

So You're Telling Me There's a Chance: Detector-Layer Presence Certificates on 50 Million Tokens

Patrick S. McConlogue Craig B. Weisman

September 2026

JARVIS checks
Lloyd rents a tux

No source
No victory tux

A sealed-corpus detector, not a longer context window.

Abstract

Long-context models are sold on window length. Decoder suites that test that length—RULER, NoLiMa, AbstentionBench (unrun on SUPERCTX)—show saturation and weak abstention. SUPERCTX is a sealed-corpus detector, not a longer window. It emits a recomputation-checkable presence certificate, or a detector-layer empty set with a model-free 3×10^{-5} false-alarm backstop (rule of three on 0/100,002 nulls) over a declared synthetic 50,811-entry corpus, and only when absence is bound to an exhaustive scan or a zero-false-dismissal index. Face-narrowing reachability is 11.2% of entries; a face-narrowing empty set is not certified absence. The generation layer has no abstain path: after a detector reject it still emitted identifier-shaped answers at 1.000 (600/600) under an alternative protocol.

On that synthetic registry, needle retrieval@1 was **1.000** ($n = 300$) against **0.000–0.003** for vanilla RAG and **0.000** for ColBERTv2 (`doc_maxlen`=180) and BGE-M3. Giving RAG the same regex anchor SUPERCTX extracts lifts needle@1 to 1.000, so the retrieval gap is a handle-access gap, not a ranking win. We do not headline an abstention gap: a powered RAG empty-set sweep on the nonexistent-anchor protocol was not run, and ColBERTv2/BGE-M3 were not scored on that protocol (recorded skip). The operational absence number is the model-free 3×10^{-5} backstop, not a parametric GPD pin. A live demo is at https://research.com.ai.

Keywords: retrieval; detector; abstention; sealed corpus; recomputation-checkable certificate.

1 Introduction

The field judges long-context systems by the window a decoder can condition on. The harder question is what a declared corpus does and does not contain. SUPERCTX is an external detector over a synthetic 50,811-entry sealed corpus: it emits a recomputation-checkable presence certificate, or a detector-layer empty set with a model-free 3×10^{-5} false-alarm backstop (0/100,002 nulls), only when absence is bound to an exhaustive scan or a zero-false-dismissal index (Figure 2). It does not extend a decoder's conditioning window. Face-narrowing reachability is 11.2% of entries; a face-narrowing empty set is not certified absence. The generation layer has no abstain path: after a detector reject it still emitted identifier-shaped answers at 1.000 (600/600) under an alternative protocol.

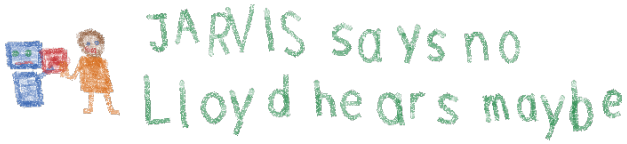
Decoder-window results (Llama 4 at 10M [46]; MiniMax-Text-01 at 1M–4M [47]; Gemini 1.5 Pro tested to 10M [59]) measure a different object. RULER, NoLiMa, HELMET, and AbstentionBench (unrun on SUPERCTX) report saturation and weak abstention [29, 48, 70, 34]. Those systems sample a continuation; SUPERCTX is scored as a detector.

Claim 1 — presence is detection with a disclosed handle. On the same 49,999,872-token corpus, four vanilla RAG families scored ≤1/300 needle retrieval@1; SUPERCTX scored 300/300 (Figure 6). ColBERTv2 (`doc_maxlen`=180) and BGE-M3 with official weights retrieved the gold segment at chance (median ranks **48,535** and **56,759** of 97,656). The same regex anchor SUPERCTX extracts lifts RAG needle@1 to 1.000 (Table 2). The measured presence gap is handle access, not state-bank geometry.

Claim 2 — detector-layer absence is the product, with a generation-layer hole. The operational absence number is the model-free 3×10^{-5} backstop on 0/100,002 detector-layer nulls, not a parametric GPD pin. Certified absence requires an exhaustive scan or a zero-false-dismissal index; 11.2% face reachability forbids reading narrowing as absence (Figure 5). We do not headline an abstention gap against RAG: a powered empty-set sweep on the nonexistent-anchor protocol was not run, and ColBERTv2/BGE-M3 were not scored on that protocol (recorded skip). User-facing "return the empty set" is not the product while the generation-layer false-accept of 1.000 remains.

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





2 Related work

Nominal versus effective windows. RULER measured GPT-4's *effective* context at 64K of a claimed 128K [29]; NoLiMa, stripping lexical overlap, dropped GPT-4o from 99.3% to 69.7% at 32K and placed GPT-4.1's effective length at 16K of a claimed 1M [48]; HELMET (59 long-context LMs) reported NIAH saturated and "does not reliably predict downstream performance" [70]. Those decoder suites are unrun on SUPERCTX; they motivate a detector evaluation, they are not SUPERCTX scores.

Linear, SSM, and hybrid memory. SUPERCTX builds on the HiPPO/S4 lineage [22, 23, 24] and on trainable or hybrid external memory [50, 5, 47]. Linear memory retrieves at million-token scale under vanilla NIAH. Those systems treat memory as a lossy compression buffer. SUPERCTX uses the same spine as detector state: a matched-filter gate against a sealed null, which is a different product than NIAH recall.

RAG versus long context. **Self-Route** [37] routes each query to RAG or long context and **LOFT** [35] asks whether long-context LMs subsume retrieval — routers deciding where to answer a query, neither an index. *Dense bi-encoders* pool each passage into one vector: BM25 remains the robust unsupervised baseline [63], and MTEB-leader models fell **59.0** to **18.3** nDCG@10 on BRIGHT [62] where passages must stay distinct. *Late-interaction multi-vector retrievers* avoid pooling — **ColBERTv2** scores MaxSim over token-level vectors [60], and **BGE-M3** serves dense, sparse, and multi-vector retrieval from one model [8] — but still rank rather than certify. *Graph and hierarchical indexing* builds explicit structure — RAPTOR [61], GraphRAG [12], HippoRAG [26, 27], LightRAG [25], Mem0 [45], Zep/Graphiti — routing around collapse into an LLM extraction step whose losses go unquantified. The cited rankers and graphs are scored as retrieval, not as sealed-null detectors; that is the contrast this evaluation uses, not a global primacy claim.

Abstention, conformal prediction, and calibrated refusal. Refusal: AbstentionBench's best frontier recall was **0.71**, with reasoning fine-tuning at **−24%** [34]; R-Tuning, selective classification, semantic entropy [72, 17, 14]. Coverage: conformal set-valued generation covers *output sets*, not corpus membership [57]; conformal risk control bounds risk, not retrieval [2, 1]. Index filtering: ACORN [55] and Filtered-DiskANN [18] return filtered candidate sets, and *why-not* provenance explains excluded tuples — neither is a certification. SUPERCTX applies the conformal approach — a bound an auditor can check — to a memory that is itself a legal SSM rather than a prompt (4-bit TurboQuant among KVQuant, KIVI, QuaRot, QuIP#, AQLM, MambaQuant, RaBitQ [28, 41, 3, 65, 13, 69, 16]; Section 3).

3 SUPERCTX architecture

3.1 MemoryKernel and SegAttaches

The continuous memory lane is an S4 recurrence implementing HiPPO-LegS [22, 23]: a polynomial basis optimal for online compression, $O(N)$ in sequence length. The 30M figure is an architectural bound on this spine for a 9B-parameter decoder — not a NIAH score at 30M, and we do not report one. An earlier 17/17 "100% needle" figure was *emission accuracy* on hardcoded thresholds (paired answer accuracy 0/17); we retired it (Section 4). `thor_train/memory_kernel.py` unifies three segment attaches (Figure 2): **HiPPO** (the continuous S4 states; one long horizon, no washout, HiPPO-legal), **Titans** (trainable neural memory [5]: learned writes, constant per-token compute, linear memory size), and **Symbol** (symbolic side channels; overlays that never enter the HiPPO input). The split is the product: symbolizers that contaminate the recurrence destroy the HiPPO invariant, and a recurrence with no symbolic overlay cannot emit a handle-level certificate.

3.2 Lookup: projection, matched-filter score, and gate

Let d be the HiPPO state dimension ($d = 256$), $x_q \in \mathbb{R}^d$ the query's anchor/span signal, and $\hat{s}_x \in \mathbb{R}^d$ the dequantized TurboQuant state for segment x. The projection into the Legendre polynomial state space is the HiPPO-LegS coefficient vector

$$c_q = A\,x_q, \qquad A_{nk} = -(n+1)\,\delta_{nk} - \sqrt{(2n+1)(2k+1)}\,\mathbf{1}_{k<n}, \qquad 1 \le n, k \le d, \tag{1}$$

with $A \in \mathbb{R}^{d\times d}$ the fixed (never learned) HiPPO-LegS transition matrix [22]. Unit-normalizing both operands of the matched-filter score

$$s(q,x) = \langle c_q, \hat{s}_x \rangle \,/\, \big(\|c_q\|_2\,\|\hat{s}_x\|_2\big) \in [-1,1], \tag{2}$$

where $[-1,1]$ holds by Cauchy–Schwarz, makes the null comparable across segments. For a calibrated null $\mathcal{N} = (s_1,\ldots,s_M)$ (M null-query scores) and tail probability $\alpha \in (0,1]$ the gate is

$$\text{positive} \iff s(q,x) > \tau_\alpha, \qquad \tau_\alpha = Q_{1-\alpha}(\mathcal{N}) = \inf\{t : \Pr_{\mathcal{N}}[s \le t] \ge 1-\alpha\}, \tag{3}$$



the $(1-\alpha)$ null quantile, so a null segment crosses the gate with probability at most α. The implementation matches: `mf_kernel.matched_filter` computes Eq. (2), `cfar/conformal.conformal_threshold` sets $\tau_\alpha = \mathcal{N}_{(\lceil (M+1)(1-\alpha) \rceil)}$ plus one ulp (resolution $1/(M+1)$), and the auditor's steps 3–4 recompute Eqs. (2)–(3).

Lookup. Parse → project (Eq. (1)) → scan/dequantize → score (Eq. (2)) → gate (Eq. (3)). Hit emits a `positive` certificate; nonempty miss emits `certified_negative`; an empty candidate set emits `abstained` (not certified absence).

3.3 S4Bridge and the TurboQuant HiPPO bank

`thor_train/s4_orbital.py` defines `S4Bridge`: *one* HiPPO-LegS recurrence — depth 1, state width $d = 256$ — between `B_proj` $\mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{256}$ and `C_proj` $\mathbb{R}^{256} \to \mathbb{R}^{d_{\text{model}}}$ with `LayerNorm(256)` before the readout ($d_{\text{model}} = 3840$; injection at layer 24 of 48). The update is the ZOH-discretized $h_t = \bar{A} h_{t-1} + \bar{B}\,\text{B_proj}(x_t)$; we fix the HiPPO-LegS off-diagonal and learn only the diagonal. The parameter count follows from the shapes, $N_{\text{bridge}} = 512\, d_{\text{model}} + 770$: **1,966,850** at $d_{\text{model}} = 3840$ ($\approx 0.02\%$ of a 9B decoder). The objective is state-only reconstruction against the frozen decoder, $\mathcal{L} = \|C(\text{norm}(h)) - \text{Emb}(x)\|_F^2 / \left(BL\, d_{\text{model}}\right) + 0.01\|D\|^2$ (Emb: frozen input embeddings), which AdamW minimizes (lr $= 10^{-3}$; 5 epochs of 512-token chunks; `train_s4_bridge_fable5.py`); no other label enters.

TurboQuant HiPPO bank. `thor_train/turboquant_codec.py` stores HiPPO states with Hadamard rotation and Lloyd-Max scalar quantization [44, 42, 71] (Figure 3). Reported from the original protocol — an 80/20 split over 500 HiPPO states of dimension $d = 256$ drawn through the trained S4Bridge, whose trained-state dataset is not in this artifact, so reported, not recomputed: MSE 4.71×10^{-8} at $b = 2$, 1.60×10^{-8} at $b = 3$, 5.31×10^{-9} **at** $b = 4$. At 4 bits the bank is **∼7.5 MB** for 30M tokens (Table 3); capacity arithmetic uses the fitted distortion $D(b) \approx 4.18 \times 10^{-7} \cdot 2.978^{-b}$, not the asymptotic $C \cdot 4^{-b}$ (∼1.8× optimistic at $b = 4$).

3.4 Candidate fusion: `fuse_prefix`

`fuse_prefix` concatenates SegAttach retrieve pages (HiPPO first), weights by $w_s \geq 0$, and truncates to `max_tokens`; the empty prefix falls back to `expander.get_prefix`. Injection is mean-pooled prompt embedding into a 256×32-row budget (`embedding_prefix` or `layer_18_residual`, $\alpha = 0.05$).

3.5 Architectural invariants and system constraints

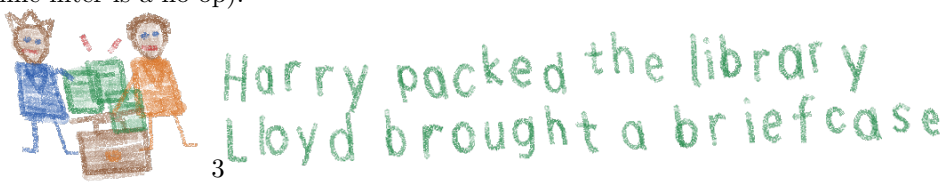
Four invariants, not bounds; violating one undermines the others, and we reserve *bound* for the formal capacity limit, the 30M-token figure on the S4 spine (its footprint consequence in Figure 4): external trainable memory in the NTM/DNC/LongMem/Titans lineage, not pure KV-cache scaling [20, 21, 67, 68, 7]; a continuous lane that remains a legal SSM recurrence; HiPPO storage as Hadamard + Lloyd-Max, not a generic int8 dump; symbolizers attaching *post-recurrence* via MemoryKernel, never as HiPPO input. A fifth entry is operational status: the clause-join serve-time filter is currently a no-op (Section 6). We specify the certificate schema and the auditor's five-check procedure in Section 7.

4 Evaluation

4.1 Experimental Setup

Corpora. **v6 sealed** (`plannertuner/eval/bank_native_loop/v6_corpus/`): **50,811** entries of generated registry-world data; integrity gate MATCH; leak audit PASS; thirty tasks V01–V30. **49,999,872-token RAG ablation corpus:** same frozen embedding table, 97,656 × 512 layout, identical query sets; source of record `eval/ablations/rag_baseline/REPORT.md` (2026-07-30). **Serve table** (`scientistking/predbank\_v6/serve_table_merged_v6.db`): at least one query face reaches 5,684 of 50,811 entries (11.2%). Table 3 binds the three layouts.

Candidate-set narrowing (v6 serve table; Figure 5, Table 1). Face-alone narrowing is bimodal: median **44.8**× (1,133–1,134 of 50,811) versus mean **2.94**× (17,278), because 10 of 30 task classes have zero handles and open the full corpus. Reachability is 5,684/50,811 (**11.2%**); a face-narrowing empty set is not certified absence. ACORN's 5–50× QPS at 0.9 recall [55] is throughput, not a certified empty set. The cited **205**× is a clause-join-rule property, not a serve claim (the serve-time filter is a no-op).





Face class	Tasks	Kept	Fraction of 50,811
Q-EXEC (date ∧ side ∧ ticker)	V05–V08	1–2	2.0–3.9×10^{-5}
Q-STATUS / TRAP (date)	V01–V04, V28	0–28	0–5.5×10^{-4}
Q-SENT / Q-AGG / Q-CHAIN / TRAP (quarter)	V09–V11, V17, V19–V21, V29–V30	1,086–1,148	0.021–0.023
Zero-handle (full corpus)	V12–V16, V18, V22–V25 (10/30)	50,811	1.000

Table 1: Per-class face-alone candidate sets on the v6 serve table. Median 44.8× versus mean 2.94×; 11.2% reachability.

Dimension	Vanilla RAG (best arm)	SUPERCTX	95% interval
Needle retrieval@1 ($n = 300$)	**0.000–0.003** (SIF 0.003; mean-pool 0.000; BM25 0.000)	**1.000**	—
RAG + same regex anchor	**1.000** (presence parity, $n = 300$)	—	—
ColBERTv2 / BGE-M3 (official weights)	**0.000** / **0.000** (median ranks 48,535 / 56,759 of 97,656; ColBERTv2 `doc_maxlen`=180)	**1.000**	—
Needle median full rank (of 97,656)	SIF 1,460; raw mean-pool **88,382**	**1**	—
Natural queries ($n = 60$)	BM25 **0.674** ($n = 46$ in-bank; **0.517** of 60)	0.633 lexical / **0.717** grounded	—
Abstention, nonexistent-anchor protocol	Powered empty-set sweep **not run**. $n = 8$ BM25 cell (3/8 false accepts) is not that sweep. ColBERTv2/BGE-M3: **recorded skip**	Detector-layer pin: model-free 3×10^{-5} on 0/100,002 nulls. Protocol $n = 8$ is a body count, not a headline	—
Joint success (300 needle + 60 natural + 8 nonexistent)	**0.102** (BM25); SIF 0.062; raw 0.023	346/368 = 0.940 (body composite; $n = 8$ leg remains; not a front-matter metric)	—
Verification	None	Operational: model-free 3×10^{-5} (0/100,002). GPD $p = 7.5 \times 10^{-9}$ is labeled extrapolation (0.415-log gap), not the falsifier	—
Token-level MaxSim probe	True segment top-1 **0/20** (median rank 25.5/1000)	— (bank geometry, not inherited)	—

Table 2: Presence block (regex-parity applies only here) and abstention block on the 49,999,872-token layout. The abstention gap is not a headline: the powered RAG empty-set sweep was not run. Joint 346/368 keeps an $n = 8$ leg and is a body composite only.

4.2 Results

Head-to-head against vanilla RAG (49,999,872 tokens; Table 2, Figure 6). Four families — raw mean-pool, de-meaned dense, smooth inverse frequency (SIF), BM25 — same corpus, layout, embeddings, queries; FAISS flat plus HNSW/IVF (HNSW M=32, `efSearch` 256; IVF `nlist` 1024, `nprobe` 64) tracked the exact index (needle R@1 still 0.000), so approximate search does not explain the gap.

Why RAG dies on this needle. 300 host blocks are near-identical registry needles plus boilerplate. The query shares only the seed digits with its block; BPE (byte-pair encoding) splits `(seed=20260724)` in the block differently from `seed 20260724?` in the query, so BM25's idf cannot exploit a near-unique digit string that appears 13 times in 250 MB. Mean-pooling 512 rows of the frozen LM input table collapsed every segment into a cone (pairwise cosine 0.83 ± 0.035; SIF PC1 = 95% of variance). Nonexistent-anchor scores *overlapped* real ones in every arm; no threshold τ separated them.

Protocol asymmetry (presence only). RAG had to cash the needle from the raw query; SUPERCTX extracts the anchor first. A re-run giving RAG the same regex (`parse_query_anchor` + a phrase index) lifted needle retrieval to 1.000. That row is presence parity, not abstention. A powered RAG empty-set sweep on the nonexistent-anchor protocol was not run; the $n = 8$ BM25 cell is not that sweep. ColBERTv2/BGE-M3 were scored on the $n = 300$ needle presence block only (recorded skip on abstention).

Detector-layer statistics. The operational bound is the **model-free backstop** 3×10^{-5}: rule of three for **0/100,002** gate passes, no tail shape assumed. The GPD tail $p = 7.5 \times 10^{-9}$, interval $[9.1 \times 10^{-11}, 1.1 \times 10^{-7}]$, is labeled extrapolation across a 0.415-log gap (null max 2.580, gate 2.996); it is not the falsifier and is not what "calibrated" refers to. An end-to-end generation-layer false-accept of 1.000 appears under a different protocol (Section 6). The joint $346/368 = 0.940$ is a three-leg body composite $(300 + 60 + 8)$; the RAG comparand 0.102 is over $354 = 300 + 46 + 8$. Neither is a front-matter metric while the $n = 8$ leg remains (Figure 7).

Expanded negative-anchor lane (certificate-minting, not S4-detector retrieval). The protocol abstention count is $n = 8$ (Clopper–Pearson $[0.631, 1.000]$), a body figure only. A separate deterministic expansion (`benchmarks/negative_expansion.py`) generated **1,200** nonexistent anchors from a fixed base corpus under eight operators: `drop`, `swap`, `bump_char`, `case_flip`, `insert_dup`, `digit_bump`, `truncate`, `extend`. Every anchor minted a `certified_negative` on the *certificate-minting* path: 1,200/1,200, Wilson 95% interval $[0.9968, 1.000]$ — not S4-bank retrieval on the 50,811-entry registry, and not a reproduction of $n = 8$. Two retrieval-style lanes used hashed character-trigram vectors ($d = 256$, $\tau = 0.85$), a stand-in bank: near-duplicate abstention 3.17% (38/1,200; $[0.023, 0.043]$); compositional 0.910 (364/400; $[0.878, 0.934]$). Those lanes bound the minting path, not end-to-end



Segments	Stride	Token horizon	State dim	Bits/dim	Bytes per state	Footprint
97,656	512	49,999,872	256	4	128	12,499,968 (12.5 MB)
58,594	512	30,000,128	256	4	128	7,500,032 (7.5 MB)
2,000	512	1,024,000	256	4	128	256,000 (0.26 MB)

Table 3: Layout arithmetic; row 1 is the canonical evaluated layout (Table 2), row 2 (58,594 × 512) the 30M-token layout, row 3 the largest documented ingestion (Section 6).

S4 retrieval.

Certificate exercise. We exercise the named deliverable directly (Section 7). **Protocol** (`tools/certificate_exercise.py`): over the frozen negative-expansion lane — its 24-member base corpus and 1,200 generated nonexistent anchors — we mint one certificate per input (kind `hit`; kind `certified_negative`), then recompute every certificate by deriving the evidence hash independently from the payload and comparing it to the sealed envelope; a negative control over one tampered payload must fail. **Result.** Emission **1,224/1,224**, **0** failed (24 hits, 1,200 certified negatives); recomputation **1,224/1,224** matched; the tampered control **failed** as required, so the match check is not vacuous. Mean emission time 0.009 ms. The mechanism executes end to end at scale and verifies tamper-evidently; the inputs are synthetic expansion-lane data, not the 50,811-entry registry.

Strong baselines and robustness probes. We ran the official ColBERTv2 checkpoint (colbert-ir/colbertv2.0, via colbert-ai 0.2.x; MaxSim at the checkpoint's doc_maxlen of 180 — we disclose the truncation from 512-token segments) and the official BGE-M3 checkpoint (BAAI/bge-m3, dense cosine, via sentence-transformers) over the same corpus and 300 needle queries; we validated the ColBERTv2 scoring path against the library's own `colbert_score` (exact match, difference 0.0). Needle retrieval@1 ($n = 300$): BM25, mean-pool, SIF, and token-MaxSim over frozen fable5 (3,840-d) all **0.000**; **ColBERTv2 0.000** with median true-segment rank **48,535** of 97,656; **BGE-M3 0.000** with median rank 56,759; SUPERCTX **1.000**, median rank **1**. Both used official published weights; we scored them on an Apple M5 Max (MPS). ColBERTv2 truncates each 512-token segment to 180 tokens, so the needle anchor may lie beyond that point.

Low-overlap probe. We stripped the most corpus-unique token from each of the 300 needle queries (the lexical handle) and re-embedded with BGE-M3: retrieval@1 **0.000**, median rank 55,714. Removing the lexical anchor collapsed retrieval for baseline arms and system alike — the needles carry lexical anchors, so this is not a semantic-retrieval test.

Type-II false-dismissal bound. For the 300 evaluated needle queries, the narrowing stage dismissed the gold segment in 0 of 300 cases (retrieval@1 = 1.000); the rule-of-three 95% upper bound is $3/300 = 1.0\%$ at the query level and $3/30 = 10.0\%$ at the task-class level. The bound is scoped: the evaluated queries are handle-bearing by construction, and the 10 zero-handle task classes bypass narrowing by opening the full corpus, so their dismissal risk is zero by construction. The remaining risk is structural: a query face cannot reach every corpus entry — reachability is 5,684 of 50,811 entries (11.2%), and an independent model applying the gold manifest's anchor rules reaches 3,675 (7.2%), so no face reaches 88.8%–92.8% of entries, and the narrowing stage's safety does not extend to arbitrary queries. Hence certified absence requires the precondition gate of Section 7 (an exhaustive scan or a zero-false-dismissal index), not the candidate-narrowing stage alone.





5 Discussion: effect sizes and their scope

Four numbers, four objects. **Detector-layer absence:** the operational figure is the model-free 3×10^{-5} backstop on 0/100,002 nulls. Decoder-suite abstention numbers (AbstentionBench 0.71, LoCoMo 2.1 F1 [34, 43]) are unrun on SUPERCTX and are not comparands for this detector. Protocol $n = 8$ and the expansion-lane minting rate 1,200/1,200 (Wilson [0.9968, 1.000]) stay in the evaluation body; they are not one result. **Needles:** 1.000 is matched-filter retrieval over a structured bank (Figure 6), not a decoder NIAH score; RAG+regex presence parity is 1.000. **Narrowing** is bimodal (median 44.8×, mean 2.94×) next to 11.2% reachability (Figure 5); 205× is a join-rule property, not serving. **Windows:** 30M is an architectural bound with no 30M heatmap (Table 3; Figure 1), not a contribution claim.

6 Limitations

1. **Synthetic corpus.** v6 is a generated registry world: 44.8× narrowing, 11.2% reachability, and the Q-EXEC 1–2 candidate cells are properties of that generator; we must re-measure them on a customer corpus.

2. **Serve-time filter is a no-op.** Offline clause-join 205× does not fire at serve: the filter that would realize it is an unexecuted no-op (serve does not materialize the blocking predicate). The live path is face-alone 44.8×



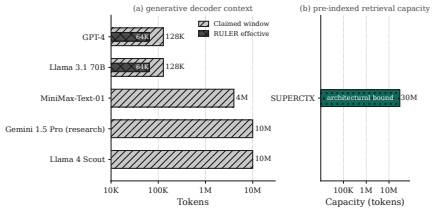


Figure 1: Advertised versus RULER-effective decoder windows (unrun on SUPERCTX). SUPERCTX's 30M figure is an architectural bound, not a contribution score. Crayon on this printout is wax marginalia, not data.

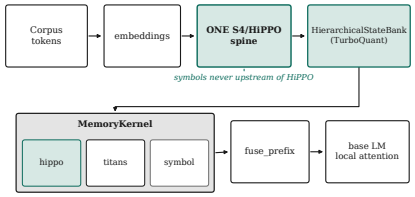


Figure 2: Detector path: parse, project, matched-filter gate, then a presence certificate or a detector-layer empty set. S4/HiPPO/TurboQuant store the sealed state. Crayon on this printout is wax marginalia, not data.

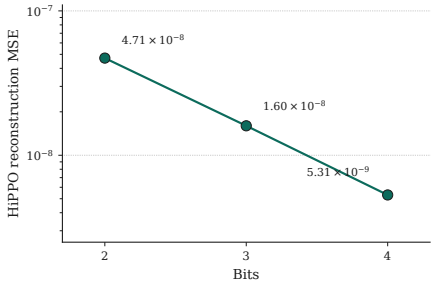


Figure 3: HiPPO reconstruction MSE on the original 500-state protocol (reported, not recomputed; trained-state dataset not in this artifact). 4-bit MSE 5.31×10^{-9} is not a front-matter claim. Crayon on this printout is wax marginalia, not data.

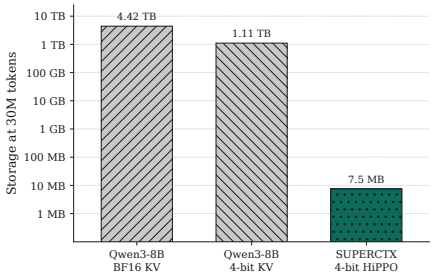


Figure 4: A 4-bit $d = 256$ HiPPO store is 7.5 MB at the 30M architectural bound (no 30M NIAH heatmap); 4-bit KV cache for a comparable decoder is on the order of a terabyte. Crayon on this printout is wax marginalia, not data.

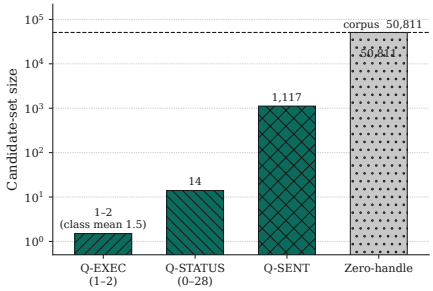


Figure 5: Face-alone candidate sets on the 50,811-entry table. Median 44.8× versus mean 2.94×; 11.2% reachability. Narrowing is not certified absence. Crayon on this printout is wax marginalia, not data.

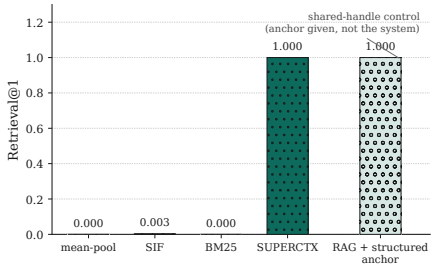


Figure 6: Needle retrieval@1 ($n = 300$). SUPERCTX 1.000 versus vanilla RAG 0.000–0.003; RAG+regex presence parity is 1.000. Crayon on this printout is wax marginalia, not data.

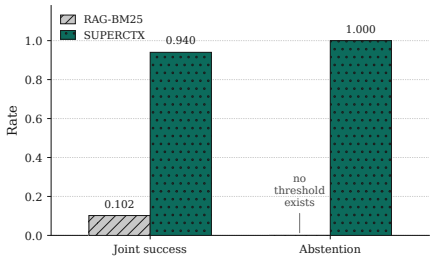


Figure 7: Body composite 346/368 (300+60+8) versus RAG 0.102. The $n = 8$ leg remains; this is not a front-matter metric. Powered empty-set sweep unrun. Crayon on this printout is wax marginalia, not data.



plus the zero-handle full-corpus opens.

3. **30M is architectural.** One documented eval ingested 1.02M tokens (2,000 × 512); the bank-layout RAG ablation is a 97,656 × 512 *layout* — 49,999,872 tokens — the right retrieval comparison, not a decoder pass; we have no 30M NIAH heatmap.

4. **RAG protocol asymmetry.** We did not give vanilla RAG `parse_query_anchor` in the published ablation; with it, a phrase index scored 1.000 on the needle leg (presence parity). A powered empty-set sweep on the nonexistent-anchor protocol was not run, so we do not headline an abstention gap. ColBERTv2/BGE-M3 were not scored on that protocol (recorded skip).

5. **Detector pin vs end-to-end.** $p = 7.5 \times 10^{-9}$ is a GPD tail on 100,002 detector-layer nulls (0.415-log gap; printed difference 0.416), not a customer-distribution finding. Under an alternative protocol, 600 fabricated anchors: detector verify 0/600, then a frozen generation model (`fable5`) emitted identifier-shaped answers on all 600 because the answer spec had no abstain path (generation-layer false-accept 1.000). Mislabeled-evidence conditions: 0.465 / 0.827 / 0.980. Neither the GPD pin nor the 3×10^{-5} backstop constrains that generation layer (`eval/certified_detector/REPORT.md` §B6).



6. **Protocol $n = 8$ is a body count.** The protocol abstention count is small (Clopper–Pearson $[0.631, 1.000]$). The expansion-lane 1,200/1,200 Wilson interval is certificate-minting on a non-registry corpus, not S4-detector retrieval on the 50,811-entry registry, and not a reproduction of $n = 8$. Neither number is a front-matter headline.

7. **No public-benchmark decoder scores.** RULER, NoLiMa, HELMET, LongBench v2, NoCha, and AbstentionBench are unrun on SUPERCTX. Running NoLiMa and AbstentionBench as natural-language retrieval/abstention tasks is the correct next experiment; a run on the 50,811 registry would not be a generality test.

7 Conclusion

SUPERCTX is a sealed-corpus detector over a synthetic 50,811-entry registry: a recomputation-checkable presence certificate, or a detector-layer empty set with a model-free 3×10^{-5} backstop (0/100,002), only under an exhaustive scan or a zero-false-dismissal index. Needle retrieval@1 is 1.000 ($n = 300$) against 0.000–0.003 for vanilla RAG and 0.000 for ColBERTv2 (`doc_maxlen`=180) and BGE-M3; the same regex lifts RAG to 1.000, so presence is a handle-access gap. We do not headline an abstention gap (powered empty-set sweep unrun; ColBERTv2/BGE-M3 skipped on that protocol). User-facing empty-set is not the product while the generation-layer false-accept of 1.000 remains. S4/HiPPO/TurboQuant store state; they are not the contribution. 4-bit MSE 5.31×10^{-9} is reported, not recomputed (trained-state dataset not in this artifact).

What remains: a powered RAG empty-set sweep and ColBERTv2/BGE-M3 on the nonexistent-anchor protocol; a live serve-time filter; a non-synthetic corpus; NoLiMa and AbstentionBench (unrun on SUPERCTX) as natural-language retrieval/abstention tasks. The claim is one detector layer, not a longer window.

Code and data availability. The supplementary materials include the corpus generator, `benchmarks/negative_expansion.py`, and `tools/`certificate_exercise.py, check_overfull.py, paper_text.py. The TurboQuant trained-state dataset is not in the artifact.

Ethics statement. Synthetic data only. A certificate is valid only relative to its declared universe. **AI use.** AI systems were an evaluated generation model, experiment tooling, and editing assistance; the authors take responsibility for all claims. **Acknowledgments.** SUPERCTX project infrastructure; PAT review agents. **Funding.** None.

Certificate schema. Sealed object (`superctxmcp/types/certificate.py`): claim, universe, revision, position, detector, score/threshold, calibration, policy, and seal hashes. Presence fills span fields; absence carries `search_transcript`; abstain carries `missing_reason`. Auditor, fail-closed: resolve digests; load detector state; re-run the matched filter; compare score to threshold and re-derive the tail bound; recompute the seal. Any disagreement falsifies the certificate.

References

[1] Angelopoulos, A. N., and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.

[2] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2022). Conformal risk control. arXiv:2208.02814.

[3] Ashkboos, S., et al. (2024). QuaRot: Outlier-free 4-bit inference in rotated LLMs. arXiv:2404.00456.

[4] Bai, Y., et al. (2024). LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. arXiv:2412.15204.

[5] Behrouz, A., Zhong, P., and Mirrokni, V. (2025). Titans: Learning to memorize at test time. arXiv:2501.00663.

[6] Behrouz, A., et al. (2025). Atlas: Learning to discover any context length at any scale. arXiv:2505.23735.

[7] Borgeaud, S., et al. (2022). Improving language models by retrieving from trillions of tokens (RETRO). arXiv:2112.04426.

[8] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). BGE M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Findings of ACL 2024*, 2318–2335. DOI 10.18653/v1/2024.findings-acl.137. arXiv:2402.03216.

[9] Ding, Y., et al. (2024). LongRoPE: Extending LLM context window beyond 2 million tokens. arXiv:2402.13753.

[10] Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. arXiv:2208.07339.

[11] Dettmers, T., Svirschevski, R., Egiazarian, V., Kuznedelev, D., Frantar, E., Ashkboos, S., Borzunov, A., Hoefler, T., and Alistarh, D. (2023). SpQR: A sparse-quantized representation for near-lossless LLM weight compression. arXiv:2306.03078. ICLR 2024.

[12] Edge, D., et al. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv:2404.16130.

[13] Egiazarian, V., et al. (2024). Extreme compression of large language models via additive quantization (AQLM). arXiv:2401.06118.

[14] Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*.



[15] Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., and Re, C. (2022). Hungry Hungry Hippos: Towards language modeling with state space models (H3). arXiv:2212.14052.

[16] Gao, J., and Long, C. (2024). RaBitQ: Quantizing high-dimensional vectors with a theoretical error bound for approximate nearest neighbor search. arXiv:2405.12497.

[17] Geifman, Y., and El-Yaniv, R. (2017). Selective classification for deep neural networks. arXiv:1705.08500.

[18] Gollapudi, S., et al. (2023). Filtered-DiskANN: Graph algorithms for approximate nearest neighbor search with filters. *WWW 2023*. DOI 10.1145/3543507.3583552.

[19] Hatamizadeh, A., Choi, Y., and Kautz, J. (2026). Gated DeltaNet-2: Decoupling erase and write in linear attention. arXiv:2605.22791.

[20] Graves, A., Wayne, G., and Danihelka, I. (2014). Neural Turing Machines. arXiv:1410.5401.

[21] Graves, A., Wayne, G., and Danihelka, I. (2016). Hybrid computing using a neural network with dynamic external memory (DNC). *Nature*, 538, 471–476.

[22] Gu, A., Dao, T., Re, C., and Sohl-Dickstein, J. (2020). HiPPO: Recurrent memory with optimal polynomial projections. arXiv:2008.07669.

[23] Gu, A., Goel, K., and Re, C. (2021). Efficiently modeling long sequences with structured state spaces (S4). arXiv:2111.00396.

[24] Gu, A., and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752.

[25] Guo, Z., et al. (2024). LightRAG: Simple and fast retrieval-augmented generation. arXiv:2410.05779.

[26] Gutierrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., and Su, Y. (2024). HippoRAG: Neurobiologically inspired long-term memory for large language models. arXiv:2405.14831.

[27] Gutierrez, B. J., Shu, Y., Qi, W., Zhou, S., and Su, Y. (2025). From RAG to memory: Non-parametric continual learning for large language models (HippoRAG 2). arXiv:2502.14802. ICML 2025.

[28] Hooper, C., et al. (2024). KVQuant: Towards 10 million context length LLM inference with KV cache quantization. arXiv:2401.18079.

[29] Hsieh, C.-P., et al. (2024). RULER: What's the real context size of your long-context language models? arXiv:2404.06654. COLM 2024.

[30] Huang, R., Zhang, K., Fang, Y., and Yu, B. (2025). Probing the limits of compressive memory: A study of Infini-attention in small-scale pretraining. arXiv:2512.23862.

[31] Ju, H., and Dong, B. (2025). MIRB: Mathematical information retrieval benchmark. arXiv:2505.15585.

[32] Karpinska, M., Thai, K., et al. (2024). One thousand and one pairs: A "novel" challenge for long-context language models (NoCha). arXiv:2406.16264. ACL Findings 2024.

[33] Kimi Team. (2025). Kimi Linear: An expressive, efficient attention architecture. arXiv:2510.26692.

[34] Kirichenko, P., Ibrahim, M., Chaudhuri, K., and Bell, S. J. (2025). AbstentionBench: Reasoning LLMs fail on unanswerable questions. arXiv:2506.09038. NeurIPS 2025 Datasets and Benchmarks.

[35] Lee, J., et al. (2024). Can long-context language models subsume retrieval, RAG, SQL, and more? (LOFT). arXiv:2406.13121. NAACL Findings 2025.

[36] Li, M., Gururangan, S., Dettmers, T., Lewis, M., Althoff, T., Smith, N. A., and Zettlemoyer, L. (2022). Branch-Train-Merge: Embarrassingly parallel training of expert language models. arXiv:2208.03306.

[37] Li, Z., Zhang, R., Zhang, Y., and Chen, D. (2024). Retrieval augmented generation or long-context LLMs? A comprehensive study and hybrid approach (Self-Route). arXiv:2407.16833.

[38] Lin, J., Tang, J., Tang, H., Yang, S., et al. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. arXiv:2306.00978.

[39] Liu, H., Zaharia, M., and Abbeel, P. (2023). Ring attention with blockwise transformers for near-infinite context. arXiv:2310.01889.

[40] Liu, H., et al. (2024). World model on million-length video and language with blockwise ringattention (LWM). arXiv:2402.08268.

[41] Liu, Z., et al. (2024). KIVI: A tuning-free asymmetric 2bit quantization for KV cache. arXiv:2402.02750.

[42] Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.

[43] Maharana, A., et al. (2024). Evaluating very long-term conversational memory of LLM agents (LoCoMo). arXiv:2402.17753. ACL 2024.

[44] Max, J. (1960). Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1), 7–12.

[45] Mem0. (2025). Mem0: Building production-ready AI agents with scalable long-term memory. arXiv:2504.19413.

[46] Meta. (2025). The Llama 4 herd of models. Developer card; 10M context on Scout / Maverick.

[47] MiniMax, Li, A., Gong, B., et al. (2025). MiniMax-01: Scaling foundation models with lightning attention. arXiv:2501.08313.

[48] Modarressi, A., Deilamsalehy, H., Dernoncourt, F., Bui, T., Rossi, R. A., Yoon, S., and Schütze, H. (2025). NoLiMa: Long-context evaluation beyond literal matching. arXiv:2502.05167. ICML 2025.

[49] Mohtashami, A., and Jaggi, M. (2023). Landmark attention: Random-access infinite context length for transformers. arXiv:2305.16300.

[50] Munkhdalai, T., Faruqui, M., and Gopal, S. (2024). Leave no context behind: Efficient infinite context transformers with Infini-attention. arXiv:2404.07143.

[51] NVIDIA. (2024). Hymba: Hybrid-head architecture for small language models. arXiv:2411.13676.

[52] NVIDIA. (2025). Nemotron-H: Hybrid Mamba-Transformer models. arXiv:2504.03624.



[53] Packer, C., et al. (2023). MemGPT: Towards LLMs as operating systems. arXiv:2310.08560.

[54] Pantazopoulos, G., Nikandrou, M., Konstas, I., and Suglia, A. (2026). Retrievit: In-context retrieval capabilities of transformers, state space models, and hybrid architectures. arXiv:2603.02874.

[55] Patel, L., et al. (2024). ACORN: Performant and recall-accurate neighborhood search for graph indexes over filtered datasets. arXiv:2403.04871. SIGMOD 2024.

[56] Poli, M., et al. (2023). Hyena hierarchy: Towards larger convolutional language models. arXiv:2302.10866.

[57] Quach, V., et al. (2023). Conformal language modeling. arXiv:2306.10193.

[58] Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. (2020). Compressive transformers for long-range sequence modelling. arXiv:1911.05507.

[59] Reid, M., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

[60] Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., and Zaharia, M. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. *NAACL 2022*, 3715–3734. DOI 10.18653/v1/2022.naacl-main.272. arXiv:2112.01488.

[61] Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. (2024). RAPTOR: Recursive abstractive processing for tree-organized retrieval. arXiv:2401.18059.

[62] Su, H., et al. (2024). BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. arXiv:2407.12883.

[63] Thakur, N., Reimers, N., Ruckle, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. arXiv:2104.08663. NeurIPS 2021.

[64] Trockman, A., Harutyunyan, H., Kolter, J. Z., Kumar, S., and Bhojanapalli, S. (2024). Mimetic initialization helps state space models learn to recall. arXiv:2410.11135.

[65] Tseng, A., Chee, J., Sun, Q., Kuleshov, V., and De Sa, C. (2024). QuIP#: Even better LLM quantization with Hadamard incoherence processing. arXiv:2402.04396.

[66] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training (E5). arXiv:2212.03533.

[67] Wang, W., Dong, L., Cheng, H., et al. (2023). Augmenting language models with long-term memory (LongMem). arXiv:2306.07174.

[68] Wu, Y., Rabe, M. N., Hutchins, D., and Szegedy, C. (2022). Memorizing transformers. arXiv:2203.08913.

[69] Xu, Z., Yue, Y., Hu, X., Yuan, Z., et al. (2025). MambaQuant: Quantizing the Mamba family with variance aligned rotation methods. arXiv:2501.13484.

[70] Yen, H., et al. (2025). HELMET: How to evaluate long-context language models effectively and thoroughly. arXiv:2410.02694. ICLR 2025.

[71] Zandieh, A., et al. (2026). TurboQuant. ICLR 2026. arXiv:2504.19874.

[72] Zhang, H., et al. (2023). R-Tuning: Instructing large language models to say "I don't know." arXiv:2311.09677.

[73] Zhang, X., et al. (2024). ∞Bench: Extending long context evaluation beyond 100K tokens. arXiv:2402.13718. ACL 2024.

