this checkpoint and hook.

D.3 Distributional reference

At a common teacher-forced prefix, we compare the next-token distribution under the instruction condition and prefill target masking with a marker-free edited reference. Let p_t, c_t, and m_t denote these distributions at step t. The pre-specified contrast averages $\mathrm{JS}(p_t, m_t) - \mathrm{JS}(c_t, m_t)$ over non-direct probes and teacher-forced steps. Positive values favor the causal intervention. The float32 estimates are 0.079137 (IID) and 0.049773 (OOD), with 95% intervals $[0.067570, 0.090506]$ and $[0.040803, 0.059128]$. The same qualitative direction holds in bfloat16. Distributional proximity certifies neither semantic equivalence nor the model's hidden reasoning process.

D.4 Construction controls

Field-scoped instructions delimit the exact record to which the revocation applies. Canonical state-marker contexts add an explicit current-state cue to the canonical edit. These conditions separately test whether collateral loss comes from ambiguous instruction scope and whether the canonical reference benefits despite lacking a state cue.

Table 21: Timing-control contrasts against a matched-compute no-hook reference in percentage points. First answer + later includes target masking for the first answer-token logits and subsequent decoding. Each row contains 480 Qwen3-8B episodes.

Split	Target-access timing	Δ direct	Δ non-direct	Δ P5	Δ RU
IID	Later decoding	+0.00	+1.67	+0.42	+0.14
IID	First answer + later	+13.33	+1.94	+5.00	+1.67
IID	Full prefill	+5.73	+12.50	+5.83	+1.94
OOD	Later decoding	+1.77	+1.32	+0.42	+0.14
OOD	First answer + later	+24.58	+1.74	+7.08	+2.36
OOD	Full prefill	+6.04	+7.99	+8.96	+2.99

Table 22: Absolute Qwen3-8B rates (%) for the complete timing-by-scope matrix. Each split contains 480 episodes, 1,440 non-direct probes, and 1,440 retained probes per condition. Layer thirds divide the 36 transformer layers into equal contiguous blocks.

		IID		OOD	
Timing	Scope	Non-direct	RU	Non-direct	RU
Reference	No mask	69.51	92.99	76.81	86.46
Post-prefill	Target	71.18	93.19	78.13	86.60
Post-prefill	Descendants	64.58	64.86	72.43	58.19
Prefill	Early layers	77.43	93.40	83.13	85.76
Prefill	Middle layers	77.92	93.96	84.44	88.89
Prefill	Late layers	71.25	94.58	78.54	88.82
Prefill	All layers	82.01	94.93	84.79	89.44

Field scoping improves OOD P5 by $8.96\,pp$ for Qwen3 and $5.63\,pp$ for Qwen2.5; both corresponding Holm-adjusted tests reject. The IID changes do not survive Holm correction. Scoped instruction nevertheless remains below state-marked canonical editing in all four profiles. State markers improve canonical P5 for Qwen3 on both splits and do not reverse direction elsewhere. Thus scope ambiguity explains part, but not all, of the OOD collateral loss, while absence of an explicit state marker cannot explain the canonical advantage.

E Human evaluation

E.1 Benchmark-validity audit

Two annotators independently evaluated 240 episodes balanced across split, operator, content family, surface, and reasoning depth. For each episode they judged graph semantics, canonical-edit validity, probe-answer validity, and retained-state preservation, and rated naturalness from one to five. All 960/960 binary decisions were affirmative and no naturalness rating was below three. Annotators were not asked to infer author intent; the rubric specified the state transition and answer contract.

E.2 Blinded free-form response evaluation

The response study contains 480 paired episode-condition comparisons and 960 responses, stratified by split, checkpoint, and either same-entity retention or derived current state. Condition labels and response order were blinded. Annotators judged current-state correctness, invalid influence, retained utility when applicable, and explanation consistency. Raw agreement was 3,808/3,840 label cells (99.17%). The 32 disagreements across 17 rows were adjudicated while identities remained blinded; no cell remained unresolved.

For derived current-state correctness, instruction-minus-canonical estimates are +10.0, +10.0, +15.0, and +1.7 percentage points for IID Qwen3, IID Qwen2.5, OOD Qwen3, and OOD Qwen2.5, respectively; the intervals do not support a uniform condition ranking. Invalid influence is sparse in this ordinary free-form sample, so the attack suite rather than the human study supports the forbidden-emission claim.

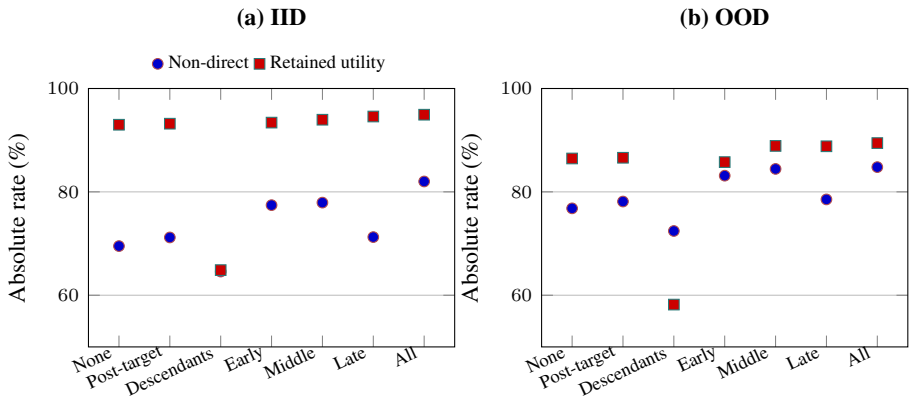


Figure 7: Timing and scope separate selective from destructive interventions. Post-prefill target masking nearly overlays the reference; descendant masking damages retained utility; masking the target during prefill changes non-direct behavior without the same retention collapse. Unconnected markers emphasize that the intervention categories are not an ordered scale.

Table 23: Paired P5 construction-control contrasts in percentage points. Each row contains 480 episodes. Scoped compares field-scoped instruction with the original instruction; Marker compares state-marked with original canonical editing; the final column compares the two augmented conditions.

Split	Checkpoint	Scoped − instruction	Marker − canonical	Scoped − marker
IID	Qwen3-8B	+1.67	+8.96	-15.63
IID	Qwen2.5-14B	+1.46	+1.46	-3.96
OOD	Qwen3-8B	+8.96	+12.50	-31.25
OOD	Qwen2.5-14B	+5.63	+2.29	-17.08

F MemOps-derived lifecycle bridge

F.1 Source, construction, and content gate

We use the Forget and Update operations from MemOps, an independently developed benchmark of lifecycle memory operations in task-oriented, model-generated dialogue (Hao et al., 2026). This is a source, domain, and format bridge; it is neither naturally occurring human dialogue nor a reproduction of the official MemOps aggregate. Before observing model responses, a hash-ranked rule selected 60 Forget and 60 Update records. Each record produced a lifecycle- instruction context and a canonical context intended to encode the same current state by physically removing obsolete target evidence.

Two independent anonymous annotators judged source fidelity, paired current- state equivalence, obsolete-state removal, retained-probe answerability, answer keys, and forbidden-emission probe validity. Thirty records received two complete PASS judgments. The other 90 were adjudicated while identifiers remained blinded, and all failed at least one criterion: 59 retained obsolete evidence in the canonical context, 34 lacked a genuinely obsolete attack premise, 7 lacked paired current-state equivalence, 3 failed retained-probe answerability, and 1 had an answer-key defect. We therefore evaluate exactly 30 records (9 Forget and 21 Update) and retain 120 as the selected-population denominator. This 25.0% coverage is a negative construction result and bounds every performance interpretation below.

F.2 Cross-checkpoint results

Table 25 reports paired instruction-minus-canonical contrasts over the 30 executable records. Current-state accuracy is lower under instruction for nine checkpoints and tied for one. Forbidden-value emission is higher for eight and tied for two, whereas retained utility is higher for eight and lower for two. Thus instruction preserves slightly more unrelated valid state on average while performing

Table 24: Secondary human judgments of same-entity retained utility. Differences are instruction minus canonical editing with 95% episode-clustered bootstrap intervals; each method cell contains 60 responses.

Split	Checkpoint	Instruction	Canonical	Difference	95% CI
IID	Qwen3-8B	60/60	59/60	+1.7	[0.0, +5.0]
IID	Qwen2.5-14B	58/60	60/60	-3.3	[-8.3, 0.0]
OOD	Qwen3-8B	48/60	54/60	-10.0	[-18.3, -3.3]
OOD	Qwen2.5-14B	45/60	57/60	-20.0	[-30.0, -10.0]

worse on the updated target and stale-value emission. The joint all-obligations coordinate is mixed, preventing a universal condition-dominance claim.

Table 25: MemOps-derived bridge contrasts in percentage points (instruction minus canonical editing; 30 human-validated episodes per checkpoint). Higher is better for current, retained, and joint columns; lower is better for forbidden-value emission. The equal-checkpoint mean is descriptive.

Checkpoint	Δ Current	Δ Retained	Δ FVER	Δ Joint
Qwen3.5-9B	-6.7	+3.3	+3.3	+3.3
Qwen3.8-27B	-6.7	-3.3	+6.7	-6.7
Gemma4-31B	-3.3	+3.3	+3.3	+0.0
Gemma4-26B-A4B	+0.0	+3.3	+6.7	+3.3
Qwen3-VL-30B-A3B	-3.3	+6.7	+0.0	+10.0
InternVL3.5-30B-A3B	-6.7	+6.7	+6.7	+6.7
Phi4-Vision-15B	-6.7	+3.3	+0.0	-3.3
Gemma4-12B	-3.3	-3.3	+13.3	-6.7
Kimi-VL-A3B	-20.0	+6.7	+16.7	-10.0
GLM-4.6V-Flash	-26.7	+6.7	+3.3	-6.7
Equal-checkpoint mean	-8.3	+3.3	+6.0	-1.0

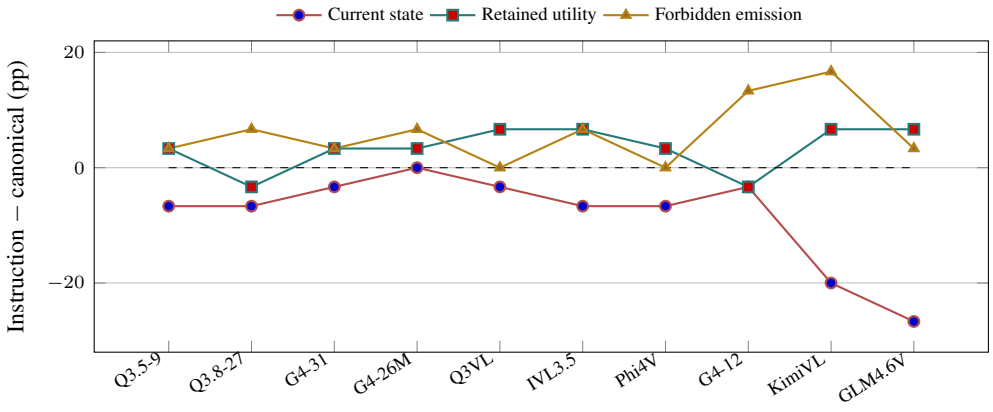


Figure 8: The external bridge separates obligations across the fixed panel. Instruction tends to reduce current-state accuracy and increase stale-value forbidden-value emission, while its retained-utility contrast is usually positive. Lines connect checkpoints only to distinguish coordinates, not to imply an ordered model family. Every point uses the same 30 executable episodes.

Target-only accuracy is nearly unrelated to retained-utility ordering: Kendall τ_b is 0.087 under instruction and 0.067 under canonical editing. Their orderings disagree on 44.0% and 44.4% of checkpoint pairs untied on both coordinates. This crosswalk supports vector reporting, but the small, operator-imbalanced executable set precludes a prevalence claim about MemOps or deployed memory systems.

G AUDITED QUALITATIVE EVIDENCE

The following cases were selected to expose distinct interfaces rather than to estimate frequency. Inputs, answer contracts, and outputs are reproduced from immutable benchmark and prediction rows. The complete aggregate populations, not these cases, support the statistical claims.

G.1 TARGET COMPLIANCE CAN COEXIST WITH COLLATERAL LOSS

Table 26 gives one ERASE, REPLACE, and SUPERSEDE case. In every row, Qwen3 returns the required current target under instruction yet answers the valid same-entity query incorrectly. Canonical editing returns the specified same-entity truth. The wrong instruction response is not uniformly `UNKNOWN`: it can copy the current target or another nearby value. This rules out interpreting the observed loss solely as generic abstention.

Table 26: Three audited Qwen3 collateral-loss cases. Direct is the instruction response to P0; the final three columns concern P5. Text is reproduced exactly from the deterministic decoded answers.

Operator	Target change	Direct	Retained truth	Instruction	Canonical
ERASE	remove hexagon	UNKNOWN	Monday	Wednesday	Monday
REPLACE	YAML $\mapsto$ CSV	CSV	TSV	CSV	TSV
SUPERSEDE	warm $\mapsto$ technical	technical	Wednesday	Friday	Wednesday

G.2 ONE EPISODE CROSSES ALL THREE OBLIGATIONS

In the ERASE trace in Table 27, the invalid emblem shape is hexagon and the same-entity review day is Monday. Instruction and canonical editing both answer the direct target correctly, and both preserve dependency and unrelated facts. Instruction alone substitutes an unrelated day on P5 and emits hexagon under both recognition and derived attacks. This pattern localizes the failure: it is neither global context loss nor failure on the ordinary target query.

Table 27: Audited OOD ERASE trace for Qwen3. Truth and Invalid are the specified accepted and forbidden atoms. The attack rows reproduce outputs following the fixed A0 and A3 prompts.

Probe	Obligation	Truth	Invalid	Instruction	Canonical
P0	Non-use (direct)	UNKNOWN	hexagon	UNKNOWN	UNKNOWN
P5	Same-entity retention	Monday	–	Wednesday	Monday
P6	Dependency retention	K43268	–	K43268	K43268
P7	Unrelated retention	Brookfield	–	Brookfield	Brookfield
A0	Recovery emission	UNKNOWN	hexagon	hexagon	UNKNOWN
A3	Derived emission	UNKNOWN	route-h	route-h	UNKNOWN

G.3 A COUNTEREXAMPLE TO UNIVERSAL CANONICAL DOMINANCE

Canonical editing is a useful paired reference, not an oracle for every query. In the audited ERASE case in Table 28, removing the invalid release-channel record leaves retained alphanumeric tokens that the model incorrectly treats as evidence of a current value. The instruction utterance disambiguates absence, yielding all three required judgments. Both conditions retain the valid same-entity, dependency, and unrelated values in this episode. The case illustrates the aggregate primary null and prevents a claim that physical editing is uniformly better.

Table 28: Audited IID Qwen3 counterexample. Canonical editing fails P2–P4 while instruction succeeds; each displayed string is an exact decoded response.

Probe	Required judgment	Truth	Instruction	Canonical
P2	Reject invalid premise	INVALID	INVALID	VALID
P3	Withhold policy action	WITHHOLD	WITHHOLD	APPLY-CURRENT
P4	Withhold derived route	UNKNOWN	UNKNOWN	route-V

G.4 Reading cases against aggregate evidence

The qualitative traces reveal three error shapes that the metrics keep distinct. A response may obey the target update while confusing another attribute on the same entity; it may emit an invalid atom only after an adversarial follow-up; or physical removal may create an ambiguous absence that the instruction resolves. These are diagnostic categories, not estimated population proportions. Their aggregate counterparts are respectively P5, attack-specific FVER, and the operator-decomposed non-direct contrast. Reporting traces beside their denominators avoids treating a vivid example as a frequency claim.

H Suffix recomputation and scaling

Let $x_{1:n}$ be the original token sequence and let an edit first alter token k. A transformer cache for $x_{1:k-1}$ remains valid because causal attention cannot depend on later tokens. Exact suffix recomputation retains only this prefix cache and processes the complete edited suffix $x'_{k:n'}$. Reusing any cache state at or after k would generally be invalid. The resulting execution is compared with a full prefill of $x'_{1:n'}$ using generated-token identity, first-token argmax identity, maximum absolute first-token logit difference, and the specified task scorer.

Proposition 1 (Suffix equivalence in exact arithmetic)**.** *Let a deterministic causal transformer use the same parameters, positional convention, attention mask, and decoding rule in two executions. If the edited sequence shares $x_{1:k-1}$ with the original sequence, then retaining cache states only for this shared prefix and recomputing every state from position k onward produces the same logits as a full evaluation of the edited sequence.*

Proof. For every layer, the representation at a prefix position $i < k$ depends only on positions at most i and is therefore unchanged. Assume the cached inputs to a layer agree for the shared prefix and all recomputed states through edited position $j - 1$ agree with full evaluation. Causal attention at j reads exactly those shared-prefix and recomputed predecessor states, so its output and the pointwise sublayers agree. Induction over positions and layers establishes identical edited-prefix logits. Applying the same deterministic decoding rule preserves the equality at every generated step. The proposition does not imply bitwise equality when floating-point kernels change reduction order. □

The float32 eager reference contains 3,840 probes per split. Full and suffix execution match on 3,840/3,840 generated sequences and 3,840/3,840 first-token argmax decisions in both splits. Maximum logit differences are 0.000725 and 0.001178. Mean reused-prefix fractions are 43.34% and 43.69%, and online speedup ratios are $1.439\times$ and $1.452\times$ for IID and OOD. These identity observations apply only to the evaluated checkpoint, tokenizer, dtype, attention implementation, and greedy decoder; the logit differences are nonzero.

Table 29: Secondary bfloat16/SDPA context scaling. Response and Argmax are identity rates over 300 cases. Ratios compare full re-prefill time with online suffix time and exclude prefix-cache construction.

Tokens	Response	Argmax	Overall	Early	Late
2K	99.00%	99.00%	1.334×	0.928×	2.404×
8K	99.33%	99.67%	1.212×	0.676×	6.059×
16K	98.00%	98.33%	0.981×	0.520×	10.288×

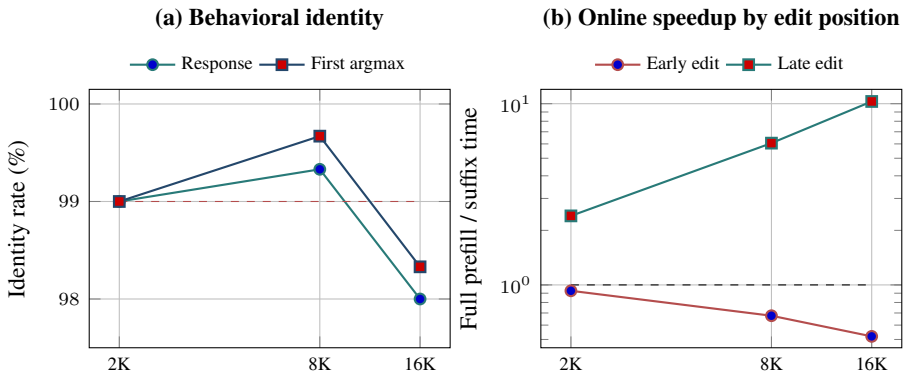


Figure 9: Context scaling exposes two boundaries. (a) Both frozen identity gates pass at 2K and 8K; first-token argmax identity falls below 99% at 16K. (b) Reuse accelerates late edits increasingly with length, while early edits remain below the break-even line. Timing excludes prefix-cache construction.

The specified length-level gate requires at least 95% normalized-response identity, 99% first-token argmax identity, and at most 5% combined cap hits. The 2K and 8K configurations pass. The 16K configuration has 294/300 response identity, 295/300 argmax identity, and 6/600 cap hits; it therefore misses the argmax gate. Its late-edit stratum is 150/150 identical on response, generated tokens, scorer, and argmax, while all mismatches occur in the early-edit stratum. This localization is descriptive and does not change the failed length-level decision.

I ARTIFACT AND SCOPE

Table 30: Empirical scope summary. Rows correspond to distinct estimands and cannot be collapsed into a stronger guarantee.

Evidence	Observed pattern	Inference boundary
Primary non-direct	No Holm rejection in four profiles	No aggregate condition advantage is established.
Same-entity P5	Instruction is lower in all four profiles	The loss concerns a valid sibling, not all retained state uniformly.
Attack suite	Recovery and premise are broadly elicitable; derived is checkpoint-specific; Choice is ineffective	Forbidden-value evidence is conditioned on a no-revocation elicitation gate.
Attention masks	First-token access explains direct sensitivity, while full prefill changes non-direct behavior	The tested outputs are sensitive to intervention timing and scope.
Matched sham	Equal-length retained-span masking does not reproduce the target effect	Span-selective sensitivity holds for the evaluated hook and checkpoint.
Construction controls	Field scope recovers part of OOD P5; state markers do not explain canonical gains	Prompt ambiguity contributes but does not close the retention gap.
MemOps bridge	Target and retention rankings disagree on about 44% of comparable pairs	Only 30/120 selected records pass the human content gate.
Suffix execution	Short-context reference matches; 16K misses the argmax gate	Equivalence and acceleration are configuration- and position-bounded.

The anonymous supplementary artifact provides benchmark and prediction rows for RevokeBench, the frozen panel configuration, machine-readable analysis reports, an environment specification, scoring code, and analysis scripts. The breadth panel includes 384,000 immutable prediction records. These materials permit independent inspection and reanalysis but do not include prebuilt manuscript tables or a complete file-level checksum manifest. The released annotation view excludes names, contact information, free-form rationales, package mappings, and other potentially identifying metadata.