HIDDEN REASONING CHANGES KV STATES MORE THAN IT CHANGES BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

When a reasoning language model deletes its hidden trace before the next conversational turn, the answer's cached key–value states become stale: they were computed at the wrong positions and under the influence of text that no longer exists. Serving systems handle this by recomputing the answer's cache from the visible transcript, at a cost linear in answer length on every follow-up. We find that much of this recomputation can be avoided. An exact RoPE rotation fixes the positional error, and the remaining content contamination—though large in norm and more behaviorally sensitive than a random perturbation of the same size—rarely changes the model's next-token decisions on the evaluated trajectories. Representational, distributional, and decision fidelity turn out to be separable: the size of the cache error does not predict which turns drift (Spearman $\rho \approx 0$), but the decision margin does, and since it is readable from the reuse forward alone it serves as an online signal for selective recomputation (AUC 0.93). The finding replicates across two dense models and a restricted hybrid-attention model (84–97% follow-up-prefill top-1 agreement), and a matched control against visible-span deletion and prompt swapping shows it is a general property of position-portable cache reuse, with hidden-reasoning deletion the hardest case. On two-turn tasks, zero-repair reuse establishes non-inferiority to recomputation on GSM8K at a five-percentage-point margin, while HumanEval accuracy is 1.2 points lower without establishing non-inferiority. In patched vLLM experiments, median next-turn time-to-first-token improves by $2.86\times$ and $3.85\times$ at concurrent batch sizes 8 and 24, respectively, excluding prior-turn compaction.

1 INTRODUCTION

Serving systems cache the key–value (KV) states of repeated prefixes so that a conversational follow-up pays only for new tokens (Kwon et al., 2023; Zheng et al., 2024). Hidden-reasoning deletion breaks this abstraction. The visible answer was generated while attending to a private trace that is absent from the next turn's transcript. Its cached states therefore have both the wrong positions and a different causal history (fig. 1). Recomputing the answer from the visible transcript restores the reference cache, but repeats answer-length prefill work on every follow-up.

We ask whether this recomputation is necessary for future behavior. A constant RoPE rotation exactly corrects the positions of retained keys. It cannot remove the content dependence on deleted reasoning: the corrected answer cache still differs substantially from a fresh prefill. Nevertheless, most evaluated next-token decisions survive. This separates three notions often conflated in cache reuse: fidelity of the tensors, of the output distribution, and of the selected token. A large error under either of the first two need not change the third.

The distinction alone leaves an explanatory gap. Saying that a logit perturbation did not cross a margin describes a preserved decision but does not identify why a large cache error had a limited effect. We address this gap by decomposing a pairwise logit displacement into two internal sources: erroneous value readout and attention rerouting caused by erroneous keys. Integrating these sources gives an exact finite-amplitude attribution under stated smoothness assumptions. The resulting token scores support a controlled intervention: at a four-token repair budget, attribution restores the reference winner at 59.8% of evaluated flipped held-out positions, compared with 17.2% for attention and 11.5% for cache-error norm. This is evidence for task-specific localization of harmful error; it does not make the reference-dependent attribution an online repair algorithm.

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.

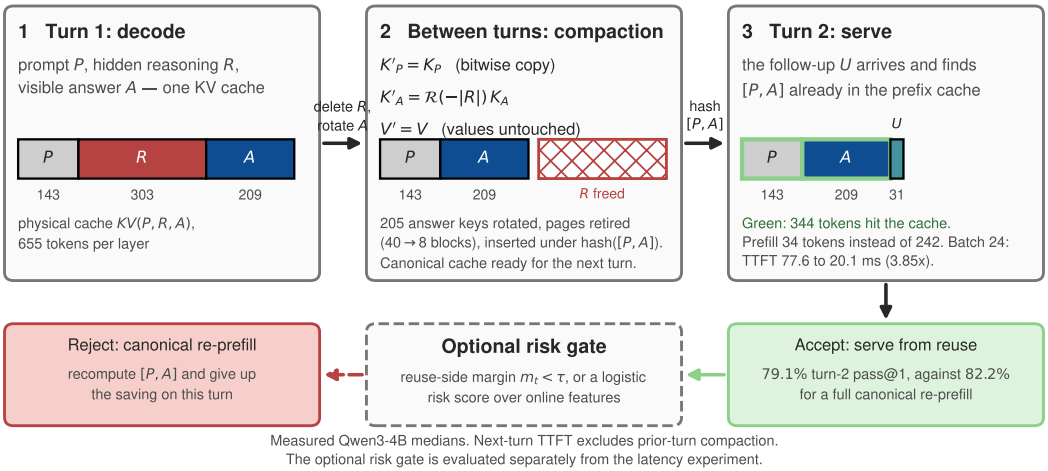


Figure 1: End-to-end serving flow. Turn 1 produces the physical $[P, R, A]$ cache. Between turns, compaction preserves prefix keys and values, rotates answer keys, registers canonical $[P, A]$ blocks, and retires physical-cache entries. Turn 2 reuses the retained answer. The optional risk gate is evaluated separately from the timing path. Span lengths and runtime annotations come from the vLLM measurements. TTFT measures the next request with the canonical cache prepared.

Our contributions connect that mechanism to its empirical and systems consequences. First, we establish that cache magnitude poorly predicts behavioral drift in the core study, and replicate the separation of representation and decision fidelity across two dense models and a restricted hybrid-attention model and matched transcript-edit controls. Second, we derive the attention-level decomposition and test reconstruction, attenuation, and targeted repair on two dense models. Third, we distinguish offline explanation from inexpensive online risk ranking: the reuse-side margin predicts token disagreement at AUC 0.93, without access to a reference forward. Finally, a patched vLLM runtime verifies the paged-KV transformation, grades substantive follow-up answers with executable tests, and reduces next-turn TTFT at concurrent batches. The systems evaluation targets the response latency of follow-up requests in multi-turn serving.

2 Background and Related Work

Prefix caching and transcript edits. PagedAttention and RadixAttention manage reusable states for matching logical prefixes (Kwon et al., 2023; Zheng et al., 2024). StreamingLLM, H2O, and SnapKV compress or select states within a causal history (Xiao et al., 2023; Zhang et al., 2023; Li et al., 2024). Here an interior span is deleted while the suffix is retained. RoPE's additive rotation structure (Su et al., 2021) permits exact relocation of a key's existing content, but does not reconstruct the content it would have had under the edited history.

Position-portable and generated-cache reuse. KVEraser studies residual suffix contamination after deletion and learns replacement states; Leyline exposes cache-edit directives (Li et al., 2026; Ma et al., 2026). CacheBlend, EPIC/LegoLink, and CacheSlide reuse shifted chunks with selected recomputation (Yao et al., 2025; Hu et al., 2025; Liu et al., 2026). RelayCaching is especially close: it reuses generated output across decode-to-prefill boundaries and rectifies selected states (Geng et al., 2026). Its zero-repair baseline already shows workload-dependent success, so neither relocation nor occasional success without repair is our novelty. "Models Take Notes at Prefill" also studies editable, portable cache content (Li, 2026). Our contribution is the separation of representation and decision fidelity, an internal attribution of finite cache errors, and controlled evidence about which retained states change decisions. Hidden-reasoning deletion is one demanding instance of this broader problem.

Attribution and runtime risk. Path integration and margin-based stability arguments are established tools (Sundararajan et al., 2017; Tsuzuku et al., 2018). We specialize the source decomposition to direct reads of edited KV, without assuming global attention contraction (Kim et al., 2021). WitCert uses runtime observability for KV quantization (Wei et al., 2026); our risk signal instead ranks errors after transcript editing. The measured AUC is an empirical ranking result, not a distribution-free certificate or conformal guarantee (Angelopoulos et al., 2024).

3 Problem and Shift-and-Reuse

Let P be the stable prefix, R the hidden reasoning-format span including delimiters, A the retained answer, and U a follow-up. Physical generation produces $C_{\text{phys}} = KV[P, R, A]$, whereas canonical recomputation produces the reference $C_0 = KV[P, A]$. We hold the visible answer tokens fixed when comparing these caches. The residual answer-block contamination after relocation is

$$\delta = (\widetilde{K}_A^{\text{phys}} - K_A^0,\ V_A^{\text{phys}} - V_A^0). \tag{1}$$

If $m = |R|$ and $\mathcal{R}$ is the model's rotary operator, a physical answer key at position $s + m + t$ is translated to $s + t$ by one constant rotation.

Lemma 1 (Exact RoPE translation)**.** *If $\mathcal{R}(x)\mathcal{R}(y) = \mathcal{R}(x + y)$, then*

$$\widetilde{K}_t^{\text{phys}} = \mathcal{R}(-m)K_t^{\text{phys}} = \mathcal{R}(s + t)\bar{k}_t^{\text{phys}}, \tag{2}$$

where $\bar{k}_t^{\text{phys}}$ is the pre-RoPE content. The equality is exact up to implementation arithmetic, including a partial rotary subspace.

Proof. Apply the group property to $\mathcal{R}(-m)\mathcal{R}(s + m + t)$. □

Shift-and-Reuse keeps the prefix KV, rotates retained answer keys, copies their values, and deletes the reasoning entries. Standard RoPE acts on keys and queries, not values. The result $C_1 = C_0 + \delta$ has the desired positions but generally different answer content. Thus our target is approximate future behavior, not tensor equality. We measure teacher-forced KL and top-1 agreement under a common continuation, autonomous rollout divergence, and downstream task accuracy separately. Figure 1 shows the serving flow; algorithm 1 gives the pseudocode.

4 From Cache Error to Decision Perturbation

Fix all future input tokens through a scored position and interpolate the answer cache as $c(s) = c_0 + s\delta$, $0 \le s \le 1$. Future states are recomputed at each s. Let $g_j(s) = z_{y^*}(c(s)) - z_j(c(s))$ be the logit gap from the reference winner to competitor j. The goal is to expose the internal sources of this gap's change before applying a margin test.

Value readout and key rerouting. Index each head and future query position that reads the edited cache by b. At the current s, write $o_b = \sum_i \alpha_{bi} v_{bi}$, with i ranging over *all* unmasked positions. Direct edits $\Delta k_{bi}, \Delta v_{bi}$ are zero outside the retained answer. Set $r_{bi} = q_b^\top \Delta k_{bi}/\sqrt{d_k}$. Differentiating only the external cache inputs at this read gives

$$h_b = \underbrace{\sum_i \alpha_{bi} \Delta v_{bi}}_{\text{value readout}} + \underbrace{\sum_i \alpha_{bi} r_{bi}(v_{bi} - o_b)}_{\text{key rerouting}}. \tag{3}$$

Let $u_{jb} = \partial g_j / \partial o_b$ be the total downstream sensitivity, $e_{bi} = u_{jb}^\top \Delta v_{bi}$, and $w_{bi} = u_{jb}^\top v_{bi}$. Reverse-mode chain rule then yields

$$g_j'(s) = \sum_b \left[\mathbb{E}_{\alpha_b} e_b + \mathrm{Cov}_{\alpha_b}(r_b, w_b)\right]. \tag{4}$$

The direct source holds upstream inputs fixed only while differentiating one read. The downstream derivative accounts for changed future queries, new KV, residual paths, and later layers in the full computation; these effects are not discarded or counted twice.

Theorem 1 (Finite attention-mediated drift)**.** *If the fixed-token forward is continuously differentiable along $c(s)$, then*

$$|g_j(1) - g_j(0)| \le B_j := \int_0^1 \sum_b \left[\mathbb{E}_{\alpha_b}|e_b| + \sigma_{\alpha_b}(r_b)\sigma_{\alpha_b}(w_b)\right] ds. \tag{5}$$

If $g_j(0) > B_j$ for every $j \neq y^*$, the winner is preserved.

The proof applies triangle and Cauchy–Schwarz inequalities to eq. (4) and integrates; appendix I gives full details. The value term is also bounded by $\beta_b \max_{i\in A} |e_{bi}|$, where $\beta_b = \sum_{i\in A} \alpha_{bi}$. Large cache errors can have a small effect when their signed task projections cancel or score changes weakly align with differences in task evidence. Errors aligned with that evidence can instead be amplified. The result is conditional and does not assume that every deletion is benign.

A token-level, falsifiable consequence. The covariance gives each retained token an explicit signed contribution. With $\bar{w}_b = \sum_i \alpha_{bi} w_{bi}$, define

$$\tau_i = \int_0^1 \sum_b \alpha_{bi} \left[e_{bi} + r_{bi}(w_{bi} - \bar{w}_b)\right] ds, \qquad \sum_{i\in A} \tau_i = \Delta g_j. \tag{6}$$

Negative τ_i identifies a token contributing against the reference winner along this path. It need not coincide with a large KV norm or large attention weight: both omit the downstream task projection and its sign. This suggests a testable intervention ranking. If these scores identify a meaningful source of decision error, repairing the most negative tokens should improve decision recovery more than simpler rules at the same token budget. This prediction is not itself an identity. Replacing selected entries changes the path and nonlinear interactions, so the finite repair effect need not equal the sum of their original scores. Section 6 measures that effect directly.

With fixed attention and an affine readout, value errors with zero weighted task projection preserve a gap regardless of their norm; aligned errors can flip it. Thus the theory predicts conditional stability, not unconditional robustness (appendix I).

What this explains and what it does not. Equal answer tokens provide a common first-layer anchor in standard RoPE-only decoders (lemma 2); higher layers can still differ substantially. Equation (4) specifies where that difference enters a decision, while a norm or a final flip identity cannot. Its path integral is an offline attribution, not an independent predictor of the endpoint gap. Numerical integration does not certify a bound without controlling quadrature error, and one runner-up is insufficient to certify the full argmax. Hard expert-routing jumps require extra terms; the derivation does not solve hybrid recurrent-state editing. The local Fisher approximation (eq. (7)) remains useful for small perturbations but is not extrapolated to full deletion.

5 Empirical Separation of Three Fidelities

Setup. We compare gold recomputation, shifted reuse, reuse without rotation, the physical cache retaining reasoning, and a clean-splice numerical control. The validated cross-model comparison covers dense Qwen3-4B and Qwen3-14B and restricted hybrid Qwen3.5-27B. Follow-ups include answer-dependent questions, corrections, reasoning probes, and unrelated turns. Reported uncertainty groups observations by originating prompt where applicable.

Large representation errors need not change decisions. Relative value contamination grows from about 0.17 in an early layer to 0.66 in middle and late layers. In the intervention study, the full answer-cache error norm nevertheless has approximately zero rank correlation with follow-up KL (fig. 2). A reasoning-budget sweep from 64 to 1024 tokens increases contamination strongly ($\rho = 0.83$), while next-turn top-1 agreement changes only modestly (appendix A.9). Longer retained answers do not systematically increase drift. This is evidence against a simple cache-distance explanation.

Nor is the error innocuous because it resembles random noise or lies in insensitive directions. Across 24 prompts and five follow-up types, equal-norm interventions give the true contamination a median $9.3\times$ the local Fisher energy of a random direction. The output distribution moves: median follow-up-prefill KL is about 0.36 nats, and mean generated-position KL is about 0.08. A small local quadratic

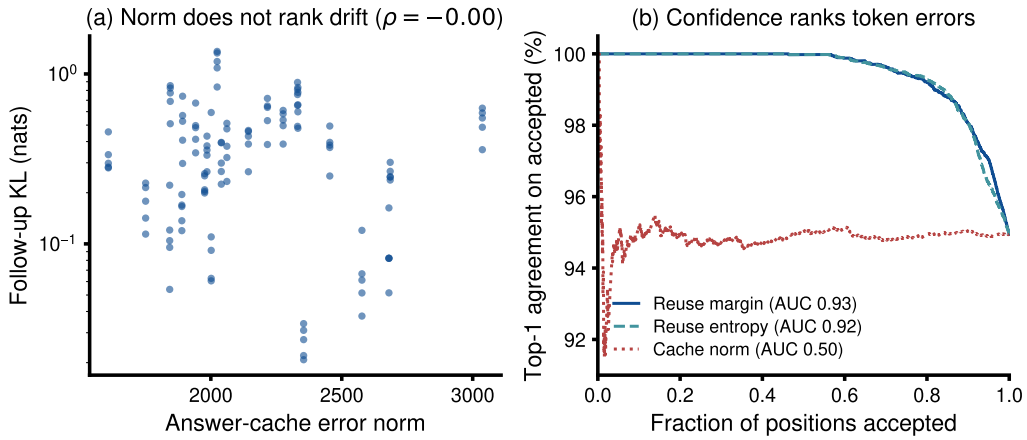


Figure 2: Two distinct cohorts show why cache size is an inadequate behavioral diagnostic. Left: answer-cache error norm versus follow-up KL in the 120-condition intervention study. Right: on 9,600 reference-generated positions, accepting positions in confidence order improves one-step agreement; cache norm provides little ranking information. Coverage is a fraction of scored positions, not a measured serving policy.

Table 1: Follow-up-prefill top-1 agreement (%), distinct from generated-position metrics. Hybrid†: full-attention reuse with recomputed linear state, with no implied whole-model speedup.

	Qwen3-4B	Qwen3-14B	Qwen3.5-27B†
Reusable layers	36/36	40/40	16/64
Shifted reuse	84.3	83.7	96.8
No rotation	48.3	69.8	95.4
Keep physical	81.6	79.2	83.0
Clean-splice control	98.2	98.6	—

accurately describes only small fractions of the perturbation; the measured-to-predicted KL ratio reaches about 2.1 at full amplitude (fig. 16). Thus neither small norm nor negligible distributional change explains the observed decision agreement.

Decision-local confidence separates stable and fragile tokens. On 9,600 positions from 150 reference greedy continuations, reuse preserves the one-step top-1 token at 95.0% of positions. The reference margin averages 13.4 at preserved positions and 1.5 at flips. A paired logit-displacement test identifies flips exactly, but this is an endpoint identity. The inexpensive, deployable result is that the *reuse-side* margin ranks disagreement at AUC 0.93; reuse entropy is comparable (0.92), while cache norm is near chance (0.50). Restricting to positions up to the first divergence retains AUC 0.94. These measurements use a common token prefix and do not imply 95% exact-sequence fidelity.

A KL drop at the assistant boundary can reflect a predictable formatting token rather than state recovery (appendix A.10 and proposition 2). We therefore separate follow-up-prefill metrics, reference-generated decisions, substantive-token interventions, and executable answer quality. Agreement on the `<think>` opener alone is not answer-quality evidence.

Architectural and transcript-edit scope. The positional correction improves agreement by 36.0 and 14.0 percentage points on the two dense models (table 1). Hybrid recurrent state cannot generally delete an interior span, so its restricted result is reported separately. Keeping reasoning is more often farther from gold than shifted reuse on the two dense models; whether reuse itself leans toward gold or the physical cache is model-dependent. Figures 10 and 11 report both KL orientations and eight-turn measurements on Qwen3-4B.

A length-matched control holds the answer fixed while deleting hidden reasoning, deleting visible text, or swapping the preceding prompt. All three produce similar relative cache error (about 0.20) and 85–88% prefill agreement. Hidden-reasoning deletion has larger behavioral KL than the prompt-swap control in 75% of cells and the visible-span control in 97%. This is a general position-portable-cache phenomenon with workload-dependent difficulty, rather than evidence that hidden reasoning is especially easy to remove.

6 Testing the Internal Contributions

The theory suggests three different tests: check the decomposition, measure its attenuation factors, and intervene on the states it identifies. These tests serve different purposes; exact reconstruction by itself is only a check of the instrumentation.

Protocol and reconstruction. We use 100 originating Qwen3-4B prompts, split into 20 development and 80 held out, each with two follow-ups. We request four substantive generated positions per follow-up where available, yielding 788 sampled positions, and add 88 distinct first-divergence positions. The combined diagnostic cohort has 876 positions and 153 flips. It is enriched for divergence and is not used to estimate an unconditional rollout failure rate. For each position we fix the reference winner and runner-up, recompute all future states along $c(s)$, and integrate the two sources in fp32. Sixteen-node quadrature meets $10^{-3} + 0.01|\Delta g|$ reconstruction tolerance at $876/876$ positions; median absolute error is 4.4×10^{-5}. Qwen3-14B repeats the protocol on 24 prompts; 16-node reconstruction passes at $212/215$ positions, increasing to $214/215$ when the stored 32-node refinements are used. Full-vocabulary argmax and post-hoc winning rivals are checked separately (appendix A.6).

Attenuation is measurable, but not a substitute for the margin. The aggregate signed value source divided by its absolute contribution mass is 0.013 at stable positions and 0.018 at flips. The ordering persists in separate margin and norm strata. Aggregate routing alignment is about 0.002 in both groups; it does not distinguish them. These are endpoint diagnostics, not isolated causal effects. Reference-endpoint factors alone attain held-out AUC 0.596, versus 0.880 for the reference margin and 0.886 for factors plus margin. The bound is also loose: median $|\Delta g|/B$ is 0.003. The useful new evidence is therefore localization and intervention, not a strong new certificate or a better online classifier.

Matched-budget repair tests localization. Integrating signed source contributions separately for each answer token ranks the states that reduce the reference gap. Starting from reuse, we replace $k \in \{1, 4, 8\}$ selected tokens' KV across every layer with their reference entries. Controls use the same number of entries, selected by largest KV error, highest attention exposure, earliest position, or random choice (three seeds). Evaluation uses held-out positions with $|\Delta g| > 0.05$: 661 positions in total, of which 122 are flipped. At $k = 4$, attributed repair restores the reference winner on 59.8% of those flips (prompt-bootstrap 95% interval $[51.6, 68.0]\%$), against 17.2% for attention, 13.9% for earliest, 11.5% for norm, and 2.5% for random selection (fig. 3). At $k = 8$ it reaches 80.3%. The Qwen3-14B check reaches 71.4% at $k = 4$, but includes only 21 qualifying flips and has a wide interval ($[50, 90]\%$).

Across all 661 qualifying positions, four attributed tokens remove a median 4.5% of squared cache-error norm, versus 15.3% for the norm rule. Reverse insertion into the reference cache at those four tokens reproduces a median 61.6% of the full gap displacement. These interventions support a concentrated, task-dependent contribution to the decision error. They do not imply that every cache error is sparse or that these reference-dependent scores can be obtained cheaply in serving. Recovery ratios are heavy-tailed, and a direction-only test barely separates selection rules: all improve the gap at 61–66% of qualifying positions. Sampled-only and added-first-divergence results are separated in appendix A.6.

In particular, restricting to the 54 qualifying sampled flips gives 61.1% recovery for attribution versus 14.8% for attention or norm selection at $k = 4$. The 68 added-first-divergence positions give 58.8%, 19.1%, and 8.8%, respectively. Thus the localization comparison is not solely a consequence of adding divergence positions. The reported intervals resample originating prompts rather than

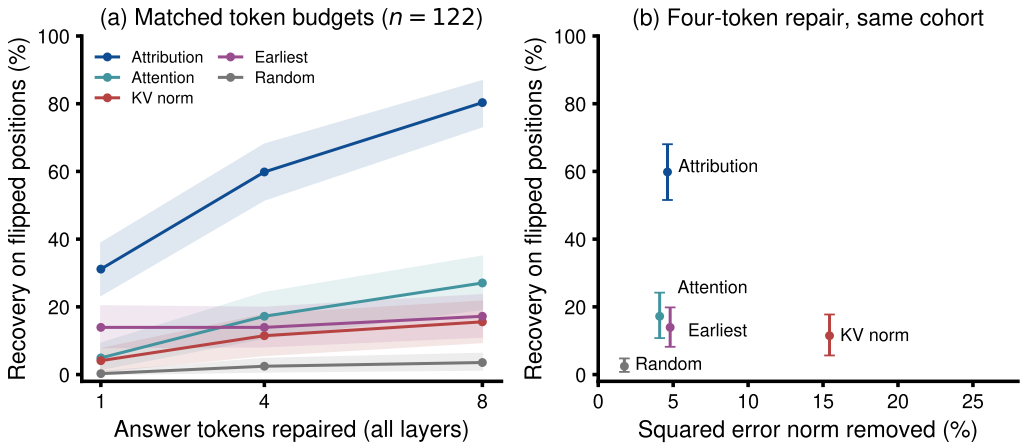


Figure 3: Offline targeted repair, Qwen3-4B. Both panels use the same 122 flipped held-out positions with $|\Delta g| > 0.05$, pooling sampled and added first-divergence positions. Left: replace k answer tokens' KV across all layers; bands bootstrap originating prompts. Right: four-token recovery versus the median squared error norm removed on that same cohort. The attribution requires the reference cache.

Table 2: Two-turn task accuracy (%). Differences are reuse minus recomputation, in percentage points, with paired 95% profile-likelihood intervals. GSM8K uses carry accuracy; HumanEval uses executable turn-2 tests. The five-point non-inferiority margin is met only by GSM8K here. vLLM excludes one task without an eligible answer span.

Task / implementation	n	Recompute	Reuse	Difference [95% CI]
GSM8K / HF	60	83.3	85.0	$+1.7$ $[-2.0, 7.1]$
HumanEval / HF	164	81.1	79.9	-1.2 $[-6.5, 4.0]$
HumanEval / vLLM	163	82.2	79.1	-3.1 $[-7.2, +0.5]$

individual tokens, preserving within-prompt dependence. They quantify uncertainty in conditional recovery; they do not estimate how often a production request would need repair.

Keys and values interact. The endpoint factorial contrast $I = g_{11} - g_{10} - g_{01} + g_{00}$ has median magnitude 0.395, with median $|I|/|\Delta g| = 0.557$. Thus key-only and value-only endpoint effects cannot simply be added. For the same four attributed tokens, keys-only repair restores 34.4% of flips and values-only repair 31.1%, versus 59.8% when both are repaired. The exact path decomposition accounts for their joint evolution; it does not assert endpoint additivity.

7 Answer Quality and Serving Latency

We evaluate two-turn follow-up tasks involving the stored turn-1 answer. In HumanEval, the function specification remains in P, so success does not establish that the model must read A rather than solve the task again. These tests measure executable answer quality, distinct from formatting-token or teacher-forced agreement. Table 2 separates the model-level and vLLM execution settings.

Quality is close but not uniformly non-inferior. On GSM8K, turn 2 multiplies the stored numerical answer by seven. Reuse establishes non-inferiority at the specified five-point margin; a parse failure is not silently removed from a policy's score. On HumanEval, the model-level reuse result is 1.2 points below recomputation and its interval crosses the non-inferiority margin. The EPIC-style first-eight-token repair and keep-physical baselines both score 83.5% there (appendix D). Statistical compatibility with zero difference is not evidence of equivalence.

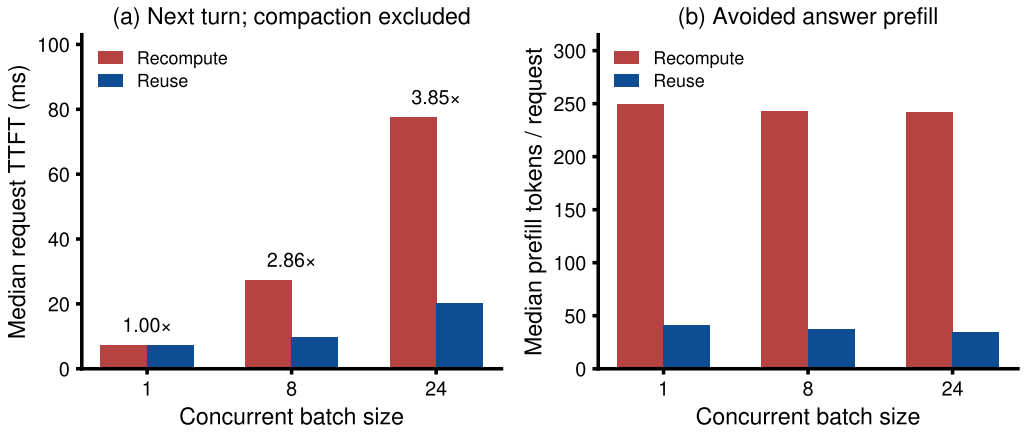


Figure 4: Next-turn latency and prefill work in patched vLLM, Qwen3-4B on H200. Left: median request TTFT, excluding prior-turn compaction, with recompute/reuse ratios. Right: median tokens requiring prefill per request; reuse avoids processing the retained answer again. Each setting has three repetitions. Preparation costs and full-conversation timings are reported in appendix H.

The vLLM quality check consumes decode-produced answer pages and generates substantive follow-up code graded by HumanEval tests. Among 163 eligible tasks, reuse passes 129 and recomputation 134; the no-rotation control passes only 49 (30.1%), and keep-physical passes 137 (84.0%). Pass/fail agreement with recomputation is 94.5%, distinct from exact token-sequence agreement. One of 164 tasks is excluded by the shared trace-eligibility check. These results materially strengthen the serving check, but do not establish non-inferiority or production reliability.

The no-rotation control detects positional corruption; recomputation isolates residual content effects, with two paired wins and seven losses for reuse. Keep-physical changes the conditioning history and retains more KV. The quality baseline starts from an empty canonical cache, while the latency baseline warms its stable prefix; their timings are not interchangeable.

The serving path implements the intended transformation. Our vLLM 0.29.0 patch copies prefix keys and all values, rotates only answer keys, stops at the computed-KV frontier, and registers complete canonical pages. A GPU page-readback experiment on 24 traces finds bitwise prefix-key and value equality; every trace's rotated-key error fits the stated bf16 arithmetic tolerance. Unrotated-answer and wrongly-rotated-prefix controls differ substantially. Canonical pages still show content contamination relative to fresh prefill, as intended. Retiring physical prefix-cache entries reduces physical-transcript hits from a median 40 blocks to the eight shared-prefix blocks while retaining 22 canonical blocks. This measures cache-index retirement, not secure erasure of information encoded in retained states.

Next-turn response latency. The latency workload has 24 coding traces, median answer length 209 and deleted span length 303. Isolated canonical cache hits rise from 136 to 344 tokens, removing approximately the answer length from prefill. Matched follow-up token IDs and varied measurement order control policy differences. At concurrent batches 8 and 24, median request TTFT falls from 27.44 to 9.60 ms and from 77.62 to 20.14 ms: 2.86× and 3.85× speedups (fig. 4). The corresponding batch-wall ratios are 2.51× and 3.16×; they should not be labeled TTFT ratios. Batch 1 is neutral. The longer-answer HuggingFace forward microbenchmark is reported separately in fig. 6.

Timing scope and between-turn preparation. For interactive serving, TTFT measures how long a follow-up request waits for its first response token. Our measurements use a prepared canonical cache, so they quantify the latency benefit available when the between-turn update completes before the request arrives. Any unfinished preparation must instead be included in that request's latency. The prototype's serial preparation takes about 9.7 ms per request; separately timed complete two-turn

conversations take 1855 ms with reuse and 1842 ms with recomputation. Detailed preparation-inclusive costs and timing aggregation are in appendix H. Repeated batch trials restore the physical cache by prefill; the full-conversation trial consumes decode-produced pages.

Risk ranking is useful, but fallback has a price. A reuse-side margin needs no reference forward and supports selective recomputation. Its token-level AUC is not a task-success guarantee. On 140 GSM8K conditions, a multi-feature reuse-only model reaches five-fold out-of-fold final-answer disagreement AUC 0.79 (appendix B). A sequential margin gate raises exact-sequence fidelity from 17.5% to 95%, but triggers recomputation on 99% of turns (appendix A.11). That operating point restores fidelity without preserving the compute saving. We therefore report risk ranking, realized quality, and latency separately rather than infer a validated serving policy from any one of them.

8 Limitations and Conclusion

Cross-model replication measures follow-up-prefill decisions on three Qwen models; generalization beyond the Qwen family is untested. Generated-trajectory and end-task studies cover fewer models. The transport mechanism is tested on two dense models, with a small 14B replication and one unresolved quadrature error after refinement. The causal repair comparison is offline, conditioned on a selected diagnostic cohort, and does not establish that trained models always produce small attention-mediated error. Hybrid state editing and discontinuous MoE routing require additional treatment.

The workloads do not cover broad tool-use, adversarial, safety, or long-horizon agent behavior. Deleting explicit reasoning entries does not remove every trace of that reasoning from retained KV. Cache reuse can also change downstream judgments (Liang et al., 2026). The vLLM latency and executable answer-quality studies cover one model and node. The reported TTFT scope assumes cache preparation has completed; mixed-load scheduling and background resource contention remain to be evaluated.

The intervention identifies consequential states, but does not explain why training produces cancellation or supply a reference-free repair ranking. Three-repetition timing summaries also cannot establish tail-latency performance. Online repair and loaded-system SLA attainment remain deployment questions.

Hidden-reasoning deletion exposes a useful distinction: cache states can be wrong in representation while remaining adequate for many decisions. Exact RoPE relocation removes positional error; value readout and key rerouting explain how residual content error reaches the logits. Targeted interventions show that decision-relevant contributions can be concentrated in states that norm-based repair misses. Together with the vLLM measurements, this supports position-portable reuse as a practical direction for reducing follow-up response latency, with workload-specific quality validation and online risk monitoring.

USE OF LARGE LANGUAGE MODELS

We used a large language model coding assistant for research-subtask support: implementing the experiment and analysis code (cache surgery, RoPE re-rotation, the four-way quality and systems harnesses, statistical analysis, and figure generation), assisting with the design of the non-inferiority and risk-ranking protocols, and drafting and editing the text of this paper. We did not use generative AI tools to fabricate, alter, or synthesize any experimental data or numerical results; every number, figure, and table in this paper is produced by the released code from raw per-condition result files, and the authors remain responsible for reviewing the AI-assisted code and claims. The present editing pass reanalyzed the supplied raw results and rebuilt the figures and document; it did not rerun GPU experiments. We take full responsibility for the final content of this work.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast large language model serving for RAG with cached knowledge fusion. In *EuroSys*, 2025.

Yang Liu et al. CacheSlide: Unlocking cross position-aware KV cache reuse for accelerating LLM serving. In *USENIX FAST*, 2026.

Junhao Hu et al. EPIC: Efficient position-independent caching for serving large language models. In *ICML* (PMLR v267), 2025. arXiv:2410.15332.

Yingsheng Geng, Yuchong Gao, Weihong Wu, Guyue Liu, and Jiang Liu. RelayCaching: Accelerating LLM collaboration via decoding KV cache reuse. *arXiv:2603.13289*, 2026.

Bojie Li. Models take notes at prefill: KV cache can be editable and composable. *arXiv:2606.17107*, 2026.

Fanzhe Wei, Li Liu, Ziyang Wang, and Chenyu Wang. WitCert: Sound runtime risk observability and gating for KV-cache quantization. *arXiv:2607.28699*, 2026.

Mufei Li, Shikun Liu, Dongqi Fu, Haoyu Wang, Yinglong Xia, Hong Li, Hong Yan, and Pan Li. KVEraser: Learning to steer KV cache for efficient localized context erasing. *arXiv:2606.17034*, 2026.

Bole Ma, Jan Eitzinger, and Harald Köstler. Leyline: KV cache directives for agentic inference. *arXiv:2606.01065*, 2026.

Sichu Liang, Zhenglin Wang, Jiajia Chu, Pengfei Xia, Hui Zang, and Deyu Zhou. When KV cache reuse fails in multi-agent systems: Cross-candidate interaction is crucial for LLM judges. In *Proceedings of ACL*, 2026. arXiv:2601.08343.

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 *SOSP*, 2023.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. *arXiv:2404.14469*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv:2104.09864*, 2021.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv:2309.17453*, 2023.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. *arXiv:2306.14048*, 2023.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. *arXiv:2312.07104*, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *ICML*, PMLR 139:5562–5571, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, PMLR 70:3319–3328, 2017.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *NeurIPS*, 2018. arXiv:1802.04034.

Input: stable prefix cache C_P, physical answer blocks (K_A, V_A), deleted length m.

1. For each layer, compute $\widetilde{K}_A \leftarrow \mathcal{R}(-m)K_A$ on the rotary subspace.
2. Return $C_{\text{reuse}} \leftarrow [C_P; \widetilde{K}_A, V_A]$.

Algorithm 1: Shift-and-Reuse cache surgery (approximate behavioral reuse, not exact reconstruction)

A Extended Methodology and Additional Results

This appendix contains methodology, secondary result panels, and predictor details relocated from the main text for length. All figures and tables regenerate from the same released pipeline.

A.1 Cache-surgery pseudocode

A.2 Future work

Several extensions would strengthen the claims. The evidence is currently confined to single-turn and two-turn probes, so agentic, tool-use, and long-horizon task correctness would test whether margin protection survives over many dependent turns. Broader model families and scales are needed, as is a serving-viable treatment of hybrid linear-attention layers, which our restricted study leaves recomputed. On the mechanism, an explicit active-direction intervention along the top eigenvector of the pullback-Fisher matrix would complement the equal-norm null test of section 5, and extending the generated-trajectory measurement beyond Qwen3-4B would let table 1 report the decision-relevant metric rather than the prefill proxy. The executable-coding comparison is underpowered at our 5-point margin (appendix D); more seeds and more tasks, clustered by problem, would settle it. Finally, a genuine train/calibration/test conformal risk protocol with leave-one-model-family-out validation, integrated into a paged-attention runtime and measured under mixed load, would turn the risk-ranking heuristic into a deployable guarantee.

A.3 Full methodology, model, and prompt counts

The initial representation study captured pre-RoPE key and value projections on the full-attention layers of a hybrid Qwen reasoning model. Exact cache surgery was then run on the models above. Thinking mode was enabled, reasoning was generated under a fixed budget, a closing reasoning delimiter was inserted when necessary, and a visible answer was generated afterward. A non-reasoning instruction model (gemma-4-31B-it) was excluded because it emits no hidden-reasoning span.

The scaled experiment uses up to 30 reasoning prompts and five follow-up classes: fresh task, answer-dependent, reasoning-probing, correction, and unrelated. On Qwen3-4B the teacher-forced and rollout matrix covers all 150 conditions (30 prompts $\times$ 5 follow-ups), as does the three-way comparison. The Qwen3.5-27B study is restricted (section 5) and covers 120 conditions (24 prompts). Scripts run eager attention to suppress sequence-length-dependent SDPA numerical variation. Statistics are paired by originating prompt.

Numerical controls. RoPE relocation must be checked against the model's own rotary operation, including its rotary subspace and any key normalization. A clean splice (reference content, identical bookkeeping) additionally checks cache manipulation and kernel numerics. In the archived follow-up matrix, clean-splice top-1 agreement is 98.2% on Qwen3-4B and 98.6% on Qwen3-14B. These are observed numerical-control discrepancies, not an attribution of their cause.

A.4 First-divergence, three-way disagreement, and multi-turn detail

Token identity is a strict metric: after the first wording bifurcation, two semantically valid greedy rollouts may never rejoin.

We keep the two KL orientations distinct. The first comparison asks whether $D_{\mathrm{KL}}(\text{gold}\|\text{physical}) > D_{\mathrm{KL}}(\text{gold}\|\text{reuse})$; it holds in $102/150$ conditions (68.0%) on Qwen3-4B and $115/150$ (76.7%) on Qwen3-14B. The second asks whether $D_{\mathrm{KL}}(\text{reuse}\|\text{gold}) < D_{\mathrm{KL}}(\text{reuse}\|\text{physical})$; it holds in

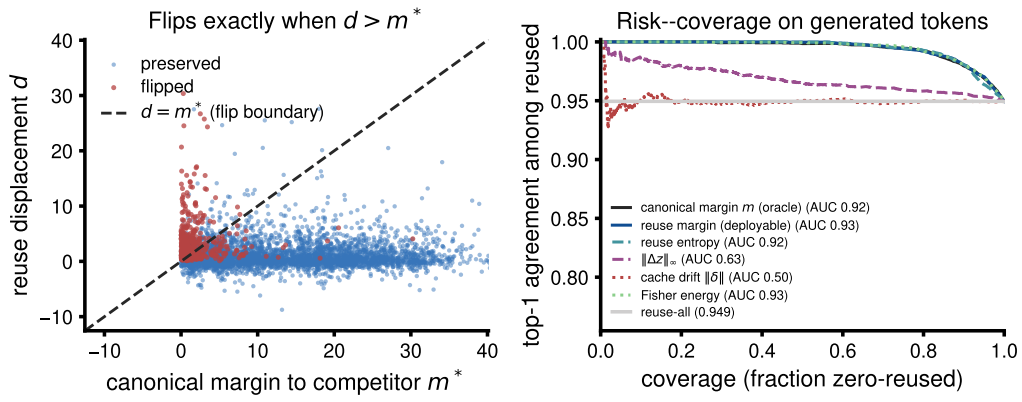


Figure 5: The margin explains and gates reuse on the generated trajectory (Qwen3-4B, 9,600 teacher-forced positions of the reference greedy assistant response). Left: a token flips exactly when the reuse displacement d toward its strongest competitor exceeds the reference margin $m^\star$; all flips lie above $d = m^\star$. Right: risk–coverage. The deployable reuse-side margin (blue) ranks flips as well as the offline oracle margin (black) and gold-cache Fisher energy, while the cache-drift norm (red, dotted) is at chance.

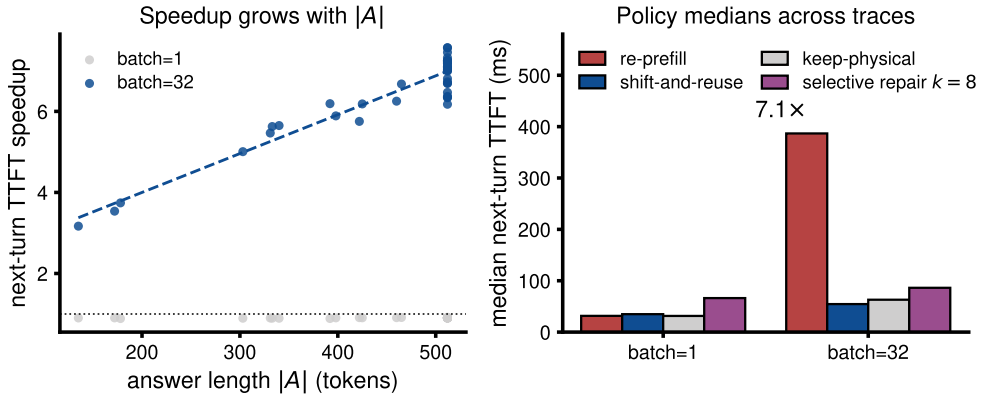


Figure 6: HuggingFace next-turn TTFT microbenchmark on coding traces (Qwen3-4B, $n = 40$). Left: per-trace recompute/reuse ratios versus answer length. Right: policy-wise median latency across traces; the displayed $7.1\times$ is the ratio of medians at batch 32 ($386.73/54.55$ ms). The median of the 40 paired per-trace ratios is instead $6.7\times$. Reuse includes the measured rotation component; selective repair adds an eight-token repair component (appendix E).

$76/150$ (50.7%) and $37/150$ (24.7%), respectively. Thus reuse is more often closer to gold than the physical cache is on the two dense models, but reuse's own lean is model-dependent. The eight-turn experiment on Qwen3-4B accumulates shifted answer blocks while holding answer tokens to a reference trajectory. The mean-KL trend is $+0.0005$ nats/turn over this measured horizon. This fixed-reference experiment measures drift over eight turns, not correctness on autonomous long-horizon tasks.

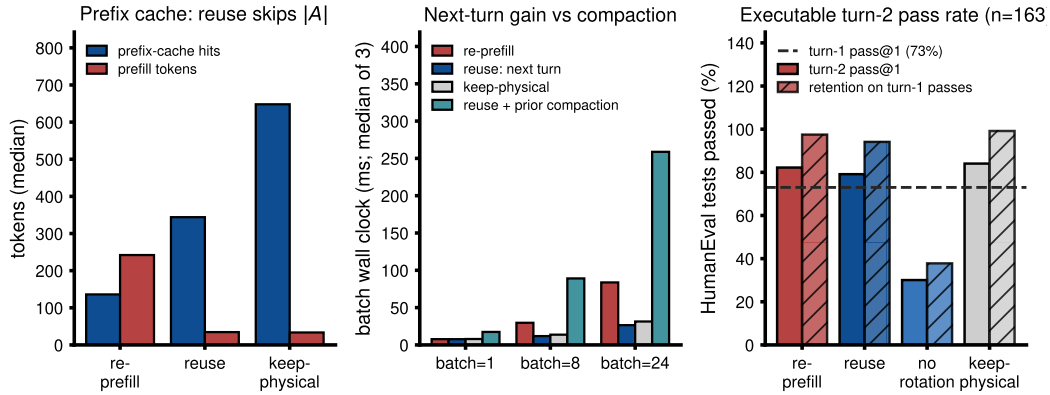


Figure 7: Shift-and-Reuse prototype in vLLM 0.29.0 on Qwen3-4B. Left: prefix-cache hits and follow-up prefill tokens on 24 coding traces; reuse skips the answer block. Middle: next-turn batch wall time and preparation-inclusive time are shown separately; matched follow-up token ids and randomized measurement order are used. Right: executable turn-2 correctness on HumanEval ($n = 163$), graded by the real unit tests; the no-rotation control collapses, which is what shows the check can detect a cache error. The separately timed full-conversation comparison is reported in appendix H.

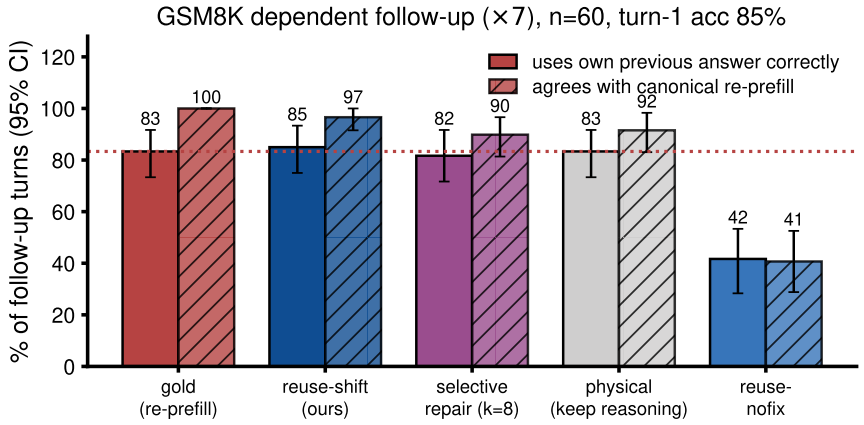


Figure 8: Downstream task quality under the dependent two-turn protocol (Qwen3-4B, $n = 60$). Solid: correct use of the model's own turn-1 answer (carry). Hatched: same final answer as the gold recomputation, with parse failures counted as disagreements. Reuse-shift (blue) is non-inferior to gold (red), physical-full (grey), and selective repair (purple) at the pre-specified 5-point margin, while reuse-nofix (the hatched pale-blue ablation of reuse-shift) collapses.

A.5 Top-1 stability diagnostic

Let the reference winner be unique, with margin $\mu_t = z_{t,y^*} - \max_{j \neq y^*} z_{t,j} > 0$. The exact sufficient condition $2\|\Delta z_t\|_\infty < \mu_t$ preserves that winner. If a validated linearization-remainder bound $\|\Delta z_t - J_t\delta\|_\infty \leq r_t$ is available, a sufficient condition is

$$2\big(\|J_t\delta\|_\infty + r_t\big) < \mu_t.$$

The stabilized score $\rho_t = 2\|J_t\delta\|_\infty/(\mu_t + \epsilon)$, with $\epsilon > 0$, is only a *heuristic risk score*. The denominator offset relaxes the margin threshold and the score omits the remainder, so $\rho_t < 1$ does not certify top-1 invariance, even when the linearization is exact. We do not estimate a certified remainder bound in these experiments.

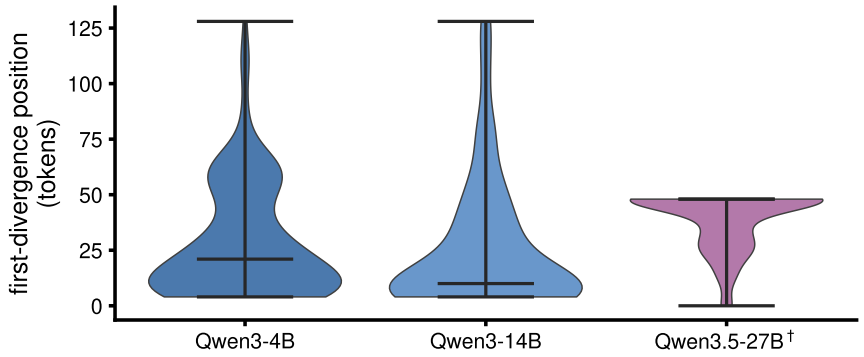


Figure 9: Greedy first-divergence position from the reference rollout (Qwen3-4B: median 23, p95 66). This token-level metric is distinct from task correctness. The hybrid uses restricted reuse.

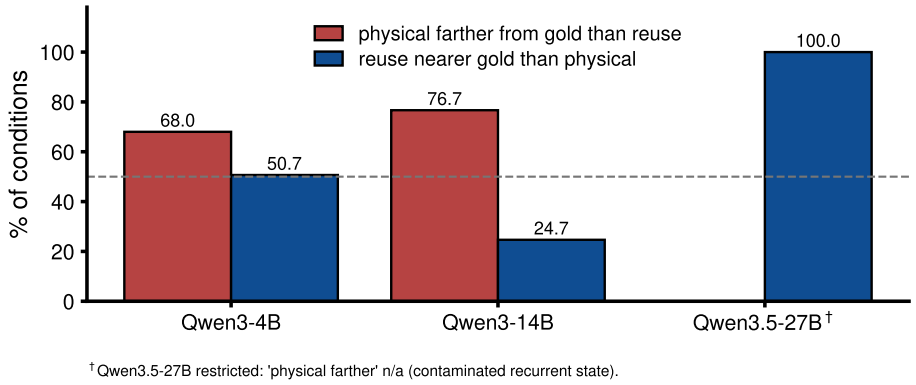


Figure 10: Two distinct three-way comparisons, with KL orientations defined in the text. Physical is farther from gold than reuse in a majority of conditions on the two dense models. Reuse's own lean is model-dependent. The restricted hybrid's first comparison is omitted because physical retains contaminated recurrent state.

A.6 Transport attribution: full protocol and diagnostics

This appendix expands section 6.

Protocol. The dataset is 100 independent originating prompts (34 numeric, 33 code prompts that force an invented identifier so the follow-up can only be answered from the stored answer, 33 word problems), split into 20 development and 80 held-out prompts; no threshold or fitted model uses held-out prompts. Each prompt receives two follow-ups: answer-dependent retrieval and correction/refactoring. From each follow-up's reference greedy continuation we sample four substantive generated-token positions, and add first-divergence positions not already sampled. The archived summaries pool these two cohorts; cohort-specific intervention results are reported below. With $C_0 = KV(P, A)$ and $C_1 = C_0 + \delta$, each evaluated position contributes $g(s) = z_{y^\star}(C_0 + s\delta) - z_j(C_0 + s\delta)$ for the reference winner $y^\star$ and its reference runner-up j. The preceding token sequence is held fixed, but at every s the follow-up and the preceding generated tokens are recomputed from the intervened cache. Everything is fp32 and deterministic. All uncertainty is a bootstrap over originating prompts, since positions and follow-ups from one prompt are correlated. Qwen3-4B yields 788 sampled and 88 added first-divergence positions (876 total, 153 flips); Qwen3-14B, run on 24 prompts with a longer answer budget, yields 215 positions and 21 held-out flips.

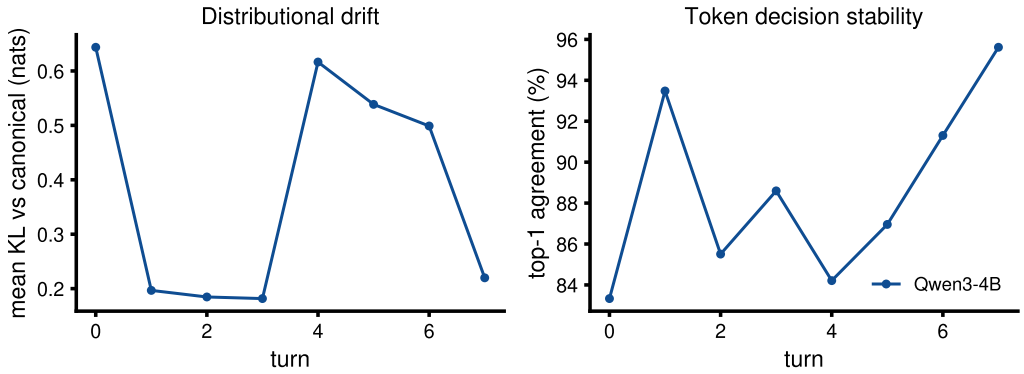


Figure 11: Eight-turn measurements on Qwen3-4B with fixed reference answer tokens. Left: mean KL from recomputation. Right: top-1 agreement. Mean KL grows little over the observed horizon; this is not an autonomous long-horizon task evaluation.

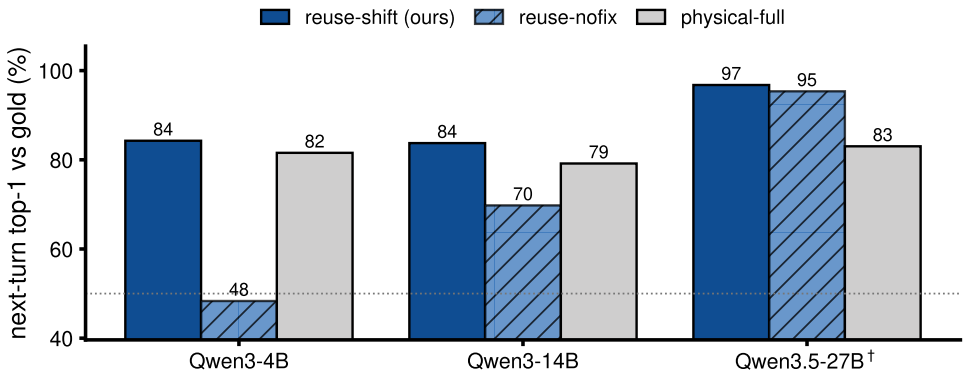


Figure 12: Follow-up-prefill top-1 agreement by cache policy. Rotation improves agreement on both dense Qwen3 models. The restricted hybrid reuses full-attention layers with a reference linear state.

Reconstruction, quadrature, and the bound. With 16-node Gauss–Legendre quadrature, $A_V + A_K$ (the path integrals of the two terms of eq. (4)) matches the independently measured Δg at $876/876$ positions within the pre-specified development tolerance $10^{-3}+0.01|\Delta g|$, with median absolute error 4.4×10^{-5}, 95th percentile 1.4×10^{-4}, and worst case 1.7×10^{-2} (fig. 14). A second, independent total derivative obtained by autograd through the cache leaves agrees to the same precision, which is what rules out an error in the block bookkeeping. Quadrature convergence is not automatic: 8 nodes suffice at only 98.7% of positions, so the harness escalates to 32 wherever the 16-node reconstruction differs from the endpoint gap beyond tolerance. Zero escalations were needed on Qwen3-4B and nine on Qwen3-14B. The stored primary 14B reconstruction passes at $212/215$ positions (98.6%); using the nine refined estimates increases the pass count to $214/215$, with one remaining failure at error 0.25. The pairwise gap is not the whole decision: a negative final runner-up gap agrees with a full-vocabulary argmax flip at 98.2% of positions, and on flipped positions we also analyze the competitor that actually wins under reuse, always labeled as a post-hoc diagnostic (29 such rows). Finally, the Cauchy–Schwarz bound implied by eq. (4), $B = \int_0^1 \sum_b (\mathbb{E}_{\alpha_b}[|e_b|] + \sigma_{\alpha_b}(r_b)\sigma_{\alpha_b}(w_b))\, ds$, is loose by about two orders of magnitude: median $|\Delta g|/B$ is 0.003 $[0.002, 0.003]$. It explains why attenuation is structurally possible but certifies nothing.

Keys and values are not separable at full amplitude. Switching the key and value perturbations on independently gives the four corners $g_{00}, g_{10}, g_{01}, g_{11}$ and the interaction $I = g_{11} - g_{10} -$

Table 3: Cross-model follow-up-prefill measurements. The generated-position study is separate. Hybrid†: only full-attention layers are reused with reference linear state. The two three-way rows use different KL orientations.

	Qwen3-4B dense, full RoPE	Qwen3-14B dense, full RoPE	Qwen3.5-27B† hybrid, 16/64 full
Reusable layers	36/36	40/40	16/64
reuse-shift top-1 (%)	84.3	83.7	96.8
reuse-nofix top-1 (%)	48.3	69.8	95.4
physical-full top-1 (%)	81.6	79.2	83.0
reuse-shift first-token KL	$\sim 10^{-5}$	$\sim 10^{-5}$	0.013
clean-splice control top-1 (%)	98.2	98.6	—
physical farther from gold than reuse (%)	68.0	76.7	—
reuse nearer gold than physical (%)	50.7	24.7	100
multi-turn KL slope (nats/turn)	$+0.0005$	—	—

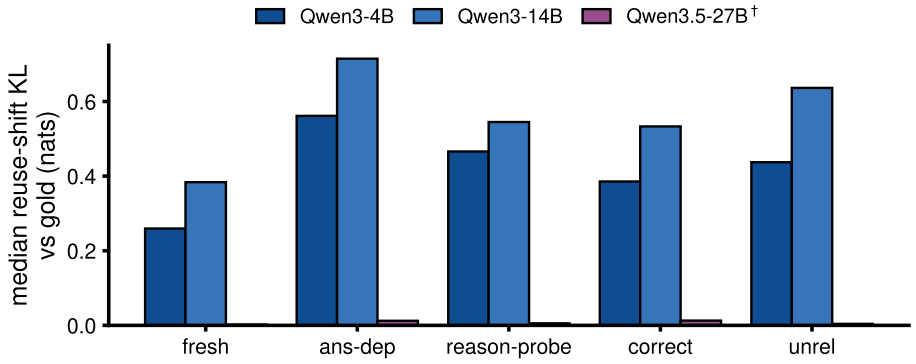


Figure 13: Median teacher-forced reuse-shift KL from gold by follow-up type. Answer-dependent and correction turns carry the largest typical residual; reasoning-probing turns are not uniquely pathological on any model. The restricted Qwen3.5-27B has the smallest residual.

$g_{01} + g_{00}$, whose median magnitude is 0.395 $[0.354, 0.452]$: 0.557 $[0.507, 0.606]$ of $|\Delta g|$ and 0.541 $[0.489, 0.611]$ of the sum of the two single-side effects (Qwen3-14B: 0.372 $[0.306, 0.502]$ of $|\Delta g|$). Wrong keys and wrong values are comparably large and strongly non-additive at the deletion amplitude, so we report I rather than assuming additivity anywhere.

Attenuation factors. Rather than presume one mechanism, we measure five candidates separately at the reference endpoint: value errors cancelling when read, key errors weakly affecting scores, rerouted values carrying similar evidence, routing changes weakly aligned with evidence, and contributions cancelling across blocks. Comparing stable and flipped positions, the largest and most consistent difference is aggregate value cancellation $|\sum_b \mathbb{E}_{\alpha_b} e_b| / \sum_b \mathbb{E}_{\alpha_b} |e_b|$: 0.013 $[0.012, 0.014]$ stable versus 0.018 $[0.014, 0.021]$ flipped (fig. 15). The raw quantities show this is a numerator effect (0.49 versus 0.76) and not a difference in how much value error is available to cancel (40.8 versus 45.3); the ordering survives within reference-margin terciles (mid tercile 0.012 versus 0.027) and cache-norm terciles (top tercile 0.014 versus 0.021), which checks each stratification separately but does not establish independence from joint confounding. Cross-block cancellation behaves similarly (0.014 versus 0.018); aggregate routing/evidence alignment $|\sum_b \mathrm{Cov}_{\alpha_b}(r_b, w_b)| / \sum_b [\sigma_{\alpha_b}(r_b)\sigma_{\alpha_b}(w_b)]$ is small (≈ 0.002) and does not separate the groups. By token type, formatting tokens cancel least (0.027), then numeric (0.017), then word tokens (0.013). We deliberately do not use $A_V + A_K$ as a predictor of Δg, since it reconstructs that quantity by construction. Using only reference-endpoint source features, with all thresholds fixed on development prompts, a logistic flip predictor reaches held-out AUC 0.596 from the factors alone, against 0.880 for the reference margin alone, 0.820 for

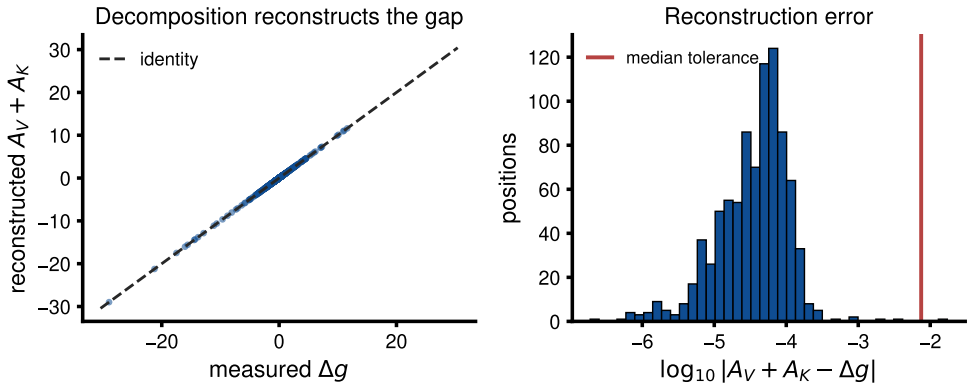


Figure 14: Transport reconstruction on Qwen3-4B. Left: measured Δg against the reconstructed $A_V + A_K$, with the identity line. Right: distribution of reconstruction error against the median tolerance, including the tail.

margin plus cache norm, and 0.886 for factors plus margin (Qwen3-14B: 0.580 versus 0.932). These reference-dependent factors describe attenuation but add little predictive information in this study.

Separate intervention cohorts. For held-out flips with $|\Delta g| > 0.05$, the sampled cohort contains 54 positions and the added-first-divergence cohort 68. At $k = 4$, attributed repair recovers 61.1% and 58.8%, respectively. Attention recovers 14.8% and 19.1%; norm selection 14.8% and 8.8%; earliest selection 14.8% and 13.2%; random selection 2.5% in each. The ordering thus survives removing the extra divergence enrichment, although the smaller cohorts have greater uncertainty. Figure 3 reports both norm removal and decision recovery on the same 122 flipped positions; the all-position norm summaries in the text use the 661 qualifying positions. These are descriptive conditional interventions, not population flip-rate estimates.

Targeted repair and reverse insertion. Selection rules spend an identical budget of repaired KV entries, $k \in \{1, 4, 8\}$ answer tokens across all layers: most negative theory attribution, largest KV-error norm, earliest answer tokens, highest attention exposure, and random tokens over three seeds. On held-out positions with $|\Delta g| > 0.05$ ($n = 661$, 122 flips), reference top-1 recovery at $k = 8$ reaches 0.803 for theory against 0.270, 0.172, 0.156, and 0.036 for attention, earliest, norm, and random. Across all 661 qualifying held-out positions, the median fraction of squared cache-error norm removed at $k = 8$ is 9.0% for theory versus 25.5% for the norm rule. The right panel of fig. 3 instead shows $k = 4$ on the 122 flipped positions; the budgets and cohorts are distinct. Reverse insertion—starting from the reference cache and contaminating only the selected tokens—gives 0.616 and 0.888 of the full displacement at $k = 4$ and $k = 8$ for theory, against 0.075 and 0.178 for the norm rule. The continuous gap-recovery ratio is heavy-tailed and we report it as a median with an interquartile range rather than a mean: $+0.46$ (IQR $[-0.39, +0.91]$) for theory at $k = 4$ against $+0.04$ for the norm rule, $+0.14$ for attention, $+0.08$ for earliest, and $+0.01$ for random. Applying the identical protocol to keys and to values separately, keys matter somewhat more than values (top-1 recovery on flips at $k = 4$: 0.344 keys, 0.311 values, 0.598 both), and at these small budgets the two sides are close to additive per position (median $(K+V)/KV = 1.07$ $[1.03, 1.10]$) even though the full-amplitude interaction above is large—localized repair is a different regime from whole-block replacement. Had targeted repair performed no better than the random and norm-based controls, we would have reported the mechanistic account as unsupported; the reviewer criterion we set in advance was that the three findings connect, namely accurate reconstruction, specific attenuation factors that characterize stable cases after accounting for margins, and interventions selected using those factors that change decisions more reliably than matched simpler controls.

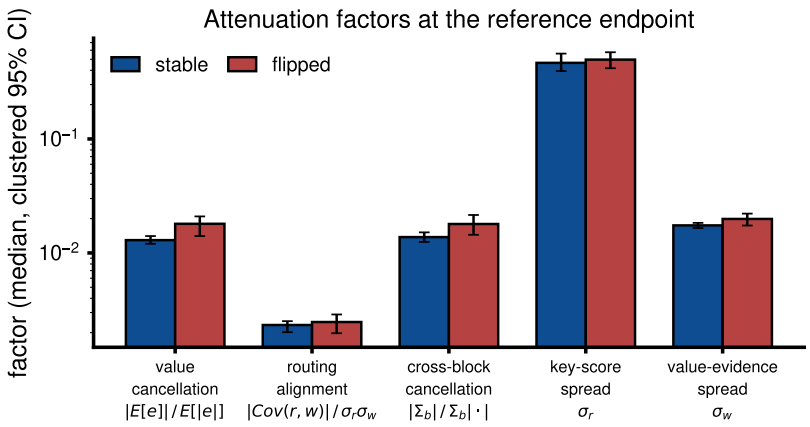


Figure 15: Attenuation factors on Qwen3-4B, measured at the reference endpoint for stable versus flipped positions; medians with prompt-clustered 95% bootstrap intervals.

A.7 Predictors and mechanistic validation detail

The near-zero magnitude correlation in the abstract and fig. 2 refers to the full answer-cache error norm in the 120-condition intervention study (Spearman $\rho = -0.0008$). Different magnitude summaries and cohorts are not interchangeable: the mid-layer contamination statistic in `results/matrix.jsonl` has a modest positive correlation with follow-up-prefill KL ($\rho = 0.267$). We do not claim that every norm statistic has exactly zero correlation.

We tested a causal decomposition risk $\approx$ (reasoning influence on answer state) $\times$ (future query readout) $\times$ (downstream logit gain) $\div$ (logit margin). Oracle predictors, which have offline access to both the gold and reuse caches, are directional rather than norm-based: $S_{\mathrm{JVP}} = \|J\delta\|$, $S_{\mathrm{Fisher}} = \delta^\top J^\top F J \delta$, and $S_{\mathrm{margin}} = 2\|J\delta\|_\infty/(\mu+\epsilon)$, all computable via JVPs or finite differences without a full Jacobian. Deployable predictors need no gold cache: source-side features during answer generation (reasoning attention mass, value-weighted reasoning contribution, per-head output contrast), readout features at the next turn (query-to-answer attention, attention entropy, key-routing sensitivity), sensitivity proxies, and logit margins. Splits are grouped by originating prompt; a genuine leave-one-model-family-out protocol remains future work (appendix B). Interpolation along $c(\lambda) = c_0 + \lambda\delta$ shows that $\|\Delta z(\lambda)\|$ grows nearly linearly and that the KL departs above the Fisher line only at large λ. For layer attribution, restoring one layer's answer block to gold shows that the raw per-layer norm barely predicts which layer matters ($\rho = 0.13$) while a local per-layer Fisher score is somewhat more predictive ($\rho = 0.25$). This is a within-condition ranking, distinct from the across-follow-up prediction in section 5, where no magnitude summary is informative.

A.8 Local Fisher diagnostic and proof

Let c vectorize the reusable answer cache, with $c_0 = C_{\mathrm{gold}}$ and $c_0 + \delta = C_{\mathrm{reuse}}$. For a fixed future continuation position t, let $z_t(c)$ be the logits and $p_t(c) = \mathrm{softmax}(z_t(c))$. Define the cache-to-logit Jacobian $J_t = \partial z_t/\partial c|_{c_0}$ and the categorical Fisher matrix $F_t = \mathrm{diag}(p_t) - p_t p_t^\top$.

Proposition 1 (Local behavioral drift)**.** *Assume $z_t(c)$ is three-times continuously differentiable near c_0. Then*

$$D_{\mathrm{KL}}\left(p_t(c_0)\|p_t(c_0+\delta)\right) = \tfrac{1}{2}\delta^\top J_t^\top F_t J_t \delta + O(\|\delta\|^3). \tag{7}$$

The proof Taylor-expands the logits and uses the fact that the categorical KL has zero first derivative and Hessian F_t in logit space (appendix A). For a future turn, define the pullback-Fisher matrix $G_U = \sum_t w_t J_t^\top F_t J_t$, with behavior-sensitive energy $\frac{1}{2}\delta^\top G_U \delta$ and behavioral null space $\mathcal{N}_U = \ker(G_U)$.

Write $z(c_0+\delta) = z_0 + J\delta + O(\|\delta\|^2)$ with $J = \partial z/\partial c|_{c_0}$. For categorical $p = \mathrm{softmax}(z)$, $D_{\mathrm{KL}}(p(z_0)\|p(z_0+u))$ has zero gradient in u at $u = 0$ and Hessian equal to the Fisher matrix

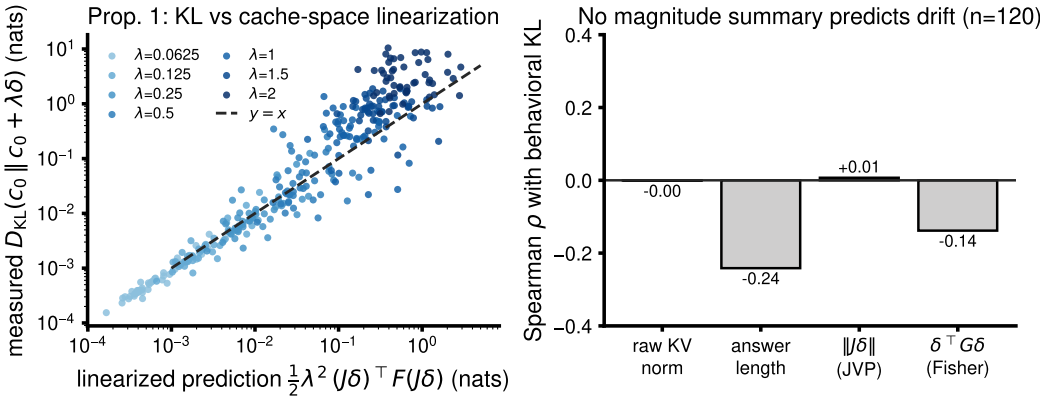


Figure 16: Mechanistic validation on Qwen3-4B (24 prompts, fp32). Left: measured next-turn KL versus the local Fisher-quadratic prediction $\frac{1}{2}\lambda^2(J\delta)^\top F(J\delta)$, extrapolated by λ^2 from one central-difference directional derivative. Points lie on $y = x$ for small λ (Prop. 1) and rise above it as higher-order terms enter. Right: Spearman correlation of candidate magnitude predictors with per-follow-up behavioral KL, all uninformative ($\rho \approx 0$).

$F = \mathrm{diag}(p) - pp^\top$. Substituting $u = J\delta + O(\|\delta\|^2)$ and expanding to second order gives $D_{\mathrm{KL}} = \frac{1}{2}\delta^\top J^\top F J\delta + O(\|\delta\|^3)$, which is eq. (7). The cross term from the $O(\|\delta\|^2)$ logit remainder enters only at third order because the KL gradient vanishes at $u = 0$. □

A.9 Reasoning- and answer-length sweep

Because the main matrix force-closes almost every trace at its budget, we run a dedicated sweep on Qwen3-4B: 24 prompts at reasoning budgets $\{64, 128, 256, 512, 1024\}$ with a generous answer budget so that answers terminate naturally, for 600 turn conditions (fig. 17). Achieved reasoning length n_R spans 64–1024 and answer length n_A spans 132–512. The result is a length-resolved restatement of the representation–decision decoupling. Reasoning length drives representational contamination strongly (mid-layer relative drift rises 0.28 to 0.40 to 0.50 across n_R terciles; Spearman $\rho(n_R, \text{contam}) = 0.83$) but moves the next-turn decision only marginally (top-1 agreement 90.6% to 88.2%; $\rho(n_R, \text{KL}) = 0.18$). Answer length does not degrade next-turn reuse: within every reasoning budget, the reused-block length correlates negatively with next-turn KL ($\rho \approx -0.4$ to -0.7), so longer stored answers are at least as reusable. This sweep separates reasoning length, retained-answer length, and generation length. It does not establish the mechanism behind the HumanEval quality difference (appendix D) or estimate a first-divergence hazard.

A.10 Turn-boundary output-KL collapse

The teacher-forced next-turn KL reported elsewhere (median about 0.36 nats) is dominated by positions that predict the next user token while the model consumes the fixed follow-up string. That is predictive drift on tokens sampling never emits, not drift in the model's own generation. To separate the two, we build, for each of the 150 conditions on Qwen3-4B, the reference greedy assistant response (up to 64 tokens) from the gold cache, fix it, and teacher-force $S = [U, y]$ on both the gold and reuse caches, recording $D_{\mathrm{KL}}(\text{gold}\|\text{reuse})$ at every position (fig. 18).

The output-KL decrease is abrupt rather than a smooth decay across the user turn. Mean KL is 0.45 at offset -4 and 0.30 at -3, then drops to about 7×10^{-6} by -2; the first generated token has median KL 1.8×10^{-6}. A template ablation associates this decrease with role-boundary tokens. It persists without `<think>` and with extra separators. A token-length-matched neutral replacement of `<|im_start|>assistant` leaves first-response KL about 0.18, whereas the user-role delimiter gives about 5×10^{-6}; withholding the boundary also leaves KL elevated (fig. 18). These controls rule out mere token consumption as a sufficient explanation within these templates. They do not hold

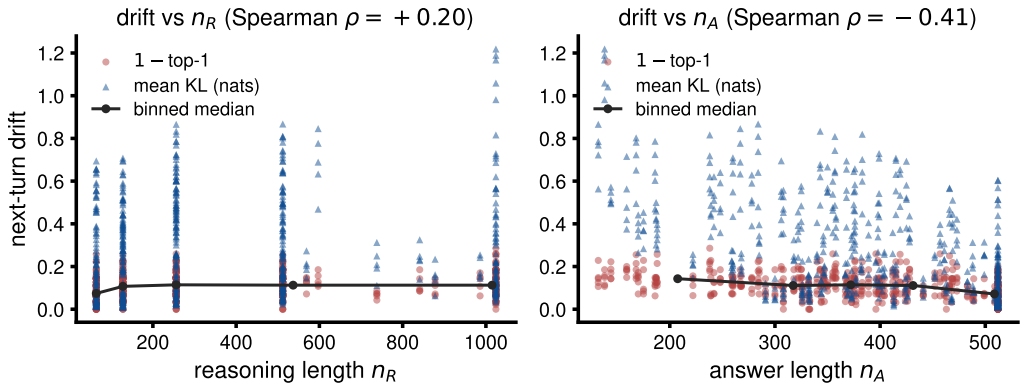


Figure 17: Length sweep on Qwen3-4B (600 turn conditions). Next-turn drift ($1-$top-1 in red, mean KL in blue) versus reasoning length n_R (left) and answer length n_A (right); black is the binned-median $1-$top-1. Reasoning length raises contamination and mildly raises the KL but leaves the top-1 decision nearly flat; longer answers do not increase drift.

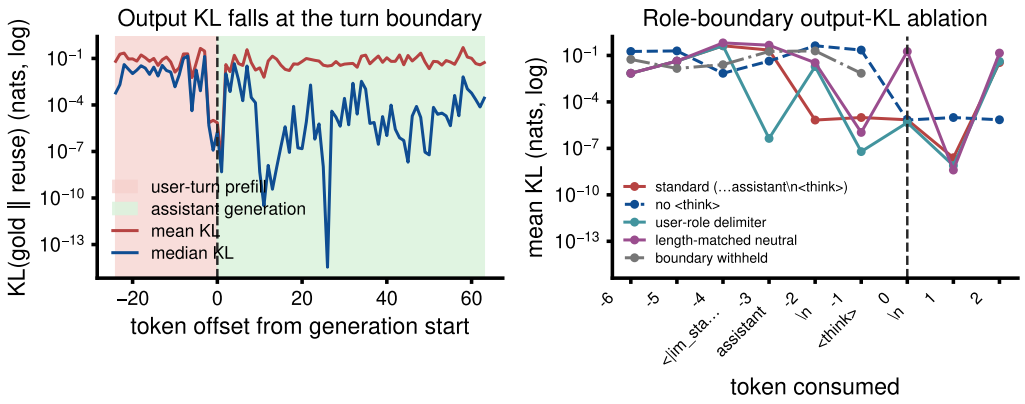


Figure 18: Boundary-associated output-KL collapse on Qwen3-4B. Left: per-position $D_{\mathrm{KL}}(\mathrm{gold}\|\mathrm{reuse})$ over 150 conditions. Right: template ablations associate the decrease with role-boundary tokens. The measurements concern output probabilities and do not establish hidden-state reset; output concentration can also produce tiny KL (proposition 2).

output entropy or centered-logit drift fixed: a predictable formatting token can yield tiny KL despite substantial internal perturbation. As proposition 2 shows, this is compatible with output concentration without hidden-state contraction. We therefore interpret the finding as boundary-associated output-KL collapse, not evidence of a learned state reset. Establishing such a mechanism would require hidden-state or centered-logit measurements and controls for output concentration.

After the boundary, mean KL over generated positions is about 0.08 nats and 14% of generated positions exceed 0.05 nats. The first generated token agrees in $150/150$ conditions, but later positions can diverge; agreement on a predictable opener is not evidence of substantive answer fidelity. Small first-response KL also occurs on Qwen3-14B, while the restricted hybrid has about 0.013 nats. These are differences in output distributions; they do not establish architecture-dependent hidden-state contraction.

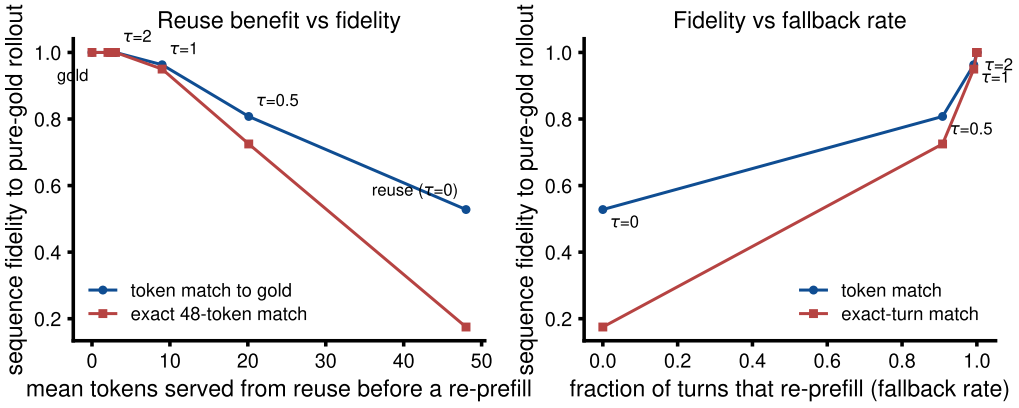


Figure 19: Deployable sequential margin-triggered fallback (Qwen3-4B, 120 turns, 48-token responses). Reuse runs on the reuse-side margin alone; the first time $m_t < \tau$, the history so far is recomputed and generation continues from the reference cache. Left: sequence fidelity to the pure-gold rollout versus tokens served from reuse before a recomputation. Right: fidelity versus fallback rate. At $\tau = 1$ the trigger recovers 95% exact-sequence fidelity but 99% of turns recompute, so this is fidelity recovery, not a compute saving.

A.11 A DEPLOYABLE SEQUENTIAL MARGIN-TRIGGERED FALLBACK

The token-level risk–coverage analysis in section 5 and fig. 2 ranks positions of a fixed reference trajectory. A deployed server instead sees tokens one at a time and cannot consult a reference branch for free. We therefore simulate the actual policy the trigger enables on Qwen3-4B (24 prompts $\times$ 5 follow-ups, 120 turns, 48-token greedy responses, mean answer 198 tokens). We run reuse generation using only the reuse-side margin m_t (top-1 minus top-2 of the reuse logits, no gold cache); the first time $m_t < \tau$, we perform one recomputation of the history emitted so far and continue from the reference cache for the rest of the turn. We compare the resulting sequence to the pure-gold greedy rollout and sweep τ (fig. 19); $\tau = 0$ is pure reuse (never recomputes) and large τ recomputes at the first token, which is exact by construction.

The teacher-forced study reports about 95% one-step agreement. In this separate sequential experiment, pure reuse matches the full 48-token reference sequence on 17.5% of turns (token-level agreement 52.8%). Small one-step disagreement can coexist with substantial sequence divergence; the pooled teacher-forced disagreement rate is not a first-divergence hazard (appendix I.7). The reuse-side margin ranks disagreement before first divergence at AUC 0.94. At $\tau = 1$, fallback occurs after a mean of 9 reuse-served tokens, giving 95.0% exact-sequence and 96.3% token fidelity; at $\tau = 0.5$, the corresponding values are about 20 tokens, 72.5%, and 80.8%.

The trigger observes only whether the reuse-side margin falls below τ; it does not detect a wording difference from a reference branch. Low-margin positions occur on almost every evaluated turn at $\tau = 1$, so the gate recomputes on 99% of turns. Fallback re-prefills the answer block that reuse had skipped. This operating point recovers sequence fidelity but retains little opportunity to avoid that prefill. Zero-repair reuse avoids answer-prefill work and reduces prepared-cache next-turn TTFT in the concurrent vLLM measurements (section 7); preparation-inclusive timing and complete-conversation timing do not show a total-processing-cost reduction (appendix H). Workload-specific answer quality remains a separate requirement. An integrated evaluation of the gate's batching, paging, and per-fallback latency (appendix C) remains future work.

B RISK-RANKING PROTOCOL AND ITS GUARANTEES

The risk study of section 7 trains a logistic model on reuse-only features: reasoning and answer lengths, first next-turn-token entropy and top-1-minus-top-2 margin, and query-to-answer attention

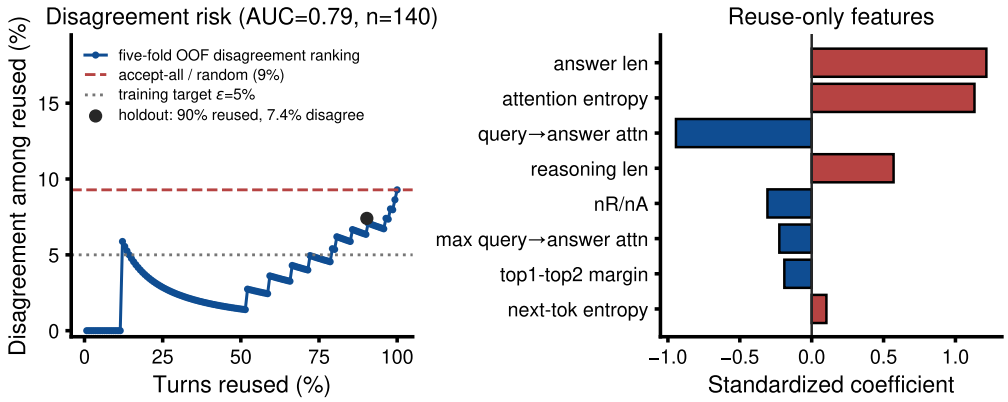


Figure 20: Final-answer disagreement risk on GSM8K ($n = 140$, 13 disagreements). Left: risk–coverage from stratified five-fold out-of-fold scores (AUC 0.79); the black point is a separate 200-split holdout heuristic with a 5% training-disagreement target. Neither analysis measures task correctness or provides a conformal guarantee. Right: standardized coefficients fitted on all conditions for interpretation.

statistics. The target is *final-answer disagreement with recomputation*, not task failure: `fail = int(a2_reuse != a2_gold)` after numerical-answer extraction. All 140 archived labels follow this rule; 13 disagree (9.3%). An extraction failure on one branch is a disagreement, but the one condition where both extractors return `None` is labeled agreement. Agreement can therefore include two incorrect or unparseable outputs, and disagreement can include a correct reuse answer. Ground-truth correctness is not the label of this experiment.

AUC and risk–coverage. We use stratified five-fold cross-validation with shuffling and seed 0. Each fold fits its scaler and class-balanced logistic regression on the other four folds, then scores the held-out fold. Concatenating one out-of-fold score per condition gives AUC 0.7862 (rounded to 0.79). The curve sorts these scores and reports observed disagreement among the accepted conditions. This is an exploratory estimate with only 13 positive labels, not a task-success or selective-risk guarantee.

Separate holdout heuristic. For 200 random 50/50 splits, the code fits the scaler and classifier on the first half and selects the largest score threshold accepting at least five training conditions with observed disagreement at most 5%. It then reports coverage and disagreement on the other half, averaging over nonempty accepted test sets. Mean test coverage is 90.3% and mean accepted-test disagreement is 7.4%, exceeding the 5% training target. The threshold-selection data are also the training data, so this is not independent calibration; the black point is distinct from the five-fold curve and AUC. A deployment guarantee would require a specified target, disjoint training/calibration/test data, and finite-sample risk control. We make no such guarantee here.

C Feature-extraction and fallback overhead

Our current risk features are computed from a probe forward that requests full attention maps (`output_attentions=True`). Optimized serving kernels avoid materializing attention matrices, so a deployable implementation would need to accumulate the query-to-answer attention statistics inside the attention kernel. We therefore do not claim the present feature extractor is production-cheap. Moreover, a rejected turn wastes the speculative reuse forward and must fall back to a recomputation, so the expected next-turn latency is

$$\mathbb{E}[\text{latency}] = P(\text{accept})\, T_{\text{reuse}} + P(\text{reject})\, \big(T_{\text{probe}} + T_{\text{fallback}}\big), \tag{8}$$

where T_{probe} is the risk-feature forward and T_{fallback} the answer recomputation. The HuggingFace speedup in fig. 6 measures T_{reuse} on accepted turns only. The vLLM experiment in fig. 7 reports next-turn latency, prior-turn preparation, and a directly timed whole conversation separately. A

runtime evaluation of the risk gate should report this accept-weighted expectation, including T_{probe}, fallback work, and any preparation unfinished when a request arrives. Wiring the risk fallback into the runtime remains future work.

D EXECUTABLE CODING FIDELITY (HUMANEVAL, TEST-GRADED)

To connect the fidelity evaluation to the long-answer coding workload that drives the systems result, we grade a two-turn HumanEval experiment with the real unit tests. In turn 1, Qwen3-4B reads a function signature and docstring and writes a full implementation under a hidden-reasoning budget (sampled, $|R| \leq 320$), yielding the physical $[P, R, A]$. In turn 2, a dependent follow-up ("give your final, complete implementation of the function you were just working on") is generated greedily from each of the five cache policies. The assistant reply is prefilled with an opening code fence so decoding writes the function directly, and the extracted function is executed against the HumanEval tests in an isolated subprocess. We report turn-2 pass@1 over all $n = 164$ problems, retention among the 128 problems whose turn-1 solution already passed, and agreement with the gold recomputation's pass/fail outcome (fig. 21). Greedy turn-2 decoding makes the cache the only source of variation across policies.

Three findings. (i) The position correction is essential: reuse-nofix collapses to 28.0% pass@1 (35.2% retention, 44.5% agreement), the coding analogue of the GSM8K ablation. (ii) Zero-repair reuse-shift is very close to gold on these longer-form answers (median $|A| = 152$ tokens, minimum 28, maximum 448): 79.9% $[73, 85]$ versus 81.1% $[74, 86]$ pass@1, with a paired-trinomial profile CI on the difference of -1.2 points $[-6.5, +4.0]$ (reuse wins 8, gold wins 10, ties 146; retention 92.2% versus 96.1%). The point estimate is a net difference of about 2 of 164 problems and the interval includes zero, but its lower end (-6.5) crosses the pre-specified 5-point non-inferiority margin, so we do not claim non-inferiority to gold on this workload. Increasing power—more seeds and more executable tasks, clustered by problem—is the way to settle the 5-point question, and we do not move the margin post hoc. This is a weaker non-inferiority outcome than the short-answer GSM8K carry; these task-level results do not identify a per-token failure mechanism. (iii) EPIC-style selective repair of the first 8 answer tokens (83.5% $[77, 88]$, 96.9% retention) and keeping the physical cache (83.5%) are each non-inferior to gold at the 5-point margin (selrep $+2.4$ points, 95% CI $[-1.4, +6.7]$; physical $+2.4$, $[-0.4, +6.0]$), and both are strictly better than zero-repair reuse here (reuse minus selrep: -3.7 points, $[-7.9, -0.0]$). These results make zero-repair reuse a candidate operating point subject to workload-specific quality validation. Its HumanEval result does not establish non-inferiority. Selective repair has a better observed quality result on this workload, with additional latency: in the separate HuggingFace batch-32 microbenchmark, the median per-trace selective-repair/reuse TTFT ratio is 1.58. The quality–latency trade-off must be evaluated for the intended workload; it does not justify a universal zero-repair default.

E HARDWARE AND SOFTWARE ENVIRONMENT

All absolute timings (figs. 6 and 7, the 3.4 ms HuggingFace rotation, the HuggingFace batch-32 comparison) are measured on a single NVIDIA H200 GPU (140 GB, Hopper, compute capability 9.0; driver 580.126.09), one GPU per measurement (`cuda:0`). Software: `PyTorch 2.10.0+cu130` (CUDA runtime 13.0); `transformers 4.57.6` for the dense Qwen3-4B core (the behavioral matrix, GSM8K quality, HumanEval coding fidelity, and risk study); `numpy 2.2.6`, `scipy 1.15.3`, `scikit-learn 1.7.2`; model weights in `bfloat16`. The restricted Qwen3.5-27B hybrid study uses `transformers 5.9.0`. The attention backend is `sdpa` (PyTorch scaled-dot-product attention), except that the behavioral matrix uses `eager` attention to suppress sequence-length-dependent SDPA numerical variation. Each HuggingFace forward or rotation component is summarized by the median of 10 timed repetitions after model warmup; forward caches are rebuilt outside the timed interval. Per-trace reuse cost adds the median follow-up-forward and rotation components, and selective repair additionally adds its median repair-forward component. Across the 40 traces, the batch-32 ratio of policy medians is 7.09, while the median paired speedup is 6.71 (fig. 6). The batch-32 numbers replicate a single request's cache to batch size 32 and time one batched forward; this is a compute-saturating microbenchmark of the HuggingFace path, not an integrated serving runtime (no continuous batching, scheduler, paged-attention allocator, or eviction). The vLLM numbers in fig. 7 use released vLLM $0.29.0$

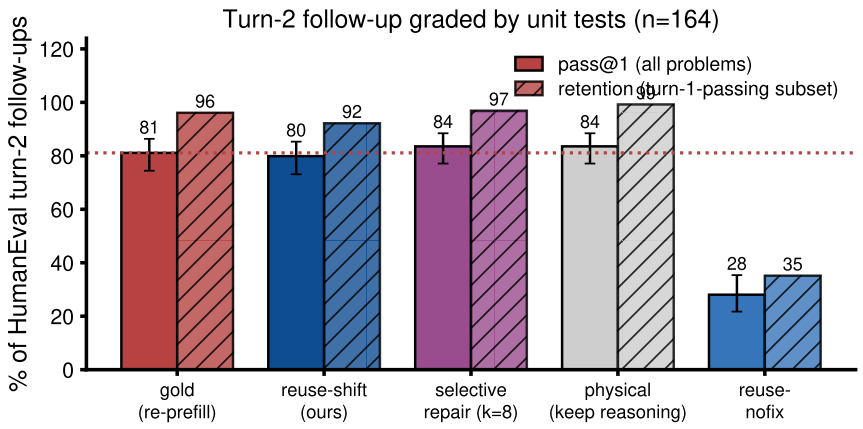


Figure 21: Executable coding fidelity on HumanEval ($n = 164$), turn-2 follow-up graded by the real unit tests. Solid bars: pass@1 over all problems (Wilson 95%). Hatched: retention among the 128 turn-1-passing problems. Zero-repair reuse-shift is within 1.2 points of gold, a difference whose interval includes zero but crosses the 5-point margin, so non-inferiority is not established here. Selective repair and the physical cache are each non-inferior to gold, and the position correction is essential.

(FlashAttention 3, prefix caching on, `PyTorch 2.13.0+cu130`) with the Shift-and-Reuse patch in `code/vllm_v0.29.0_shift_and_reuse.patch`, in-process engine core, and the same H200. The archived vLLM batch experiment uses three timed repetitions per policy at each batch size, not the ten repetitions of the HuggingFace component microbenchmark. Within each batch size the three policies are given byte-identical follow-up token ids (the per-slot nonce depends on the batch slot only) and are measured in a randomized order; the whole-conversation timing decodes turn 1 for real and consumes the cache that decode produced, so it is a single measured interval rather than a sum of separately timed parts. The transport-attribution study of section 6 runs in `float32` with `eager`-equivalent attention through a registered capture interface that reproduces the eager forward exactly (maximum logit difference 0.0), on two H200s in parallel; at about 48 s per prompt for Qwen3-4B the full 100-prompt study takes roughly 40 minutes per GPU. Absolute numbers will differ on other hardware.

F Reproducibility Statement

The submission includes the experiment code and archived per-condition results. Main-paper data figures and the vLLM paired interval regenerate with `python code/gfigs_main.py`; the serving overview uses `python code/gfig_system_main.py`. Detailed figures use the corresponding analysis scripts. The current editing pass reanalyzes these files and does not re-run model inference. The cache-surgery procedure is stated in full in algorithm 1; the models, cache conditions, prompt counts, and follow-up classes are given in appendix A, alongside the numerical controls. The local Fisher result is proved in appendix A, and the finite attention bound in appendix I, and the scope of the claim it supports is delimited in appendix G. The two-turn GSM8K and HumanEval protocols, including the pre-specified non-inferiority margin and the interval construction, are described in section 7 and appendix D; the risk-ranking protocol and its limits are in appendix B. Hardware, driver, library versions, and the timing methodology behind every reported latency are listed in appendix E. The transport-attribution study of section 6 states its development/held-out split, its pre-specified reconstruction tolerance, its quadrature-escalation rule, and its prompt-clustered bootstrap in the main text; the instrumentation, the 100-prompt dataset, and the intervention harness are `code/transport_core.py`, `code/transport_prompts.py`, `code/transport_repair.py`, and `code/measure_transport.py`, with analysis in `code/gfigs_transport.py`. Appendix F.1 maps each experiment to the script that produces it.

F.1 Reproducibility details

The accompanying code includes scripts for phase-one KV capture, exact cache surgery, scaled prompt–follow-up evaluation, numerical checks, multi-turn accumulation, three-way reference-versus-physical comparisons, the four-way downstream-quality and systems experiments, the vLLM Shift-and-Reuse patch and TTFT harness (`code/bench_vllm_ttft.py`), the direct paged-KV verification and page-retirement probe (`code/verify_vllm_kv.py` → `results/vllm_kv_verify.json`), the executable turn-2 grading under vLLM (`code/quality_vllm.py` → `results/vllm_quality_*.json`), the transport-attribution study of section 6 (`code/measure_transport.py`, analyzed by `code/gfigs_transport.py`, raw rows in `results/transport_q4b_{a,b}.jsonl` and `results/transport_q14b.jsonl`, console summaries in `results/transport_summary_q{4,14}b.txt`), non-inferiority analysis, and aggregation. Every figure and table regenerates from the raw per-condition result files (shipped in `results/`) via the released `g*.py` scripts; analysis scripts read from `out/` when present and otherwise fall back to `results/`. Environment-specific patches used for the hybrid Qwen3.5-27B model remain separate from the dense-model core experiment.

G Scope of the theoretical claims

The local KL approximation and the finite attention-mediated bound have distinct scopes. Equation (7) (proved in appendix A) approximates the next-turn behavioral KL by the pullback-Fisher quadratic $\frac{1}{2}\delta^\top J^\top F J\delta$ to second order in the cache perturbation, with an $O(\|\delta\|^3)$ remainder; it is an asymptotic expansion, not a bound. Figure 16 validates this linearization directly: we estimate a single local directional derivative $J\delta$ by central finite difference, extrapolate the pure quadratic $\frac{1}{2}\lambda^2(J\delta)^\top F(J\delta)$, and compare it against independently measured $D_{\mathrm{KL}}(c_0 \,\|\, c_0 + \lambda\delta)$. We make no global-equivalence claim: the rollout divergences at large λ show the expansion is a small-perturbation diagnostic, not a guarantee of behavioral identity. Section 4 supplies the attention-level decomposition of key-routing versus value-content perturbations that eq. (7) does not, and unlike eq. (7) it is exact at finite amplitude because the perturbation enters the forward pass only through injected answer-block keys and values; the price is that it is a per-position identity for one pairwise logit gap, not a statement about the output distribution. We still do not derive closed-form smoothness constants, and the Cauchy–Schwarz bound that eq. (4) implies is loose by about two orders of magnitude in practice, so neither result yields a usable certificate.

H Additional vLLM measurements

We also implement the mechanism in vLLM $0.29.0$. At request completion the patch copies prefix keys unchanged, RoPE-translates only answer keys, preserves values, stops at the computed-KV frontier, registers complete canonical blocks under the visible transcript's hashes, and retires the prefix-cache entries for the physical blocks holding the deleted trace (`code/vllm_v0.29.0_overlay/`). Rather than infer correctness from cache hits, we read the pages back out of the GPU after compaction and compare them against the intended operation ($n = 24$; `code/verify_vllm_kv.py`). Prefix keys and all values are bitwise identical to their sources on $24/24$ traces, and the rotated answer keys deviate by at most 6.2×10^{-2} against a KV magnitude of 318, a relative error of 3.3×10^{-4} within the one-ULP bf16 budget of 3.9×10^{-3} on $24/24$. Two controls establish that this is not a vacuous test: the materialized answer keys differ from their *unrotated* sources by 32.1, and the prefix keys differ from a wrongly rotated version by 31.1. Against a genuine $[P, A]$ prefill the compacted pages carry the expected contamination—relative drift 0.101 on answer keys and 0.260 on answer values, with the prefix bitwise identical—so the serving path reproduces the representational error the rest of the paper studies. Retirement is measured, not asserted: after compaction a request presenting $[P, R, A]$ hits 8 blocks rather than 40, exactly the $\lfloor |P|/16 \rfloor$ blocks it legitimately shares with the canonical sequence, while $[P, A]$ still hits all 22.

Behavioral fidelity is graded by execution rather than by token agreement. Our earlier serving check compared a single follow-up token, which on Qwen3 is always the `<think>` opener and therefore verifies only the formatting prefix. We now generate turn 2 through the visible answer and run the resulting function against the real HumanEval unit tests ($n = 163$; `code/quality_vllm.py`),

with turn 1 decoded for real so the reused pages are the ones decoding produced. Against a canonical re-prefill at 82.2% turn-2 pass@1, Shift-and-Reuse reaches 79.1% with 94.5% pass/fail agreement and 94.1% retention on problems turn 1 already solved, and keep-physical reaches 84.0%. The no-rotation ablation collapses to 30.1% pass@1 and 37.8% retention, which is what makes the check informative: it detects a cache error that first-token agreement cannot see.

On the same Qwen3-4B coding traces ($n = 24$, median $|A| = 209$, $|R| = 303$), isolated lookups show re-prefill hitting 136 tokens and prefilling 242, while reuse hits 344 and prefills 34, matching keep-physical's 34-token prefill. With three timed repetitions per policy, byte-identical follow-up token ids across policies, and randomized measurement order, the next-turn-only wall clock is $1.00\times$, $2.51\times$, and $3.16\times$ at concurrent batches 1, 8, and 24 (7.7 vs 7.7 ms, 29.7 vs 11.8 ms, and 83.8 vs 26.5 ms; fig. 7). Batch 1 is exactly neutral once the prompts are matched. These are next-turn gains only. Charging prior-turn compaction at about 9.7 ms per request gives reuse 17.4, 89.1, and 258.8 ms, slower than re-prefill. Timing a complete two-turn conversation as a single interval—turn-1 decode, compaction, and the follow-up—rather than summing parts, reuse takes 1855 ms against re-prefill's 1842 ms, a whole-conversation ratio of $0.99\times$, because the 1834 ms turn-1 decode dominates everything the follow-up can save at this answer length. These preparation-inclusive measurements characterize total processing cost; the next-turn TTFT experiment isolates the response-latency benefit of a prepared cache.

Eligibility and paired uncertainty. The archived HumanEval rows cover 163 unique tasks; HumanEval/152 is absent. The harness skips a trace if it has no usable hidden span or fewer than 32 answer tokens, and the archive does not identify which branch caused this omission. Against recomputation, reuse has two paired wins and seven losses, giving -3.1 points with paired profile-likelihood 95% interval $[-7.2, +0.5]$. This does not establish non-inferiority at a five-point margin. The quality baseline recomputes its canonical input from an empty cache; it is not the prefix-warmed latency baseline. All timing and quality numbers are therefore reported as separate experiments.

Preparation budget. The next-turn batch-wall saving permits approximately 2.23 ms per request at batch 8 and 2.39 ms at batch 24 for compaction to break even on total batch wall time, compared with about 9.7 ms for the current serial implementation. This arithmetic budget concerns total processing time. The request-TTFT benefit reported in the main text assumes the between-turn update has completed before the follow-up arrives. Preparation still consumes GPU resources; mixed-load interference and tail-latency SLA attainment are not measured here.

Timing aggregation. The three saved batch wall-clock samples are reduced by a median; charged time is the median of the per-repetition wall-plus-compaction sums, not necessarily the sum of two marginal medians. The TTFT field is a summary of request TTFTs within the measured batch runs; the archive does not retain every request's timing. Median TTFT ratios are 1.00, 2.86, and 3.85 at batches 1, 8, and 24, whereas batch-wall ratios are 1.00, 2.51, and 3.16. No independent-request confidence interval is inferred from these summaries.

I Finite Cache Perturbations: Derivations and Falsifiable Predictions

I.1 Scope, notation, and the common token sequence

All comparisons in this section fix U and the continuation tokens before the scored prediction. The gold and reuse forwards recompute those future tokens independently from their respective caches. Hence their future hidden states and queries may differ. At a first greedy divergence they still share this token prefix, so the analysis applies; after divergence it describes a counterfactual fixed-token comparison, not the two autonomous trajectories.

The external variables are the post-RoPE answer keys and answer values at every reusable layer. The stable prefix is fixed. Let $c(s) = c_0 + s\delta$, where c_0 is the answer block in $KV(P, A)$ and δ is the translated physical block minus that reference block. The mask and logical positions are fixed along this path. Cache interpolation need not correspond to a tokenizable prompt; it is a well-defined intervention on continuous model inputs. For grouped-query attention a shared cached

key/value is used by several heads, and the sum over those uses is required. The factor $1/\sqrt{d_k}$ in the equations is replaced by the actual attention scale when different. Nonstandard score functions require differentiating their own score function.

Lemma 2 (An exact first-layer anchor)**.** *Consider a deterministic causal decoder whose inputs are token embeddings, whose pre-attention operations in the first layer are tokenwise, and whose positional dependence before caching consists only of RoPE on keys. For fixed answer tokens at their translated logical positions, the first-layer answer keys and values in reuse equal those in the reference cache, in exact arithmetic.*

Proof. The input embedding and any tokenwise normalization/projection are identical for each answer token in the two histories. Thus their first-layer pre-RoPE keys and values agree. Lemma 1 puts the keys at the same rotary phase, while the values do not depend on that phase. No preceding-token attention has yet entered these projections. □

The first-layer *attention outputs* need not agree: during physical generation they attended to R. The lemma therefore does not propagate equality to later layers. Additive absolute positions, cross-token operations before the first projection, dropout, recurrent preprocessing, and finite-precision kernels require separate treatment. It supplies an exact shared source of token identity, not a proof that deeper perturbations are small.

I.2 Differentiating one attention read

Suppress b and consider one head with scores $a_i = q^\top k_i/\sqrt{d_k}$, $\alpha = \mathrm{softmax}(a)$, and $o = \sum_i \alpha_i v_i$. For the *direct* cache derivative hold q and all upstream-computed inputs fixed at their current values and differentiate only the edited cached k_i, v_i. Writing $r_i = q^\top \Delta k_i/\sqrt{d_k}$, zero for untouched keys, yields

$$\partial_s \alpha_i\big|_{\mathrm{direct}} = \alpha_i \left(r_i - \sum_j \alpha_j r_j \right), \tag{9}$$

$$\partial_s o\big|_{\mathrm{direct}} = \sum_i \alpha_i \Delta v_i + \sum_i \alpha_i r_i (v_i - o) = h. \tag{10}$$

This identity holds at every point of the full-amplitude interpolation. For a fixed vector u, define $e_i = u^\top \Delta v_i$ and $w_i = u^\top v_i$. Then

$$u^\top h = \mathbb{E}_\alpha e + \mathrm{Cov}_\alpha(r, w). \tag{11}$$

The first term is a weighted, signed value error. The second measures whether score perturbations preferentially redirect attention toward different *task-projected* evidence. Attention weights can move substantially while this covariance remains small. Neither conclusion follows from $\|\delta\|$ or attention mass alone.

A useful controlled special case. If only values change, the head's incoming query and keys are fixed, and its downstream readout is affine, then u and α are constant and the gap change is exactly $u^\top \sum_i \alpha_i \Delta v_i$. Arbitrarily large value perturbations whose weighted task projection is zero have no effect on that gap. By contrast, choosing $\Delta v_i = -Mu/\|u\|^2$ on an answer block of positive mass reduces the gap by $M\beta$ and can flip the decision. These are mathematical examples, not claims about the measured cache distribution. They show why a norm-based universal robustness conclusion is impossible.

I.3 Lifting the source term through the full decoder

Unroll the fixed-token computation as a directed acyclic graph. Each external answer-cache entry is read at the attention heads of its own layer. Let b enumerate every such head and future query position that can affect z_t. Reverse-mode chain rule expresses the derivative of g_{tj} with respect to these external cache inputs as the sum over their direct uses:

$$g'_{tj}(s) = \sum_b u_{tjb}(s)^\top h_b(s), \qquad u_{tjb} = \frac{\partial g_{tj}}{\partial o_b}. \tag{12}$$

The derivative u_{tjb} includes all descendant operations, including subsequent token states and their KV computations. Query changes caused by an earlier layer's cache edit are accounted for through that earlier source and its descendants. Adding an extra total-query derivative to each h_b would double-count these paths.

Proof of theorem 1. Apply eq. (11) to each direct use. Since $e_i = 0$ outside A, triangle inequality also gives $|\mathbb{E}_{\alpha_b} e_b| \le \mathbb{E}_{\alpha_b}|e_b|$. The looser exposure form is

$$|\mathbb{E}_{\alpha_b} e_b| \le \beta_b \epsilon^V_{tjb}, \qquad |\operatorname{Cov}_{\alpha_b}(r_b, w_b)| \le \sigma_{\alpha_b}(r_b)\sigma_{\alpha_b}(w_b). \tag{13}$$

Triangle inequality bounds $|g'_{tj}(s)|$ by the summed integrand in eq. (5). The fundamental theorem of calculus then bounds $|g_{tj}(1) - g_{tj}(0)|$. If each initial gap exceeds its bound, all final gaps remain positive. No assumption that $\|\delta\|$ is small is needed. □

One can retain $|\mathbb{E}_\alpha e|$ instead of $\mathbb{E}_\alpha|e|$ for a sharper bound, or retain the signed covariance in an attribution. These progressively preserve cancellation; all remain offline quantities. Evaluating only the original runner-up does not certify top-1 preservation because a different competitor can overtake. Numerical quadrature estimates an integral and is not a certified upper bound unless its integration error is bounded.

Non-smoothness and architecture. The result also holds for an absolutely continuous fixed-token map with the derivative defined almost everywhere. Smooth dense decoders with positive normalization epsilon satisfy the stated path regularity. Hard top-k expert routing may create jumps when the selected experts change; it is not enough to invoke an almost-everywhere derivative in that case. For piecewise-smooth g with jumps at s_r, the endpoint difference includes $\sum_r [g(s_r^+) - g(s_r^-)]$, and the bound must add the sum of their absolute sizes. The present instrumentation targets the dense Qwen3 core; it does not certify discontinuous expert routing or solve hybrid recurrent-state editing.

I.4 Three factors, and why the random-direction result does not settle them

Equation (5) separates exposure (β and the distribution of r), downstream readout (u), and evidence contrast (w). Value error cancels when its task projection has a small weighted mean; key error is attenuated when the query does not see it, when score changes are common across the attended positions, or when rerouted values carry similar task evidence. Downstream gains may amplify as well as attenuate these effects. Higher answer attention is not universally protective: it can expose more error, or improve access to useful answer content. The protective coefficient in the existing task-risk regression does not identify which of these effects dominates.

The existing equal-norm control randomizes directions separately for each layer's key block and value block. Its Fisher-energy ratio compares the measured perturbation with that particular random ensemble. It does not estimate the fraction of cache norm in a null space, and it does not establish the relative orientation to every token-pair decision direction. In particular, Fisher energy averages squared logit differences under the reference probabilities, while a greedy decision concerns the signed differences to competitors. A perturbation can affect probabilities more than random noise and still preserve the winner.

I.5 A confidence proxy is not a decision certificate

Let $\Delta z = z(1) - z(0)$ and $\operatorname{osc}(\Delta z) = \max_j \Delta z_j - \min_j \Delta z_j$. The standard sufficient condition

$$\mu_0 > \operatorname{osc}(\Delta z) \quad \Leftarrow \quad \mu_0 > 2\|\Delta z\|_\infty \tag{14}$$

preserves the reference winner. The same conclusion follows if the *reuse* margin exceeds this oscillation: if a different token won under reuse, its advantage over the reference winner could be no larger than the oscillation. Both tests require perturbation information unavailable to a reuse-only gate. A large reuse margin can coexist with a confidently wrong winner if the perturbation is larger still. The measured margin AUC is therefore an empirical ranking result, not a distribution-free safety guarantee.

If $B_t = \max_{j \ne y^*} B_{tj}$, then for any $b > 0$ and any prompt distribution,

$$\Pr(\text{flip at } t) \le \Pr(\mu_t \le b) + \Pr(B_t > b). \tag{15}$$

This union bound needs no independence assumption. It separates the frequency of fragile decisions from the tail of the attention-mediated perturbation. Bounding either term empirically requires suitable held-out data. A margin-only classifier estimates their joint regularity on a workload; it does not control the second term by construction.

I.6 Output concentration can mimic a state reset

Proposition 2 (Finite output-KL identity and concentration bound)**.** *For arbitrary finite logits* $z, z + \Delta z$ *with* $p = \mathrm{softmax}(z)$*,*

$$D_{\mathrm{KL}}(p \| \mathrm{softmax}(z + \Delta z)) = \log \mathbb{E}_p e^{\Delta z} - \mathbb{E}_p \Delta z. \tag{16}$$

If a token y *has* $p_y = 1 - \varepsilon$ *and* $|\Delta z_j - \Delta z_y| \le a$ *for every* j*, then*

$$D_{\mathrm{KL}}(p \| \mathrm{softmax}(z + \Delta z)) \le \varepsilon(e^a - 1 - a). \tag{17}$$

Proof. Substitute the two log-softmax expressions into KL to obtain eq. (16). Subtract the common displacement Δz_y, so $d_y = 0$. Using $\log x \le x - 1$ gives $D_{\mathrm{KL}} \le \sum_{j \ne y} p_j (e^{d_j} - 1 - d_j)$. On $[-a, a]$, the largest value of $e^d - 1 - d$ is $e^a - 1 - a$. The probabilities over $j \ne y$ sum to ε. □

This is a property of softmax, not a new model-specific mechanism. It shows that tiny KL near a predictable delimiter does not establish small hidden-state drift. Entropy, centered-logit drift, and substantive-token behavior must be measured before interpreting the boundary observations as a state contraction. The bound does not imply small KL when a grows enough to offset the small ε.

I.7 First-divergence hazard and rollout length

Let T be the first greedy divergence and $h_t = \Pr(T = t \mid T \ge t)$, over originating conditions. Then

$$\Pr(T \ge L) = \prod_{t=0}^{L-1} (1 - h_t). \tag{18}$$

This follows from conditional probabilities, without assuming independent token errors. The pooled flip rate over all teacher-forced positions, including positions after a divergence, is not h_t and cannot be exponentiated to predict sequence fidelity. The fixed-token bound applies on the surviving prefix; later autonomous histories require a new comparison. Near-parity on aggregate task accuracy also does not imply exact-sequence parity.