AFTER ERROR FEEDBACK: CONTINUED ACCESS TO FAILED REASONING CONTRIBUTES TO CONTEXTUAL DRAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Iterative LLM workflows increasingly rely on previous attempts and feedback to improve subsequent solutions. However, when earlier attempts are incorrect, retaining them may paradoxically impair later reasoning, a phenomenon known as contextual drag. While prior work has documented this effect, it remains unclear whether the interference arises from the initial exposure to the failed attempt or from the model's continued access to it during subsequent reasoning. We introduce a same-text attention intervention that isolates this mechanism by preserving the full prompt, token positions, prefill computation, and first-token logits, while blocking only later generated queries from attending to the failed draft. Across controlled evaluations, Qwen3-8B on selected Game24 problems shows that accuracy drops from 87.0% in clean context to 39.1% with full access to a failed draft, but recovers to 78.3% when continued access is removed, yielding a +39.1 percentage-point effect (95% question-cluster bootstrap interval: +23.9 to +54.3). Additional models and two MATH settings show consistent trends. In contrast, blocking verified correct history reduces accuracy, indicating that the effect depends on the retained history and its feedback. These results identify continued direct access after initial processing as a causal contributor to contextual drag.

1 INTRODUCTION

Iterative reasoning systems must decide what information to carry from one attempt to the next. Self-refinement and agents with verbal feedback retain earlier solutions and feedback to guide revision (Madaan et al., 2023; Shinn et al., 2023). This design preserves intermediate work that may be useful, but also keeps rejected reasoning available throughout the next attempt. Feedback identifies what went wrong without necessarily changing which parts of the history remain accessible. Understanding the consequences of this continued access is important for designing systems that can reuse prior work while recovering from failure.

A concrete challenge is *contextual drag*: failed attempts retained in context can lower accuracy on a subsequent solution (Cheng et al., 2026). This effect can persist even when the earlier failure is explicitly identified. The finding raises a question about how retained history affects computation. A failed trace may influence the next solution when the model first processes it, and it may remain influential through repeated access during generation. Establishing whether this later access contributes would clarify what can still be changed after the model has already read the failed attempt.

These two stages offer distinct routes for influence. During *initial exposure*, processing the failed trace in prefill can change other prompt states, which may then carry its influence into the new solution. Through *continued direct access*, later generated queries can also attend to the draft's keys and values. Both routes may contribute to contextual drag. Deleting, rewriting, or resetting the history changes initial exposure and later access together, so these prompt-level interventions cannot separate their contributions.

We separate them by holding initial exposure fixed and changing only later direct access. Our principal conditions, Full and LateBlock, contain the same failed draft, error signal, prompt tokens, and absolute positions. Both process the full prompt normally and sample the first generated token from the same full-prefill logits. Their attention rules diverge only after this point: generated queries

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.

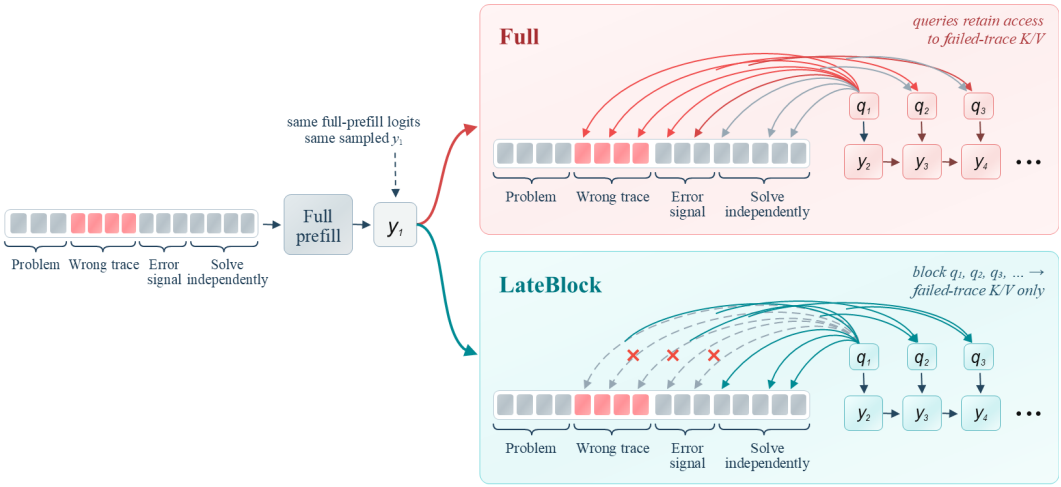

Figure 1: Same-text intervention isolating continued access after full exposure. The prompt itself is unchanged. In Full, every generated query retains ordinary causal access to the known-wrong draft. In LateBlock, the complete prompt is processed with ordinary causal attention, and y_1 is sampled from the same full-prefill logits as in Full. Only then does blocking begin: the query at the y_1 position, which predicts y_2, is the first affected query. It and all subsequent generated queries cannot directly attend to the failed-draft key–value positions. Dashed arrows and crosses denote the removed attention edges. Access to the problem, error signal, and independent-solve instruction is preserved.

retain direct access to the draft in Full and lose it in LateBlock. This comparison addresses a concrete question:

> *After the model has processed the full failed trace, does blocking direct access to it improve the next solution?*

Figure 1 illustrates the intervention boundary: full exposure and the first sampled token precede the change in access.

This comparison tests whether changing access after exposure can improve the next solution. It leaves information already carried into other prompt states available, so recovery would establish an effect of continued access even when the model has not forgotten the failed trace.

Main Findings. On 23 selected Game24 questions, Qwen3-8B accuracy falls from 40/46 in clean context to 18/46 with a failed draft and explicit error feedback. Blocking only later direct access raises accuracy to 36/46, an effect of $+39.1$ percentage points [23.9, 54.3] (Section 3). This is the primary result: a substantial part of the observed loss depends on continued access even after full prefill exposure, while the interval and residual gap to clean context do not establish complete recovery.

The same directional pattern appears for Qwen3-32B and Gemma 4 E4B on a shared 15-question subset and for Qwen3-8B on two eight-question MATH source groups. On eight Game24 questions with verified correct history, however, blocking lowers accuracy from 16/16 to 13/16. Together, these results suggest that the effect of continued access depends on the quality of the retained history.

Finally, beginning the block earlier gives no consistent additional benefit. The late intervention is therefore sufficient for the observed recovery, even though other prompt and generated states may still carry information from the draft. We use access control as an identification tool: it tests the causal contribution of continued direct access after exposure. Error detection, a unique information pathway, and the deployment value of blocking require separate evidence.

Table 1: Experimental conditions. Full, LateBlock, and PathBlock share identical prompt text, token IDs, absolute positions, draft span, error signal, and sampling configuration.

Condition	History	Prefill access	Later draft access
Clean	None	N/A	N/A
Full	Failed draft	Full	Full
LateBlock	Failed draft	Full	Blocked
PathBlock	Failed draft	Blocked after draft	Blocked

2 Access Intervention and Evaluation Setup

Let a reasoning problem be x, and let $d = (d_1, \ldots, d_m)$ be a model-generated trajectory for the same problem that has been externally verified as incorrect. The evaluation prompt contains the problem, the complete text of d, an explicit signal that d is incorrect, and an instruction to solve x independently. Let $y = (y_1, \ldots, y_T)$ denote the new generated trajectory.

Under full causal attention, a generated query at position q can attend prompt keys in the failed-draft span D. We intervene on the binary accessibility relation $A(q, k)$. Full preserves the ordinary causal mask:

$$A_{\text{full}}(q, k) = \mathbb{1}[k \leq q]. \tag{1}$$

Our primary intervention, LateBlock, preserves all prefill access and then blocks generated queries from attending to the draft. For a draft span $D = [s_d, e_d)$ and prompt length L,

$$A_{\text{late}}(q, k) = A_{\text{full}}(q, k)\mathbb{1}[q < L \ \vee \ k \notin D]. \tag{2}$$

All prompt queries, including the full prefill, therefore use ordinary causal attention. The logits used to sample y_1 are also the ordinary full-prefill logits. Blocking first affects the next query after y_1 has been appended.

The secondary PathBlock diagnostic begins earlier, at the prompt suffix following the draft:

$$A_{\text{path}}(q, k) = A_{\text{full}}(q, k)\mathbb{1}[q < e_d \ \vee \ k \notin D]. \tag{3}$$

This blocks post-draft prompt queries and generated queries from directly attending to D, while preserving computation within the draft itself. It tests intervention timing. Unblocked states may still carry draft-derived information. The equations display the rule for full causal layers. For a model with native sliding-window layers, we intersect the additional access restriction with that layer's native attention relation, preserving its ordinary computation wherever the intervention does not apply.

2.1 Experimental conditions

Table 1 summarizes the four conditions used in the main experiment.

Our primary estimands are

$$\Delta_{\text{drag}} = \text{Acc}(\text{Clean}) - \text{Acc}(\text{Full}), \tag{4}$$

$$\Delta_{\text{access}} = \text{Acc}(\text{LateBlock}) - \text{Acc}(\text{Full}). \tag{5}$$

$\Delta_{\text{drag}} > 0$ denotes an observed loss from the failed-history prompt. The primary causal contrast is LateBlock versus Full: Δ_{access} measures the accuracy change from removing later direct access at fixed initial exposure. A positive contrast represents recovery when the drag prerequisite holds. Its sampling uncertainty is reported separately. The secondary contrast between PathBlock and LateBlock probes whether beginning the block earlier improves recovery.

2.2 Method

Same-text access intervention. For each problem–draft pair, we tokenize the complete prompt once and reuse its token IDs, positions, and draft span in Full, LateBlock, and PathBlock. The error label and decoding settings are also fixed, so Full and LateBlock differ only in later draft access.

The mask changes logical attention edges while keeping the draft keys and values at their original positions. Invalid span or access metadata stops the request. Qwen uses vLLM with FlexAttention, while Gemma uses an eager Hugging Face implementation that preserves its hybrid attention layout. We analyze within-model contrasts because these backends are not bitwise equivalent. Backend-specific details, mask validation, and hash checks are reported in Appendix A.

2.3 Experimental Design

Models and decoding. We evaluate Qwen3-8B, Qwen3-32B (Yang et al., 2025), and Gemma 4 E4B (Google DeepMind, 2026), each in its recorded thinking mode. Qwen3-8B is evaluated under all four conditions. The other models use Clean, Full, and LateBlock. Within each model, conditions share the same inference implementation, decoding configuration, and scorer. Decoding settings and random seeds are provided in Appendix A.3.

Game24. Game24 requires combining four input numbers with arithmetic operations to obtain 24, using each number exactly once. It provides executable correctness and a compact solution structure, and has been used to study deliberate LLM reasoning (Yao et al., 2023). We score the final expression for arithmetic correctness and input-multiset compliance.

We evaluate 23 questions with Qwen3-8B and a shared 15-question subset with the other targets, using two paired seeds per question. Baseline screening required two verified correct and two substantive incorrect trajectories, with one question retained under an earlier rule. Each question has one frozen failed draft, reused across target models where applicable. Intervention outcomes did not determine inclusion. Appendix A.4 details screening. Table 8 in Appendix E reports eligibility and draft-content sensitivities.

Mathematical reasoning. As a cross-task check, we evaluate 16 MATH/MATH500 problems (Hendrycks et al., 2021) with Qwen3-8B, two seeds, and four conditions per problem. Eight MATH500 problems use Qwen3-1.7B source drafts, and eight MATH-train problems use Qwen3-4B source drafts. All drafts are verified substantive failures. We report the groups separately because problem source and draft-source model vary together. Scoring details are provided in Appendix A.3.

Clean comparator. Removing the entire failed-history block yields the Clean prompt. This changes both initial exposure and later access, providing a no-history reference. The primary controlled contrast is LateBlock versus Full, which holds prompt content and initial exposure fixed.

Correct-history specificity. To examine whether blocking helps regardless of history quality, we repeat the LateBlock–Full comparison with verified correct drafts on eight Game24 questions using Qwen3-8B. Correct drafts are selected without observing intervention outcomes, using character length for matching to failed drafts, and are explicitly labeled `CORRECT`. On these same eight questions and paired seeds, we define

$$\Delta_h = \mathrm{Acc}_h(\mathrm{LateBlock}) - \mathrm{Acc}_h(\mathrm{Full}), \quad h \in \{\mathrm{failed}, \mathrm{correct}\}, \tag{6}$$

$$\Delta_{\mathrm{specific}} = \Delta_{\mathrm{failed}} - \Delta_{\mathrm{correct}}. \tag{7}$$

Here Acc_h denotes accuracy with history type h. The interaction measures dependence on history quality. It does not isolate truth value, because the drafts also differ in content and potentially other features.

Metrics and uncertainty. Accuracy is computed over all paired requests. Unresolved scores remain in the denominator and count as incorrect. We also report paired wrong-to-correct and correct-to-wrong transitions. The 95% percentile intervals use 50,000 bootstrap resamples of questions, keeping both seeds for each question together. These intervals describe variation within the selected question set and do not account for uncertainty from question selection or study-design choices. Termination and score coverage are reported in Appendix B.

Table 2: Game24 accuracy with two seeds per condition: 23 questions for Qwen3-8B and a shared 15-question subset for the other targets. The primary contrast is LateBlock versus Full. Effects are percentage points. Brackets are 95% intervals resampling questions. PathBlock and generation diagnostics appear in Appendix B.

Target	Clean	Full	LateBlock	Δ_{access} [95% CI]
Qwen3-8B	40/46	18/46	36/46	$+39.1$ [$+23.9, +54.3$]
Qwen3-32B	27/30	16/30	24/30	$+26.7$ [$+3.3, +50.0$]
Gemma 4 E4B	28/30	22/30	27/30	$+16.7$ [$0.0, +30.0$]

3 Empirical Evaluation of Continued Access

3.1 Failed reasoning remains harmful after explicit feedback

We first establish the prerequisite for interpreting an access effect as recovery. With Qwen3-8B, the failed-history prompt scores 18/46 versus 40/46 in clean context, despite the explicit incorrect-answer label and instruction to solve independently (Table 2). The observed drag is 47.8 percentage points (95% interval: 34.8 to 60.9). On the shared 15-question subset, Qwen3-32B and Gemma 4 E4B also show losses, respectively 36.7 points [16.7, 56.7] and 20.0 points [3.3, 33.3]. The comparison between Clean and Full establishes the phenomenon in these selected questions. It does not isolate the contribution of later access.

3.2 Blocking continued access recovers most of the observed loss

LateBlock preserves the full prompt and its prefill computation, yet raises Qwen3-8B accuracy from 18/46 to 36/46. The 39.1-point access effect [23.9, 54.3] recovers most of the 47.8-point aggregate Clean–Full gap. Paired outcomes clarify the change: 21 trials move from wrong to correct, three from correct to wrong, 15 remain correct, and seven remain wrong. Figure 3 shows how the gain is distributed across the selected questions: LateBlock yields more correct seeds on 15 questions, the same count on seven, and fewer on one. LateBlock is 8.7 points below Clean, with an interval of $[-23.9, 4.3]$ points. This is not an equivalence test or evidence of complete restoration. Excluding question 381, whose failed draft also contains a correct intermediate alternative, leaves a 38.6-point effect [22.7, 54.5] over 22 questions. The main conclusion is unchanged when excluding the single question admitted under an earlier eligibility rule (Appendix E).

Because the conditions agree through full prefill and the first sampled token, the recovery cannot be explained solely by avoiding initial exposure to the failed trace. Continued direct access therefore remains behaviorally consequential after full exposure. Figure 2 places this primary effect alongside the additional model, task, and correct-history comparisons.

3.3 The direction extends across target models and implementations

On the shared 15 questions and source drafts, Qwen3-8B gains 46.7 points [26.7, 66.7]. Qwen3-32B gains 26.7 points [3.3, 50.0] from LateBlock versus Full, with 11 improvements and three harms. Gemma 4 E4B gains 16.7 points [0.0, 30.0], with six improvements and one harm. The Gemma interval touches zero. These results show the same directional pattern with different magnitudes and uncertainty. The experiment is not designed to support conclusions about scaling with model size.

Qwen uses the vLLM/FlexAttention implementation, while Gemma uses a separate eager Hugging Face implementation that preserves its native hybrid attention. Observing the same direction in both provides a robustness check beyond one custom backend. Since architecture and implementation change together, it does not isolate backend effects or establish cross-backend equivalence. Generation termination and unresolved-score counts are reported in Appendix B.

3.4 Blocking can harm correct history

Does late blocking help because the retained history is wrong, or could disrupting attention to any prior attempt improve performance? To examine this distinction, we apply the same intervention

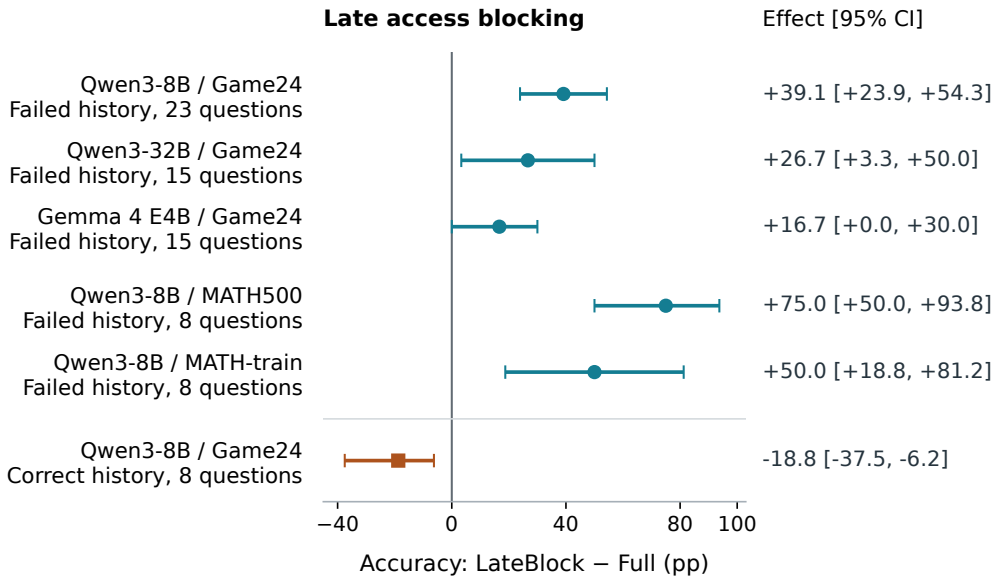

Figure 2: Accuracy change from late access blocking. Points show LateBlock minus Full accuracy in percentage points for the indicated history type. Bars are 95% percentile bootstrap intervals over questions, keeping the two seeds together. Qwen3-8B covers 23 Game24 questions. The other targets cover a shared 15-question subset, and the correct-history control covers eight questions. Each MATH row has eight questions and the Qwen3-8B target, with Qwen3-1.7B and Qwen3-4B source drafts for MATH500 and MATH-train, respectively. These rows are not a matched failed-versus-correct interaction. That comparison uses the same eight questions in Section 3.4. Unresolved scores count as incorrect.

to verified correct history on eight Game24 questions with the same Qwen3-8B target. Full access scores 16/16, while late blocking scores 13/16: a -18.8-point effect $[-37.5, -6.2]$. All three discordant pairs favor full access. Access to these correct drafts is useful in the observed sample, so blocking has no unconditional advantage.

The failed-history access effect on these *same eight questions* is $+37.5$ points. The matched difference in differences is therefore $+56.2$ points $[+37.5, +75.0]$, rather than a comparison with the full 23-question effect. One blocked correct-history output is length-truncated and unresolved. Scoring it as correct still yields a $+50.0$-point interaction. Correct and failed histories differ in content, may differ in length, and receive different truthful feedback labels. The interaction supports dependence on history quality, but it does not isolate truth value or rule out all generic effects of attention disruption.

3.5 The direction also appears in longer-form mathematical reasoning

MATH provides a cross-task check under longer-form reasoning (Table 3). Both source groups show the drag prerequisite $\mathrm{Acc}(\mathrm{Clean}) > \mathrm{Acc}(\mathrm{Full})$ and a positive access effect: 75.0 points [50.0, 93.8] for MATH500/Qwen3-1.7B drafts and 50.0 points [18.8, 81.2] for MATH-train/Qwen3-4B drafts. The groups differ simultaneously in source questions and draft-source model. Their contrast cannot identify either a source-model or draft-length mechanism.

Termination is less balanced than in Game24: of 128 MATH generations, 114 reach EOS, 14 stop at the length limit, and 13 scores are unresolved. MATH-train LateBlock and PathBlock have 4/16 and 6/16 length stops, respectively, versus none in Full. The contrasts thus concern scored success under the fixed generation budget and can reflect changes in both answer content and completion. They provide a directional cross-task check under longer-form reasoning, but do not establish a correctness effect independent of termination. Appendix B gives the condition-level counts.

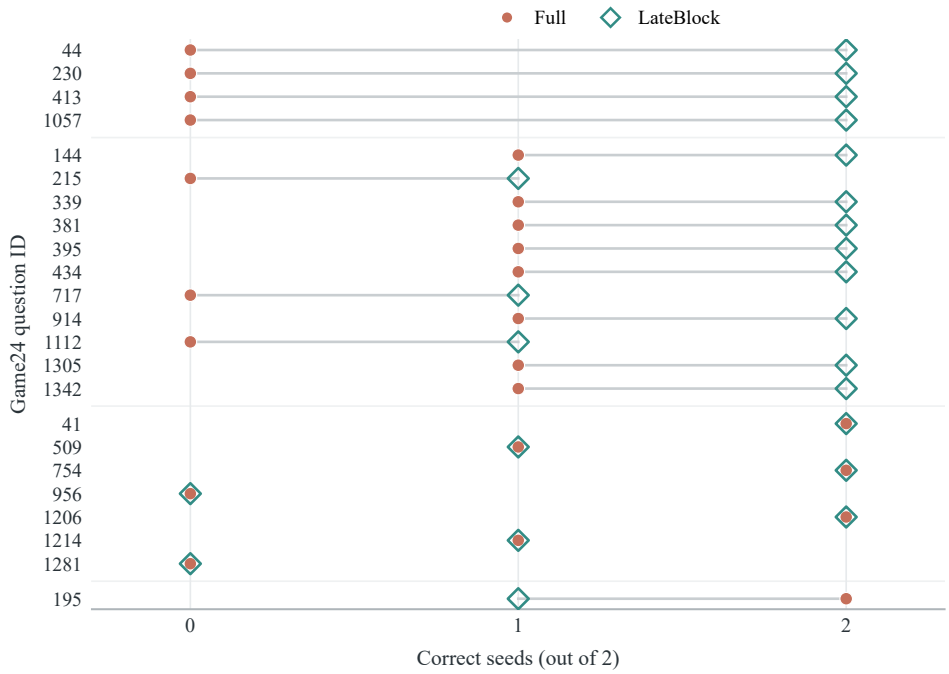

Figure 3: Question-level recovery from late access blocking on Game24 with Qwen3-8B. Each row shows one of the 23 selected questions. Points denote the number of correct generations out of two paired seeds under full access (Full) and late access blocking (LateBlock). Lines connect the paired conditions for the same question. Rows are ordered by transition type, from the largest gains to declines, and by question ID within each type. The aggregate Full-to-LateBlock effect is $+39.1$ percentage points. The question-level view shows how recovery is distributed across the selected set.

Table 3: Qwen3-8B target on MATH: eight questions per source group, two seeds per condition. Accuracy counts all requests. Unresolved scores count incorrect.

Source / draft model	Clean	Full	LateBlock	Δ_{access} [95% CI]
MATH500 / Qwen3-1.7B	16/16	2/16	14/16	$+75.0$ $[+50.0, +93.8]$
MATH-train / Qwen3-4B	7/16	0/16	8/16	$+50.0$ $[+18.8, +81.2]$

3.6 Earlier blocking provides no consistent additional benefit

PathBlock tests whether blocking post-draft prompt queries also improves on late blocking (Table 6, Appendix B). On Game24 with Qwen3-8B, it scores 35/46 versus LateBlock's 36/46, giving $\text{Acc}(\text{PathBlock}) - \text{Acc}(\text{LateBlock}) = -2.2$ points $[-17.4, +13.0]$. The corresponding MATH contrasts are $+6.2$ points $[-12.5, +25.0]$ and -12.5 points $[-43.8, +12.5]$. None shows a consistent additional benefit. These timing results constrain the conclusion: the observed recovery does not require the earlier intervention, and the data do not identify all mediated paths through which draft information may propagate.

4 Analysis

4.1 What the access contrast identifies

The comparison between Full and LateBlock shows that processing a failed trace does not fully determine its effect on the next solution. Even after identical prefill computation and the same first sampled token, changing later direct access changes accuracy. Continued direct access therefore

makes a causal contribution under the evaluated conditions. The error feedback is supplied externally, so this result concerns the model's response to known failure rather than its ability to recognize an error.

The clean baseline answers a complementary question. Comparing Clean with Full establishes a loss associated with the failed-history prompt, but changes both initial processing and later access. Comparing LateBlock with Full holds initial processing fixed and intervenes on later access. Their effect sizes are therefore not an additive decomposition of contextual drag into exposure and access components. Recovery of much of the clean-context gap does not measure the fraction of erroneous information removed from the model.

Across the 46 Qwen3-8B Game24 pairs, late blocking changes 21 incorrect outputs to correct ones and three correct outputs to incorrect ones. The remaining pairs contain 15 shared successes and seven shared failures (Table 7). Figure 3 shows their distribution across questions. The average benefit thus coexists with individual losses and persistent failures.

4.2 History dependence of the access effect

The correct-history comparison tests whether the benefit extends to useful prior reasoning. On the same eight questions, the failed-history access effect is $+37.5$ points, whereas the correct-history effect is -18.75 points. The resulting interaction is $+56.2$ points (Section 3.4). Comparing the same questions avoids attributing the reversal to a difference in question composition. The small control set nevertheless limits how precisely this dependence can be characterized.

Within each history type, Full and LateBlock receive the same feedback label. Across history types, both the draft and its truthful label change. The interaction therefore concerns the draft–feedback combination, rather than correctness as an isolated property. Matching drafts by character length does not equate their token counts, reasoning paths, or useful intermediate content. The reversal motivates selective access, but the final-answer label alone cannot locate every useful or harmful step in a draft.

An output-level audit on these eight questions provides a complementary behavioral measure. After normalization, the extracted final expression exactly repeats the frozen failed expression in 9/16 Full outputs, 2/16 LateBlock outputs, and 1/16 PathBlock outputs. Reduced reuse accompanies improved accuracy, connecting the access intervention to a visible change in the final answer. This measure does not classify algebraically equivalent expressions or shared reasoning routes. It therefore describes output persistence without establishing an internal copying mechanism or an estimate for all 23 questions.

4.3 Scope of the intervention

LateBlock preserves full prefill computation, so unblocked prompt states can carry draft-derived information into generation. Retaining these possible indirect routes makes the comparison informative about access after exposure. Once generation proceeds, the conditions can produce different tokens and states. The measured effect includes these downstream consequences of changing the access rule across layers and generation steps.

Blocking draft keys also changes the set of positions over which attention is normalized. The relative contribution of the problem, feedback, and generated prefix can therefore change even though their keys remain available. Improved accuracy may reflect reduced use of misleading draft information, increased use of other context, or both. These changes are consequences of the access intervention. The comparison identifies their combined effect on accuracy without separating their internal contributions. A matched irrelevant-history control could test whether the benefit depends on the draft's relationship to the current problem.

PathBlock additionally blocks post-draft prompt queries, changing the suffix computation during prefill as well as later access. Its comparison with LateBlock evaluates this earlier-starting access policy. Table 6 shows no consistent additional gain, although the intervals allow both benefits and harms. Recovery with LateBlock establishes that the earlier restriction is unnecessary for the observed improvement. The comparison does not establish equivalence between the policies or show that initial processing is irrelevant.

Selective-access experiments could compare useful and erroneous steps while controlling span length, position, and feedback. Spans would need to be selected without consulting later outcomes. Practical use would also require evaluating span selection, runtime, memory, and task performance.

5 Related Work

Refinement and contextual drag. Self-refinement and Reflexion retain earlier outputs or feedback to guide later attempts (Madaan et al., 2023; Shinn et al., 2023). Cheng et al. (2026) show that failed in-context attempts can lower performance, induce structurally related errors, and persist despite explicit verification. They also study denoising and targeted fine-tuning. We take contextual drag as the starting point and examine whether continued direct access contributes after the failed trajectory has already been processed.

Prior attempts across turns. Studies of multi-turn interaction report over-reliance on early answers and persistence of errors in later assistant responses (Laban et al., 2026; Huang et al., 2026). Those settings change the information revealed over time or filter what history is processed. Our same-problem comparison holds the prompt and initial exposure fixed while varying later direct access.

Context management and causal interventions. PagedAttention organizes KV-cache storage for efficient serving, while H_2O studies KV-cache eviction (Kwon et al., 2023; Zhang et al., 2023). MEMENTO studies how LLMs manage their textual context (Kontonis et al., 2026). Causal tracing, activation patching, and path patching intervene on internal computations to localize effects (Meng et al., 2022; Zhang & Nanda, 2024; Goldowsky-Dill et al., 2023; Dura et al., 2026). Our attention mask preserves prompt tokens, full prefill, and first-token computation while controlling later access to a designated prompt span. FlexAttention supplies the programmable mask interface (Dong et al., 2025). Our contribution is the controlled access comparison enabled by this interface.

6 Limitations

The 23 Game24 questions are selected for mixed baseline outcomes, and eight came from the completed portion of a screening run that ended early for runtime reasons, leaving ten candidates incomplete. The cutoff could therefore favor candidates with shorter or faster screening trajectories. The protocol also developed across collection batches, and the bootstrap intervals do not capture this selection or adaptation. These restrictions limit the population to which the reported effect sizes generalize. Broader coverage of independently selected questions is needed to characterize generalization beyond this selected sample.

The intervention additionally assumes externally verified incorrect drafts and known token spans. The additional models vary in both architecture and inference implementation, so their agreement supports robustness but does not explain model differences. Logical blocking leaves allocated KV tensors and information in unblocked states intact. MATH has condition-dependent termination and unresolved answers, and the correct-history subset does not match failed histories on every feature. The results therefore do not establish a sole information pathway, a scaling law, or a universally useful mitigation.

7 Conclusion

Continued direct access to failed reasoning can affect a new solution after full initial exposure and explicit error feedback. Our same-text intervention identifies this contribution while preserving prefill computation and the first sampled token. Late blocking recovers substantial accuracy on the selected Game24 set, with the same directional pattern in additional models and MATH settings. Blocking correct history instead harms performance, showing that the effect depends on the retained history and its feedback. These findings motivate studying when selective access can preserve useful prior work while limiting interference from failed reasoning.

Reproducibility statement

The paper specifies the model and inference configurations, random seeds, and scoring rules. Appendix A describes the intervention semantics, implementation checks, and prompt templates. The supplementary artifact contains implementation and scoring code, all 524 experimental records with infrastructure metadata removed, and an evidence manifest linking these records to the reported summaries. Its analysis scripts reconstruct the reported statistics and result figures, including question-level and seed-level tables.

AI use statement

Generative AI tools assisted with code development, debugging, discussions of experimental design, experiment documentation, literature discovery, result interpretation, and language editing. Numerical claims are tied to persisted results and auditable analysis scripts. The submitting authors are responsible for the final manuscript, its claims, and the associated artifacts.

References

Yun Cheng, Xingyu Zhu, Haoyu Zhao, and Sanjeev Arora. Contextual drag: How errors in the context affect LLM reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf?id=zpiYsPVDlV.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. In *Proceedings of the 8th MLSys Conference*, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/file/61a9278dfef5f871b5e472389f8d6fa1-Paper-Conference.pdf.

Murat Dura, Serkan Öztürk, and Selma Tekir. Mechanistic interpretability of chain-of-thought reasoning via sequential activation patching. *arXiv preprint arXiv:2608.22332*, 2026. doi: 10.48550/arXiv.2608.22332. URL https://arxiv.org/abs/2608.22332.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. doi: 10.48550/arXiv.2304.05969. URL https://arxiv.org/abs/2304.05969.

Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Jenny Y. Huang, Leshem Choshen, Wei Sun, Omar Khattab, Ramón Fernandez Astudillo, Mehul Damani, Tamara Broderick, and Jacob Andreas. Do LLMs benefit from their own words? In *Conference on Language Modeling (COLM 2026)*, 2026. URL https://colm.eventhosts.cc/Conferences/2026/AcceptedPapers. Accepted; to appear.

Vasilis Kontonis, Yuchen Zeng, Shivam Garg, Lingjiao Chen, Hao Tang, Ziyan Wang, Ahmed H. Awadallah, Eric Horvitz, John Langford, and Dimitris Papailiopoulos. MEMENTO: Teaching LLMs to manage their context. In *Conference on Language Modeling (COLM 2026)*, 2026. URL https://colm.eventhosts.cc/Conferences/2026/AcceptedPapers. Accepted; to appear.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59f6421e64707225fdf5b28840679a07-Abstract-Conference.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract.html.

A Intervention and Experimental Details

Condition names and artifact IDs. The display names Clean, Full, LateBlock, and PathBlock correspond to internal artifact IDs `R0`, `R2`, `R3`, and `R4`, respectively. These IDs are retained in code and persisted results. Unless explicitly labeled as a correct-history control, Full, LateBlock, and PathBlock use a failed draft with explicit error feedback.

A.1 Access semantics

Table 4 distinguishes KV allocation from attention access during prefill and decoding.

Table 4: Operational semantics of the access modes. "Allocated" means that the KV tensors remain present in memory. The intervention controls whether the corresponding logical attention edges participate.

Mode	Draft KV allocated	Prefill	Decode
Full	Yes	Ordinary causal	Ordinary causal
LateBlock	Yes	Ordinary causal	Draft keys blocked
PathBlock	Yes	Block after draft end	Draft keys blocked

LateBlock's first sampled token uses the same full-prefill logits as Full. In implementation terms, the first blocked query has logical position equal to the prompt length. PathBlock starts blocking at the end of the draft span and therefore also affects the prompt suffix following the draft. Neither mode physically compacts KV storage or renumbers token positions.

A.2 Implementation validation checklist

Before a scientific run, we verify:

1. equality of Full, LateBlock, and PathBlock prompt text, token IDs, length, positions, and hashes;
2. validity and minimum-span coverage of the persisted draft interval;
3. agreement of request-specific paged logical masks with a dense reference;
4. first-token agreement between full and late-block conditions in deterministic checks;
5. zero output change under mutations confined to keys blocked by the tested path mask in targeted greedy checks;
6. fail-closed behavior for missing metadata, unsupported layouts, or version mismatch.

A.3 Decoding, scoring, and the clean comparator

We use temperature 0.6, top-p 0.95, top-k 20, and a 32,768-token context limit. The paired Game24 seeds are 927001 and 927002. The MATH seeds are 930001 and 930002. All Clean outputs are generated for the reported comparisons, without substituting screening outputs. For MATH, the frozen scorer compares the last complete boxed answer with the reference answer.

Removing the complete failed-history block, including its label and delimiters, yields the Clean user content under the recorded chat wrapper. This equivalence was audited for all 32 MATH Clean requests.

A.4 Draft acquisition and leakage control

The 23 Game24 questions are distinct, and each has one frozen failed draft. Baseline screening determines draft correctness before intervention evaluation. Manual review distinguishes substantive failures from format-only errors, incomplete trajectories, and scorer failures. Twenty-two questions meet the final two-correct/two-failure rule. ID 41 was admitted under an earlier rule. Eight questions were retained from the completed portion of a larger screening run that ended early for runtime

reasons. The set was collected in successive batches and is treated as one selected sample, not as a preregistered replication. We transfer only source models' visible final responses, without private thinking traces. No intervention response is used to relabel its frozen source draft.

A.5 Statistical reporting

For each condition and contrast, we report raw numerators and denominators, accuracy, paired transitions, question-cluster bootstrap intervals, EOS rate, score coverage, and length termination. We resample questions, keeping both seed-level observations together. Effects are descriptive for the selected questions. Intervals are not adjusted for multiple contrasts and do not incorporate the uncertainty introduced by sample selection or study-design choices.

B Generation Diagnostics and Earlier Blocking

Table 5 reports all 524 completed scientific requests used in the paper. EOS and scoring coverage are distinct: one Gemma LateBlock output reaches EOS but has an unresolved score, whereas four length-stopped outputs across the full dataset have an extracted answer scored incorrect. All length-stopped outputs are either scored incorrect or unresolved. No records are dropped because of termination or parsing outcomes.

Table 5: All completed requests in the paper. Correctness uses persisted benchmark scores; unresolved answers count as incorrect. Termination is reported separately. MATH rows use the Qwen3-8B target.

Task / target	Condition	Correct	EOS	Length	Unresolved
Game24 / Qwen3-8B	Clean	40/46	44	2	1
	Full	18/46	46	0	0
	LateBlock	36/46	44	2	1
	PathBlock	35/46	45	1	0
Game24 / Qwen3-32B	Clean	27/30	30	0	0
	Full	16/30	30	0	0
	LateBlock	24/30	30	0	0
Game24 / Gemma 4 E4B	Clean	28/30	30	0	0
	Full	22/30	30	0	0
	LateBlock	27/30	30	0	1
MATH500 / Qwen3-8B	Clean	16/16	16	0	0
	Full	2/16	15	1	1
	LateBlock	14/16	15	1	1
	PathBlock	15/16	15	1	1
MATH-train / Qwen3-8B	Clean	7/16	15	1	1
	Full	0/16	16	0	0
	LateBlock	8/16	12	4	3
	PathBlock	6/16	10	6	6
Correct history / Qwen3-8B	Full	16/16	16	0	0
	LateBlock	13/16	15	1	1

Table 6 compares PathBlock with both Full and LateBlock, separating recovery under earlier blocking from its additional effect relative to late blocking.

Table 6: Earlier-blocking diagnostic with the Qwen3-8B target. The PathBlock column reports correct counts. We define $\Delta_{\text{path}} = \text{Acc}(\text{PathBlock}) - \text{Acc}(\text{Full})$ and $\Delta_{\text{early}} = \text{Acc}(\text{PathBlock}) - \text{Acc}(\text{LateBlock})$. Effects are percentage points with question-cluster 95% bootstrap intervals.

Task	PathBlock	Δ_{path} [95% CI]	Δ_{early} [95% CI]
Game24	35/46	$+37.0$ $[+21.7, +52.2]$	-2.2 $[-17.4, +13.0]$
MATH500	15/16	$+81.2$ $[+56.2, +100.0]$	$+6.2$ $[-12.5, +25.0]$
MATH-train	6/16	$+37.5$ $[+12.5, +75.0]$	-12.5 $[-43.8, +12.5]$

C Prompts

The prompt builder uses the following Game24 user-content template, with the two bracketed fields replaced by the frozen pair. Line wrapping of the final instruction below is typographic. The builder joins it into one line. The model-specific chat template surrounds this content.

```
Original problem:
[problem]

The following draft has an INCORRECT final answer.
-- beginning of the draft --
[draft]
-- end of the draft --

Solve the original problem independently and provide a
complete solution. You do not need to evaluate or refer
to any preceding draft. Conclude with an expression using
the original four numbers exactly once in \boxed{}.
```

For MATH, the last sentence is "Conclude with your final answer in \boxed{}." Clean omits the entire history block, including the incorrect-answer label and delimiters. The correct-history control changes the label to `CORRECT` and inserts a verified correct draft.

D Complete Question-Level Results

Table 7 gives all Game24 question-level counts. Full/LateBlock prompt hashes, input-token hashes, draft spans, and first generated tokens match for all 46 Qwen3-8B paired units and all 30 units per other target model. The Qwen3-8B Full/LateBlock transitions are 21 wrong-to-correct, three correct-to-wrong, 15 both correct, and seven both wrong. For Qwen3-32B they are 11, three, 13, and three. For Gemma 4 E4B they are six, one, 21, and two.

E Coverage and Draft-Content Sensitivity

Table 8 summarizes how the pooled access effect changes when either the mixed-content draft or the earlier-rule question is excluded.

The eight questions from the completed screening prefix contribute all 64 planned outputs, using fresh Clean generations and the same seeds, decoding, and scoring as the earlier evaluation. Clean, Full, LateBlock, and PathBlock score 15/16, 8/16, 12/16, and 13/16. Their LateBlock–Full effect is 25.0 points [0.0, 43.8], with five wrong-to-correct and one correct-to-wrong transition. Thus this extension agrees in direction but has a smaller and less precise effect than the shared 15-question subset. Its two length-stopped outputs are scored incorrect and remain in the denominators. All 64 extracted final expressions were checked with exact arithmetic and input-multiset verification, agreeing with their persisted scores.

Before generation, we specified a sensitivity analysis excluding question 381 because its fixed source draft contains a correct intermediate expression but explicitly selects an invalid final expression. Excluding it leaves Clean, Full, LateBlock, and PathBlock at 38/44, 17/44, 34/44, and 33/44 over 22

Table 7: Game24 correct counts per question. Each count is out of two seeds. Counts are ordered as Clean, Full, LateBlock, and PathBlock for Qwen3-8B (23 questions). The other models cover the shared 15-question subset and report the first three conditions in the same order. A dash denotes an unevaluated model–question combination.

Question	Qwen3-8B	Qwen3-32B	Gemma 4 E4B
41	2/2/2/2	2/1/2	2/2/2
44	2/0/2/2	1/2/2	2/1/2
144	2/1/2/2	—	—
195	2/2/1/2	—	—
215	2/0/1/1	—	—
230	2/0/2/2	2/1/1	2/2/2
339	2/1/2/2	2/2/2	2/2/2
381	2/1/2/2	—	—
395	2/1/2/2	—	—
413	1/0/2/1	2/0/1	2/1/0
434	2/1/2/2	—	—
509	2/1/1/1	2/1/2	2/1/2
717	1/0/1/1	2/2/1	2/2/2
754	2/2/2/2	2/1/2	2/1/2
914	2/1/2/2	2/1/1	2/2/2
956	2/0/0/1	1/1/1	2/1/2
1057	1/0/2/0	1/0/2	2/2/2
1112	2/0/1/2	2/0/2	2/2/2
1206	2/2/2/2	—	—
1214	1/1/1/2	2/1/2	2/1/2
1281	1/0/0/0	—	—
1305	1/1/2/0	2/1/2	1/0/1
1342	2/1/2/2	2/2/1	1/2/2

questions. The access effect remains 38.6 points [22.7, 54.5]. Within the eight-question extension alone, exclusion leaves 13/14, 7/14, 10/14, and 11/14, with an access effect of 21.4 points [-7.1, 42.9]. This smaller subset remains uncertain. The pooled conclusion does not depend on the mixed-content draft. No draft was replaced after observing intervention outcomes.

As a separate eligibility sensitivity, excluding question 41, which was admitted under an earlier rule, leaves 22 questions with Full/LateBlock accuracy of 16/44 and 34/44. The access effect is +40.9 points (95% question-cluster bootstrap interval: [+25.0,+56.8]), close to the pooled estimate. This confirms that the main conclusion does not depend on the single earlier-rule inclusion.

Table 8: Sensitivity of the Qwen3-8B Game24 access effect to question eligibility and draft content. Each question contributes two paired seeds. Intervals are 95% question-cluster percentile bootstrap intervals.

Analysis	Questions	Full	LateBlock	Δ_{access} [95% CI]
Pooled main sample	23	18/46	36/46	$+39.1$ [$+23.9, +54.3$]
Exclude question 381	22	17/44	34/44	$+38.6$ [$+22.7, +54.5$]
Exclude question 41	22	16/44	34/44	$+40.9$ [$+25.0, +56.8$]

F Data Availability

Public benchmark items retain their original licenses and provenance. Model weights are referenced by immutable model identifiers and are not redistributed.