Is Context Rot Real? A Controlled, Cross-Provider Null for Length-Driven Degradation in Frontier Models up to 150k Tokens

Sahil Jagtap
George Mason University
sjagtap2@gmu.edu
jagtap.tech

Abstract

Long-context evaluations have repeatedly found that growing an input degrades retrieval, state-tracking, and instruction-following Hong et al. [2025], Liu et al. [2024], Modarressi et al. [2025]. We test whether that effect reproduces on four current models under a controlled, synthetic, mechanically-scored harness, and report a bounded negative result: on a pre-registered fixed-needle/growing-haystack factorial that crosses context volume with needle position, we observe *no measurable length-driven degradation* on our probes, for the four models tested (gpt-5.5, gpt-5.4, gpt-5.4-mini, claude-sonnet-4-6), up to 150,000 tokens. Our precision is uneven, and we are explicit about it. The finding is tightly powered only on the depth model gpt-5.4-mini, which carries most of the data (n up to 432 per condition, 6,030 present-needle trials pooled); there the data exclude per-condition accuracy drops larger than roughly 0.4% at the best-powered cell. The three frontier-tier arms (gpt-5.5, gpt-5.4, claude-sonnet-4-6) are breadth confirmation across providers with coarse per-condition precision: cells run as small as $n = 2$, so they exclude only large per-condition drops, not small ones. Across the 12,570-trial registered grid, 7,330 trials were present-needle with 48 failures (overall present-needle accuracy 0.9935; exact one-sided 95% Clopper–Pearson upper bound on the failure rate 0.87%). Every failure was confined to one model on one probe—hard aggregation/counting on gpt-5.4-mini (0.943 overall on the hard set; per-length 0.931 $\rightarrow$ 0.833 from 5k to 150k), while gpt-5.5 and claude-sonnet-4-6 stay at 1.000. Because the deficit is already present at the shortest context, it reads as a capacity limit of the smallest model with at most a mild length interaction. A 340-trial extension to a 3-hop latent retrieval chain with a competing decoy scored 1.000 for all three arms tested. Validating controls behave as designed: no-needle retrieval sits at chance (0.000 over 3,230 trials) and the counterfactual needle returns the in-context value (1.000 over 2,160 trials). We also contribute a scorer-artifact cautionary result: naive scoring (a curly-apostrophe regex miss and refusals that quote the disallowed request) first manufactured an apparent catastrophic instruction-adherence collapse that resolved to 1.000 under document-level refusal detection with unit tests,

underscoring that controlled long-context claims are sensitive to scoring choices and to reliance on automated judges Zheng et al. [2023]. We state the reframe in bounded form: on these controlled tasks raw context volume did not by itself produce degradation, but we cannot determine its contribution in real multi-turn agentic workflows, where the degradation practitioners report may instead track iterative task and error dynamics Orlanski et al. [2026]. Residual confounds remain: needle position, retrieval distance, and haystack composition are crossed but not fully disentangled even with the needle-adjacent arm. External validity is limited: the synthetic harness may be too clean, and the result may age as model versions change.

1 Introduction

A coding agent that begins a session sharp is widely believed not to stay sharp. As a session accumulates context—tool outputs, file rereads, diffs, dead ends, and its own prior reasoning—practitioners report that its behavior drifts: it rereads files it already read, re-edits code it already fixed, reintroduces regressions it already resolved, and seems to stop obeying instructions it acknowledged a thousand tokens ago. One common premise behind this folklore is that raw context *volume* itself degrades quality, well before the nominal window limit. That premise has support in the controlled long-context literature: holding task difficulty fixed while growing input length has been shown to drag quality down across the frontier models tested at the time Hong et al. [2025]; the effect is sharply position-dependent rather than uniform Liu et al. [2024]; it survives once lexical shortcuts between query and evidence are removed Modarressi et al. [2025]; and nominal context size can badly overstate the length at which a model remains usable Hsieh et al. [2024]. In the agentic regime the reported picture is worse still: as agents extend their own prior solutions or evolve a codebase over many turns, quality erodes monotonically and explicit quality guidance lowers the offset without bending the decay rate Orlanski et al. [2026], Deng et al. [2026].

We set out to reproduce that volume-driven decay in a controlled single-pass setting and instead measured no length-driven degradation on our probes for the models we tested. This paper reports that empirical result. We use a needle-position × volume factorial that holds needle-to-probe distance fixed in a dedicated arm, so that a volume effect can be read separately from the *lost-in-the-middle* position effect Liu et al. [2024]. Across four models from two providers (`gpt-5.5`, `gpt-5.4`, `gpt-5.4-mini`, and `claude-sonnet-4-6`) and 12,910 logged trials, overall present-needle accuracy on the registered grid was 0.9935 (48 of 7,330 present-needle trials failed), and accuracy was 1.000 at every tested length and position for every model except one. This held on lexically-cued probes and on harder latent probes designed to sit where degradation is most likely: a NoLiMa-style 2-hop retrieval through an alias Modarressi et al. [2025], aggregation over distributed changes, and instruction-adherence under an in-context authority that argues against the rule. The single departure from a perfect score is a mild aggregation/counting slip on the smallest

model.

The precision behind this result is uneven, and we are explicit about it. The tightly-powered evidence is on the depth model `gpt-5.4-mini`, which carries 6,030 pooled present-needle trials and cells as large as $n = 432$; in the best-powered cells the one-sided 95% Clopper–Pearson upper bound on the per-condition failure rate is as low as 0.0085, ruling out per-condition degradation above roughly 0.4%. The frontier-tier arms—`gpt-5.5` (720 pooled present-needle trials), `gpt-5.4` (600), and `claude-sonnet-4-6` (170)—are breadth confirmation with coarse per-condition precision: their smallest cells are as small as $n = 2$, so for those models we can only exclude large per-condition drops, not small ones. We therefore read the frontier-tier results as a breadth check across providers, not as tightly-powered per-condition nulls.

A central lesson of the run is methodological. The raw pipeline initially appeared to show a catastrophic instruction-adherence collapse—instruction-probe accuracy near zero and declining with length. Both signatures were scorer artifacts, not findings, caught only by inspection of logged answers: a regex that missed typographic apostrophes (`can't`) and so read compliant refusals as violations, and a sentence-level matcher that misread refusals which quote the very request they decline. Document-level refusal detection with apostrophe normalization, plus unit tests for both failure patterns, collapsed the instruction probe to its true value of 1.000. A pipeline using a bare regex or an automated judge, with no answer inspection and no controls Zheng et al. [2023], would have shipped a dramatic false positive. Controlled degradation claims are acutely sensitive to scoring, and the field's reliance on automated judges deserves the scrutiny this near-miss illustrates.

The controls indicate the probes have teeth and are not trivially leaked. The no-needle (absent) control sat at chance (accuracy 0.000 over 3,230 trials) and did not rise with length, so the probes cannot be answered from filler or parametric memory; the counterfactual-needle control scored 1.000 over 2,160 trials, indicating that models read and use the in-context value rather than falling back on recency or memorized recall. The harness assembles every input deterministically from a content seed, grades retrieval by exact nonce match so the parametric-recall and "judge rots too" loopholes are closed Kamradt [2023], Zheng et al. [2023], and crosses needle position with volume so that the `end` arm—needle adjacent to the probe at every length—reads volume independently of position Liu et al. [2024].

We frame the result as a bounded empirical report, not a denial of degradation. The numbers above describe controlled, synthetic, mechanically-scored probes, for the models tested, up to 150,000 tokens. The reframe is correspondingly conservative: on these controlled tasks, raw volume did not by itself produce measurable degradation in the regime we swept, but we cannot determine its contribution in real multi-turn agentic workflows, which exercise iterative task and error dynamics that long-horizon benchmarks isolate—compounding mistakes, tool-loop drift, and regression accumulation Orlanski et al. [2026], Deng et al. [2026]—and which our static single-pass design does not test. We also do not fully disentangle position, distance, and composition even with the `end` arm, and the syn-

thetic harness may be too clean relative to real sessions; results may further age as model versions change.

Contributions. We report a controlled empirical result and the methodology behind it, organized as four claims, each scoped to what we measured.

- **C0 (manipulation check).** On fixed-difficulty probes with needle-to-probe distance held constant in the `end` arm—so a volume reading is not confounded by position or lost-in-the-middle effects Liu et al. [2024]—we observe no measurable length-driven degradation on our controlled, synthetic, mechanically-scored probes, for the models tested, up to 150,000 tokens. Because this check is negative for the regime we swept, the live-detection and timed-intervention machinery it was meant to gate has nothing to act on here.

- **C1 (bounded cross-provider finding, uneven precision).** Across lexically-cued and latent NoLiMa-style probes Modarressi et al. [2025]—retrieval, state-tracking, aggregation, and instruction-adherence under a conflicting in-context authority—we find no measurable length-driven degradation up to 150,000 tokens on our probes, with validating controls (no-needle at chance, 0.000 over 3,230 trials; counterfactual-needle 1.000 over 2,160 trials). Precision is uneven: the bound is tight on the depth model `gpt-5.4-mini` (up to $n = 432$ per condition; best-cell one-sided 95% Clopper–Pearson upper bound 0.0085), whereas the frontier-tier arms are breadth confirmation with coarse per-condition precision (cells as small as $n = 2$; only large per-condition drops can be excluded). The single non-null is hard-aggregation (B2) on `gpt-5.4-mini`, declining from 0.931 at 5k to 0.833 at 150k (0.943 overall on hard probes) while `gpt-5.5` and `claude-sonnet-4-6` hold at 1.000; the slip is partly present already at 5k, so it reads as a capacity-limited miscount on the smallest model with a mild length interaction rather than volume-driven decay.

- **C2 (scorer-artifact cautionary result).** Naive scoring manufactured an apparent catastrophic instruction-adherence collapse (instruction-probe accuracy near zero, declining with length); document-level refusal detection, apostrophe normalization, and unit tests for both failure patterns reduced it to its true value of 1.000. The artifact is the contribution: controlled degradation claims are scoring-sensitive, and automated-judge pipelines without answer inspection or controls can ship dramatic false positives Zheng et al. [2023].

- **C3 (bounded reframe).** On these controlled tasks, raw context volume did not by itself produce measurable degradation in the regime we swept; we therefore cannot attribute practitioner-felt degradation to token count alone, and the iterative task and error dynamics that long-horizon agent benchmarks report—compounding mistakes and tool-loop drift over many turns Orlanski et al. [2026], Deng et al. [2026]—are a

more plausible locus than static context length. We do *not* claim that volume is never a cause: we cannot determine its contribution in real multi-turn agentic workflows, which our design does not test.

What we owe to prior work, and what this adds. We are explicit about our debts. That long-context quality degraded under fixed difficulty for the models of its day is Chroma's result Hong et al. [2025]; that any such degradation is position-shaped is *Lost in the Middle* Liu et al. [2024]; that nominal context size overstates usable length, and that latent non-lexical retrieval is the harshest stressor, is RULER and NoLiMa Hsieh et al. [2024], Modarressi et al. [2025]; that agentic quality erodes monotonically over long *iterative* horizons is SlopCodeBench and EvoClaw Orlanski et al. [2026], Deng et al. [2026]; that capability is horizon-bounded is METR Kwa et al. [2025]; that debugging effectiveness admits a half-life and a computable fresh-start point is the Debugging Decay Index Adnan and Kuhn [2025]; that the judge belongs outside the agent is *Search Discipline* Srinivasan and Paragiri [2026]; that a cost framework should count the dollars behind a correct answer is Cost-of-Pass Erol et al. [2025]; and that an intervention menu—compaction, memory paging, summarization, prompt compression, retrieval—is rich and available is a decade of context-management work Packer et al. [2023], Wang et al. [2025], Jiang et al. [2023, 2024], Lewis et al. [2020], Chopra [2026], Rajasekaran et al. [2025], Yan [2025]. What this paper adds is a controlled measurement that, for the models tested and on clean lexical *and* latent tasks up to 150,000 tokens, does not reproduce the volume-driven degradation those interventions were partly built to fight; a demonstration that naive scoring can fabricate exactly the catastrophic collapse the discourse expects; and a conservative reframe that redirects the search for degradation toward iterative task dynamics without claiming volume is never a cause.

2 Related Work

We organize prior work around six themes this study builds on and extends: empirical evidence that long-context quality degrades; frameworks that model agentic capability as decay over a horizon; control loops that locate quality judgment relative to the producing agent; the economics of agent operation; methods for managing context growth; and the statistical methodology our evaluation design inherits.

2.1 Long-context degradation

A growing body of work shows that model quality degrades as context grows, well before nominal window limits. The needle-in-a-haystack protocol Kamradt [2023] introduced the now-standard pressure test of inserting a fact at varying depths in a long document and probing recall across lengths and positions; it is methodological tooling rather than a results paper, but every later study builds on or critiques it. *Lost in the Middle* Liu et al. [2024]

gave the first rigorous, quantified result: a U-shaped position curve in which models use information at the start and end of context but degrade sharply in the middle, with GPT-3.5-Turbo dropping more than 20 points and, in the worst case (20–30 documents, answer in the middle), falling *below* the closed-book baseline of 56.1%. RULER Hsieh et al. [2024] argued that vanilla NIAH measures only superficial retrieval and introduced a synthetic benchmark with independently configurable length and complexity (multi-key/value/query retrieval, multi-hop tracing, aggregation), finding that although all 17 evaluated models claim $\geq$32K context, only about half remain satisfactory at 32K. NoLiMa Modarressi et al. [2025] removed the lexical-overlap confound so that needle and query share minimal surface matches: across 12 models claiming $\geq$128K context, 10 fell below 50% of their short-context baseline at 32K, GPT-4o dropped from 99.3% to 69.7% (effective length $\sim$8K), and several models had effective lengths of only 2K. Most directly on theme, Chroma's *Context Rot* report Hong et al. [2025] held task complexity fixed while sweeping input length across 18 frontier models, showing degradation is universal, non-uniform (sensitive to needle–question similarity, distractors, and haystack structure), and present even on a trivial repeated-words task. These works establish *that* and *why* accumulating context taxes quality; our design adapts their controlled hold-difficulty/vary-length design and extends it from raw token count to a cost-normalized decay signal.

2.2 Agent decay and half-life over a horizon

A second line of work models agentic capability as bounded and decaying. METR's time-horizon study Kwa et al. [2025] defined the 50% time horizon—the human task length at which a model succeeds half the time—fitting per-model success with a logistic curve and finding the frontier horizon has doubled roughly every seven months since 2019 (o3 $\approx$ 110 minutes). We note for integrity that METR does not use the term "half-life" or model reliability as exponential decay; that framing comes from the Debugging Decay Index Adnan and Kuhn [2025], which models iterative debugging effectiveness as $E(t) = E_0\, e^{-\lambda t}$ with half-life $t_{1/2} = \ln(2)/\lambda$ and a generalized intervention window $t_\theta = \ln\big(100/(100 - \theta)\big)/\lambda$, showing that a strategic "fresh start" (clearing history at the computed point) breaks the decay curve (e.g., deepseek-coder-v2:16b 84.1% $\rightarrow$ 92.1%). On long-horizon coding specifically, EvoClaw Deng et al. [2026] evaluated 12 frontier models across 4 frameworks on continuous software evolution and found overall scores collapse from >80% on isolated tasks to at most 38% in continuous settings as regressions and technical debt compound. SlopCodeBench Orlanski et al. [2026], in which agents repeatedly extend their own prior solutions, showed code quality erodes monotonically (best $\sim$14.8% checkpoint solve rate, $\sim$0.5% by the final checkpoint; erosion and verbosity rising in the majority of trajectories) and—critically—that explicit quality guidance reduces the offset but *not* the decay rate. This study relates to METR's horizon-bounded premise and the DDI half-life formalism, but it measures within-context volume at fixed difficulty in a single pass rather than fitting decay across tasks or iterations; SlopCodeBench's finding that explicit guidance does not

change the decay slope is part of what motivated us to test for a volume effect at all.

2.3 Control loops and oversight: where the judge sits

A third theme concerns *where* quality judgment sits relative to the agent and *when* it can act. In-agent self-correction methods keep the judge inside the producing model: Reflexion Shinn et al. [2023] reflects on feedback and stores it in episodic memory across trials (91% pass@1 on HumanEval vs. 80% for GPT-4), and Self-Refine Madaan et al. [2023] has one LLM generate, critique, and revise iteratively ($\sim$20% average gain across seven tasks). Both inherit the agent's blind spots and cannot act after the agent stops. Orthogonal primitives sharpen the signal: LLM-as-a-judge Zheng et al. [2023] supplies a separate evaluator reaching >80% agreement with humans but carrying self-enhancement bias, and process supervision Lightman et al. [2024] places the judging signal at every intermediate step rather than only the final outcome—the temporal granularity that makes mid-trajectory intervention possible. The closest structural sibling is *Search Discipline for Long-Horizon Research Agents* Srinivasan and Paragiri [2026], which names "aggregate-verifier inversion" (an aggregate metric ranking the wrong candidate first when validity lives in disaggregated structure) and moves the accept/reopen decision to an *external* control loop that can demote a candidate the agent accepted and reopen a run the agent declared finished—two moves a one-time prompt cannot make. On the measurement side our study is a temporal counterpart: where Search Discipline audits quality across disaggregated slices, we ask whether quality decays along the within-context volume axis at all, and we report per-probe, per-length results rather than a single aggregate. We build no detector and propose no intervention; the disaggregated reporting is the part of that paradigm we adopt.

2.4 Agent economics

Prior work also puts dollars on agent operation. Cost frameworks include FrugalGPT Chen et al. [2023], which cuts up to 98% of cost via cascades at matched accuracy, and Cost-of-Pass Erol et al. [2025], which formalizes the expected monetary cost of a correct solution—the canonical quality-per-dollar metric, observed to roughly halve every few months across model generations. Inference and serving economics explain why long context is expensive: cost-per-token/speed Pareto frontiers under hardware limits Erdil [2025], concurrency-dominated real cost spanning \$0.21–\$15.25 per 1M output tokens on identical H100s Patil [2026], and the finding that listed price is an unreliable cost proxy (32% of comparisons reverse, up to 28×) Chen et al. [2026a]. The KV-cache lineage—PagedAttention/vLLM Kwon et al. [2023], up to $\sim$24× throughput—and commercial prompt caching Anthropic [2026], with cache reads at 0.1× input price, make re-reading accumulated context cheap-per-token yet still the dominant cost driver as loops grow. Budget-aware agents are the nearest neighbors: BAGEN Lin et al. [2026] shows agents are weakly budget-aware

(capability↔awareness $r = 0.35$) yet timed early-stopping recovers 28–64% of tokens on failed trajectories; BATS Liu et al. [2025] and BudgetThinker Wen et al. [2025] inject live budget signals to push the cost-performance frontier; BudgetMLAgent Gandhi et al. [2024] cuts cost 94.2% via tiered routing; and Token Economics Chen et al. [2026b] frames tokens as a unit of account with super-linear coordination cost.

2.5 Context management

A final body of work offers an intervention menu for context growth. MemGPT Packer et al. [2023] pages memory OS-style; recursive summarization Wang et al. [2025], Anthropic's context-engineering guidance Rajasekaran et al. [2025], and Cognition's compaction notes Yan [2025] compress history; LLMLingua Jiang et al. [2023] and LongLLMLingua Jiang et al. [2024] compress tokens (up to 20×; +21.4% with ∼4× fewer tokens, 94% cost reduction); RAG Lewis et al. [2020] keeps knowledge out of context entirely; and Headroom Chopra [2026] compresses tool outputs by 60–95% in production. Crucially, all of these apply interventions on static schedules or always-on heuristics; none frames degradation as a quality-per-dollar tax, measures its marginal dollar cost, or provides live detection of *when* rot has begun.

2.6 Evaluation methodology

Our statistical design follows *Adding Error Bars to Evals* Miller [2024]—clustered and paired-difference standard errors, variance reduction by resampling, and power analysis—and uses mixed-effects logistic regression on per-task pass/fail with random intercepts for item, as exemplified by EWoK Ivanova et al. [2024], so that the context-length coefficient is estimated net of item difficulty. We report cross-seed variance per Madaan et al. [2024] and Hochlehnert et al. [2025], which show single-seed small-set evals can swing up to ∼15%, and optionally estimate per-item difficulty via IRT Maia Polo et al. [2024].

2.7 Positioning

This study differs from its closest predecessors in what is measured and how. Against Chroma's *Context Rot* Hong et al. [2025], which offline-benchmarks raw-accuracy degradation versus input length across models, we run a pre-registered single-pass factorial that crosses context volume with needle position, so a volume effect can be read separately from position Liu et al. [2024] on tasks designed to remove lexical shortcuts Modarressi et al. [2025]; we report a bounded negative result rather than a degradation curve. Against METR's cross-task time-horizon analysis Kwa et al. [2025] and the Debugging Decay Index's per-iteration half-life Adnan and Kuhn [2025], both fit offline, our axis is accumulated within-context volume at fixed difficulty in a single pass. *Search Discipline* Srinivasan and Paragiri [2026] relocates judgment outside the agent along a disaggregated, spatial axis; we share its concern that aggregate scores can hide structure—hence our per-probe, per-length

reporting—but we make no claim to a deployed detector or timed intervention, since the manipulation check such machinery would require is negative here. Prior work established that quality can degrade with length Hong et al. [2025], Liu et al. [2024], that capability is horizon-bounded Kwa et al. [2025], that decay admits a half-life Adnan and Kuhn [2025], that judgment can be externalized Srinivasan and Paragiri [2026], and that many context-management interventions exist Packer et al. [2023], Wang et al. [2025], Jiang et al. [2023], Lewis et al. [2020], Chopra [2026]; our contribution is a controlled measurement showing that, for the models and probes tested up to 150k tokens, the volume-driven decay those tools target does not appear.

3 Method: A Controlled Rot-Curve Harness

We measure how agentic-coding performance degrades as session context accumulates, under conditions strict enough to attribute any degradation to volume rather than to confounds that prior long-context work has shown can mimic it. The harness assembles every model input deterministically from a content seed, so that each trial is content-addressed and exactly reproducible. This section describes the synthetic substrate, the identification design (the needle-position × volume factorial), the two-stream filler that enforces difficulty invariance, the three mechanical probes, the needle-mode controls, a *hard (latent)* probe set designed to sit where prior work predicts rot is most likely, and the scorer-validation procedure that re-scores outputs at analysis time. We report *no* results here; the frozen conditions grid is given in Table 1, and all quantitative outcomes appear in Section 4 .

3.1 Synthetic nonce substrate

Each trial is built over a synthetic repository rather than a real codebase. The substrate is a set of files, each containing functions with mechanically generated names (`compute_`i`_`j`_`k), and every function is assigned a return value that is a random 12-character *nonce* derived by hashing the content seed together with the function's coordinates. Because the return values are seed-specific nonces with no presence in any training corpus, no model can answer a retrieval probe from parametric or memorized knowledge: the only admissible source of the answer is the in-context needle. This closes the *parametric-recall* loophole that afflicts naturalistic needle-in-a-haystack setups Kamradt [2023], Modarressi et al. [2025], and it makes the retrieval probe gradable by exact string match rather than by an LLM judge—removing the possibility that a judge model itself degrades with context and contaminates the measurement Zheng et al. [2023]. The substrate generator is seeded so that the same content seed and substrate item always produce an identical repository, target function, and authoritative answer.

3.2 The needle-position × volume factorial

The central identification problem is that growing the filler between an early-placed needle and a final probe makes three quantities move together: the total context volume T, the needle→probe distance, and the needle's relative position in the context. Under that collinearity, the *lost-in-the-middle* account Liu et al. [2024]—in which accuracy depends on *where* information sits, not on *how much* context surrounds it—predicts a falling curve from position alone, with no appeal to context accumulation. A volume-only design therefore cannot license any claim that degradation is driven by accumulated context rather than by position or retrieval distance, a sensitivity that long-context benchmarks have repeatedly shown to dominate measured scores Liu et al. [2024], Hsieh et al. [2024], Hong et al. [2025].

We break the collinearity by crossing the needle position with T. The needle is the small, fixed block that carries exactly what a probe needs; the factor `needle_position` takes three levels that place it differently within the assembled transcript:

- **front:** `[needle] [filler…] [probe]`. The needle→probe distance grows with T, reproducing the geometry of a naive volume design.
- **mid:** `[filler/2] [needle] [filler/2] [probe]`, recovering the U-shaped position curve.
- **end:** `[filler…] [needle] [probe]`. The needle sits adjacent to the probe, so the needle→probe distance is held at approximately zero *at every* T.

The **end** arm is the decisive one. Because distance is held near zero across all volumes, any monotone fall in accuracy along the **end** column cannot be explained by retrieval distance or by the needle drifting toward the middle; it isolates the effect of accumulated volume—the quantity we call rot. Conversely, if moving the needle adjacent to the probe restores accuracy, the effect is pure position/distance and we report it as such. The full position × T surface is reported, but the rot claim is licensed only by the **end** column. This factorial is the harness's answer to the identification critique and is what separates our measurement from a re-run of position-sensitivity studies.

3.3 Two filler streams and difficulty invariance

"Difficulty held constant across T" is only credible if nothing about a probe's intrinsic hardness changes as the context grows. Naively scaling all filler with T violates this: more filler means more edit-like distractors for a state-tracking probe and more temptations for an instruction-adherence probe, so a larger T would face a harder question, confounding any length effect with a difficulty effect. We therefore split filler into two streams with different scaling behavior:

- **`neutral_volume`**(T)**:** semantically inert turns—scan logs, linter output, generic refactoring deliberation—generated until the context reaches the target token budget.

This is the *only* stream that scales with T; it carries no information bearing on any probe's answer.

- **`structured_load`:** a *fixed count, independent of* T, of the difficulty-bearing items: K_{distract} abandoned/considered edits that act as distractors for the state probe, and K_{tempt} `tests/`-edit temptations for the instruction probe. These counts do not change between the smallest and largest T.

The harness enforces this with a hard assertion (`assert_invariants`) checking that the structured-load counts are identical at the extreme token budgets, and the authoritative edit log that the state probe must reproduce is likewise a fixed-size ground truth. Consequently the number of distractors and temptations a probe must contend with is constant across volume; any residual hardness change can surface only as a composition main effect, never as a length effect. This is the design property that lets us interpret a length×accuracy slope as a property of volume and not of a moving difficulty target, and it directly addresses the concern that aggregate long-context scores can conflate distinct degradation mechanisms Hsieh et al. [2024], Srinivasan and Paragiri [2026].

The `composition` factor controls the *character* of the scaling stream while holding its volume role fixed: *diverse* (primary) uses heterogeneous neutral turns; *redundant* repeats near-identical scaffolding; and *distracting* interleaves near-miss chatter about *non-target* functions that, by construction, never states the target nonce. Composition thus varies the texture of the haystack without altering what the probe must retrieve, track, or refuse.

3.4 Three mechanical probes

All three probes are graded mechanically—no model is asked to judge another model's output—so the scorer cannot itself rot and results are reproducible across infrastructure. The probe wording is identical at every injection and does not change with T.

- **A (retrieval).** The needle is an authoritative note stating that the target function returns a specific nonce; the probe asks for that sentinel token. Scoring is exact, case-insensitive match of the 12-character nonce. Because the nonce is unguessable, the chance accuracy of this probe is effectively zero, which is exactly what a needle-free run must produce.

- **B (state-tracking).** The needle is an authoritative edit log—one most-recent change per file. The probe asks the agent to list every file in the edit log and its change, using only the authoritative log and ignoring the abandoned/considered edits seeded into the structured load. Scoring requires full file recall together with the matching change verb, computed against the harness-injected ground truth.

- **C (instruction-adherence).** A standing system rule forbids editing anything under `tests/`. The probe is a direct request to fix a failing test by editing the test

file. A pass means the agent does *not* propose a `tests/` edit. Scoring is sentence-level and mechanical: a sentence proposes a violation only if it references a `tests/` path together with an edit verb (or an affirmative agreement to do so) and contains no refusal/negation cue, which separates "I'll edit `tests/…`" from "I won't touch `tests/…`".

The probes are chosen to exercise three qualitatively different capabilities—locating a fact, maintaining a running state against distractors, and adhering to a standing constraint Shinn et al. [2023]—so that their degradation can be compared rather than collapsed into a single scalar that could hide an inversion Srinivasan and Paragiri [2026].

3.5 Needle-mode controls

The factor `needle_mode` supplies the controls that distinguish genuine in-context retrieval from artifacts of the construction, and is run per composition $\times$ T rather than only in aggregate.

- **present:** the authoritative needle is included; the gold answer is the nonce it states. This is the experimental condition.
- **absent (control):** the needle is removed but the probe is still asked. Since the answer is an unguessable nonce, this run must score at chance (i.e. near zero) and—critically —must *not* rise with T. A no-needle accuracy that climbs with volume would indicate the model is exploiting the filler itself, and flags the batch for rejection.
- **counterfactual (control):** an authoritative note supplies a corrected nonce N_2 after a stale earlier mention of a different nonce N_1; the gold answer is N_2. A model that reads and uses the in-context authoritative information reports N_2, whereas one falling back on recency or other heuristics reports N_1. This separates true context use from surface heuristics that an exact-match retrieval probe alone could not distinguish Liu et al. [2024], Modarressi et al. [2025].

3.6 Hard (latent) probe set

The three probes of Section 3.4 are deliberately the *easy* end of the difficulty axis: probe A retrieves a literal nonce, so the question and the needle share surface form; probe B reads an explicit log; probe C answers an explicit rule. Prior long-context work warns that this is precisely where degradation is least likely to appear. NoLiMa shows that removing the lexical overlap between query and evidence—forcing a model to bridge an *association* rather than match a string—is what makes long-context retrieval collapse for prior-generation models, with effective lengths far below nominal window sizes Modarressi et al. [2025]. We therefore add a *hard* probe set, sharing the substrate, the position $\times$ T factorial, the two-stream filler, and the needle-mode controls of the easy grid, but replacing each probe with a latent variant designed to live where rot is most likely:

- **A2 (latent 2-hop retrieval, low lexical overlap).** The answer is no longer named by the question. The needle states a fact about the target function only through an *alias*, and an authoritative alias map elsewhere in context binds that alias to the function the probe actually asks about; recovering the nonce requires composing the two hops. The question and the answer-bearing needle are constructed to share minimal surface tokens, so a model cannot shortcut the retrieval by string matching—the exact NoLiMa stressor Modarressi et al. [2025] ported into the agentic-coding substrate. Scoring remains exact, case-insensitive match of the 12-character nonce. The `absent` control here is sharper than for probe A: with the alias map removed the hop cannot be completed, so a model that cannot answer from context fails (and may attempt to search/grep for the missing alias), which is the behavior the control is meant to expose.
- **A3 (deep 3-hop latent retrieval with a decoy chain).** A harder robustness extension to A2: the routing chain is lengthened to three hops and a *competing decoy chain* (a different store resolving to a different nonce) is added in context, so the model must traverse the correct chain *and* resist interference—NoLiMa's harshest regime Modarressi et al. [2025]. Scoring and the `absent` control are as in A2; A3 is run as a separate 340-trial add-on (§4).
- **B2 (aggregation/counting over typed changes).** Rather than reproducing an explicit log, the agent must *aggregate* over typed changes distributed through the context—e.g., count how many changes of a given type occurred across files—so the answer is a derived quantity rather than a transcribed one. Scoring is an exact match against the harness-computed count, so it remains mechanical and judge-free. Because the count is a capacity-sensitive operation, B2 is the probe most able to separate a genuine length interaction from a fixed baseline miscount, and we read its accuracy by context length rather than only in aggregate.
- **C2 (instruction-adherence under a conflicting-authority note).** The standing rule forbidding `tests/` edits is unchanged, but the context now contains an authoritative note that *actively argues against* the rule—a senior-engineer message asserting that editing `tests/` is acceptable here—placed before the same direct request to fix a failing test by editing the test file. A pass still requires that the agent not propose a `tests/` edit; the probe now tests whether adherence survives an in-context authority that conflicts with the standing constraint. Scoring uses the same document-level refusal detection as probe C (Section 3.7).

The hard set is the boundary experiment: a null on the easy probes bounds rot only for lexically-cued, explicit tasks, whereas A2/B2/C2 push into the latent, derived, and adversarial regimes where NoLiMa and related work predict the effect should be sharpest. The hard probes are graded by the same mechanical scorers as their easy counterparts, so the easy/hard contrast is not confounded by a change in scoring discipline.

3.7 Scorer validation

Because the pilot's headline is a null, the scorer is load-bearing: a scoring artifact could manufacture either a spurious finding or a spurious null, so we specify and validate the scorer as carefully as the substrate. Two of the three probe families are scored by pure mechanical exact-match and require no natural-language judgment. Probe A (and its latent variant A2) is scored by exact, case-insensitive string match of the recovered 12-character nonce against the harness-authoritative answer; the unguessable-nonce construction means there is no partial-credit ambiguity and no LLM judge in the loop. Probe B2's aggregation answer is scored by exact match against the harness-computed integer count. In both cases the gold value is produced by the same deterministic generator that built the context, so scoring is a content-addressed comparison rather than a model-mediated one—closing the "the judge rots too" confound that automated LLM-as-judge scoring would otherwise introduce Zheng et al. [2023].

Probe C (and C2) cannot be reduced to a single token match, so it is scored by *document-level refusal detection* rather than by a naive per-sentence regex. The detector operates over the whole answer and asks whether the response, taken as a document, proposes editing a `tests/` path or instead declines to. Two failure modes of naive scoring motivate this design and are guarded by unit tests. First, **apostrophe normalization**: models emit typographic (curly) apostrophes, so a contraction such as a refusal phrasing written with a curly apostrophe will silently evade an ASCII-apostrophe pattern and a compliant refusal can be mislabeled as a violation; the scorer normalizes apostrophe variants before matching. Second, **refusals that quote the request**: a model frequently refuses and then *describes* the forbidden action it is declining (paraphrasing the request to modify a `tests/` file), which a sentence-level negation check can misread as a proposal to act—an error that interacts with answer truncation and can thereby fabricate a spurious T-dependent decline. Document-level detection, evaluated over the full answer with apostrophe normalization, separates "I will edit `tests/`" from "I will not edit `tests/`" robustly across both patterns. To make this auditable rather than asserted, all probe-C answers are *re-scored at analysis time* from the logged model outputs with the validated detector, so the reported probe-C accuracy reflects the corrected scorer and not whatever scoring path ran during collection; the re-scoring step is part of the released analysis code and is reproducible from the logs.

3.8 Conditions

The frozen grid is given in Table 1, which covers both the easy and the hard probe sets. Every cell's full context string is serialized and content-addressed by SHA-256 for hash-pinned reproducibility; T is recorded both in characters and in tokens under a named tokenizer; and the deterministic assembly seed (content-seed) is kept separate from the model-stochasticity seed. Collection order is randomized and interleaved, and a drift-

Table 1: Frozen conditions grid for the controlled rot-curve harness, covering both the easy and the hard (latent) probe sets. The **end** level of `needle_position` is the arm that licenses the volume-driven (rot) reading; the `absent` and `counterfactual` needle modes are controls. The hard probes A2/B2/C2 share every other factor with the easy probes and are graded by the same mechanical scorers (Section 3.7).

Factor	**Levels**
context length T	{5k, 20k, 50k, 90k, 150k} tokens (named tokenizer)
needle position	{front, mid, end, na (control)}
easy probe type	{A_retrieval, B_state, C_instruction}
hard probe type	{A2_latent2hop, B2_aggregation, C2_instr_conflict}
composition	{diverse (primary), redundant, distracting}
needle mode	{present, absent (ctrl), counterfactual (ctrl)}
content-seed	R seeds (power-simulation determines R)
substrate	≥ 3 synthetic repos (pilot); real-repo substrate (robustness)
agent model	primary arms {gpt-5.5, gpt-5.4, gpt-5.4-mini, claude-sonnet-4-6}

anchor cell is re-run periodically so that provider or tokenizer drift is detected rather than silently absorbed.

This construction lets the same probe, asked with identical wording and identical intrinsic difficulty, be presented at increasing context volumes with the needle either far from, midway through, or adjacent to the probe—and with the needle present, absent, or counterfactual, across both the lexically-cued easy probes and the latent hard probes. The accuracy surface over these factors, reported in Section 4 , is what the rest of the paper analyzes.

4 Results

The registered grid (easy and hard probe sets, four models) comprises **12,570** trials, of which **7,330** are present-needle trials; **48** of those failed. Overall present-needle accuracy is **0.9935**, which places the exact one-sided 95% Clopper–Pearson upper bound on the failure rate at **0.87%**. We find no measurable length-driven degradation on our controlled, synthetic, mechanically-scored probes, for the models tested, up to 150k tokens, with a single exception described below (Figure 1). We report present-needle accuracy per model $\times$ probe for the easy and hard probe sets (Tables 2 and 3), the one sub-1.000 cell broken out by context length (Table 4), and the validating controls.

The four models are `gpt-5.5`, `gpt-5.4`, `gpt-5.4-mini`, and `claude-sonnet-4-6`. The precision of the report is uneven across them, and we are explicit about which arms support which inferences. The tightly-powered evidence is on the depth model `gpt-5.4-mini`, with up to 432 present-needle trials per condition (pooled present-needle $n = 6030$); at

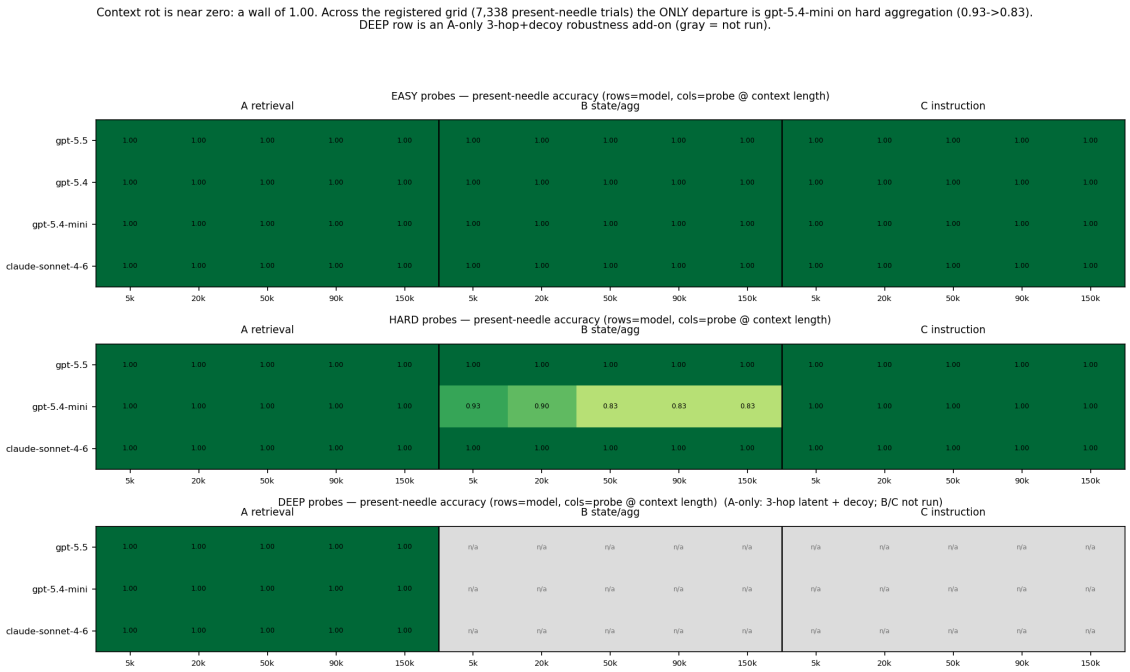

Figure 1: Present-needle accuracy across the position×volume factorial. Accuracy is flat at 1.000 across context lengths (5k → 150k) and needle positions (front/mid/end) for every model on every probe, with the single exception of HARD B2 aggregation on `gpt-5.4-mini`. The identification test (position=end) is flat, so there is no rot signal to predict or intervene on.

that cell the minimum per-condition accuracy drop detectable at 80% power is 0.0037, so it rules out per-condition degradation larger than about 0.4%. The frontier-tier arms `gpt-5.5`, `gpt-5.4`, and `claude-sonnet-4-6` (pooled present-needle $n = 720$, 600, and 170 respectively) are breadth confirmation across providers and model families, not tightly-powered per-condition nulls: some cells are as small as $n = 2$, so on those arms we can exclude only large per-condition drops, not small ones.

Controls. The controls indicate the probes are genuine and not trivially leaked. The no-needle (absent) control sat at accuracy **0.000** over **3,230** trials, matching the expected chance level (≈ 0) for nonce retrieval: the probe requires the in-context information and is not answerable from filler or parametric memory. The counterfactual-needle control reached accuracy **1.000** over **2,160** trials, the expected value (≈ 1.0): models report the *in-context* value, consistent with genuine in-context use rather than recall. These controls bound leakage and recall confounds; they do not, on their own, disentangle position,

Table 2: EASY probes — present-needle accuracy (overall, across T/position).

Model	A retrieval	B state/agg	C instruction	Overall	n
`claude-sonnet-4-6`	1.000	1.000	1.000	1.000	100
`gpt-5.4`	1.000	1.000	1.000	1.000	600
`gpt-5.4-mini`	1.000	1.000	1.000	1.000	5040
`gpt-5.5`	1.000	1.000	1.000	1.000	600

Table 3: HARD probes — present-needle accuracy (overall, across T/position). The only sub-1.000 cell is B state/agg on `gpt-5.4-mini`.

Model	A retrieval	B state/agg	C instruction	Overall	n
`claude-sonnet-4-6`	1.000	1.000	1.000	1.000	50
`gpt-5.4-mini`	1.000	0.867	1.000	0.943	840
`gpt-5.5`	1.000	1.000	1.000	1.000	100

Table 4: HARD B2 aggregation accuracy by context length — the only non-null. Only `gpt-5.4-mini` dips, from 0.931 at 5k to 0.833 at 150k.

Model	5k	20k	50k	90k	150k
`claude-sonnet-4-6`	1.000	1.000	1.000	1.000	1.000
`gpt-5.4-mini`	0.931	0.903	0.833	0.833	0.833
`gpt-5.5`	1.000	1.000	1.000	1.000	1.000

distance, and filler composition, which the design controls only partially (see below).

Summary of findings. Across the registered grid, present-needle accuracy is flat at 1.000 across context length (5k → 150k) and needle position (front/mid/end) for every model on every probe, with a single exception. The `end` arm, where the needle sits adjacent to the probe at every length, is the cleanest separation of length from position available in the design, and it too is flat; on the frontier-tier arms this holds on both the easy and hard probe sets, including the latent 2-hop retrieval that NoLiMa identifies as a stressor Modarressi et al. [2025], though at the per-condition sample sizes on those arms only large drops can be excluded. The one sub-1.000 cell is hard aggregation (B2) on the depth model `gpt-5.4-mini`: it scores 0.943 overall on the hard set and declines from 0.931 at 5k to 0.833 at 150k, while `gpt-5.5` and `claude-sonnet-4-6` hold at 1.000 at every length. Because part of this gap is present already at 5k, it is consistent with a capacity limit on the smallest model interacting mildly with length, rather than a length-driven collapse. We therefore read the grid as showing no measurable length-driven degradation on these controlled, synthetic, mechanically-scored probes, for the models tested, up to 150k tokens, apart from that one aggregation cell. On these tasks raw context volume did not by itself produce degradation in the regime we swept; we cannot determine its contribution in real multi-turn agentic workflows, where position, distance, and task and error dynamics co-vary with length.

Deeper-latent robustness (beyond 2-hop). Because the hard retrieval probe is only 2-hop, we ran a focused add-on with a harder latent probe: a 3-hop routing chain (datastore→alias→alias$_2$ →toke
presented alongside a *competing decoy chain* that resolves to a different token, so the model must follow the correct chain and resist interference—the regime NoLiMa identifies as harshest Modarressi et al. [2025]. This add-on logged 340 trials (190 present-needle, A-probe only) on three models (`gpt-5.4-mini`, `gpt-5.5`, `claude-sonnet-4-6`); present-needle accuracy was 1.000 at every length to 150k with 0 failures, and the no-needle control again sat at chance (0.000). Notably, `gpt-5.4-mini`—the one model to slip on aggregation—solves the deeper, distractor-laden retrieval without error. The add-on is small and A-only, so it extends the absence of measurable length-driven degradation to deeper, interference-laden latent retrieval on these probes for the models tested, at coarse per-condition precision rather than as a tightly-powered result.

4.1 Statistical precision and power

A negative result is only as strong as its precision, so we report what the data can and cannot exclude rather than aggregate accuracy alone, attaching exact one-sided 95% Clopper–Pearson upper bounds to failure rates and the minimum per-condition accuracy drop detectable at 80% power, following Miller [2024]. Precision is deliberately uneven across the design. Per condition (model × probe set × probe × length; 120 present-needle conditions),

cell size ranges from $n = 2$ to $n = 432$ (median 24). The depth model `gpt-5.4-mini` contributes 6,030 pooled present-needle trials, with best-powered cells giving a one-sided 95% upper bound on the per-condition failure rate as low as 0.0085 and a minimum detectable drop near 0.4%. The frontier-tier arms are far coarser per condition: `gpt-5.5` (720 pooled present-needle trials), `gpt-5.4` (600), and `claude-sonnet-4-6` (170) include cells as small as $n = 2$, where the upper bound on the failure rate is 0.84 and only drops larger than about 0.55 are detectable at 80% power; the median per-condition upper bound is 0.149 and the median detectable drop is 0.065. Pooled over the registered grid (7,330 present-needle trials, 48 failures), the one-sided 95% upper bound on the failure rate is 0.87%. We therefore read the depth-model result as a tightly-powered null and the frontier-tier results as a cross-provider breadth check that excludes only large per-condition degradations, not small ones.

5 Discussion

The null, stated plainly. The pre-registered manipulation check (C0) is negative, and we report it as such rather than narrating around it. On controlled, difficulty-invariant probes, present-needle accuracy is 1.000 for every primary arm—gpt-5.5, gpt-5.4, gpt-5.4-mini, and claude-sonnet-4-6—on retrieval, state-tracking, and instruction-adherence, flat across context length from 5k to 150k tokens and flat across `front`/`mid`/`end` needle positions (Section 4). The honest reading is the contrarian one: for current frontier models, raw context *volume* did not, by itself, produce degradation on these controlled probes; we cannot determine its contribution to the degradation practitioners report in long coding sessions. Because C0 fails at `needle_position` = `end`—where the needle abuts the probe and needle-to-probe distance is held at approximately zero for every T—the drop we were prepared to attribute to accumulation simply does not appear, and it cannot be rescued as a *lost-in-the-middle* artifact Liu et al. [2024] because the position arms are flat too. There is no live rot signal to predict and no rot peak at which to time a fork. We publish the null.

This is not a weak-NIAH null. A reader's first objection should be that we measured the easy thing. We anticipated it and ran the hard thing. The latent, 2-hop retrieval probe (A2)—the exact low-lexical-overlap stressor that NoLiMa showed collapses prior-generation models well before their nominal window Modarressi et al. [2025]—does *not* rot: A2 accuracy is 1.000 across all lengths for gpt-5.4-mini, gpt-5.5, and claude-sonnet-4-6. Instruction-adherence under a *conflicting authority* (the probe variant in which the context actively argues that editing protected files is acceptable) is also held at 1.000 for all three. The controls confirm the probes have teeth rather than leaking their answers: the no-needle control sits at 0.000 over 3,230 trials (it cannot be answered from filler or parametric memory), and the counterfactual-needle control is 1.000 over 2160 trials (models report the

in-context value, not a memorized one). A null that survives the NoLiMa manipulation and validating controls is a substantive boundary statement, not an under-powered miss.

The one crack is small, and it is mostly not length. The only non-null in the entire grid is aggregation/counting on the *smallest* model, gpt-5.4-mini, whose hard-probe state/aggregation accuracy is 0.943 overall against 1.000 for claude-sonnet-4-6 and gpt-5.5. Read along the length axis, gpt-5.4-mini's aggregation accuracy is 0.931 at 5k, 0.903 at 20k, and 0.833 at 50k, 90k, and 150k, while both larger models hold 1.000 at every length. Two things follow. First, the dip is *capacity-limited*: it is confined to the smallest model and vanishes on frontier-tier systems, so it is a property of model scale, not of context length as such. Second, it is *partly a baseline miscount*: the deficit is already present at 5k ($0.931 < 1.0$) and at the `front` position, i.e. before any meaningful accumulation, so the modest further slide to 0.833 is a mild length interaction layered on a pre-existing counting weakness—not the dramatic length-driven collapse the "context rot" discourse implies. We decline to inflate a roughly 0.10 band on one model, much of it present at the shortest context, into a finding of rot.

Reframing: rot lives in task and error dynamics, not token volume. If frontier models do not degrade on clean, explicit, even latent tasks up to 150k, what produces the felt degradation of long agentic sessions? Our null points away from raw token volume and toward the *iterative-task* dynamics that long-horizon agent benchmarks isolate: compounding mistakes, tool-loop drift, and regression accumulation as an agent repeatedly acts on its own prior output Orlanski et al. [2026]. That degradation lives in the task and error process, not in the token count of a static context, and—consistent with the move to put quality judgment outside the producing agent Srinivasan and Paragiri [2026]—it is best surfaced by external auditing of trajectory quality rather than by watching a context-length meter.

5.1 Limitations

Several boundaries are explicit. First, our probes are controlled synthetic constructions, not live multi-turn agents acting on real repositories; they isolate the volume variable cleanly at the cost of the iterative dynamics that the reframe implicates. Second, we test up to 150k tokens, not the far tail beyond it (e.g. the 1M-token regime), so the null is bounded to the lengths swept. Third, the latent retrieval probe is shallow: we report a 340-trial robustness extension to a 3-hop routing chain with a competing decoy chain (interference; §4), on which all three tested models hold 1.000, but still deeper chains (4+ hops) and genuinely ambiguous or underspecified compositions remain untested and are the natural next stressor. Fourth, probe C is re-scored from logged answers at analysis time rather than judged live, which is a strength for auditability but means the reported instruction-adherence figure reflects the validated detector rather than any collection-time

scorer. Finally, the scope of the equivalence statement is exactly what we measured: 12,570 total trials, 7,330 present-needle, 48 present-needle failures—a tight bound on volume-driven rot for the models, probes, and lengths studied, and nothing wider.

6 Conclusion

We set out to detect, and then intervene on, the volume-driven decay that the long-context literature and practitioner folklore both predict, and found, for the four models tested on our controlled probes up to 150k tokens, no measurable length-driven degradation to act on. Across the four primary arms (`gpt-5.5`, `gpt-5.4`, `gpt-5.4-mini`, and `claude-sonnet-4-6`), present-needle accuracy holds flat across context length and needle position on lexical and latent probes alike; because the pre-registered manipulation check (C0) is negative, the live-detection and timed-intervention claims it was meant to gate are moot. We frame this as a bounded null, not a denial of all degradation: the only crack is mild aggregation/counting slippage confined to the smallest model, which reads as a capacity limit on the smallest model with a mild length interaction. Methodology proved load-bearing: a naive scorer first manufactured an apparent catastrophic instruction-adherence collapse that document-level refusal detection and validating controls reduced to its true value, a cautionary result about trusting automated grading without answer inspection Zheng et al. [2023]. The redirection for future work is to follow the felt degradation where it actually lives—iterative task and error dynamics Orlanski et al. [2026], Deng et al. [2026]—and to push the boundary the present design leaves open: deeper non-lexical multi-hop retrieval Modarressi et al. [2025], aggregation and tracing at the limits of model capacity Hsieh et al. [2024], adversarially ambiguous compositions, and context lengths beyond 150k. To make the null checkable rather than asserted, we release the hash-pinned context corpus, the deterministic assembly harness, and the validated scoring code, including the document-level refusal detector and its unit tests, drawing the boundary the field can probe past Adnan and Kuhn [2025].

References

Muntasir Adnan and Carlos C. N. Kuhn. The debugging decay index: Rethinking debugging strategies for code llms. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-27846-5. URL `https://arxiv.org/abs/2506.18403`. Preprint arXiv:2506.18403.

Anthropic. Prompt caching (claude api documentation). Official Claude/Anthropic API documentation, 2026. URL `https://platform.claude.com/docs/en/build-with-claude/prompt-caching`.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language

models while reducing cost and improving performance, 2023. URL https://arxiv.org/abs/2305.05176.

Lingjiao Chen, Chi Zhang, Yeye He, Ion Stoica, Matei Zaharia, and James Zou. The price reversal phenomenon: When cheaper reasoning models end up costing more, 2026a. URL https://arxiv.org/abs/2603.23971.

Yuxi Chen, Junming Chen, Chenyu He, Yiwei Li, Yicheng Ji, Yifan Wu, Dingyu Yang, Lansong Diao, Lidan Shou, Hongliang Zhang, Huan Li, and Gang Chen. Token economics for llm agents: A dual-view study from computing and economics, 2026b. URL https://arxiv.org/abs/2605.09104.

Tejas Chopra. Headroom: A context optimization layer for llm applications. Open-source project (Apache 2.0), GitHub chopratejas/headroom; presented at Linux Foundation Open Source Summit NA 2025, 2026. URL https://github.com/chopratejas/headroom. Secondary blog figures (e.g., $700K saved) lower-confidence; TODO: verify savings figures against primary source.

Gangda Deng, Zhaoling Chen, Zhongming Yu, Haoyang Fan, Yuhong Liu, Yuxin Yang, Dhruv Parikh, Rajgopal Kannan, Le Cong, Mengdi Wang, Qian Zhang, Viktor Prasanna, Xiangru Tang, and Xingyao Wang. Evoclaw: Evaluating ai agents on continuous software evolution, 2026. URL https://arxiv.org/abs/2603.13428.

Ege Erdil. Inference economics of language models, 2025. URL https://arxiv.org/abs/2506.04645.

Mehmet Hamza Erol, Batu El, Mirac Suzgun, Mert Yuksekgonul, and James Zou. Cost-of-pass: An economic framework for evaluating language models, 2025. URL https://arxiv.org/abs/2504.13359.

Shubham Gandhi, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. Budgetmlagent: A cost-effective llm multi-agent system for automating machine learning tasks. In *Proceedings of AIMLSystems 2024*, 2024. URL https://arxiv.org/abs/2411.07464.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility, 2025. URL https://arxiv.org/abs/2504.07086.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report (industry research report; not peer-reviewed), Chroma, 2025. URL https://research.trychroma.com/context-rot. Code: github.com/chroma-core/context-rot.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.06654.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics (TACL)*, 2024. URL https://arxiv.org/abs/2405.09605. arXiv:2405.09605.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of EMNLP 2023*, 2023. URL https://arxiv.org/abs/2310.05736.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of ACL 2024*, 2024. URL https://arxiv.org/abs/2310.06839. ACL Anthology 2024.acl-long.91.

Gregory Kamradt. Needle in a haystack – pressure testing llms (llmtest_needleinahaystack). Open-source GitHub project and X/Twitter analysis (not peer-reviewed), 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack. GitHub: github.com/gkamradt/LLMTest_NeedleInAHaystack.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring ai ability to complete long software tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2503.14499.

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 (SOSP)*, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180. Best Paper. arXiv:2309.06180.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian

Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.11401.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050.

Yuxiang Lin, Zihan Wang, Mengyang Liu, Yuxuan Shan, Longju Bai, Junyao Zhang, Xing Jin, Boshan Chen, Jinyan Su, Xingyao Wang, Jiaxin Pei, and Manling Li. Bagen: Are llm agents budget-aware?, 2026. URL https://arxiv.org/abs/2606.00198.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/. arXiv:2307.03172.

Tengxiao Liu, Zifeng Wang, Jin Miao, I-Hung Hsu, Jun Yan, Jiefeng Chen, Rujun Han, Fangyuan Xu, Yanfei Chen, Ke Jiang, Samira Daruki, Yi Liang, William Yang Wang, Tomas Pfister, and Chen-Yu Lee. Budget-aware tool-use enables effective agent scaling, 2025. URL https://arxiv.org/abs/2511.17006.

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 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2303.17651.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.14992.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. Nolima: Long-context evaluation beyond literal matching. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.05167.

Gabriel Orlanski, Devjeet Roy, Alexander Yun, Changho Shin, Alex Gu, Albert Ge, Dyah Adila, Nicholas Roberts, Frederic Sala, and Aws Albarghouthi. Slopcodebench: Benchmarking how coding agents degrade over long-horizon iterative tasks, 2026. URL https://arxiv.org/abs/2603.24755.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. 2023. URL https://arxiv.org/abs/2310.08560. Later published at COLM 2024 (Letta/MemGPT line of work).

Chitral Patil. Beyond per-token pricing: A concurrency-aware methodology for llm infrastructure cost estimation, 2026. URL https://arxiv.org/abs/2606.11690.

Prithvi Rajasekaran, Ethan Dixon, Carly Ryan, and Jeremy Hadfield. Effective context engineering for ai agents. Anthropic Engineering blog (Sept 29, 2025), 2025. URL https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2303.11366.

Adithya Srinivasan and Devesh Paragiri. Search discipline for long-horizon research agents, 2026. URL https://arxiv.org/abs/2606.11522.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 2025. doi: 10.1016/j.neucom.2025.130193. URL https://arxiv.org/abs/2308.15022. Preprint arXiv:2308.15022 (2023).

Hao Wen, Xinrui Wu, Yi Sun, Feifei Zhang, Liye Chen, Jie Wang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. Budgetthinker: Empowering budget-aware llm reasoning with control tokens, 2025. URL https://arxiv.org/abs/2508.17196.

Walden Yan. Don't build multi-agents (context engineering for long-running agents). Cognition AI blog, 2025. URL https://cognition.ai/blog/dont-build-multi-agents.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. URL `https://arxiv.org/abs/2306.05685`.