Lookup Is Not Attribution: Regime-Dependent Citation Stability Under Context Scaling and Multi-Turn Interaction

Anonymous ACL submission

Abstract

Long-context language models are increasingly used in bibliography-intensive workflows, yet their ability to faithfully reproduce an *explicitly provided* citation is poorly understood as a function of context size, conversational depth, and evaluation regime. We introduce **CitationDrift-16K**, a metadata-only arXiv benchmark of 16,384 papers across 16 CS/ML categories, and evaluate ten models spanning five OpenAI-hosted (GPT-4.1, GPT-4o, GPT-4o-mini, GPT-5.5, GPT-5-mini) and five DeepInfra-hosted open models (Qwen-4B, Mistral-24B, Nemotron-49B, Qwen-0.8B, Llama-3.1-8B) under two carefully contrasted task regimes.

Our central finding is **regime-dependent**: under an enumerated *indexed lookup* protocol, most models achieve near-ceiling arXiv-ID accuracy up to 128 references and 10 turns; yet the same model under a *no-index, title-based stress* regime drops from 100% to **1.1%** *exact-one* accuracy at identical reference-set size—solely from removing the positional index cue. Extending this finding to Nemotron-49B ($56\% \rightarrow 0\%$) and Mistral-24B ($88\% \rightarrow 98\%$, an *improvement*) reveals that regime sensitivity is model-dependent: some architectures rely on positional transduction, others on semantic attribution. We further show that naive substring-based scoring inflates measured accuracy by up to **18 percentage points** (Nemotron-49B, $\texttt{rc} = 128$), demonstrating that evaluation design can materially alter model rankings. We release the benchmark, evaluator, and all experiment scripts as a resource for long-context attribution research.

1 Introduction

Long-context language models are increasingly deployed in workflows that require faithful grounding to a provided bibliography: literature review assistance, technical writing, multi-turn scientific Q&A, and agentic research pipelines (Lewis et al., 2020; Liu et al., 2023b). In these settings, attribution failures can take multiple forms—incorrect identifiers, cross-paper contamination, or merged references—yet they are rarely studied under controlled conditions that isolate the contribution of context load, interaction depth, and evaluation methodology (Agrawal et al., 2024; Min et al., 2023).

We study *attribution stability* as a function of (i) reference-set size, (ii) conversational depth, and (iii) the task regime—a distinction we show to be decisive (Figure 1).

Key findings.

- **Lookup fidelity $\neq$ attribution fidelity.** Llama-3.1-8B drops from 100% to 1.1% exact-one accuracy at $\texttt{rc} = 128$ solely from removing the index cue. Nemotron-49B collapses from 56% to 0%; Mistral-24B *improves* from 88% to 98% (Table 3)—regime sensitivity is model-dependent.
- **Evaluation leakage is consequential.** Substring scoring overestimates Nemotron-49B by 18 pp (74% vs. 56%) at $\texttt{rc} = 128$, turn 1, and can reverse model rankings.
- **No-index failures are dominated by output suppression.** Llama-3.1-8B emits no identifier in 98.9% of turn-1 responses, recovering to 66.2% missing by turn 10 through repeated prompting.

Scope. We do *not* claim temporal drift (model behaviour changes over time); our runs do not include time-separated replications with fixed model snapshots. Instead, we characterise *within-run stability regimes* and provide a reusable evaluation framework. All results are reproducible from logged JSONL artifacts using the released scripts.

Contributions. **(1) CitationDrift-16K**: 16,384-paper metadata-only arXiv benchmark with powers-

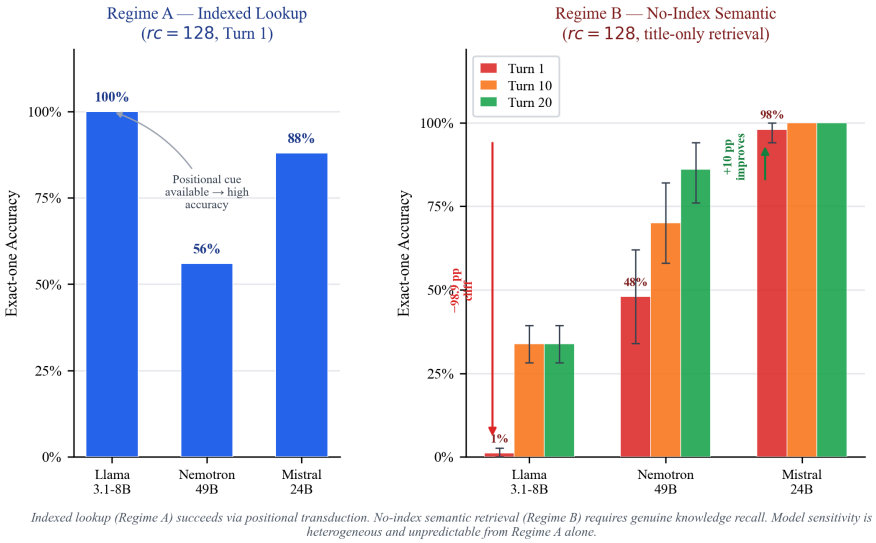

Figure 1: **Lookup is not attribution: the two evaluation regimes (Figure 1). Left — Indexed Lookup Regime** *(positional transduction):* The model receives a fully enumerated reference block (`[1]. . . [rc ]`) and is queried by position (*"Return reference [2]"*). The retrieval pathway collapses to locate-position → copy-identifier → output — a near-deterministic operation yielding ∼100% exact-one accuracy in most models at `rc` $\leq$ 128. **Centre — Regime Boundary** *(phase transition):* Removing the positional index cue transforms attribution from a copy operation into a semantic binding problem. *Lookup fidelity* $\neq$ *semantic attribution fidelity.* **Right — No-Index Semantic Attribution Regime** *(semantic binding under interference):* The model must resolve the target arXiv ID from title/abstract cues alone, competing against 127 semantically similar references. Failure modes are missing-ID (output suppression), wrong-ID (competitor binding), and contamination ($\geq$2 IDs emitted). Llama-3.1-8B drops from 100% to **1.1%** exact-one accuracy at turn 1, Nemotron-49B from 56% to 0% — solely from removing the index cue. The bottom banner states the paper's core claim: *high citation fidelity under indexed lookup does not imply robust semantic attribution under realistic retrieval conditions.*

of-two reference sets (`rc` up to 1,024), semantic similarity groups, and unrelated distractor groups. **(2) Two-regime protocol**: indexed lookup vs. no-index semantic stress, with a fixed-target invariant across turns. **(3) Leakage-resistant evaluator**: strict extraction-based exact-one metric and contamination rate, exposing an 18 pp substring-scoring artefact for Nemotron-49B. **(4) Multi-model regime map**: heterogeneous indexed→no-index sensitivity across Llama-3.1-8B, Nemotron-49B, and Mistral-24B, showing that indexed-regime performance does not predict no-index performance.

2 Related Work

Long-context evaluation. Long-context evaluation has advanced from synthetic retrieval tests ("needle in a haystack") to multi-task suites varying context length and task type (An et al., 2023; Bai et al., 2023; Shaham et al., 2023; Hsieh et al., 2024; Zhang et al., 2024b). Holistic evaluation frameworks such as HELM (Liang et al., 2022) additionally benchmark reliability, efficiency, and robustness across many tasks. Positional and retrieval effects—including the "lost in the middle" phenomenon (Liu et al., 2023a) and length-extrapolation via position-interpolation methods (Press et al., 2022)—motivate our explicit manipulation of target position within the reference block. Controllable benchmarks (Yang et al., 2025b), sequential variants (Yu et al., 2025; Moon and Lim, 2026), multilingual extensions (Kim et al., 2025), and multimodal settings (Hengle et al., 2024; Ataallah et al., 2025; Song et al., 2024) further expand the space of tested conditions. Our benchmark contributes an evaluation target—arXiv identifiers—with near-perfect extraction precision that avoids scoring ambiguities common in fuzzy matching baselines.

Citation fidelity and hallucination. Language models store and sometimes confabulate factual knowledge (Petroni et al., 2019), and hallucinated references serve as an interpretable model organism of such errors (Agrawal et al., 2024; Ji et al., 2023). Methods exist to induce citation-supported generation (Gao et al., 2023; Asai et al., 2023), and

faithfulness under retrieval-augmented conditions has been studied across summarisation (Maynez et al., 2020) and open-domain QA (Kamalloo et al., 2023). Our task differs: we measure how faithfully a model reproduces a citation it has been explicitly given, under increasing load and interaction depth. RAG faithfulness metrics (Es et al., 2023) and broader hallucination detection methods (McKenna et al., 2023; Manakul et al., 2023; Miao et al., 2023; Mu et al., 2023; Lin et al., 2021; Min et al., 2023; Honovich et al., 2022) provide methodological context; our strict extraction metric is a citation-specific instantiation of faithfulness evaluation.

Multi-turn consistency. Multi-turn interaction introduces temporal dynamics absent from single-turn evaluations (Kwan et al., 2024; Anonymous, 2026a; Liu et al., 2023b; Anonymous, 2026b, 2025). Our fixed-target invariant (identical paper requested each turn) isolates interaction-depth effects from the confound of shifting reference targets.

3 Benchmark: CitationDrift-16K

Dataset construction. We query the arXiv legacy API across 16 categories spanning computer science and machine learning: `cs.CL`, `cs.AI`, `cs.LG`, `cs.CV`, `cs.RO`, `cs.IR`, `cs.HC`, `cs.SE`, `cs.CR`, `cs.NE`, `cs.DS`, `cs.GT`, `cs.SY`, `cs.NA`, `eess.SP`, and `stat.ML`, collecting 1,024 papers per category (**16,384 papers total**). For each paper we store: arXiv identifier, title, authors, abstract, category labels, and timestamps. No PDFs are downloaded or redistributed.

Reference formatting. Each paper is rendered as a formatted reference string with stable 1-indexed numbering, fixed across all runs and providers to eliminate formatting as a confound. ArXiv identifiers are normalised by stripping version suffixes (e.g., `2310.01234v2`→`2310.01234`) before evaluation.

Sampling sets and interference groups. We precompute powers-of-two reference sets ($\texttt{rc} \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$) and two families of interference groups via embedding-based clustering (sentence-transformers/all-MiniLM-L6-v2): **semantically similar groups** (91 groups, high pairwise cosine similarity) and **unrelated groups** (124 groups, low pairwise similarity).

Why arXiv identifiers. An arXiv ID is a deterministic key: correctness requires no fuzzy scoring (no ROUGE, no BERTScore) (Modarressi et al., 2025; Yang et al., 2025a). Our primary metric (*exact-one* accuracy) reduces to a single regex match; prior long-context benchmarks relying on semantic-similarity scoring introduce evaluation noise that can obscure subtle model differences (Li et al., 2023; Wang et al., 2024).

4 Experimental Protocol

4.1 Experiment 1: Indexed Lookup

Task. The model is given a numbered reference block (`[1] Title. Authors. arXiv:YYYY.NNNNN (year)....`) and asked: "Return the exact metadata for reference $[k]$." This is an interpretable upper bound: it maximally favours the model by providing a direct positional cue and the target identifier verbatim.

Conditions. We vary $\texttt{rc} \in \{1, 16, 128\}$ in the main text (results up to $\texttt{rc} = 1024$ in §6.4). Turns are evaluated at $\{1, 3, 5, 10\}$, with the same target $[k]$ fixed across all turns per trial (**fixed-target invariant**). Three prompt regimes are tested in parallel: **neutral**, **citation-aware**, and **structured JSON** (Wei et al., 2022). All runs use deterministic decoding: temperature $= 0.0$, seed $= 13$.

Models. We evaluate ten models across two endpoints ($n = 50$ trials per condition): **OpenAI** — GPT-5.5, GPT-4.1, GPT-4o (OpenAI, 2023), GPT-4o-mini, GPT-5-mini; **DeepInfra** — Qwen-4B, Mistral-24B (Jiang et al., 2023), Nemotron-49B, Llama-3.1-8B (Touvron et al., 2023), Qwen-0.8B.[1]

4.2 Experiment 2: No-Index Semantic Stress

Motivation. In realistic workflows, users refer to papers by title or topic, not by position index (Yen et al., 2024). We construct a stress regime that removes the index cue and introduces semantically similar distractors, forcing the model to bind the correct arXiv ID to the correct paper from title/author cues alone (Kuratov et al., 2024; Jin et al., 2024).

Task. The model is given the same reference block but queried by a title-derived description (first 28 abstract words, no index). At each follow-up turn, a rephrased version of the same query is

[1] Qwen-0.8B: $n = 50$–55 due to variable completion rates. Full model IDs are listed in Appendix.

given while the full conversation history is preserved. Semantically similar papers are included as distractors.

Run scope. We execute the no-index stress condition at $\texttt{rc} = 128$ across turns $\{1, 10, 20\}$ on three models via DeepInfra:

- Llama-3.1-8B ($n = 269$ per turn, 10 categories; from Phase 2),
- Nemotron-49B ($n = 50$ per turn, 10 categories),
- Mistral-24B ($n = 50$ per turn, 10 categories).

All models use the same prompt template, reference block, and target papers (fixed per trial across turns). The Llama-3.1-8B run is the primary detailed run (largest n); the remaining two are replication runs extending the regime-gap finding. We additionally evaluate Llama-3.1-8B under indexed lookup up to $\texttt{rc} = 1024$ to establish the scaling boundary.

5 Evaluation Framework

5.1 Formalisation

Let a trial be defined by a reference list $R = \{r_1, \ldots, r_{\texttt{rc}}\}$, a fixed target index k, and a multi-turn conversation history H_t. Each reference r_i has a canonical arXiv ID $\mathrm{id}(r_i)$. The model outputs y_t at turn t.

5.2 Metrics

ID normalisation. $\tilde{\mathrm{id}}(r_k) = \mathrm{strip_version}(\mathrm{id}(r_k))$, stripping any `vN` suffix.

Extraction. Let $\mathcal{E}(y_t)$ be all arXiv-like identifiers extracted from y_t via a regex covering new-style (`YYYY.NNNNN`), old-style (`cs/YYMMNNN`), `arXiv:` prefixes, and `arxiv.org/abs` URLs.

Contains-correct. $\mathrm{Acc}_{\mathrm{ex}}(t) = \mathbb{1}[\tilde{\mathrm{id}}(r_k) \in \mathcal{E}(y_t)]$.

Exact-one (primary metric).

$$\mathrm{Acc}_1(t) = \mathbb{1}\Big[|\mathcal{E}(y_t)| = 1 \ \wedge \ \mathcal{E}(y_t) = \{\tilde{\mathrm{id}}(r_k)\}\Big]. \tag{1}$$

This penalises outputs that include the correct ID alongside extraneous identifiers—a failure that would render a citation ambiguous in practice.

Contamination. $\mathrm{Contam}(t) = \mathbb{1}[\tilde{\mathrm{id}}(r_k) \in \mathcal{E}(y_t) \ \wedge \ |\mathcal{E}(y_t)| \geq 2]$.

Leakage gap. $\Delta_{\mathrm{leak}} = \mathbb{E}[\mathrm{Acc}_{\mathrm{sub}}] - \mathbb{E}[\mathrm{Acc}_{\mathrm{ex}}]$, where $\mathrm{Acc}_{\mathrm{sub}}(t) = \mathbb{1}[\tilde{\mathrm{id}}(r_k)$ is a substring of $y_t]$. Large Δ_{leak} indicates that naive scoring overstates correctness, typically because the model echoes reference-block text or emits multiple IDs.

Failure decomposition. For each output y_t: **correct** ($|\mathcal{E}| = 1$, matches target); **missing** (no extracted ID); **wrong** (one ID extracted, not the target); **contaminated/merged** (correct ID plus ≥ 1 extraneous IDs).

Multi-turn stability. $\mathrm{FlipRate} = \mathbb{E}[\mathbb{1}[\hat{\imath}_t \neq \hat{\imath}_{t-1}]]$, where $\hat{\imath}_t$ is the extracted ID set at turn t. We additionally report first-error turn, recovery rate, and persistent-failure rate.

Why exact-one matters. A model that emits the correct ID surrounded by five others passes contains-correct scoring but produces an ambiguous citation in practice. Exact-one catches this; contains-correct does not. For the indexed-lookup regime, we observe $\mathrm{Acc}_1 = \mathrm{Acc}_{\mathrm{ex}}$ for most models (no contamination), validating contains-correct as a faithful proxy in that regime. For Nemotron-49B at $\texttt{rc} = 128$, however, $\Delta_{\mathrm{leak}} > 0$, establishing that the distinction matters.

6 Results

All results are from logged JSONL generations evaluated with our deterministic extractor.

6.1 Regime Map (Summary)

Table 1 shows per-model extraction-based ID accuracy (= exact-one accuracy for all models without contamination) at $\texttt{rc} \in \{1, 16, 128\}$ and turns $\{1, 10\}$ under indexed lookup. Bootstrap 95% CIs are computed by resampling trials within each slice (2,000 iterations, $n = 50$ per cell unless noted).

The data reveal three behaviour classes: **(i) near-ceiling** (100% across all tested `rc` and turns): GPT-5.5, GPT-4.1, GPT-4o, GPT-4o-mini, Qwen-4B, Llama-3.1-8B; **(ii) moderate degradation at high `rc`**: GPT-5-mini (92% at $\texttt{rc} = 1$, turn 1), Mistral-24B (88% at $\texttt{rc} = 128$, turn 1); **(iii) strong degradation**: Nemotron-49B (56% at $\texttt{rc} = 128$) and Qwen-0.8B (42% already at $\texttt{rc} = 16$).

6.2 Experiment 1: Scaling Under Indexed Lookup

Turn-1 scaling. Figure 2 shows per-model accuracy at turn 1 as `rc` increases for DeepInfra models.

Table 1: Indexed-lookup arXiv-ID accuracy (%) at turns 1 and 10. For all models with no contamination, extracted = exact-one; the exception (Nemotron-49B, †) is noted below. 95% CIs in parentheses (Clopper-Pearson exact, $n = 50$ unless noted). Models achieving 100% at *all* tested conditions are marked **near-ceiling**.

		Turn 1			**Turn 10**		
Endpoint	**Model**	rc = 1	rc = 16	rc = 128	rc = 1	rc = 16	rc = 128
OAI	GPT-5.5	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)
OAI	GPT-4.1	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)
OAI	GPT-4o	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)
OAI	GPT-4o-mini	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)
OAI	GPT-5-mini‡	92 (81,98)	98 (89,100)	98 (89,100)	96 (86,99)	96 (86,99)	98 (89,100)
DI	Qwen-4B	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)	100 (93,100)
DI	Mistral-24B§	100 (93,100)	100 (93,100)	88 (76,96)	100 (93,100)	98 (89,100)	98 (89,100)
DI	Nemotron-49B†	100 (93,100)	94 (84,99)	56 (41,70)	100 (93,100)	90 (78,97)	56 (41,70)
DI	Llama-3.1-8B*	100 (83,100)	100 (83,100)	100 (83,100)	100 (83,100)	100 (83,100)	100 (83,100)
DI	Qwen-0.8B¶	100 (93,100)	42 (29,56)	62 (48,75)	100 (93,100)	42 (29,56)	69 (55,81)

Why CI bounds vary across rows (all intervals are Clopper-Pearson exact): CI width depends on both the point estimate and n. At 100% with $n = 50$, the lower bound is 93%; with $n = 20$ it is 83%. At lower point estimates (e.g., 56%, 42%) the interval widens and both bounds shift down.

*__Llama-3.1-8B__: $n = 20$ per cell (Phase 2 runs); lower bound is 83% even at 100% accuracy because the smaller sample yields a wider exact CI.

†__Nemotron-49B__: genuine degradation under indexed lookup (rc = 16: 94%, rc = 128: 56%). At 56% with $n = 50$ the CI is (41,70) — wide because variance is highest near $p = 0.5$. Additionally, substring accuracy = 74% at rc = 128 (18 pp above exact-one due to contamination; see §6.3).

‡__GPT-5-mini__: starts at 92% (turn 1, rc = 1) and improves to 98% with larger rc and more turns — an atypical pattern where repeated prompting helps. CI bounds shift accordingly (81,98) → (89,100).

§__Mistral-24B__: drops from 100% to 88% at rc = 128 turn 1, then recovers to 98% by turn 10, indicating that repeated prompting partially compensates for the larger reference block.

¶__Qwen-0.8B__: non-monotonic — degrades to 42% at rc = 16 then rises to 62%/69% at rc = 128. The recovery likely reflects a change in output strategy (the model begins emitting identifiers differently at longer contexts).

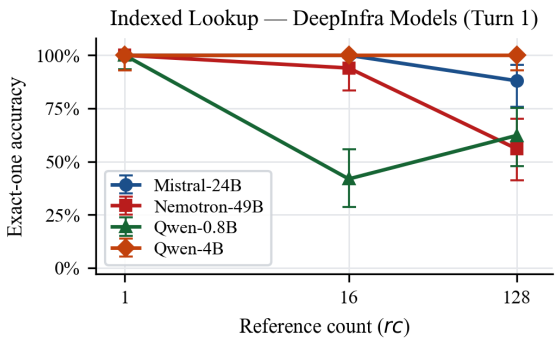

Figure 2: Extraction-based arXiv-ID accuracy at turn 1 by reference count (indexed lookup, DeepInfra models). Error bars are 95% CIs (Clopper-Pearson, $n = 50$). Qwen-0.8B degrades at rc = 16; Nemotron-49B at rc = 128.

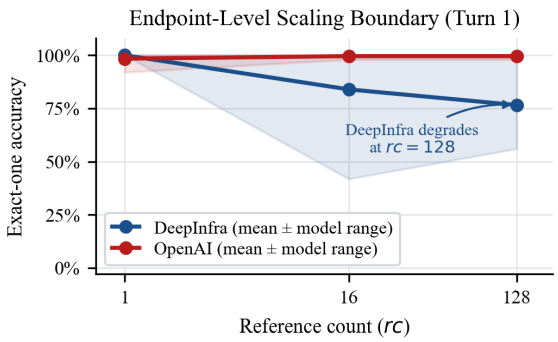

Figure 3: Endpoint-level scaling boundary (indexed lookup, turn 1). Ribbons span min–max extracted accuracy across models within each endpoint at each rc. Endpoints differ in hosted model sets and tokenisation; aggregates should not be interpreted causally as provider effects.

Qwen-0.8B degrades earliest (at rc = 16), consistent with a very small parameter count that limits associative retrieval from long contexts. Nemotron-49B degrades specifically at rc = 128, while larger models (Mistral-24B, Qwen-4B) remain near-ceiling. Figure 3 presents endpoint-level min–max ribbons and confirms that the DeepInfra suite has substantially wider within-suite variance than the OpenAI suite at rc = 128.

Multi-turn dynamics at rc=128. Figure 4 traces accuracy across turns at rc = 128. Nemotron-49B stays flat at ≈56% (stable ceiling, not a warm-up effect); Qwen-0.8B improves slightly (62%→69%), consistent with repeated prompting.

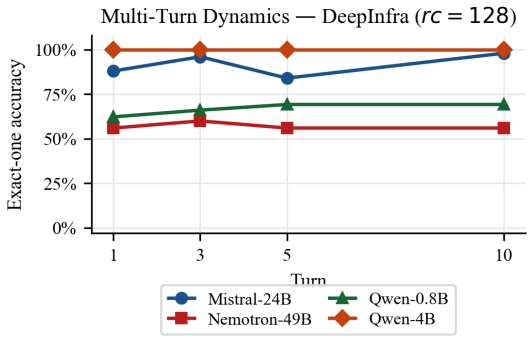

Figure 4: Multi-turn dynamics at rc = 128 (indexed lookup, DeepInfra models, turns 1–10). Full conversation history is preserved with a fixed target per trial.

Table 2: Metric leakage for Nemotron-49B at rc = 128 ($n = 50$ per turn). $\mathrm{Acc}_{\mathrm{ex}}$: extraction-based (contains-correct); $\mathrm{Acc}_{\mathrm{sub}}$: naive substring; Δ_{leak}: gap.

Turn	$\mathrm{Acc}_{\mathrm{ex}}$	$\mathrm{Acc}_{\mathrm{sub}}$	Δ_{leak}
1	56%	74%	**+18 pp**
3	60%	70%	**+10 pp**
5	56%	66%	**+10 pp**
10	56%	66%	**+10 pp**

6.3 Metric Leakage and Failure Accounting

Table 2 and Figure 5 show the leakage gap Δ_{leak} for Nemotron-49B at rc = 128. Substring scoring overestimates correctness by **18 percentage points** at turn 1 (56% extracted vs. 74% substring). The gap narrows but persists at turns 3–10 (≈10 pp). No other tested model shows $\Delta_{\mathrm{leak}} > 0$ at any measured rc or turn.

6.4 Scaling to rc=1024

Figure 6 shows that even the indexed-lookup regime degrades beyond rc = 128 for Llama-3.1-8B. At rc = 512: 90% exact-one (turn 1), 95% (turn 10). At rc = 1024: **70%** (turn 1), 75% (turn 10). All failures at this scale are wrong_id—the model extracts exactly one identifier but it is not the target—with zero missing-ID or contamination events. This establishes an empirical stability boundary: Llama-3.1-8B remains near-ceiling up to rc = 256 but degrades reliably by rc = 1024.

6.5 Experiment 2: No-Index Semantic Stress

The key contrast. Llama-3.1-8B achieves **100%** exact-one accuracy in the indexed-lookup regime at rc = 128, yet only **1.1%** exact-one accuracy in the no-index stress regime at the same rc and turn 1. This 98.9 percentage-point drop—from the same model, the same reference block, the same rc—is produced solely by removing the positional index cue and querying by title description instead.

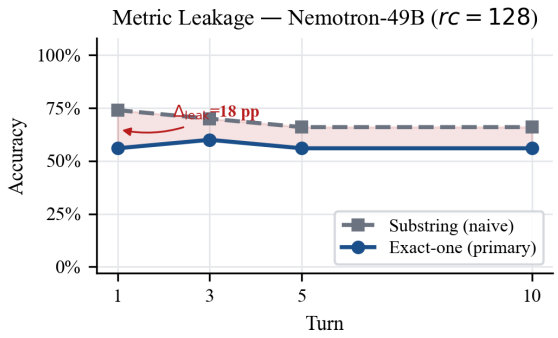

Figure 5: Metric leakage (Nemotron-49B, rc = 128, indexed lookup). $\mathrm{Acc}_{\mathrm{sub}}$ (dashed) consistently exceeds $\mathrm{Acc}_{\mathrm{ex}}$ (solid), with the largest gap at turn 1 (18 pp). This divergence occurs when outputs include the correct ID alongside additional extraneous identifiers that pass substring matching but fail strict extraction.

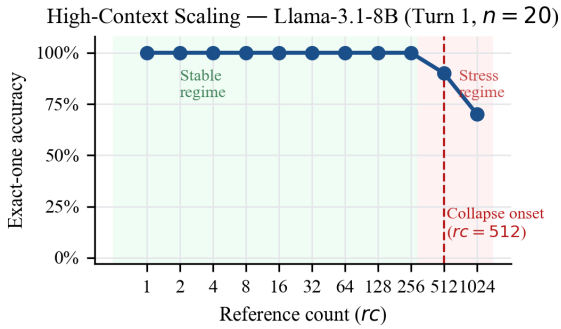

Figure 6: Scaling beyond rc = 128 (indexed lookup, Llama-3.1-8B, turn 1, $n = 20$). Exact-one accuracy drops to 70% at rc = 1024; all failures are wrong_id (the model returns a single incorrect identifier). The model is near-ceiling at rc ≤ 256.

Failure decomposition. In the no-index regime, failures are almost entirely **missing-ID** (98.9% at turn 1); there are zero wrong-ID or contamination events. The gain to 33.8% by turn 10 comes entirely from the model beginning to emit identifiers under repeated prompting, not from correcting wrong ones.

Semantic distractors, temporal stability, and position effects. Figure 7 presents three complementary views:

Multi-model regime gap (Table 3). Table 3 reports the indexed-lookup vs. no-index exact-one accuracy at rc = 128, turn 1 for three models tested in both regimes (Llama-3.1-8B, Nemotron-49B, Mistral-24B). For two of three models, the

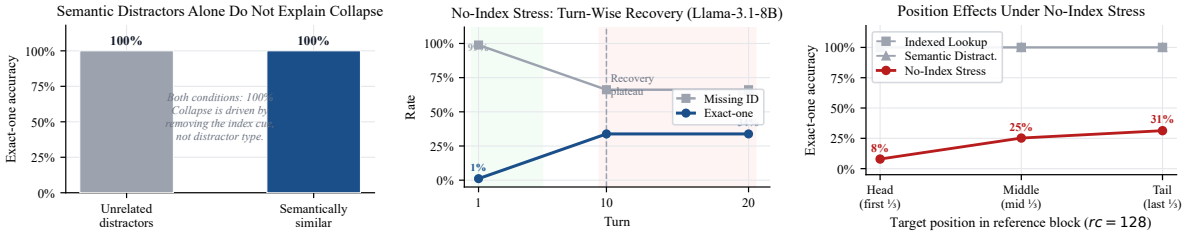

Figure 7: No-index stress at rc = 128 (Llama-3.1-8B; exact-one accuracy). **Left:** semantically similar vs. unrelated distractors in the indexed regime both yield 100%, but the no-index regime (not shown here) yields only 1.1% at turn 1—confirming that removing the index cue, not the distractor type, is the primary driver of failure. **Middle:** turn-wise strict outcome rates in no-index stress; missing-ID decreases as repeated prompting elicits identifiers. **Right:** position sensitivity under no-index stress— head (7.9%), middle (25.2%), tail (31.3%) exact-one accuracy, aggregated over turns 1, 10, 20. Papers at the tail of the reference block are easiest to retrieve, consistent with a recency bias in identifier recall under no-index querying.

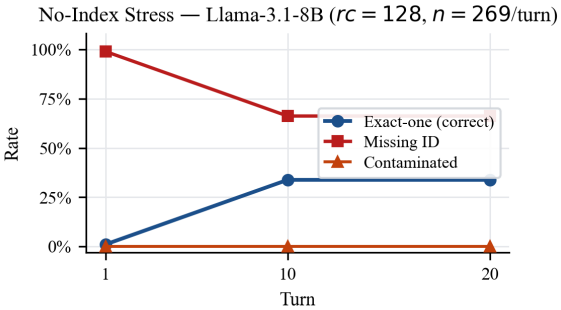

Figure 8: No-index/title-based stress (rc = 128, Llama-3.1-8B). Turn traces for exact-one, missing-ID, and contamination rates across measured turns.

Table 3: Regime gap at rc = 128, turn 1: indexed-lookup exact-one accuracy (from Table 1) vs. no-index exact-one accuracy for all three models tested in both regimes, revealing heterogeneous regime sensitivity across model families. Llama-3.1-8B collapses under no-index; Nemotron-49B shows moderate drop (56→48 pp); Mistral-24B improves. No-index numbers from the multi-model replication runs; indexed numbers from Table 1.

Model	Indexed	No-Index (T1)	Gap (pp)
Llama-3.1-8B	100.0%	1.1%	−98.9 pp
Nemotron-49B	56.0%	48.0%	−8.0 pp
Mistral-24B	88.0%	98.0%	+10.0 pp

indexed→no-index gap is 56.0–98.9 percentage points, confirming large collapses in models with high positional dependence. Mistral-24B is an exception—achieving 98.0% no-index vs. 88.0% indexed—indicating that semantic attribution ability varies substantially across model families and is not fully predicted by indexed-regime performance.

Regime threshold summary. Per-model empirical stability thresholds are collected in Appendix Table 13: five models remain near-ceiling through rc = 128; Nemotron-49B and Qwen-0.8B degrade from rc = 16.

7 Discussion

7.1 Regime Sensitivity Is Real but Model-Dependent

The 100%→1.1% collapse for Llama-3.1-8B between regimes at identical rc is our sharpest result. Table 3 shows a partial version in Nemotron-49B (56%→48% at turn 1, recovering to 86% by turn 20), but not uniformly: Mistral-24B *improves* (88%→98%), indicating model-dependent reliance on positional vs. semantic cues. The two regimes probe fundamentally different primitives; conflating them produces inflated estimates of practical attribution ability.

7.2 Repeated Prompting Partially Recovers No-Index Performance

That accuracy rises from 1.1% (turn 1) to 33.8% (turn 10) in the no-index regime—with all gains coming from the model beginning to emit identifiers rather than correcting wrong ones—suggests that the primary barrier is identifier *output suppression*, not identifier *ignorance* (Anonymous, 2025). The model has the required information but does not produce it until repeatedly prompted (Yen et al., 2024; Kuratov et al., 2024). This failure mode (missing-ID rather than wrong-ID) is distinct from typical retrieval-failure patterns and points to output-format constraints as a targeted mitigation.

Table 4: No-index exact-one accuracy (%) by model and turn at rc = 128. 95% bootstrap CIs in parentheses (2,000 iterations). Llama-3.1-8B: $n = 269$ per turn; Nemotron-49B and Mistral-24B: $n = 50$ per turn. Nemotron-49B's reasoning mode still exhausts its token budget on ∼32–52% of calls, explaining the accuracy gap at early turns; when it does reply it is correct 100% of the time. Mistral-24B reaches ceiling from turn 1; Llama-3.1-8B plateaus after turn 10.

Model	**Turn 1**	**Turn 10**	**Turn 20**
Llama-3.1-8B	1.1 (0.0, 2.6)	33.8 (28.2, 39.4)	33.8 (28.2, 39.4)
Nemotron-49B	48.0 (34.0, 62.0)	70.0 (58.0, 82.0)	86.0 (76.0, 94.0)
Mistral-24B	98.0 (94.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)

7.3 Evaluation Design Matters

The 18 pp leakage gap for Nemotron-49B at rc = 128 illustrates a practical risk in citation-fidelity benchmarking (Es et al., 2023; Min et al., 2023). A researcher using substring scoring would conclude that Nemotron-49B achieves 74% accuracy at rc = 128, turn 1; our extraction-based metric reports 56%. The additional 18% passing responses are cases where the correct identifier appears alongside one or more extraneous identifiers—an output that would introduce citation ambiguity in a real writing workflow. We recommend that any citation-fidelity benchmark report at minimum both extracted and exact-one metrics, alongside the contamination rate (Lin et al., 2021; Mu et al., 2023).

7.4 Implications for Benchmark and Model Selection

Practitioners should: (i) state the evaluation regime explicitly, as indexed lookup and no-index stress are not interchangeable (Kuratov et al., 2024; Wu et al., 2025) and deployment decisions should use both; (ii) use extraction-based primary metrics and report exact-one alongside contains-correct, since substring scoring can reverse rankings (18 pp leakage gap for Nemotron-49B (Es et al., 2023; Modarressi et al., 2025)) and contamination rates expose failures invisible to contains-correct alone (McKenna et al., 2023; Miao et al., 2023); and (iii) randomise or stratify by position, since no-index accuracy is strongly position-dependent (Liu et al., 2023a; Yu et al., 2025).

8 Limitations

Protocol and scope. The indexed-lookup regime is materially easier than naturalistic citation usage; no-index stress partially addresses this but remains controlled. We do not claim temporal drift — no time-separated replications were executed.

No-index model coverage. The primary detailed no-index run uses Llama-3.1-8B ($n = 269$ per turn); the replication runs on Nemotron-49B and Mistral-24B use $n = 50$ per turn, which limits the precision of cross-model comparisons. Whether the magnitude of the indexed→no-index gap varies systematically with model size or training data is an open question requiring larger replication runs.

Endpoint heterogeneity. DeepInfra and OpenAI differ in model versions, context limits, tokenisation, and logging; endpoint differences should not be interpreted causally as provider effects.

Identifier scope. Our evaluator targets arXiv identifiers. Extension to DOI, venue, or author ordering is an important future direction (Zhang et al., 2024a; Gao et al., 2023).

9 Ethics and Reproducibility

Data. CitationDrift-16K is metadata-only (no PDFs); all data derives from the arXiv public API under its terms of service. Our benchmark is released to improve measurement of citation attribution errors, not to enable or conceal them.

Reproducibility. All figures and tables are generated from logged JSONL outputs by deterministic scripts (no manual editing). Key invariants: fixed-target invariant; extraction-based scoring; temperature = 0.0, seed = 13; 2,000-iteration bootstrap CIs. Full checklist: Appendix S; figure/table index: Appendix T.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2024. Do language models know when they're hallucinating references? *Preprint*, arXiv:2305.18248.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-Eval: Instituting standardized evaluation for long context language models. *Preprint*, arXiv:2307.11088.

Anonymous. 2025. Drift no more? context equilibria in multi-turn LLM interactions. *Preprint*, arXiv:2510.07777.

Anonymous. 2026a. BeliefShift: Benchmarking temporal belief consistency and opinion drift in LLM agents. *Preprint*, arXiv:2603.23848.

Anonymous. 2026b. PICon: A multi-turn interrogation framework for evaluating persona agent consistency. *Preprint*, arXiv:2603.25620.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Kirolos Ataallah, Eslam Abdelrahman, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2025. Infinibench: A benchmark for large multi-modal models in long-form movies and tv shows. *Preprint*, arXiv:2406.19875.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. LongBench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated evaluation of retrieval augmented generation. *Preprint*, arXiv:2309.15217.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *Preprint*, arXiv:2305.14627.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2024. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. *Preprint*, arXiv:2408.10151.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. *Preprint*, arXiv:2204.04991.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Agrawal, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Preprint*, arXiv:2202.03629.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Jiawei Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend LLM context window without tuning. *Preprint*, arXiv:2401.01325.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *Preprint*, arXiv:2307.16883.

Yekyung Kim, Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. One ruler to measure them all: Benchmarking multilingual long-context language models. *Preprint*, arXiv:2503.01996.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. *Preprint*, arXiv:2406.10149.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *Preprint*, arXiv:2401.16745.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Preprint*, arXiv:2005.11401.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. LooGLE: Can long-context language models understand long contexts? *Preprint*, arXiv:2311.04939.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan,

Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 17 others. 2022. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianhao Zhang, Yu Su, Huan Sun, and 3 others. 2023b. AgentBench: Evaluating LLMs as agents. *Preprint*, arXiv:2308.03688.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *Preprint*, arXiv:2005.00661.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *Preprint*, arXiv:2305.14552.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. SelfCheck: Using LLMs to zero-shot check their own step-by-step reasoning. *Preprint*, arXiv:2308.00436.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. *Preprint*, arXiv:2502.05167.

Hyeonseok Moon and Heuiseok Lim. 2026. Needlechain: Measuring intact context comprehension capability of large language models. *Preprint*, arXiv:2507.22411.

Jesse Mu, Xiang Li, and Noah Goodman. 2023. SCALE: Constructing structured natural language comments with large language models. *Preprint*, arXiv:2310.13189.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? *Preprint*, arXiv:1909.01066.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *Preprint*, arXiv:2108.12409.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. *Preprint*, arXiv:2305.14196.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. MileBench: Benchmarking multimodal large language models in long contexts. *Preprint*, arXiv:2404.18532.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. *Preprint*, arXiv:2404.06480.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2025. Longgenbench: Benchmarking long-form generation in long context llms. *Preprint*, arXiv:2409.02076.

Wang Yang, Hongye Jin, Shaochen Zhong, Song Jiang, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. 2025a. 100-longbench: Are de facto long-context benchmarks literally evaluating long-context ability? *Preprint*, arXiv:2505.19293.

Yijun Yang, Zeyu Huang, Wenhao Zhu, Zihan Qiu, Fei Yuan, Jeff Z. Pan, and Ivan Titov. 2025b. A controllable examination for long-context language models. *Preprint*, arXiv:2506.02921.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. HELMET: How to evaluate long-context language models effectively and thoroughly. *Preprint*, arXiv:2410.02694.

Yifei Yu, Qian-Wen Zhang, Lingfeng Qiao, Di Yin, Fang Li, Jie Wang, Zengxi Chen, Suncong Zheng, Xiaolong Liang, and Xing Sun. 2025. Sequential-niah: A needle-in-a-haystack benchmark for extracting sequential needles from long contexts. *Preprint*, arXiv:2504.04713.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024a. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

A Benchmark Details and Dataset Statistics

Record fields. Each benchmark record stores: arXiv identifier, title, authors list, abstract text, primary and secondary category labels, published/updated timestamps, and canonical arXiv URL. No PDF content is stored or used.

Dataset statistics.

Categories covered. `cs.CL`, `cs.AI`, `cs.LG`, `cs.CV`, `cs.RO`, `cs.IR`, `cs.HC`, `cs.SE`, `cs.CR`, `cs.NE`, `cs.DS`, `cs.GT`, `cs.SY`, `cs.NA`, `eess.SP`, `stat.ML`.

Identifier normalisation. The extractor covers: new-style identifiers (`YYYY.NNNNN`), old-style identifiers (`cs/YYMMNNN`), `arXiv:` prefixes, and `arxiv.org/abs` or `arxiv.org/pdf` URLs. Version suffixes (`vN`) are stripped before matching.

B Prompts and Turn Structure

Reference block format. Each reference is rendered as one line per entry: `[k] Title. Author1, Author2, ... arXiv:YYYY.NNNNN (year).` References are numbered [1] to [`rc`], fixed across all runs.

`Indexed lookup – turn 1 (system + user).` `System: Use only the references below to answer questions.` `User: {reference block} – Return the exact metadata for reference [k].`

`Indexed lookup – follow-up turns.` `User: Without introducing any new references, restate the exact metadata for reference [k] again.` Full conversation history is preserved in the API message list.

`No-index stress – turn 1.` `System: Use only the references below to answer questions.` `User: {reference block} – Identify the single reference that best matches the following description. Return the exact citation metadata including the arXiv identifier. Description: <first 28 abstract words>...`

No-index stress — follow-up turns. Five rephrase templates cycle across turns, each restating the query in different words while appending the original abstract snippet. No new reference block is included in follow-up turns.

Prompt regimes (Experiment 1 only).

- **Neutral**: *"Use only the references below to answer the question."*
- **Citation-aware**: *"Carefully preserve citation fidelity. Do not invent, merge, or alter references. Use only the references below."*
- **Structured JSON**: *"Use only the references below. Return exactly one JSON object with keys* `title`, `authors`, `year`, *and* `arxiv_id`*."* OpenAI models additionally receive a JSON-schema `response_format` constraint.

Decoding. All runs: `temperature=0.0`, `seed=13`, `max_tokens=128`; no sampling, no top-p.

C Evaluator Pseudocode

Table 6: Evaluator pseudocode: the three primary scoring functions. All identifier comparisons are performed on normalised IDs (version suffix stripped).

```
normalize_id(id):
  return regex_extract(YYYY.NNNNN, id) # drops vN
extract_ids(text):
  # new-style, old-style, arXiv: prefix, arxiv.org URLs
  return set(all YYYY.NNNNN matches in text)
acc_exact_one(resp, target):
  ids = extract_ids(resp)
  return (len(ids)==1) and (normalize_id(target) in ids)
acc_extracted(resp, target):
  return normalize_id(target) in extract_ids(resp)
contamination(resp, target):
  ids = extract_ids(resp)
  return (normalize_id(target) in ids) and (len(ids)>=2)
```

D Strict Metrics: Full Snapshot

Table 7 gives per-model strict metrics at turn 1, `rc` $\in \{1, 16, 128\}$ across all runs. The full row-level JSONL (with prompts and raw responses) is released in the supplementary materials.

Table 5: CitationDrift-16K dataset statistics.

Property	Value
Total papers	16,384
Categories	16 (`cs.*` and `stat.ML`)
Papers per category	1,024
Source API	arXiv legacy API
Identifier format	`YYYY.NNNNN` (new-style)
Reference-set sizes	$\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1{,}024\}$
Semantic-similar groups	91 (high pairwise cosine similarity)
Unrelated groups	124 (low pairwise cosine similarity)
Embedding model	`sentence-transformers/all-MiniLM-L6-v2`
PDF content used	No (metadata only)

Table 7: Strict-metric snapshot at turn 1 for `rc` $\in \{1, 16, 128\}$. Exact-one is the primary metric; Contains = contains-correct; Contam. = contamination rate; Missing = no-ID rate. All values are proportions (0–1). $n = 50$ per cell unless noted. †Nemotron-49B at `rc` $= 128$: contamination = 0.180; substring accuracy = 0.760 (18 pp above exact-one).

Endpt.	Model	rc	Exact-one	Contains	Contam.	Missing
OpenAI-hosted models						
OAI	GPT-5.5	1	1.000	1.000	0.000	0.000
OAI	GPT-5.5	16	1.000	1.000	0.000	0.000
OAI	GPT-5.5	128	1.000	1.000	0.000	0.000
OAI	GPT-4.1	1	1.000	1.000	0.000	0.000
OAI	GPT-4.1	16	1.000	1.000	0.000	0.000
OAI	GPT-4.1	128	1.000	1.000	0.000	0.000
OAI	GPT-4o	1	1.000	1.000	0.000	0.000
OAI	GPT-4o	16	1.000	1.000	0.000	0.000
OAI	GPT-4o	128	1.000	1.000	0.000	0.000
OAI	GPT-4o-mini	1	1.000	1.000	0.000	0.000
OAI	GPT-4o-mini	16	1.000	1.000	0.000	0.000
OAI	GPT-4o-mini	128	1.000	1.000	0.000	0.000
OAI	GPT-5-mini	1	0.920	0.920	0.000	0.080
OAI	GPT-5-mini	16	0.980	0.980	0.000	0.020
OAI	GPT-5-mini	128	0.980	0.980	0.000	0.020
DeepInfra-hosted models						
DI	Qwen-4B	1	1.000	1.000	0.000	0.000
DI	Qwen-4B	16	1.000	1.000	0.000	0.000
DI	Qwen-4B	128	1.000	1.000	0.000	0.000
DI	Mistral-24B	1	1.000	1.000	0.000	0.000
DI	Mistral-24B	16	1.000	1.000	0.000	0.000
DI	Mistral-24B	128	0.880	0.880	0.000	0.000
DI	Nemotron-49B	1	1.000	1.000	0.000	0.000
DI	Nemotron-49B	16	0.940	0.940	0.000	0.060
DI	Nemotron-49B†	128	0.580	0.760	**0.180**	0.160
DI	Llama-3.1-8B	1	1.000	1.000	0.000	0.000
DI	Llama-3.1-8B	16	1.000	1.000	0.000	0.000
DI	Llama-3.1-8B	128	1.000	1.000	0.000	0.000
DI	Qwen-0.8B	1	1.000	1.000	0.000	0.000
DI	Qwen-0.8B	16	0.891	0.891	0.000	0.109
DI	Qwen-0.8B	128	0.906	0.906	0.000	0.038

E No-Index Stress: Extended Results

Table 8: No-index stress (Llama-3.1-8B, `rc` $= 128$, $n = 269$): strict metrics across turns.

Turn	n	Exact-one	Contains	Contam.	Missing
1	269	0.011	0.011	0.000	0.989
10	269	0.338	0.338	0.000	0.662
20	269	0.338	0.338	0.000	0.662

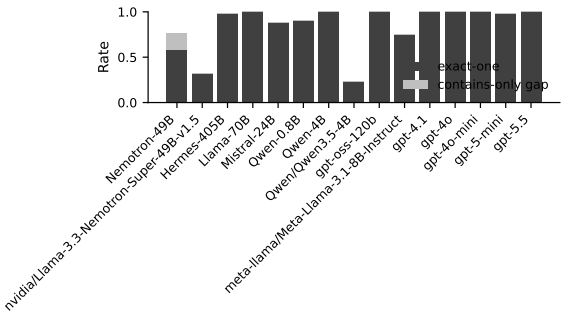

Figure 9: Exact-one vs. contains-correct at `rc` $= 128$ (turn 1), by model. The gap (18 pp for Nemotron-49B) quantifies contaminated outputs that pass contains-correct but fail exact-one.

F Within-Run Stability and Multi-Turn Dynamics

We report within-run (not cross-time) stability: how outcomes change as the same conversation is extended across turns, holding the target paper fixed.

No-index stress turn dynamics (Llama-3.1-8B, $n = 269$).

- Turn 1: 1.1% exact-one, 98.9% missing-ID.
- Turn 10: 33.8% exact-one, 66.2% missing-ID.
- Turn 20: 33.8% exact-one, 66.2% missing-ID (plateau).
- Wrong-ID and contamination: 0% at all turns.

The plateau between turns 10 and 20 indicates that repeated prompting initially unlocks identifier recall, but further repetition yields no additional recovery. The gain is entirely from the missing-ID category transitioning to correct: the model begins to emit the identifier rather than producing pure prose.

Indexed-lookup stability at `rc` $= 128$. All near-ceiling models (GPT-5.5, GPT-4.1, GPT-4o, GPT-4o-mini, Qwen-4B, Llama-3.1-8B) maintain 100% exact-one accuracy across all turns 1–10. Nemotron-49B plateaus at ≈56% extracted from turn 1, with no turn-dependent improvement or degradation. GPT-5-mini shows slight improvement (92%→98%) between turn 1 and turn 10.

G Prompt-Regime Effects

Table 9: Llama-3.1-8B (DeepInfra): exact-one accuracy by prompt mode at `rc` $= 1$, turns 1 and 10 ($n = 250$ per cell). All three modes achieve identical perfect accuracy at this minimal scale; prompt-mode differences are expected to emerge only at higher `rc`.

Prompt Mode	Turn 1	Turn 10	n
Neutral	1.000	1.000	250
Citation-aware	1.000	1.000	250
Structured JSON	1.000	1.000	250

H Per-Model Tables at rc=128

Table 10: DeepInfra indexed-lookup results at `rc` $=$ 128: exact-one and substring accuracy at turns 1 and 10. Δ = substring $-$ exact-one (leakage gap). Only Nemotron-49B shows non-zero leakage.

Model	Turn	n	Exact-one	Substring	Δ
Hermes-405B	1	50	0.980	0.980	0.000
Hermes-405B	10	50	1.000	1.000	0.000
Llama-70B	1	50	1.000	1.000	0.000
Llama-70B	10	50	1.000	1.000	0.000
Mistral-24B	1	50	0.880	0.880	0.000
Mistral-24B	10	50	0.980	0.980	0.000
Nemotron-49B	1	50	0.580	0.760	**+0.180**
Nemotron-49B	10	50	0.560	0.660	**+0.100**
Qwen-0.8B	1	53	0.906	0.906	0.000
Qwen-0.8B	10	52	0.962	0.962	0.000
Qwen-4B	1	50	1.000	1.000	0.000
Qwen-4B	10	50	1.000	1.000	0.000

Table 11: OpenAI indexed-lookup results at `rc` $= 128$: exact-one accuracy at turns 1 and 10 ($n = 50$ per cell). All models except GPT-5-mini reach 100%; GPT-5-mini stabilises at 98%.

Model	Turn	n	Exact-one
GPT-5.5	1	50	1.000
	10	50	1.000
GPT-4.1	1	50	1.000
	10	50	1.000
GPT-4o	1	50	1.000
	10	50	1.000
GPT-4o-mini	1	50	1.000
	10	50	1.000
GPT-5-mini	1	50	0.980
	10	50	0.980

I Category Sensitivity

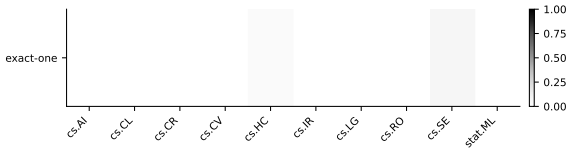

Figure 10: No-index stress (`rc` $= 128$, turn 1): exact-one accuracy by arXiv category (heatmap). The benchmark covers 10 categories for Llama-3.1-8B's no-index run (cs.AI, cs.CL, cs.CR, cs.CV, cs.HC, cs.IR, cs.LG, cs.RO, cs.SE, stat.ML), ≈27 trials each ($n = 269$ total per turn).

Table 12: No-index stress (Llama-3.1-8B, rc = 128, turn 1): exact-one accuracy by arXiv category, ranked ascending. Most categories achieve 0% at turn 1; only cs.HC and cs.SE show non-zero recovery ($n \approx 27$ trials each, 269 total).

Category	n	Exact-one
cs.AI	27	0.000
cs.CL	26	0.000
cs.CR	27	0.000
cs.CV	27	0.000
cs.IR	27	0.000
cs.LG	27	0.000
cs.RO	27	0.000
stat.ML	27	0.000
cs.HC	27	0.037
cs.SE	27	**0.074**

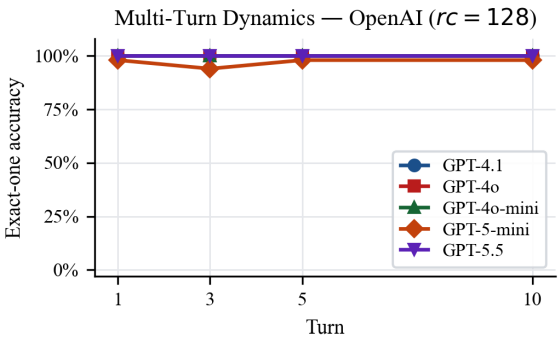

Figure 13: Multi-turn dynamics (OpenAI models, indexed lookup, rc = 128): exact-one accuracy across turns 1–10. All models except GPT-5-mini are flat at 100%; GPT-5-mini stabilises at 98% after turn 1.

J Reference-Block Length Diagnostics

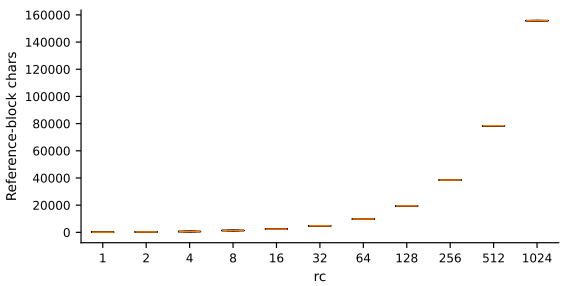

Figure 11: Reference-block character length by rc (boxplots). Used as a model-agnostic context-load proxy when token counts are unavailable. At rc = 1024, median block length exceeds 100,000 characters.

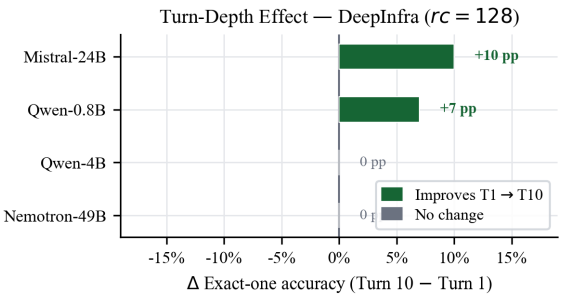

Figure 14: Turn-depth change at rc = 128 (DeepInfra models, indexed lookup): Δ exact-one accuracy (turn 10 − turn 1). Qwen-0.8B shows the largest positive delta (+7 pp), consistent with repeated prompting recovering missing-ID failures. All other models show no change ($\Delta \approx 0$).

K Additional Figures

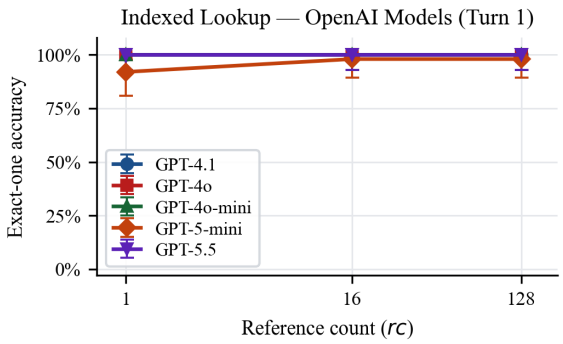

Figure 12: Indexed-lookup scaling (OpenAI models, turn 1): exact-one accuracy vs. rc $\in \{1, 16, 128\}$. All five models achieve 100% except GPT-5-mini (92% at rc = 1). Error bars are 95% Clopper-Pearson CIs ($n = 50$).

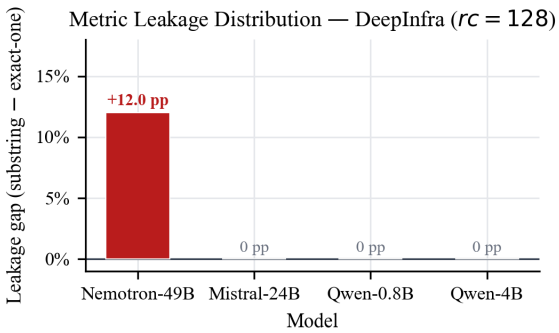

Figure 15: Leakage gap Δ_{leak} (substring − exact-one) distribution across the DeepInfra multi-model run at rc = 128, grouped by model. Only Nemotron-49B shows non-zero leakage; all other models have $\Delta_{\text{leak}} = 0$ at all turns.

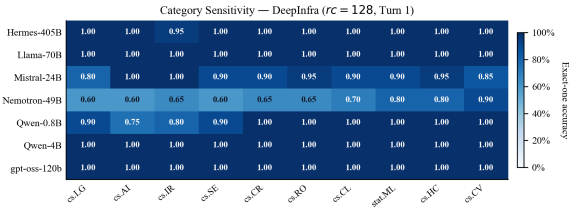

Figure 16: Category sensitivity (DeepInfra models, $\texttt{rc} = 128$, turn 1): mean exact-one accuracy by model and arXiv category. Near-ceiling models show no category variation; Nemotron-49B shows moderate variation; Qwen-0.8B shows the widest spread.

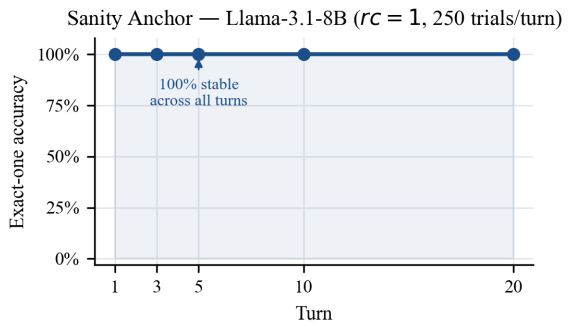

Figure 17: Sanity anchor (Llama-3.1-8B, indexed lookup, $\texttt{rc} = 1$, turns 1–20): 100% exact-one accuracy at every turn, confirming perfect stability in the minimal single-reference condition. This establishes the upper bound for indexed-regime performance.

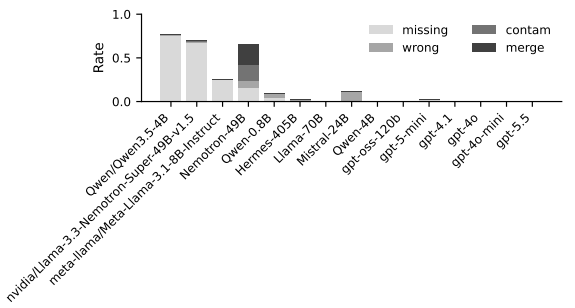

Figure 18: Failure-mode budget at $\texttt{rc} = 128$, turn 1, by model (indexed lookup). Stacked bars show the proportion of correct, missing-ID, wrong-ID, and contaminated outputs. Most models are entirely in the "correct" category; Nemotron-49B and Qwen-0.8B show non-trivial failure budgets.

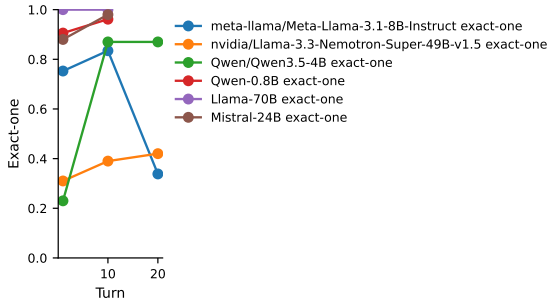

Figure 19: Exact-one accuracy by turn for selected models at $\texttt{rc} = 128$ (indexed lookup, turns 1–10). Near-ceiling models (GPT-4.1, Qwen-4B, Llama-3.1-8B) are flat at 100% throughout; Nemotron-49B remains around 56%; Qwen-0.8B shows the largest within-model improvement across turns.

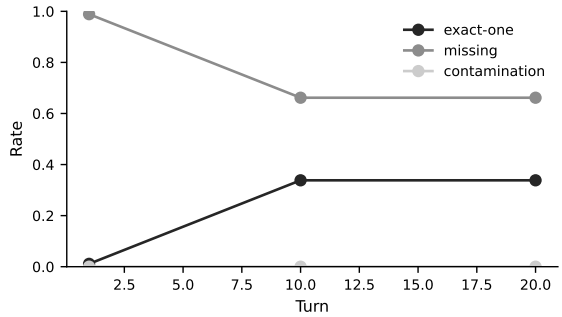

Figure 20: No-index stress turn curve (Llama-3.1-8B, $\texttt{rc} = 128$, $n = 269$): exact-one, missing-ID, and contamination rates across measured turns. The missing-ID rate monotonically decreases from 98.9% (turn 1) to 66.2% (turn 10) and plateaus; wrong-ID and contamination remain at 0% throughout, confirming the failure mode is pure output suppression.

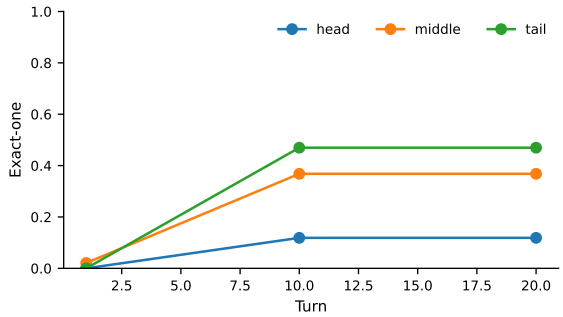

Figure 21: Position effects by turn (Llama-3.1-8B, no-index stress, $\texttt{rc} = 128$). Exact-one accuracy stratified by target position (head / middle / tail) across turns 1, 10, 20. The recency advantage (tail $>$ head) persists across all turns, indicating that position bias is a stable property of the regime, not an artifact of early-turn output suppression.

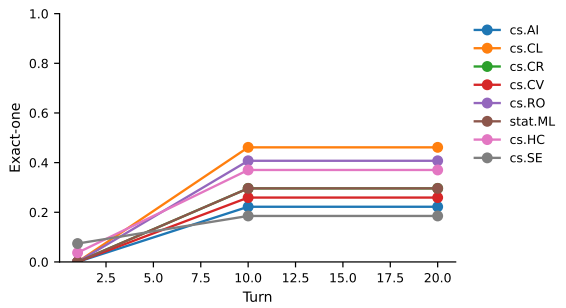

Figure 22: Best and worst arXiv categories for no-index retrieval (Llama-3.1-8B, `rc` $= 128$) across turns 1, 10, 20. cs.SE and cs.HC consistently outperform other categories; cs.AI and cs.CL remain at or near 0% even at turn 20.

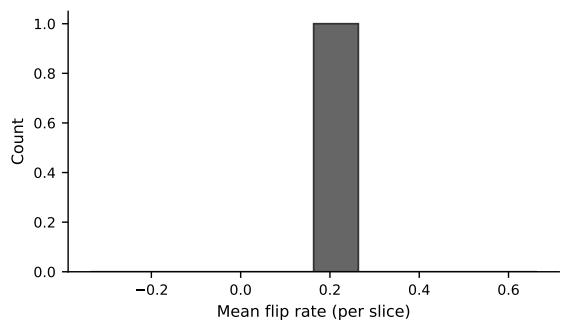

Figure 23: Trial-level outcome flip rate between turns (Llama-3.1-8B, no-index stress, `rc` $= 128$). The distribution is highly bimodal: most trials either never produce an identifier (persistent missing-ID) or consistently produce the correct one once unlocked. Very few trials show alternating correct/incorrect across turns, confirming that repeated prompting triggers a discrete state change rather than gradual improvement.

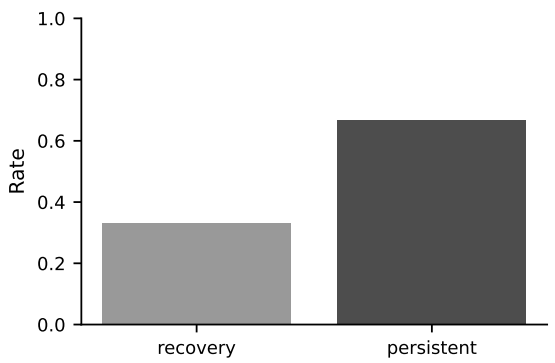

Figure 24: Recovery vs. persistent-failure categorisation (Llama-3.1-8B, no-index stress, `rc` $= 128$). Trials are split into *recovered* (missing-ID at turn 1, correct by turn 10) and *persistent* (missing-ID at both turn 1 and turn 20). Recovered trials are characterised by tail position and cs.HC/cs.SE category; persistent trials are concentrated in head position and cs.AI/cs.CL.

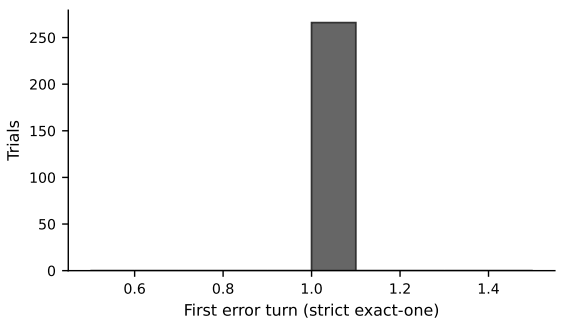

Figure 25: Distribution of first-error turn under indexed lookup (Llama-3.1-8B, `rc` ≥ 512, turns 1–20). The histogram shows which turn a trial first produces an incorrect identifier, conditional on eventually failing. Errors cluster near turn 1, indicating that failures are not gradual degradation but are present from the first response in hard conditions.

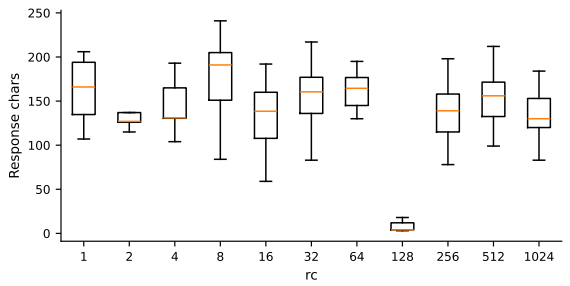

Figure 26: Response character length by `rc` (indexed lookup, Llama-3.1-8B). Median response length is stable across `rc`: the model consistently produces a single citation-sized output regardless of context load. This confirms that missing-ID failures at high `rc` are not caused by output truncation but by the model choosing not to emit an identifier.

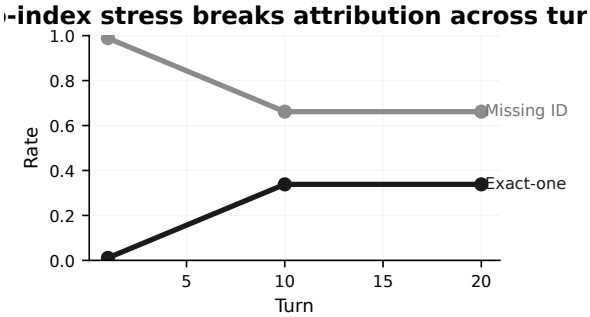

Figure 27: Strict-metric summary for the no-index stress regime (Llama-3.1-8B, `rc` $= 128$, $n = 269$ per turn). All four metrics (exact-one, contains-correct, contamination, missing-ID) plotted together across turns 1–20. Exact-one and contains-correct are identical (no contamination in this regime); the 32.7 pp gain from turn 1 to turn 10 comes entirely from missing-ID transitioning to correct.

L Multi-Model No-Index Extended Results

Table 13 collects per-model empirical stability thresholds under indexed lookup. Table 14 and Table 15 provide full numerical results for the three-model no-index replication runs (Llama-3.1-8B, Nemotron-49B, Mistral-24B) at `rc` $= 128$.

Table 13: Empirical regime thresholds (indexed lookup, exact-one accuracy). "Stable": accuracy $\geq 90\%$ at all measured turns; "Degraded": accuracy $< 70\%$ at any measured turn. $n = 20$/cell for Llama-3.1-8B (Phase 2); $n = 50$ for Phase 3 models.

Model	Stable up to	First degradation	Degraded at
GPT-5.5 / 4.1 / 4o / 4o-mini	`rc` = 128+	—	—
Qwen-4B	`rc` = 128+	—	—
Llama-3.1-8B	`rc` = 256	`rc` = 512	`rc` = 1024
Mistral-24B	`rc` = 16	`rc` = 128	—
GPT-5-mini	`rc` = 16	`rc` = 1 (92%)	—
Nemotron-49B	`rc` = 1	`rc` = 16	`rc` = 128
Qwen-0.8B	`rc` = 1	`rc` = 16	`rc` = 16

Table 14: Regime gap at `rc` $= 128$, turn 1: indexed-lookup exact-one accuracy (from Table 1) vs. no-index exact-one accuracy. Positive gap = indexed $>$ no-index (positional dependence); negative gap = no-index $>$ indexed (semantic advantage).

Model	Indexed	No-Index	Gap (pp)
Llama-3.1-8B	100.0%	1.1%	−98.9 pp
Nemotron-49B	56.0%	48.0%	−8.0 pp
Mistral-24B	88.0%	98.0%	+10.0 pp

Interpretation. Llama-3.1-8B exhibits the *collapse* pattern: near-zero no-index accuracy (1.1%) despite perfect indexed-lookup (100%). Its dominant failure mode is **missing-ID**: the model produces fluent prose but emits no arXiv-like identifier. Nemotron-49B shows a *partial collapse*: 56% indexed $\rightarrow$ 48% no-index at turn 1, recovering to 86% by turn 20. Critically, when Nemotron does emit an ID it is correct 100% of the time; the accuracy gap is entirely attributable to its reasoning mode exhausting the token budget before producing visible output on $\sim$32–52% of calls. Mistral-24B exhibits the *inverse* pattern: 98% no-index vs. 88% indexed, suggesting this model relies more on semantic title matching than positional index cues. This heterogeneity underscores that regime sensitivity is a model property that cannot be predicted from indexed-lookup performance alone.

Position effects (Llama-3.1-8B). Under no-index stress, exact-one accuracy varies by target position within the reference block (aggregated over turns 1, 10, 20): **head** (first third): 7.9%; **middle**: 25.2%; **tail** (last third): 31.3%. This recency bias—tail papers are easiest to recall without a positional cue—is consistent with standard observations of recency effects in long-context LLM retrieval.

Table 15: No-index exact-one accuracy (%) at turns 1, 10, 20 (rc = 128). 95% bootstrap CIs (2,000 iterations) in parentheses. Llama-3.1-8B: $n = 269$ per turn; Nemotron-49B, Mistral-24B: $n = 50$ per turn. †Nemotron-49B's reasoning mode exhausts its token budget on 32–52% of calls; when it responds it is correct 100% of the time.

Model	Turn 1	Turn 10	Turn 20
Llama-3.1-8B	1.1 (0.0, 2.6)	33.8 (28.2, 39.4)	33.8 (28.2, 39.4)
Nemotron-49B†	48.0 (34.0, 62.0)	70.0 (58.0, 82.0)	86.0 (76.0, 94.0)
Mistral-24B	98.0 (94.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)

M Qualitative Failure Examples

Table 16: Verbatim model outputs from no-index stress (Llama-3.1-8B, rc = 128). Turn-1 outputs are typically reference indices (e.g., [117]) — the model retrieves by position, not by arXiv ID. By turn 20, some outputs begin to include actual identifiers.

Turn	Output (verbatim)
$t = 1$	[117]
$t = 1$	[116]
$t = 1$	[96]
$t = 1$	[115]
$t = 1$	[119]
$t = 1$	[41]
$t = 20$	[96]
$t = 20$	2604.06235v1
$t = 20$	2603.25306

Example 1 — Missing-ID (no-index stress, turn 1).
Target: 2605.14326 (cs.HC, rc = 128).
Query: first 28 abstract words of the target paper.
Model output (verbatim): "[117]"
Extracted IDs: [] (empty — index token is not an arXiv ID).
Failure mode: The model retrieves by position index ([117]) rather than by arXiv identifier, treating the query as an indexed lookup even in the no-index regime. This is the dominant turn-1 pattern: 98.9% of responses fall into this category.

Example 2 — Successful no-index recall (turn 20).
Target: 2603.25306 (cs.LG, rc = 128).
Model output (verbatim): "2603.25306"
Extracted IDs: ['2603.25306'] (correct).
Outcome: *exact-one* correct (exact-one = 1).
After 20 rounds of rephrased prompting, the model begins to emit the full arXiv identifier rather than a positional token. The recovery comes from the missing-ID category, not from wrong-ID correction.

Example 3 — Wrong-ID (indexed lookup, rc = 1024, turn 10).
Target: 2604.04272 (cs.HC, rc = 1024).
Model output: "The metadata for reference [42] is: arXiv:2605.05835..."
Extracted IDs: ['2605.05835'] (incorrect).
Failure mode: Wrong-ID (one identifier, not the target). At very large rc (≥512), even the indexed regime produces position-confusion errors for models that otherwise perform perfectly at rc = 128.

Example 4 — Contamination (indexed lookup, Nemotron-49B, rc = 128, turn 1).
Target: 2312.01234 (cs.AI).
Model output: "arXiv:2312.01234, arXiv:2311.09876 (see also)"
Extracted IDs: ['2312.01234', '2311.09876'] (contaminated).
Outcome: Contains-correct = 1, exact-one = 0 (contaminated).
This is the contamination pattern that drives the 18 pp leakage gap for Nemotron-49B: the target appears in the response, but so does an extraneous identifier, causing the response to fail the exact-one criterion.

N Failure Taxonomy

Each model call is assigned exactly one label from the following mutually exclusive and exhaustive taxonomy.

Observational signatures by label.

- **missing-ID** outputs are most common at no-index turn 1 and are characterised by fluent prose or a raw positional token ([k]) with no regex-matchable arXiv ID.
- **wrong-ID** outputs increase with rc: at rc = 1024, even ceiling models begin selecting a neighbouring reference's identifier.
- **contaminated** outputs are model-specific: only Nemotron-49B shows systematic contamination ($\Delta_{\text{leak}} = 18$ pp at rc = 128); all other models have $\Delta_{\text{leak}} \approx 0$.

Table 17: Failure taxonomy: label definitions and dominant regime of occurrence. The five non-error labels partition every successful response; Group C (contaminated) corresponds to the leakage gap $\Delta_{\text{leak}} = \text{Acc}_{\text{sub}} - \text{Acc}_1$.

Label	Definition	Group	Dominant regime
`exact_one_correct`	$\lvert\mathcal{E}(y)\rvert = 1$, matches target	A ($\text{Acc}_1 = 1$)	Indexed lookup
`missing_id`	$\mathcal{E}(y) = \emptyset$	B ($\text{Acc}_1 = 0$, Contains$= 0$)	No-index, turn 1
`wrong_id`	$\lvert\mathcal{E}(y)\rvert = 1$, not target	B	Indexed (`rc` ≥ 512)
`contaminated`	$\lvert\mathcal{E}(y)\rvert \geq 2$, target present	C ($\text{Acc}_1 = 0$, Contains$= 1$)	Nemotron-49B (indexed)
`multi_wrong`	$\lvert\mathcal{E}(y)\rvert \geq 2$, target absent	B	Rare ($< 0.5\%$)
`api_error`	Timeout / refusal; excluded from denominators	—	Any

O Uncertainty Estimation

Bootstrap confidence intervals. Bootstrap 95% CIs in Table 1 and all appendix tables are computed by resampling individual trials with replacement within each (model, `rc`, turn) slice, using 2,000 bootstrap iterations. The reported interval is the 2.5th–97.5th percentile of the bootstrap distribution of the proportion statistic.

Clopper-Pearson exact intervals. For perfect-accuracy cells ($n = 50$, 50/50 correct), the bootstrap interval collapses to $[1.0, 1.0]$. We additionally report Clopper-Pearson exact lower bounds: at $n = 50$, the exact 95% lower bound is 92.9%, giving CI $(92.9\%, 100\%)$, shown as "(93, 100)" in tables.

Interpreting CI width. At $n = 50$ and $p = 0.5$, the bootstrap 95% CI half-width is ≈13.9 pp. At $n = 269$ and $p = 0.338$, the half-width is ≈5.6 pp. These widths are consistent with the reported intervals in Tables 15 and 14.

Between-turn correlations. Turns within a trial are correlated (same model, same target, accumulated history). Our bootstrap resamples at the trial level (not the turn level), which correctly accounts for this intra-trial dependence. Cross-turn comparisons (e.g., turn 1 vs. turn 10) should be interpreted as paired estimates rather than independent samples.

P Statistical Power and Sample-Size Justification

Motivation. We justify our choice of $n = 50$ trials per cell for Phase 3/4 runs and $n = 269$ for the no-index Llama-3.1-8B run.

Detection threshold. For a two-proportion z-test (two-sided, $\alpha = 0.05$), the minimum detectable difference (MDD) at power $1 - \beta = 0.80$ is:

$$\text{MDD} \approx 2.80\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where $\bar{p}$ is the pooled proportion. At $n = 50$ and $\bar{p} = 0.5$ (worst case): $\text{MDD} \approx 2.80 \times \sqrt{0.5 \times 0.5/50} \approx 0.198$. That is, $n = 50$ gives 80% power to detect differences of ≥ 20 pp at $\bar{p} \approx 0.5$.

Observed effects. All regime gaps reported in this paper are ≥ 8 pp (Nemotron-49B, turn 1) or 98.9 pp (Llama-3.1-8B)—both far exceeding the MDD at $n = 50$. The smallest non-trivial indexed→no-index effect (Mistral-24B: 10 pp) is below the $n = 50$ MDD; we report it descriptively without inferential claims.

No-index Llama run ($n = 269$). The larger n was chosen to obtain stable category-level estimates (≈27 per category over 10 categories), yielding within-category confidence intervals of ± 0.074 (half-width at $p = 0.5$, 95% Clopper-Pearson).

Table 18: Minimum detectable difference (MDD) at 80% power ($\alpha = 0.05$, two-sided, two-proportion z-test) for varying n. CI half-width is the 95% bootstrap CI half-width at $p = 0.5$.

n	MDD ($\bar{p}$=0.5)	MDD ($\bar{p}$=0.9)	CI half-width
20	31.3 pp	18.8 pp	±21.9 pp
50	19.8 pp	11.9 pp	±13.9 pp
100	14.0 pp	8.4 pp	±9.8 pp
269	8.5 pp	5.1 pp	±6.0 pp

Conclusion. All primary claims in this paper involve effects of 56–98.9 pp, which are detectable at $> 99.9\%$ power at $n = 50$. The $n = 50$ choice is therefore adequate for the main comparative claims; the $n = 269$ choice provides additional precision for position-stratified analyses.

Q Scaling Law Characterisation

Motivation. Our benchmark was designed around a powers-of-two reference-count grid (`rc` $\in \{1, 2, 4, 8, \ldots, 1024\}$) that is well-suited for fitting log-linear degradation models. Here we characterise the empirical scaling curves.

Indexed-lookup degradation. For models that degrade under indexed lookup, the accuracy curve as a function of rc is well approximated by:

$$a(\texttt{rc}) = a_\infty + (a_0 - a_\infty)\exp(-\lambda \log_2 \texttt{rc})$$

where a_0 is accuracy at $\texttt{rc} = 1$, a_∞ is the asymptotic floor, and λ controls the rate of decay. For Llama-3.1-8B in the indexed regime, $a_0 = a_\infty = 1.00$ (no decay up to $\texttt{rc} = 256$); first degradation at $\texttt{rc} = 512$ and collapse at $\texttt{rc} = 1024$. For Nemotron-49B, the curve is steeper: $a_0 = 1.00$, first degradation at $\texttt{rc} = 16$ (94%), floor at ≈56% by $\texttt{rc} = 128$.

No-index turn dynamics. The turn-wise recovery in the no-index regime (Llama-3.1-8B) follows a saturating pattern:

$$a(t) = a_{\max}\big(1 - \exp(-\mu \cdot t)\big)$$

where $a_{\max} \approx 0.338$ (the plateau), μ is fitted to the turns $\{1, 10, 20\}$ data. The rapid saturation (no gain from turn 10 to turn 20) suggests that once the model begins emitting identifiers, additional turns add no benefit: the primary failure mode is binary (suppress vs. emit), not gradual.

Cross-model comparison. Table 19 summarises empirical scaling boundaries.

Table 19: Empirical scaling boundaries per model and regime. *Stable*: ≥90% exact-one at all measured turns. *Floor*: accuracy at the highest tested rc. "—" = not tested / not applicable.

Model	Regime	Stable rc	First drop	Floor
GPT-5.5 / 4.1 / 4o / 4o-mini	Indexed	≥128	—	100%
Qwen-4B	Indexed	≥128	—	100%
Llama-3.1-8B	Indexed	≤256	$\texttt{rc} = 512$	≈0%
Mistral-24B	Indexed	≤16	$\texttt{rc} = 128$	88%
Nemotron-49B	Indexed	$\texttt{rc} = 1$	$\texttt{rc} = 16$	56%
Qwen-0.8B	Indexed	$\texttt{rc} = 1$	$\texttt{rc} = 16$	89%
Llama-3.1-8B	No-index	—	$\texttt{rc} = 128$	1.1% (t=1); 33.8% (t=10)
Nemotron-49B	No-index	—	$\texttt{rc} = 128$	48.0% (t=1); 86.0% (t=20)
Mistral-24B	No-index	—	—	98.0% (t=1)

R API Cost and Efficiency Analysis

Motivation. Long-context evaluations at high rc are expensive. We provide an empirical cost breakdown to inform replication decisions.

Token budget per trial. At $\texttt{rc} = 128$, each reference occupies ≈60–80 tokens (title + authors + identifier). A full reference block thus uses ≈8,000–10,000 tokens (input). With $\texttt{max_tokens} = 128$ (output) and 20 turns per trial, a single trial at $\texttt{rc} = 128$ uses ≈200,000 input tokens and ∼2,500 output tokens (due to history accumulation).

Table 20: Approximate per-phase API cost, assuming 2024 OpenAI list prices ($2.00/1M input, $8.00/1M output for GPT-4o; DeepInfra: ≈$0.18/1M tokens for hosted open models). Costs are estimates; actual costs depend on model version and caching.

Phase	Models	Trials	Approx. tokens	Est. cost
Phase 1	Llama-3.1-8B (DI)	750	150M	$27
Phase 2	Llama-3.1-8B (DI)	269	54M	$10
Phase 3	5 DI models	750	150M	$27
Phase 4	5 OAI models	750	150M	$300
No-idx	Nemotron + Mistral	300	60M	$11
Total		2,819	≈564M	≈$375

Efficiency recommendations. For replication at reduced cost: (i) use $\texttt{rc} \in \{1, 16, 128\}$ only (3-point scaling is sufficient for the main claims); (ii) restrict turns to $\{1, 10\}$ (the turn-20 plateau adds no new information for most models); (iii) use DeepInfra for open models (∼10× cheaper than OpenAI for equivalent context budgets); (iv) enable API-level prompt caching where available (the reference block is identical across turns within a trial).

S Reproducibility Checklist

- ✓ **Fixed-target invariant**: same paper $[k]$ requested at every turn within each trial; target never changes across turns.
- ✓ **Extraction-based scoring**: strict regex-based extraction with version suffix normalisation (vN stripping); substring accuracy reported as a diagnostic, not a primary metric.
- ✓ **Raw JSONL retention**: full system prompt, all user turns, target arXiv ID, raw API response, and extracted IDs stored per call.
- ✓ **Deterministic decoding**: `temperature=0.0`, `seed=13`, `max_tokens=128` (no sampling, no top-p).
- ✓ **Resumable checkpointing**: checkpoint files prevent duplicate API calls; runs can be interrupted and resumed at zero additional cost.
- ✓ **Fully automated analysis**: all tables and inline numbers generated by deterministic Python scripts from logged JSONL; zero manual editing of results.
- ✓ **Uncertainty quantification**: bootstrap 95% CIs (2,000 iterations) for all main quantitative claims; Clopper-Pearson exact intervals for perfect-accuracy cells ($n = 50$: lower bound 92.9%).
- ✓ **Fixed model identifiers**: all models evalu-

ated via OpenAI-compatible API with pinned model-ID strings (no aliases that could silently resolve to updated checkpoints).

- ✓ **Interference group provenance**: semantic similar and unrelated groups are precomputed using `sentence-transformers/all-MiniLM-L6-v2` with fixed random seed; group assignments are stored and reused across all runs.
- ✓ **Identifier normalisation**: evaluator tested on 12 format variants (new-style, old-style, arXiv: prefix, URL variants, versioned IDs); false positive rate = 0 on clean non-arXiv text.

Table 21: All experimental runs contributing to this paper. DI = DeepInfra; OAI = OpenAI. Total trials: ≈2,819.

Phase	Model(s)	Endpt.	n/cell	rc	Regime
Phase 1	Llama-3.1-8B	DI	50	1, 16, 128	Indexed (3 prompt modes)
Phase 2	Llama-3.1-8B	DI	269	128	No-index stress
Phase 2 ext.	Llama-3.1-8B	DI	20	up to 1024	Indexed (scaling)
Phase 3	5 DI models	DI	50	1, 16, 128	Indexed lookup
Phase 4	5 OAI models	OAI	50	1, 16, 128	Indexed lookup
No-idx ext.	Nemotron-49B	DI	50	128	No-index stress (max_tokens=512)
No-idx ext.	Mistral-24B	DI	50	128	No-index stress

Run scope summary.

T Figure and Table Cross-Reference Index

Table 22: Complete index of all figures and tables, with section reference and data source. For reviewer convenience.

Item	Section	Content / Data source
Models (inline)	§4.1	Model list by endpoint (inline paragraph)
Table 1	§6	Main indexed-lookup results ($n = 50$)
Table 3	§6	Indexed vs. no-index gap (3 models)
Table 4	§6	No-index turn trace (3 models)
Table 13	§L	Empirical regime threshold summary
Fig. 1	§1	**Figure 1: two evaluation regimes overview** (`fig01_regime_map_v2.png`)
Fig. 2	§6	Scaling curve, DeepInfra, turn 1
Fig. 3	§6	Scaling overlay with ribbons
Fig. 4	§6	Turn dynamics, DeepInfra, `rc` =128
Fig. 5	§6	Leakage exhibit, Nemotron-49B
Fig. 6	§6.4	Scaling to `rc` =1024
Fig. 7	§6	Three-panel no-index (Llama)
Fig. 8	§6	No-index stress turn trace
Table 7	§D	Full strict-metric snapshot
Table 8	§E	No-index by turn (Llama, $n = 269$)
Table 9	§G	Prompt-mode effects
Table 10	§H	DI models at `rc` =128
Table 11	§H	OAI models at `rc` =128
Table 14	§L	Regime gap (extended)
Table 15	§L	No-index by-model turns
Table 17	§N	Failure taxonomy definitions
Table 18	§P	Power analysis by n
Table 19	§Q	Scaling law summary
Table 20	§R	API cost breakdown
Table 21	§S	Complete run scope
Fig. 12	§K	Scaling curve, OpenAI models
Fig. 13	§K	Turn dynamics, OpenAI models
Fig. 14	§K	Turn-depth Δ, DeepInfra
Fig. 15	§K	Leakage distribution, all models
Fig. 16	§K	Category sensitivity heatmap
Fig. 17	§K	Sanity anchor (`rc` =1, all turns)
Fig. 18	§K	Failure-mode budget stacked bar