GOLD PANNING: Strategic Context Shuffling for Closed-Box Needle-in-Haystack Reasoning

Adam Byerly **Daniel Khashabi**
Department of Computer Science
Johns Hopkins University
{abyerly2, danielk}@jhu.edu

Abstract

Large language models (LLMs) exhibit pronounced position bias in long-context retrieval, systematically prioritizing information location over relevance. Existing mitigations modify positional encodings or attention internals, but in the common output-only deployment setting these are inapplicable. We introduce GOLD PANNING, a Bayesian framework for inference-time active search that (i) reorders documents to concentrate high-belief items in highly diagnostic positions (signal anchoring) and (ii) updates relevance beliefs from model outputs. Unlike active learning, which prioritizes uncertainty reduction, GOLD PANNING exploits anchoring—once flagged, keep it in sight—to preserve weak cues. A greedy assignment derived from the diagnosticity profile identifies the target in $O(\log N)$ rounds once a positive belief margin is established, under the detector assumptions; we do not bound this lock-in time theoretically. On Gemma-3-12B with 100 documents, a positive margin is reached within two rounds in 72–85% of instances and within four in 90–98%. Across open-weight and closed-source models on multi-document QA, GOLD PANNING can match PSC's accuracy with 50–87.5% fewer queries by exploiting positional heterogeneity, while achieving comparable accuracy when positional profiles are nearly flat. Calibration transfers across tasks, while static-placement ablations show the additional value of sequential belief updating. These results show that inherent model biases need not be failures, but can serve as exploitable signals.

1 Introduction

Reasoning tasks often require identifying relevant signals in large contexts, including scientific synthesis, legal analysis, and agentic fact retrieval. LLMs have emerged as practical tools for these tasks, enabling knowledge synthesis in retrieval-augmented generation (RAG) [15, 7] and agentic systems [30, 34], but a critical bottleneck is *position bias*, where LLMs often prioritize information based on *where* it appears in context rather than its intrinsic relevance [31, 39, 16]. Even state-of-the-art retrieval systems [12, 37] cannot guarantee optimal placement. When critical information falls into these "lost in the middle" zones, reasoning degrades substantially [1].

Current approaches to mitigate this bias fall into two categories. **Open-box mitigations** modify model architectures to reduce bias [22, 9, 32, 38], but require model access or extensive fine-tuning, rendering them inapplicable for output-only deployment settings. **Closed-box alternatives**, such as Permutation Self-Consistency (PSC; Tang et al. [28]), use inference-time ensembling by randomly shuffling documents across multiple queries and aggregating results via majority vote. However, PSC has two structural weaknesses. First, PSC uses nonadaptive placements independent of earlier outputs, while citation aggregation retains evidence across queries. Second, PSC is bias-agnostic, treating all positions as exchangeable rather than exploiting known differences in positional reliability.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

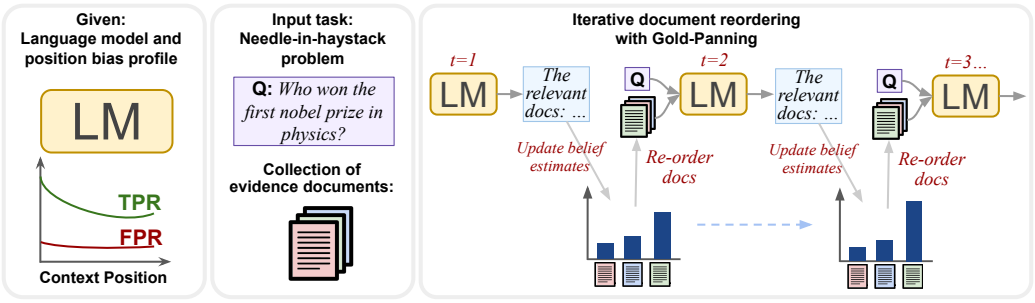

Figure 1: Overview of GOLD PANNING. **Given:** a language model with a measured positional diagnosticity profile (TPR/FPR by position; §2.4). **Task:** identify relevant information from a multi-document collection. **Method:** iteratively query the model, update Bayesian beliefs based on observed responses (§2.3), and reorder documents to place high-belief items in high-diagnosticity positions (§2.5). The updated beliefs then determine the document arrangement for the next query.

Rather than treating position bias as noise to average away, we propose to *model it explicitly* and *learn from observations*. We introduce GOLD PANNING, a closed-box, inference-time framework that maintains Bayesian beliefs over document relevance and iteratively reorders high-belief documents towards positions of high discriminative power. Position bias has a consistent structure (Fig. 1, left pane), with positions near the beginning and end reliably surfacing relevant content while middle positions (the "lost in the middle" zone [16]) obscure it. We characterize this structure by estimating a position's *diagnosticity* (§2.4), a measure of how reliably it distinguishes relevant from irrelevant documents. A lightweight, per-model calibration on synthetic needle-in-haystack instances (§B) suffices to estimate this diagnosticity profile. Empirically, this profile transfers across Multi-Document QA (MDQA) tasks under a fixed protocol (§4.3). A controlled attention probe provides mechanistic grounding: position dominates document identity, document length, and lexical overlap with the query as a driver of attention mass on the gold document (§B.6).

Equipped with this profile, we view each context position as a noisy detector with known reliability, replacing the traditional view of a passive input buffer. By adopting a Bayesian lens (Fig. 1, right pane), we maintain a running probabilistic belief about each document's relevance, treating every LLM output as a *noisy measurement* rather than a final answer. This enables *sequential Bayesian inference*, where each query extracts partial evidence that shapes the next input configuration, allowing the system to "pan for gold." Crucially, document retrieval is fundamentally a *search* problem: we do not need perfect belief calibration for all documents, only efficient identification of the relevant ones. Reducing uncertainty about irrelevant distractors yields only marginal gain toward this goal, and worse, as we show in §3.2, the plain entropy policy demotes confident positives once their entropy collapses, starving the signal it just found.

The mechanism behind this is *signal anchoring*: high-belief documents placed at highly diagnostic positions. Beliefs in long-context reasoning are inherently fragile. Anchoring increases the probability of receiving highly diagnostic evidence early, which accelerates belief separation and reduces the number of rounds to confidently select the needles. This principle—that **once identified, a needle must be kept in view**—maximizes recall with fewer queries. GOLD PANNING operationalizes these insights via greedy matching at each round. We evaluate on MDQA across *five* state-of-the-art models at scales up to $N{=}800$ documents. Signal anchoring outperforms the plain entropy objective; a stronger entropy variant closes most of the gap by protecting confident positives (§D.3). GOLD PANNING can match PSC's accuracy with **50–87.5% fewer queries** by exploiting positional heterogeneity, while achieving comparable accuracy when positional profiles are nearly flat.

Contributions. (1) **Framework:** We formalize position bias as a *diagnostic signal* and introduce GOLD PANNING, a closed-box inference-time method for Bayesian active search over long contexts via calibrated position diagnosticity. (2) **Theory:** Under the stated detector assumptions, signal anchoring sustains informative drift after a positive belief margin is established, yielding a conditional $O(\log N)$ rate. We characterize time to the first positive margin empirically. (3) **Empirical:** Across open-weight and closed-source models at scales up to $N{=}800$, we measure query savings, calibration transfer, the contribution of sequential updating, and robustness under realized layouts, multiple needles, and semantic distractors.

2 The GOLD PANNING Framework

Here we present our algorithm (§2.1), then each component: documents and positions (§2.2), belief dynamics (§2.3), positional diagnosticity (§2.4), and scoring objective (§2.5).

2.1 Overview and Algorithm

GOLD PANNING operates in rounds: (1) we score documents based on current beliefs, (2) construct an assignment matching high-scoring documents to high-diagnosticity positions, (3) query the LLM with this arrangement, and (4) update beliefs based on which documents were cited. Algorithm 1 summarizes this procedure: **Lines 3-4** compute positional diagnosticity and establish a fixed ordering of positions by discriminative power (§2.4); **Line 5** initializes uniform prior beliefs; the main loop (**Lines 6-13**) iterates: **Line 7** scores documents for determining their placement in the next iteration; we will introduce two belief-aware variants alongside a non-adaptive baseline for scoring documents (§2.5). **Lines 8-9** construct the greedy assignment matching document and position orderings, **Lines 10-11** format and query the LLM, and **Line 12** performs the Bayesian belief updates (§2.3).

Algorithm 1 GOLD PANNING

1: **Input:** Query Q, Documents $\mathcal{D}$, Detector Profile $\{(\mathrm{TPR}_j, \mathrm{FPR}_j)\}_{j=1}^{N}$, Rounds T
2: **Output:** Final beliefs $\mathbf{b}_T$
3: Compute $d_{\text{diag}}(j) \leftarrow |\mathrm{TPR}_j - \mathrm{FPR}_j|$
4: $\pi_{\text{pos}} \leftarrow \text{argsort}(\mathbf{d}_{\text{diag}}, \text{descending})$
5: Initialize $b_{0,i} \leftarrow 0.5$
6: **for** $t = 1$ **to** T **do**
7: $\quad s_i \leftarrow \textsc{Score}(b_{t-1,i})$
8: $\quad \pi_{\text{doc}} \leftarrow \text{argsort}(\mathbf{s}, \text{descending})$
9: $\quad \sigma_t \leftarrow \text{zip}(\pi_{\text{doc}}, \pi_{\text{pos}})$
10: $\quad P_t \leftarrow \text{Format}(Q, \text{Permute}(\mathcal{D}, \sigma_t))$
11: $\quad \mathbf{O}_t \leftarrow \text{QueryLLM}(P_t)$
12: $\quad$ Update $\mathbf{b}_t$ via Eq. (2) using citations $\mathbf{O}_t$
13: **end for**
14: **return** $\mathbf{b}_T$

2.2 Documents and Positions

We consider N documents indexed by $i \in \{1, \ldots, N\}$, each with an *unknown* binary relevance $Z_i \in \{0, 1\}$. For each query, we also have N context positions indexed by $j \in \{1, \ldots, N\}$. We model each position as a *binary detector* with true/false positive rates $\mathrm{TPR}_j, \mathrm{FPR}_j$, a measurement device whose reliability we can characterize empirically (see §4.1 and §B). Letting $O_{ij} \in \{0, 1\}$ denote if the LLM cites document i at position j, we have:

$$\Pr(O_{ij} = 1 \mid Z_i = z) = z \cdot \mathrm{TPR}_j + (1 - z) \cdot \mathrm{FPR}_j, \tag{1}$$

with $\Pr(O_{ij} = 0 \mid \cdot) = 1 - \Pr(O_{ij} = 1 \mid \cdot)$.

Empirical validity. This position-only observational model rests on two claims: (1) The detection rates depend more on position than on the specific document occupying that position, conditioned on document relevance ($Z_i = z$). We validate this with a retrieval-head attention probe that finds that position alone accounts for approximately 20% of variance in retrieval attention ($p < 10^{-20}$), while content identity ($p = 0.466$), fact length ($p = 0.444$), and lexical overlap ($p = 0.185$) show null effects (full analysis in §B.6). (2) Conditional on document relevance and the assigned permutation, observations at distinct positions are approximately independent, despite the autoregressive nature of LLM decoding. Checks under realized GP-BELIEF layouts support the use of calibrated marginals on Gemma-3-12B, while also revealing within-read dependence (§B.5).

2.3 Actions, Observations, and Beliefs

Observations. At each round t, we choose a *permutation* $\sigma_t \in S_N$ (the symmetric group on $\{1, \ldots, N\}$) specifying which document is placed in which position; $\sigma_t(i) = j$ means "document i is placed at position j." Given a formatted prompt containing permutations of the documents (lines 10-11), the language model returns a set of cited document IDs $\mathbf{O}_t$, yielding one observation per document, $O_{t,i} = \mathbb{I}[i \in \mathbf{O}_t]$. These are output-behavior measurements: citations need to be statistically informative about relevance, not faithful traces of the model's internal evidence use.

Belief updates. We maintain posterior beliefs $\mathbf{b}_t = (b_{t,1}, \ldots, b_{t,N})$, where $b_{t,i} = \Pr(Z_i = 1 \mid \mathcal{F}_t)$ is our belief that document i is relevant, given $\mathcal{F}_t = \{\mathbf{O}_1, \ldots, \mathbf{O}_t, \sigma_1, \ldots, \sigma_t\}$ which represents the observations and document orderings up to and including round t. Starting from a uniform prior,

$b_{0,i} = 0.5$, we update these beliefs via Bayes' theorem:

$$b_{t,i} = \frac{b_{t-1,i}\mathcal{P}_1}{b_{t-1,i}\mathcal{P}_1 + (1 - b_{t-1,i})\mathcal{P}_0}, \tag{2}$$

where $\mathcal{P}_z = \Pr(O_{t,i} \mid Z_i = z, \sigma_t(i) = j)$ (Eq. (1)) is the Bernoulli likelihood under label z for $j = \sigma_t(i)$. When upstream scores are available, these could inform a non-uniform prior. In Algorithm 1, beliefs are initialized uniformly (**Line 5**) and updated after each query via Eq. (2) (**Line 12**).

2.4 Diagnosticity

We quantify how informative position j is via Youden's index [21, 35], a standard measure of diagnostic discriminative power. Youden's J-statistic is defined as $J_j = \mathrm{TPR}_j - \mathrm{FPR}_j$. Under this measure, a position is maximally informative when $J_j = 1$, uninformative when $J_j = 0$, and maximally *anti*-informative when $J_j = -1$. Crucially, when $J_j < 0$, a citation from position j is more likely under irrelevance than relevance, and therefore provides strong *negative* evidence. Because we primarily need to decide *which positions yield the most information*, we rank positions by *magnitude*, $d_{\text{diag}}(j) = |J_j|$. The sign of J_j is preserved in the likelihoods $(\mathrm{TPR}_j, \mathrm{FPR}_j)$ used by the Bayesian update (Eq. (2)), so positions with $J_j < 0$ contribute correctly-signed evidence despite being ranked by magnitude. We present the balanced case with N documents and N positions. Asymmetric cases reduce to this setting via dummy positions or documents; we describe these details in Appendix F.

2.5 Scoring Objectives

At each round t, we score each document i by a scalar $s_{t,i}$ as defined below **Line 7**, sort documents in descending order, and greedily match the highest-scoring document to the most diagnostic position, second-highest to second-most, and so on. The scoring function reflects a fundamental choice between active learning [24, 25] and active search [8], which we instantiate as two variants, alongside our non-adaptive baseline:

$$s_{t,i}^{\text{PSC}} \sim \mathcal{U}(0,1) \qquad s_{t,i}^{\text{Entropy}} = \mathcal{H}(b_{t-1,i}) \qquad s_{t,i}^{\text{Belief}} = b_{t-1,i} \tag{3}$$

PSC (Permutation Self-Consistency). Scores are drawn i.i.d. from a uniform distribution, independent of the belief state $\mathbf{b}_{t-1}$. Sorting i.i.d. continuous random keys yields a uniformly random permutation, so each round independently re-shuffles the documents and the procedure reduces to averaging detector outputs over T random orderings. The evaluated PSC baseline aggregates citation frequencies rather than Bayesian posteriors.

GP-ENTROPY (uncertainty reduction). High scores correspond to high uncertainty, anchoring ambiguous documents at diagnostic positions to maximize information gain—standard active-learning intuition: probe the most uncertain cases to learn fastest.

GP-BELIEF (signal anchoring). High scores correspond to likely relevant documents, anchoring them at diagnostic positions to maximize recall—the active search objective.

We formalize both in §3, showing that uncertainty reduction is aligned with maximizing expected entropy decrease under our detector model, while signal anchoring maximizes a belief-weighted sum of position-wise detection rates. The algorithm returns final posterior beliefs $\mathbf{b}_T$, mapped to discrete predictions via standard decision rules (§2.6).

2.6 Decision Rules

GOLD PANNING outputs continuous beliefs $\mathbf{b}_T \in [0,1]^N$, which need to be mapped to binary predictions via a *decision rule*. Crucially, this rule is a parameter of *evaluation*, not of the algorithm. The same belief trajectory supports different selection criteria without re-running inference. We consider two canonical rules. **Top-k** selects the k highest-belief documents (natural when k is known), while **Threshold-τ** selects all documents with belief above the threshold (natural when k is unknown). Our primary experiments use Top-1; ablations include Top-5 and Threshold-0.5 (§D). With one gold document and exactly one prediction, F_1 equals the Top-1 hit rate (recall@1).

2.7 Computational Complexity

The per-round complexity of GOLD PANNING is $O(N \log N)$ for sorting documents by score, compared to $O(N)$ for PSC's random shuffle. However, this is negligible CPU overhead, as the dominant cost in practice is LLM inference. The $O(\log N)$ call bound in §3.1 applies after positive-margin lock-in under the stated drift assumptions; it does not bound the total time from initialization. We measure query savings against PSC at matched accuracy (§4.2).

3 Theoretical Properties

We analyze GOLD PANNING as sequential Bayesian inference and characterize how the number of LLM queries scales with the collection size N and the diagnostic profile of the available positions.

To facilitate this analysis, we convert belief updates to the log-odds scale. We define the log-odds $\lambda_{t,i}$, with $\ell_j(o)$ as the position-dependent log-likelihood ratio:

$$\lambda_{t,i} = \log \frac{b_{t,i}}{1 - b_{t,i}}, \qquad \ell_j(o) = \log \frac{\Pr(O = o \mid Z = 1, \text{pos} = j)}{\Pr(O = o \mid Z = 0, \text{pos} = j)}. \tag{4}$$

This admits an additive Bayesian update of the form $\lambda_{t,i} = \lambda_{t-1,i} + \ell_{\sigma_t(i)}(O_{t,i})$ (Eq. (2)). The log-odds thus evolve as a random walk with increments determined by the assigned position, a structure that underlies our analysis. Full proofs appear in §A.

3.1 Convergence and Sample Complexity

Drift. For a relevant document at position j, the expected per-round increment of the log-odds is the KL divergence

$$\mu_j \;=\; D_{\mathrm{KL}}\left(\mathrm{Bern}(\mathrm{TPR}_j) \,\|\, \mathrm{Bern}(\mathrm{FPR}_j)\right), \tag{5}$$

strictly positive whenever $|J_j| > 0$. Drift is the velocity at which evidence accumulates; everything that follows is a statement about how a policy controls this quantity.

Consistency (Thm. A.1 in §A). If the policy does not permanently starve a document of informative positions, beliefs converge to ground truth almost surely: $b_{t,i} \to Z_i$. PSC satisfies the coverage condition unconditionally; GP-BELIEF and GP-ENTROPY satisfy it under a structural assumption on the sorting rule. Consistency is necessary but not sharp—the operative question is *how fast?*

Sample complexity (Thm. A.3 in §A). With a per-round drift floor $\underline{\mu}$, Azuma–Hoeffding on the centered log-odds martingale plus a union bound over documents gives

$$T \;=\; \mathcal{O}\left(\frac{L^2}{\underline{\mu}^2} \log \frac{N}{\delta}\right)$$

rounds for confidence $1 - \delta$ recovery of all N labels. The structure is logarithmic in both N and $1/\delta$; the constant $\underline{\mu}^{-2}$ is where policy quality enters.

Strategic placement vs. random shuffling (Cor. A.6 in §A). PSC satisfies the same bound but with the *position-averaged* drift $\bar{\mu} = \frac{1}{N}\sum_j \mu_j$ in place of $\underline{\mu}$. When only K of N positions are informative (the rest near-deadzone), $\bar{\mu} \approx Kc/N$ for typical informative drift c, while a targeted policy that maintains informative exposure has $\underline{\mu} \approx c$. Under this best-case exposure assumption:

$$T_{\text{GOLD PANNING}} \;=\; \mathcal{O}\Big(\tfrac{L^2}{c^2} \log \tfrac{N}{\delta}\Big), \qquad T_{\text{PSC}} \;=\; \mathcal{O}\Big(\tfrac{L^2}{c^2} \cdot (N/K)^2 \log \tfrac{N}{\delta}\Big). \tag{6}$$

The $(N/K)^2$ factor is a best-case comparison of sufficient bounds, conditional on maintaining the targeted drift floor; it is not an unconditional speedup guarantee. With both policies evaluated using realized drifts, the predicted round ratio is more modest (median $\approx 1.6\times$, maximum $\approx 13\times$; §C.4). When diagnosticity is uniform ($K = \Theta(N)$), the gap collapses to a constant factor.

This shifts the central question from "does strategic placement help over random shuffling," to "*which* scoring rule actually sustains $\underline{\mu} \geq c$?"

3.2 Drift Collapse vs. Drift Persistence

§3.1 reduced the central question to which scoring rule actually sustains $\mu \geq c$ over time. GP-BELIEF and GP-ENTROPY answer this in opposite directions: GP-ENTROPY's drift collapses as confidence grows, while GP-BELIEF's drift persists once the correct ranking is achieved. Throughout we work in the sparse-search regime ($k \ll N$) under a lost-in-the-middle profile, where the analysis is sharpest and the empirical gap is largest.

GP-Entropy: drift collapse (Prop. A.7 in §A). GP-ENTROPY ranks documents by entropy. As a relevant document accumulates evidence, $b_{t-1,i} \to 1$ and its entropy $H(b_{t-1,i}) \to 0$, while ambiguous distractors near $b \approx \frac{1}{2}$—those occupying deadzone positions where minimal evidence accumulates— retain near-maximal entropy. The relevant document is therefore demoted past every undecided distractor; once all k true positives become confident, they occupy the bottom ranks where the matching scheme assigns the lowest-drift positions. The policy effectively says "we are sure about this one, so stop checking it," starving the document of signal at exactly the moment it most needs to cross the confidence threshold.

GP-Belief: drift persistence via lock-in (Prop. A.8 in §A). GP-BELIEF ranks by belief, so high-confidence true positives are assigned to the highest-drift positions. Once the k true positives occupy the top-k belief ranks at some round τ, each is matched to a position with drift in $\{\mu_{(1)}, \ldots, \mu_{(k)}\}$, all strictly greater than the best distractor drift $\mu_{(k+1)}$. The margin

$$\Delta_t \;=\; \min_{i:Z_i=1} \lambda_{t,i} - \max_{j:Z_j=0} \lambda_{t,j}$$

then has uniformly positive expected drift, and a Lundberg-type supermartingale argument shows the favorable configuration persists for all $t > \tau$ with probability at least $1 - e^{-\gamma \Delta_\tau}$ for some $\gamma > 0$ (proof in §A). We do not formally bound the time-to-lock-in τ; on Gemma-3-12B at N=100, the first positive margin occurs within two rounds in 85% / 72% of MonoRel / PIR trials, and within four in 98% / 90%. Other regimes have substantial never-locked mass within the eight-round budget (§C.1). The configuration is self-reinforcing: positive margin drift pushes Δ_t away from the boundary, making future deviations exponentially less likely.

4 Empirical Validation

4.1 Setup

Benchmarks. We evaluate on two FLenQA benchmarks [14]: **MonoRel** (monotonic transitive reasoning) and **PIR** (People in Rooms; compositional reasoning). We selected these tasks because they offer unambiguous, deterministic ground-truth labels. Unlike open-domain QA, where ambiguity and answer-granularity mismatches can confound evaluation [19, 11], FLenQA permits precise measurement of retrieval recall. While these benchmarks are natively single-fact QA, we convert each instance into a controlled MDQA task by augmenting the ground-truth fact with $N - 1$ distractors from other instances in the same dataset, yielding instances with one relevant document exactly.

Models. We test two model families at multiple scales: **Gemma-3** [29] (12B, 27B) and **OLMo-3** [20] (OLMo-3-7B, OLMo-3.1-32B). We additionally evaluate **GPT-5.4-Nano**, a closed-source API model. These models exhibit varying degrees of position bias as measured by their calibrated diagnosticity profiles (§B), allowing us to isolate when position-aware reordering helps.

Default configuration. Unless stated otherwise, we use $N = 100$ documents with $k = 1$ relevant document and run $T = 8$ rounds. We follow model-recommended sampling parameters: temperature 1.0 with top-p 0.95 for Gemma, temperature 0.6 with top-p 0.95 for OLMo, and API defaults for GPT. All methods request reasoning followed by a boxed JSON answer and citation list; a tolerant parser extracts this output without constrained decoding. Parse success is at least 0.95 except in one cell (0.948; §E). Each round returns a set of citation IDs, which we map to per-document indicators $O_{t,i}$ (Eq. (1)) for belief updates.

Baselines. We compare GOLD PANNING against Permutation Self-Consistency (PSC) [28], which queries the model on independent random permutations and aggregates predictions via majority vote. To represent agentic reranking, we also compare against **RankGPT** [27] using the GPT-5.4-Nano API. RankGPT-Single uses one call over all 800 documents. RankGPT-Sliding uses eight calls (window

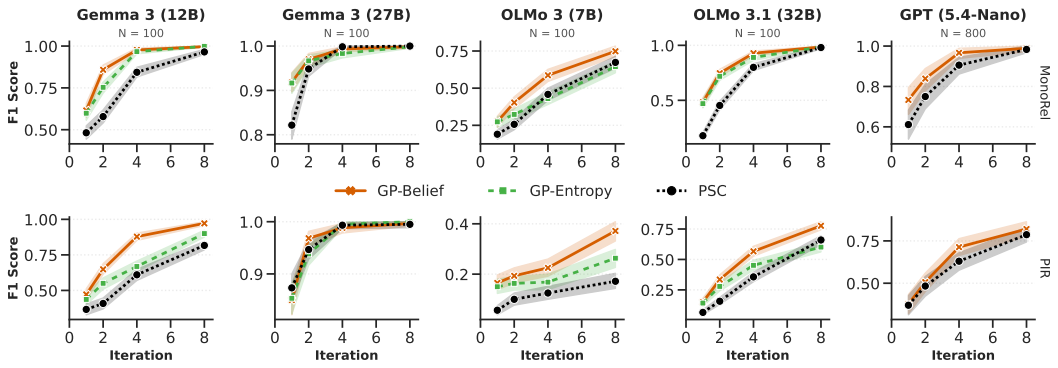

Figure 2: **Main results across models and tasks.** F_1 over 8 iterations for five models, including closed-source GPT-5.4-Nano evaluated at 8× scale (N=800). GP-BELIEF (orange) generally outperforms GP-ENTROPY (green) and PSC (black), especially with pronounced position bias. Gemma-3-27B (negative control) shows little separation with its flat profile. GP-Entropy omitted for GPT-5.4-Nano due to cost. Y-axis ranges vary per panel. Shaded regions indicate 95% CIs.

200, stride 86), matching the call count of eight-round GOLD PANNING and PSC; windowed and full-context calls need not have equal token cost.

Evaluation. GOLD PANNING maintains continuous beliefs $\mathbf{b}_t \in [0,1]^N$, converted to binary predictions via decision rules (§2.6). Our primary experiments use Top-1 selection; ablations on Top-5 and Threshold-0.5 appear in §D. We report mean F_1 with 95% confidence intervals.

Calibration of diagnosticity. For each model, we run a lightweight calibration to extract its diagnosticity profile. At position j, we place a known-relevant "gold" document, fill the remaining $N-1$ positions with distractors, and query the model. Citing the gold at j contributes to TPR_j, while citing position j when gold is elsewhere contributes to FPR_j.

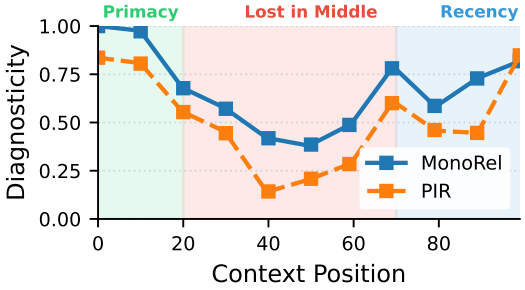

Figure 3: **Diagnosticity profile for Gemma-3-12B on MonoRel and PIR (N=100).** Both tasks show primacy and recency effects, i.e., diminished diagnosticity in the middle.

Estimating diagnosticity at all N positions is unnecessary. Smooth diagnosticity trends (Fig. 3) allow sparse fixed-grid sampling with linear interpolation. We evaluate K=11 positions with 50 trials each (550 queries total) and linearly interpolate to recover the full N-position profile; the resulting estimates are stable under both grid densification (§B.2) and per-position sample resplits (§B.3). Full details in §B. The 550-call calibration is paid once per deployment configuration. Saving four calls per request amortizes it after $550/4 \approx 138$ requests (about 140); prompt, model-version, or context-length changes may warrant recalibration.

The Gemma-3-12B profile (Fig. 3) exhibits the characteristic "lost in the middle" pattern, with high diagnosticity at primacy/recency positions and a depressed middle. GP-BELIEF exploits this heterogeneity to accelerate belief separation.

4.2 Main Results

Fig. 2 compares GP-BELIEF, GP-ENTROPY, and PSC across five models—four open-weight (Gemma-3, OLMo-3) at N=100 and closed-source GPT-5.4-Nano at N=800. We test (i) adaptive Bayesian placement versus random shuffling, (ii) signal anchoring versus uncertainty reduction, and (iii) transfer to closed-source APIs at scale.

Signal anchoring versus plain entropy placement. In our experiments, GP-BELIEF generally outperforms plain GP-ENTROPY with pronounced positional heterogeneity. On Gemma-3-12B (PIR), GP-BELIEF achieves $F_1 \approx 0.88$ at iteration 4 versus 0.72 for GP-ENTROPY, a gap that persists

across model scales. This supports protecting likely signals in sparse retrieval, rather than demoting them when their entropy falls. In simulation on the calibrated profiles, exempting confident positives from demotion largely closes the gap (§D.3); the result concerns this placement mechanism, not every active-learning policy. GP-ENTROPY can underperform because it prioritizes documents near $b \approx 0.5$—predominantly irrelevant distractors not yet ruled out. By contrast, GP-BELIEF anchors likely signals in high-diagnosticity positions to accelerate separation.

Adaptive placement can improve on random shuffling. Even with the plain entropy objective, the Bayesian framework generally outperforms PSC's random-placement baseline. On OLMo-3-7B (PIR), GP-ENTROPY attains $F_1 \approx 0.25$ at iteration 8 compared to PSC's 0.17. The gains come from *adaptivity*: PSC often re-tests resolved negatives in high-value positions, whereas GOLD PANNING accumulates evidence, deprioritizing low-belief documents to make room for promising ones. The theoretical rate comparison in Cor. A.6 additionally requires the stated informative-exposure assumptions.

Table 1: **Comparison of Gold Panning against RankGPT** baselines on the gold-document identification task (F_1@1) with GPT-5.4-Nano over $N = 800$ candidate documents. RankGPT-Single issues one ranking call over the full context; RankGPT-Sliding, PSC, and GP-Belief each use eight calls; this matches calls, not token cost. Best in bold.

Method	MonoRel	PIR
RankGPT-Single	0.778	0.628
RankGPT-Sliding	0.867	0.700
PSC	0.983	0.787
GP-Belief	**0.989**	**0.820**

Query efficiency. Across the 24 open-weight model×task×N settings, GP-BELIEF saves 50–87.5% of calls at matched PSC accuracy in nine settings. Other informative settings show +4–12 percentage points at matched budget, while flat or saturated settings offer little saving; Tab. 10 gives model-specific examples. On Gemma-3-12B (MonoRel), GP-BELIEF reaches ≈ 0.97 in just 4 iterations, matching PSC's performance at iteration 8. On PIR, the gains are even larger: GP-BELIEF's iteration-4 performance (0.88) strictly exceeds PSC's final iteration-8 result (0.82), making it highly viable for latency-constrained applications.

Nearly flat profiles yield little gain. Gemma-3-27B (Fig. 2, column 2) serves as a control as this model exhibits a fairly flat diagnosticity profile (Fig. 5). Reordering yields little gain here; the methods reach comparable final accuracy. This control links the gain to positional heterogeneity. Flatness here means similar diagnosticity across positions, not necessarily zero diagnosticity: uniformly informative reads can lock in quickly without benefiting from reordering. Calibration can therefore guide whether to use reordering; capability alone does not determine applicability.

Closed-box generalization at scale. GP-BELIEF extends cleanly to closed-source APIs at deployment scale. On GPT-5.4-Nano (N=800, $>$ 100K tokens, an 8× scale-up from the open-weight default), GP-BELIEF outperforms PSC across all iteration budgets, with the largest gap at intermediate rounds: at T=4, GP-BELIEF reaches $F_1 \approx 0.73$ versus PSC's 0.63 (ΔF_1=0.10). The gap narrows by T=8 as PSC accumulates sufficient evidence, but GP-BELIEF still leads. Against RankGPT at matched query budgets (Tab. 1), GP-BELIEF leads on both tasks: 0.989 vs 0.867 (sliding-window) on MonoRel and 0.820 vs 0.700 on PIR. Two takeaways: (i) the framework operates strictly via API-observable citations, and (ii) the advantage persists at production scale.

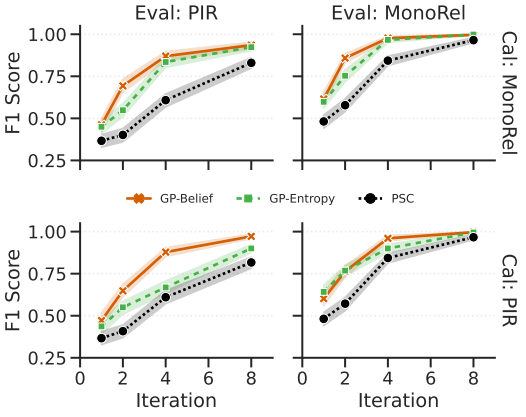

Figure 4: **Cross-task transferability for Gemma-3-12B ($N = 100$).** Rows show calibration (MonoRel, PIR); columns show evaluation (PIR, MonoRel). Matched calibration is on the anti-diagonal; cross-calibration is on the diagonal. GP-BELIEF retains its advantage under cross-calibration with modest degradation. The profile transfers despite imperfect rate matching; the static ablation (Tab. 2) separately tests the value of sequential updating.

4.3 Cross-dataset Transferability

If position bias is a structural model property (as suggested by Fig. 3) rather than dataset-specific, diagnosticity profiles should transfer across datasets. We test this via cross-evaluation: *applying the diagnosticity profile calibrated on MonoRel to guide GP on PIR, and vice-versa.* As

shown in Fig. 4, **GP-BELIEF retains its advantage under cross-calibration**, with only modest degradation relative to matched calibration. This shows tolerance to imperfect rate estimates and positional rankings. It does not establish that ordering alone accounts for the gain: §4.4 compares adaptive and static placement using the same profile.

4.4 Mechanism and Robustness Checks

Static placement. Three Gemma-3-12B ablations share the calibration without sequential belief updates: STATIC-FIXED repeats one layout; STATIC-ROTATE sweeps documents through diagnosticity-ranked slots; STATIC-FILTERANCHOR pins round-1 citations to informative slots thereafter. At $N{=}100$ (600 trials per arm), GP-BELIEF reaches the best static arm's round-8 F_1 by round 4 on both tasks (Tab. 2), halving the query budget. On mid-context PIR starts, STATIC-FIXED ends at 0.15–0.32, versus 0.95–1.00 for GP-BELIEF at every start (§C.2). This isolates the value of sequential updating and adaptive placement over the tested static policies.

Table 2: Static-placement ablation on Gemma-3-12B, $N{=}100$, 600 trials per arm. Entries are Top-1 F_1 (MonoRel / PIR). FilterAnchor uses round 1 as its filter read; predictions start at round 2.

Placement	Round 1	Round 2	Round 4	Round 8
GP-Belief	0.61 / 0.49	0.85 / 0.77	0.99 / 0.91	1.00 / 0.98
Static-Rotate	0.42 / 0.38	0.52 / 0.40	0.80 / 0.62	0.98 / 0.86
PSC	0.45 / 0.36	0.56 / 0.42	0.82 / 0.62	0.96 / 0.82
Static-Fixed	0.46 / 0.36	0.52 / 0.43	0.57 / 0.47	0.60 / 0.49
Static-FilterAnchor	– / –	0.70 / 0.56	0.75 / 0.65	0.78 / 0.68

Protecting confident positives. GP-Entropy-Withhold retains entropy scoring but exempts documents with belief above 0.9 from demotion. In simulation on Gemma-3-12B's calibrated PIR profile, round-4 Top-1 performance reaches 0.95, matching GP-Belief versus 0.78 for plain GP-Entropy. Protecting found signal largely closes this gap (§D.3).

Realized layouts. On Gemma-3-12B ($N{=}100$), visitation-weighted TPR under GOLD PANNING closely matches calibration (0.977 vs. 0.983 MonoRel; 0.910 vs. 0.908 PIR), and false-citation rates fall below the profile mean. Residual citation dependence remains on MonoRel; these checks support the measured operating regime without proving posterior calibration (§B.5).

Beyond one random needle. Real-model $k{=}5$ runs on Gemma-3-12B ($N{=}100$, 300 trials per setting) improve answer accuracy with both distinct and redundant golds. With five distinct question–gold pairs, round-8 PIR answer accuracy is 0.700 for GOLD PANNING versus 0.612 for PSC. With five redundant copies, at least one copy is cited in 96–97% of GOLD PANNING reads across the two tasks, versus 62% for PSC. PSC nevertheless leads on PIR identify-all-five set recovery. A nearest-neighbor semantic-distractor probe (Gemma-3-12B, MonoRel, $N{=}50$) preserves profile heterogeneity: mean diagnosticity is 0.852 versus 0.797 with random distractors, and the stale random-distractor calibration retains the benefit. Initial-layout sensitivity recovers in later rounds (§C.3). These controlled probes extend the evidence within single-hop retrieval; they do not establish open-domain or multi-hop performance.

5 Related Work

Exploiting systematic biases. Systematic biases are typically viewed as obstacles, but prior work in human–computer interaction has leveraged cognitive biases to optimize interfaces [5, 6] and guide user behavior [3, 18]. In LLMs, predictable regularities enable similar efficiency gains, as in speculative decoding and early-exit methods [13, 23, 2]. In a similar spirit, we leverage position bias as a *diagnostic signal* to improve in-context NIAH retrieval.

Position bias in LLMs. Position bias in long-context LLMs is well-studied [31, 39, 16, inter alia], with mitigations broadly divided into open-box and closed-box approaches. *Open-box methods* suppress bias by modifying model internals, including alternative positional encodings [38] and attention manipulations [22, 9, 32]. *Closed-box methods* instead treat the model as fixed, such as Permutation Self-Consistency (PSC) [28], which averages predictions over random permutations. PSC retains citation counts across queries but uses nonadaptive placements. While Liu et al. [17]

explores relevance via internal model states, we operate purely on API outputs while accumulating evidence across queries.

Active search and iterative retrieval. We frame context reordering as a *Bayesian active search* problem [8, 10], distinct from conventional active learning [24, 25]. Bandit-based methods [26] have been applied to optimize retrieval in RAG systems [4], but these primarily operate at training time. At inference time, GOLD PANNING is most closely related to iterative reranking methods such as RankGPT [27]. RankGPT prompts the model to explicitly sort candidates, which is itself subject to position bias; GOLD PANNING instead infers relevance from the model's natural responses under structured layouts, requiring no instruction tuning. We compare directly in §4.2.

6 Discussion, Limitations, and Conclusion

Recursive scaffolds and agentic reranking. Recent inference-time methods address long-context reasoning by treating the prompt as an external object to be explicitly manipulated. For instance, RankGPT [27] prompts the model to explicitly sort candidates, while Recursive Language Models (RLMs) [36] programmatically decompose inputs via recursive sub-queries. However, explicit sorting instructions are themselves vulnerable to position bias. GOLD PANNING bypasses this by inferring relevance from natural responses under structured layouts; as shown in Tab. 1, our signal-anchoring method consistently outperforms RankGPT in both single-shot and sliding-window configurations. Furthermore, while RLMs require models capable of reliable code execution, GOLD PANNING sits at a highly accessible point on the cost-capability frontier, requiring only parsable citations. Crucially, these approaches compose cleanly: layering GOLD PANNING onto the inner calls of an RLM or a sliding-window agent could tighten per-chunk reliability while preserving outer scaling.

Limitations. While highly effective, GOLD PANNING relies on a few core assumptions. Foremost is the conditional independence approximation of our observation model (Eq. (1)), which treats citation events as independent given position and relevance. While autoregressive decoding can introduce within-round dependencies (e.g., implicit citation budgets), the realized-layout checks in §B.5 measure their size on Gemma-3-12B; fixed marginal rates remain an approximation. Additionally, our current evaluation focuses on controlled multi-document QA with deterministic labels. The real-model multi-needle and nearest-neighbor probes broaden this controlled evidence, while open-domain evaluation, graded relevance, adversarial retrieval, and multi-hop evidence remain outside the demonstrated scope. Multi-hop retrieval would require beliefs over jointly sufficient document sets and a corresponding observation model. A comprehensive discussion of limitations, including recalibration triggers and unconstrained decoding behavior, is provided in §G.

Anchoring's benefits depend on both positional heterogeneity and detector reliability. Heterogeneity determines where to place documents; reliability determines how much evidence each read provides. On Gemma-3-12B PIR, the fraction of trials without a positive belief margin within eight rounds rises from 2% at $N{=}100$ to 41% at $N{=}200$ (§C.1). This illustrates why the empirical time-to-margin distribution matters alongside the conditional logarithmic rate: anchoring can reinforce found signal without guaranteeing that every instance enters the favorable regime.

Conclusion. We introduce GOLD PANNING, a Bayesian framework that exploits position bias as a diagnostic signal for long-context retrieval. By treating positions as noisy detectors with calibrated reliability, and actively anchoring high-belief documents in high-diagnosticity positions, signal anchoring improves query efficiency where profiles are heterogeneous. The theory is conditional on positive-margin lock-in; static-placement and entropy-withhold ablations identify the value of updating and protecting found signal. Experiments include open-weight and closed-source models at scales up to $N = 800$ documents. More broadly, our results suggest that predictable inference-time biases can be leveraged as information rather than treated as noise. Natural extensions include graded relevance via multinomial beliefs, evaluation under adversarially-constructed semantic distractors, and open-box variants that manipulate positional encodings directly while reusing the calibrated diagnosticity profile.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers and area chairs for their comments and feedback, as well as Cecily Agu, Hannah Collison, Tianjian Li, Andrew Wang, and Jack Zhang for their

constructive discussions which have helped to improve this work. This work was carried out in part at the Advanced Research Computing at Hopkins (ARCH) core facility, which is supported by the National Science Foundation (NSF) grant number OAC1920103. Adam Byerly is supported by the Johns Hopkins University Applied Physics Laboratory (APL).

References

[1] Adam Byerly and Daniel Khashabi. Self-consistency falls short! The adverse effects of positional bias on long-context problems. *Transactions of the Association for Computational Linguistics*, 14:292–317, 2026. doi: 10.1162/tacl.a.625. URL https://aclanthology.org/2026.tacl-1.15/.

[2] Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. EE-LLM: Large-scale training and inference of early-exit large language models with 3D parallelism. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7163–7189. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24ae.html.

[3] Robert B Cialdini. *Influence: The psychology of persuasion*, volume 55. Collins New York, 2007.

[4] Shaohua Duan, Xinze Li, Zhenghao Liu, Xiaoyuan Yi, Yukun Yan, Shuo Wang, Yu Gu, Ge Yu, and Maosong Sun. Chunks as arms: Multi-armed bandit-guided sampling for long-context LLM preference optimization. *arXiv preprint arXiv:2508.13993*, 2025.

[5] Krzysztof Gajos and Daniel S Weld. Preference elicitation for interface optimization. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 173–182, 2005.

[6] Krzysztof Z Gajos, Jacob O Wobbrock, and Daniel S Weld. Improving the performance of motor-impaired users with automatically-generated, ability-based interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 1257–1266, 2008.

[7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[8] Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, and Richard Mann. Bayesian optimal active search and surveying. In *International conference on machine learning*, page 843–850. PMLR, 2012.

[9] Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14982–14995, 2024.

[10] Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic active search. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1714–1723. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jiang17d.html.

[11] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.307. URL https://aclanthology.org/2023.acl-long.307/.

[12] Wenjun Ke, Yifan Zheng, Yining Li, Hengyuan Xu, Dong Nie, Peng Wang, and Yao He. Large language models in document intelligence: A comprehensive survey, recent advances, challenges and future trends. *ACM Transactions on Information Systems*, 2025.

[13] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

[14] Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[16] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 2024.

[17] Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, and Hanghang Tong. Selfelicit: Your language model secretly knows where is the relevant evidence. *arXiv preprint arXiv:2502.08767*, 2025.

[18] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–32, 2019.

[19] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

[20] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[21] Charles S Peirce. The numerical measure of the success of predictions. *Science*, (93):453–454, 1884.

[22] Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

[23] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. *Advances in Neural Information Processing Systems* (NeurIPS), 35:17456–17472, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6fac9e316a4ae75ea244ddcef1982c71-Paper-Conference.pdf.

[24] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[25] Burr Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

[26] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

[27] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? Investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL https://aclanthology.org/2023.emnlp-main.923/.

[28] Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Türe. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340, 2024.

[29] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[30] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2023.

[31] Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. Primacy effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115, 2023.

[32] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*, 2024.

[33] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

[34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[35] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[36] Alex L Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025.

[37] Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Yi Chang, et al. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*, 2025.

[38] Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, et al. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37:60755–60775, 2024.

[39] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Appendix Contents

A Theoretical Analysis and Proofs **15**

A.1 Setup and Assumptions 15

A.2 Consistency 15

A.3 Sample Complexity and Policy Comparison 16

A.4 Entropy Demotion and Belief Anchoring 18

B Calibration and Detector Validation **19**

B.1 Calibration Protocol 19

B.2 Sparse-Grid Validation 20

B.3 Held-Out Validation 20

B.4 Cross-Task Profile Stability 21

B.5 Calibration and Citation Dependence Under Realized Layouts 21

B.6 Retrieval-Head Attention Probe 22

C Additional Real-Model Experiments **22**

C.1 Time to First Positive Belief Margin 22

C.2 Static Placement and Sequential Updating 24

C.3 Multiple Needles and Semantic Distractors 24

C.4 Model-Specific Efficiency 24

D Simulation Ablations **25**

D.1 Single-Needle Ablations ($k = 1$) 25

D.2 Multiple-Needle Ablations ($k = 5$) 26

D.3 Protecting Confident Positives in an Entropy Policy 26

E Implementation Details and Compute **26**

E.1 Output Parsing and Reliability 26

E.2 Recalibration and Amortized Cost 27

E.3 Hardware and Compute Accounting 27

F Unequal Numbers of Documents and Positions **27**

G Extended Limitations **27**

A Theoretical Analysis and Proofs

A.1 Setup and Assumptions

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying the relevance vector $Z \in \{0,1\}^N$ and observations $\{O_{t,i}\}_{t\geq 1, i\in[N]}$. Let

$$\mathcal{F}_t = \sigma\big(Z,\, O_1, \ldots, O_t,\, \sigma_1, \ldots, \sigma_{t+1}\big),$$

be the filtration after round t, with Z adjoined so we may freely condition on relevance, and σ_{t+1} included because the policy commits to round $t{+}1$'s assignment from history available at the end of round t. The log-odds $\lambda_{t,i} = \log \frac{b_{t,i}}{1-b_{t,i}}$ satisfy the additive recursion

$$\lambda_{t,i} = \lambda_{t-1,i} + \ell_{\sigma_t(i)}(O_{t,i}), \qquad \ell_j(o) = \log \frac{\Pr(O = o \mid Z = 1, \mathrm{pos} = j)}{\Pr(O = o \mid Z = 0, \mathrm{pos} = j)}, \tag{7}$$

which is Eq. (2) on the log-odds scale.

Assumptions.

- **(A1)** *Bounded log-likelihood ratios.* There exists $L < \infty$ with $|\ell_j(o)| \leq L$ for all j, o. This is automatic whenever $\mathrm{TPR}_j, \mathrm{FPR}_j \in (\varepsilon, 1{-}\varepsilon)$ for some $\varepsilon > 0$, which our calibration ensures (§B).
- **(A2)** *Persistent informative coverage.* There exist $\delta_J, \rho > 0$ such that, for each document i, almost surely

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\big\{|J_{\sigma_s(i)}| \geq \delta_J\big\} \;\geq\; \rho.$$

Remark A.2 discusses coverage under each policy.

A.2 Consistency

Theorem A.1 (Consistency of GOLD PANNING)**.** *Under (A1) and (A2), $b_{t,i} \xrightarrow{a.s.} Z_i$ for every document i.*

Proof. Condition on $Z_i = 1$. Set $X_s = \ell_{\sigma_s(i)}(O_{s,i})$ and $\mu_s = \mathbb{E}[X_s \mid \mathcal{F}_{s-1}]$. The observation law gives

$$\mu_s = D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{TPR}_{\sigma_s(i)}) \,\big\|\, \mathrm{Bern}(\mathrm{FPR}_{\sigma_s(i)})\big) \;\geq\; 0, \tag{8}$$

with equality exactly when $J_{\sigma_s(i)} = 0$. Thus $M_t = \sum_{s=1}^{t}(X_s - \mu_s)$ is a martingale and

$$\lambda_{t,i} = \lambda_{0,i} + \underbrace{\sum_{s=1}^{t} \mu_s}_{\text{drift}} + \underbrace{M_t}_{\text{martingale}}. \tag{9}$$

By (A1), $|X_s|, |\mu_s| \leq L$ and $|M_s - M_{s-1}| \leq 2L$. Azuma–Hoeffding yields

$$\mathbb{P}\big(|M_t| \geq \varepsilon t\big) \;\leq\; 2\exp\!\left(-\frac{\varepsilon^2 t}{8L^2}\right).$$

This bound is summable in t; Borel–Cantelli, applied for arbitrary $\varepsilon > 0$, gives

$$\frac{M_t}{t} \;\xrightarrow{a.s.}\; 0. \tag{10}$$

For informative positions, define

$$\mu_\star \;=\; \min_{j\,:\,|J_j|\geq\delta_J} D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{TPR}_j) \,\big\|\, \mathrm{Bern}(\mathrm{FPR}_j)\big).$$

Pinsker gives $\mu_\star \geq 2\delta_J^2$. By (A2), these positions are visited with limiting lower frequency at least ρ; elsewhere $\mu_s \geq 0$. Hence

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mu_s \;\geq\; \rho\,\mu_\star \;\geq\; 2\rho\delta_J^2 \;=\; \underline{\mu}_{\mathrm{avg}} > 0 \qquad \text{a.s.} \tag{11}$$

Combining Eqs. (9), (10), and (11), with $\lambda_{0,i} = 0$, gives

$$\liminf_{t\to\infty} \frac{\lambda_{t,i}}{t} \;\geq\; \underline{\mu}_{\text{avg}} > 0 \quad \text{a.s.}$$

Consequently $\lambda_{t,i} \to +\infty$ and $b_{t,i} \to 1$. For $Z_i = 0$,

$$\mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = -D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{FPR}_{\sigma_s(i)}) \,\big\|\, \mathrm{Bern}(\mathrm{TPR}_{\sigma_s(i)})\big) \;\leq\; 0,$$

whose magnitude is at least $2J^2_{\sigma_s(i)}$. Reversing signs in the same argument gives $\lambda_{t,i} \to -\infty$ and $b_{t,i} \to 0$. □

Remark A.2 (On Assumption (A2)). For PSC, the SLLN applied to informative-position indicators gives $\rho = |\{j : |J_j| \geq \delta_J\}|/N$. For GP-Belief and GP-Entropy, coverage is a structural assumption on the joint belief-and-sorting dynamics; we do not prove it. Uniform initialization creates ties, and subsequent evidence can change the ranking. The observed convergence is consistent with coverage, under which Theorem A.1 applies.

A.3 Sample Complexity and Policy Comparison

For finite-time recovery, strengthen (A2) to a per-round drift floor:

(A2′) *Per-round drift lower bound.* There exists $\underline{\mu} > 0$ such that, almost surely, $\mu_s \geq \underline{\mu}$ for every $s \geq 1$ and every document i (with the analogous lower bound on $-\mathbb{E}[X_s \mid \mathcal{F}_{s-1}, Z_i = 0]$).

(A2′) is implied by every position having diagnosticity at least δ_J, since by Pinsker $\mu_s \geq 2J^2_{\sigma_s(i)} \geq 2\delta_J^2$ in both signs of Z_i. For PSC the per-round drift is random rather than deterministically bounded below; we recover the same sample complexity rate without (A2′) in Remark A.5.

Theorem A.3 (Sample Complexity of GOLD PANNING)**.** *Under (A1) and (A2′), fix $\delta \in (0, 1)$. If the number of rounds satisfies*

$$T \;\geq\; \frac{8L^2}{\underline{\mu}^2} \log\frac{2N}{\delta}, \tag{12}$$

then, with probability at least $1 - \delta$, $\mathrm{sign}(\lambda_{T,i}) = \mathrm{sign}(2Z_i - 1)$ for every document $i \in \{1, \ldots, N\}$ simultaneously. Equivalently, the threshold-0.5 rule classifies all N documents correctly with probability at least $1 - \delta$.

Proof. For $Z_i = 1$ and $\lambda_{0,i} = 0$, the decomposition and (A2′) give

$$\lambda_{T,i} = \sum_{s=1}^{T} \mu_s + M_T \;\geq\; T\underline{\mu} + M_T.$$

Thus misclassification, $\lambda_{T,i} \leq 0$, requires $M_T \leq -T\underline{\mu}$. Since $|M_s - M_{s-1}| \leq 2L$, Azuma–Hoeffding gives

$$\mathbb{P}\big(M_T \leq -T\underline{\mu}\big) \;\leq\; \exp\left(-\frac{(T\underline{\mu})^2}{2 \cdot T \cdot (2L)^2}\right) \;=\; \exp\left(-\frac{T\underline{\mu}^2}{8L^2}\right). \tag{13}$$

Under Eq. (12),

$$\mathbb{P}\big(\lambda_{T,i} \leq 0 \,\big|\, Z_i = 1\big) \;\leq\; \frac{\delta}{2N}.$$

The sign-reversed argument for $Z_i = 0$ contributes another $\delta/(2N)$. A union bound over documents yields

$$\mathbb{P}\big(\exists\, i : \mathrm{sign}(\lambda_{T,i}) \neq \mathrm{sign}(2Z_i - 1)\big) \;\leq\; N \cdot \frac{\delta}{N} \;=\; \delta.$$ □

Reading the rate. The policy enters the bound through the drift floor $\underline{\mu}$. Substituting the Pinsker bound $\underline{\mu} \geq 2\delta_J^2$ where $\delta_J = \min_j |J_j|$ gives

$$T \;=\; \mathcal{O}\!\left(\frac{L^2 \log(N/\delta)}{\delta_J^4}\right),$$

Using Bernoulli KL directly can tighten this constant. Corollary A.6 compares targeted and random exposure under a heterogeneous profile.

Remark A.4 (Belief margin). If the desired guarantee is a belief margin $b_{T,i} \geq 1-\eta$ (rather than only sign-correctness), the threshold becomes $\tau_\eta = \log\frac{1-\eta}{\eta}$, and the tail event becomes $\{\lambda_{T,i} < \tau_\eta\} \subseteq \{M_T < \tau_\eta - T\underline{\mu}\}$. The same concentration and union-bound argument yields

$$T \;\geq\; \frac{32L^2}{\underline{\mu}^2}\log\frac{2N}{\delta} \;+\; \frac{2\tau_\eta}{\underline{\mu}}.$$

The additive term accounts for reaching the confidence threshold; the $\log(N/\delta)$ dependence is unchanged.

Remark A.5 (Unconditional bound for PSC). Under PSC, σ_t is uniform on S_N independently of history, so the drifts $\{\mu_s\}_{s\geq 1}$ are i.i.d. with common mean

$$\bar{\mu} \;=\; \frac{1}{N}\sum_{j=1}^{N} D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{TPR}_j) \,\big\|\, \mathrm{Bern}(\mathrm{FPR}_j)\big).$$

Although (A2′) need not hold pointwise, Hoeffding's inequality on the i.i.d. sequence $\{\mu_s\}$ gives $\sum_s \mu_s \geq T\bar{\mu}/2$ with probability at least $1-\exp(-T\bar{\mu}^2/(2L^2))$. Combining this with the Azuma bound (13) (with $T\bar{\mu}/2$ in place of $T\underline{\mu}$) and a union bound over the two failure modes yields the unconditional sample complexity bound

$$T \;\geq\; \frac{32L^2}{\bar{\mu}^2}\log\frac{4N}{\delta}$$

for PSC, replacing the drift floor with position-averaged drift.

Corollary A.6 (GOLD PANNING vs. PSC under heterogeneous diagnosticity)**.** *Suppose the position profile contains K* informative *positions with $D_{\mathrm{KL}}(\mathrm{Bern}(\mathrm{TPR}_j) \,\|\, \mathrm{Bern}(\mathrm{FPR}_j)) \geq c$ and $N-K$* deadzone *positions with $D_{\mathrm{KL}}(\mathrm{Bern}(\mathrm{TPR}_j) \,\|\, \mathrm{Bern}(\mathrm{FPR}_j)) = o(c)$. Suppose further that* GOLD PANNING *satisfies (A2′) with $\underline{\mu} \geq c$, i.e. its policy places the document at an informative position on every round. Then the sample complexities to recover all labels with probability at least $1-\delta$ satisfy*

$$T_{\text{GOLD PANNING}} \;=\; \mathcal{O}\!\left(\frac{L^2}{c^2}\log\frac{N}{\delta}\right), \qquad T_{\text{PSC}} \;=\; \mathcal{O}\!\left(\frac{L^2}{c^2}\cdot\left(\frac{N}{K}\right)^2\log\frac{N}{\delta}\right). \tag{14}$$

Proof. The GOLD PANNING bound follows from substituting $\underline{\mu} \geq c$ into Theorem A.3. For PSC, the average drift satisfies

$$\bar{\mu} \;=\; \frac{1}{N}\sum_{j=1}^{N} D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{TPR}_j) \,\big\|\, \mathrm{Bern}(\mathrm{FPR}_j)\big) \;=\; \Theta\!\left(\frac{Kc}{N}\right),$$

since informative positions contribute $\Theta(Kc)$ in aggregate and deadzone positions contribute $o(Kc)$ by hypothesis. Substituting into Remark A.5 yields the second bound. □

Interpretation. Equation (14) compares sufficient bounds under the corollary's distinct exposure assumptions. Its $(N/K)^2$ factor is a best-case remark: it requires GOLD PANNING to maintain the informative drift floor c, whereas PSC receives position-averaged drift. It does not bound discovery time or establish a lower bound on PSC's actual complexity. For uniform diagnosticity, $K = \Theta(N)$ and the factor is constant.

Section C.4 compares the policies using empirical exposure.

A.4 Entropy Demotion and Belief Anchoring

Throughout this subsection we work in the *sparse-search lost-in-the-middle* regime: there are k relevant documents and $N-k$ distractors, only K of the N positions are informative ($|J_j| \geq \delta_J$), and the remaining $N-K$ positions are deadzone ($|J_j| = o(\delta_J)$), with $K+k \leq N$. We write $\mu_{(1)} \geq \mu_{(2)} \geq \cdots \geq \mu_{(N)}$ for the position drifts in descending order.

A.4.1 GP-Entropy: Drift Collapse

Proposition A.7 (GP-ENTROPY demotes confident positives)**.** *Under (A1), suppose at round t at least $N-k$ distractors at deadzone positions retain $b_{t-1,j} \approx \frac{1}{2}$, so $H(b_{t-1,j}) \approx \log 2$. As the beliefs of all k true positives approach certainty, $\min_{i:Z_i=1} b_{t-1,i} \to 1$, the entropy ranking under* GP-ENTROPY *relegates the k true positives to ranks $\{N-k+1,\ldots,N\}$. The greedy matching scheme then assigns each true positive a position with drift bounded above by $\mu_{(N-k+1)}$. In the lost-in-the-middle regime ($N-k+1 > K$), $\mu_{(N-k+1)} = o(\delta_J)$, so the per-round drift μ_s on each true positive collapses to $o(\delta_J)$ asymptotically.*

Proof. For each true positive, $H(b_{t-1,i}) \to 0$, whereas the $N-k$ hypothesized ambiguous distractors retain entropy near $\log 2$. Descending entropy order therefore puts the k true positives in the bottom k ranks. Algorithm 1, lines 8–9, assigns them drifts $\mu_{(N-k+1)},\ldots,\mu_{(N)}$, each at most $\mu_{(N-k+1)}$. Since $N-k+1 > K$, this upper bound is $o(\delta_J)$. □

A.4.2 GP-Belief: Drift Persistence via Lock-in

Proposition A.8 (GP-BELIEF lock-in via Lundberg bound)**.** *Under (A1), suppose at some round τ the k true positives occupy the top-k belief ranks under* GP-BELIEF*, with margin*

$$\Delta_\tau \;=\; \min_{i:Z_i=1} \lambda_{\tau,i} - \max_{j:Z_j=0} \lambda_{\tau,j} \;>\; 0,$$

and suppose the top $k+1$ positions (by diagnosticity) are informative ($|J_{(\ell)}| \geq \delta_J$ for $\ell \leq k+1$). Then there exists $\gamma > 0$ depending only on δ_J and L such that the favorable configuration persists for all $t > \tau$ with probability at least

$$1 \;-\; k(N-k)\, e^{-\gamma \Delta_\tau}.$$

Proof. For each gold–distractor pair, set $\nu_{i,j,t} = \lambda_{t,i} - \lambda_{t,j}$. Then

$$\Delta_t \;=\; \min_{(i,j):Z_i=1,Z_j=0} \nu_{i,j,t}, \tag{15}$$

because the minimum pairs the lowest-scoring gold with the highest-scoring distractor. While the favorable configuration holds, their positions have drifts in $\{\mu_{(1)},\ldots,\mu_{(k)}\}$ and $\{\mu_{(k+1)},\ldots,\mu_{(N)}\}$, respectively. Both are informative by hypothesis, so Pinsker gives

$$\begin{aligned}
\mathbb{E}[\nu_{i,j,t+1} - \nu_{i,j,t} \mid \mathcal{F}_t] &= \mathbb{E}[X_{i,t+1} \mid \mathcal{F}_t] - \mathbb{E}[X_{j,t+1} \mid \mathcal{F}_t] \\
&= D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{TPR}_{\sigma(i)}) \,\big\|\, \mathrm{Bern}(\mathrm{FPR}_{\sigma(i)})\big) \\
&\quad + D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathrm{FPR}_{\sigma(j)}) \,\big\|\, \mathrm{Bern}(\mathrm{TPR}_{\sigma(j)})\big) \\
&\geq 2J_{\sigma(i)}^2 + 2J_{\sigma(j)}^2 \;\geq\; 4\delta_J^2,
\end{aligned}$$

Write $\xi = 4\delta_J^2$. By (A1), $|\Delta\nu| = |X_i - X_j| \leq 2L$; Hoeffding's lemma gives

$$\mathbb{E}\big[e^{-\gamma\Delta\nu} \mid \mathcal{F}_t\big] \;\leq\; \exp\big(-\gamma\xi + \tfrac{1}{2}\gamma^2(4L)^2\big) \;=\; \exp(-\gamma\xi + 8\gamma^2 L^2).$$

Taking $\gamma = \xi/(16L^2)$ makes this bound $\exp(-\xi^2/(32L^2)) \leq 1$. Thus $W_{i,j,t} = e^{-\gamma\nu_{i,j,t}}$ is a nonnegative supermartingale while the favorable configuration holds. Pair failure is $\nu_{i,j,t} \leq 0$, equivalently $W_{i,j,t} \geq 1$. Doob's maximal inequality yields

$$\mathbb{P}\left(\sup_{t>\tau} W_{i,j,t} \geq 1 \,\middle|\, \mathcal{F}_\tau\right) \;\leq\; \mathbb{E}[W_{i,j,\tau} \mid \mathcal{F}_\tau] \;=\; e^{-\gamma\nu_{i,j,\tau}} \;\leq\; e^{-\gamma\Delta_\tau},$$

where Eq. (15) gives $\nu_{i,j,\tau} \geq \Delta_\tau$. The configuration fails exactly when one of the $k(N-k)$ pairs fails, so

$$\mathbb{P}\big(\exists\, t > \tau : \Delta_t \leq 0 \,\big|\, \mathcal{F}_\tau\big) \;\leq\; \sum_{(i,j)} \mathbb{P}\left(\sup_{t>\tau} W_{i,j,t} \geq 1 \,\middle|\, \mathcal{F}_\tau\right) \;\leq\; k(N-k)\, e^{-\gamma \Delta_\tau}.$$

The complement is the lock-in event. □

Remark on the lock-in time τ. We do not formally bound τ. Section C.1 reports the empirical time to first positive margin across models, tasks, and context sizes. A first positive margin alone does not certify permanent lock-in or a negligible ruin probability.

A.4.3 Summary of the Rate Comparison

After lock-in, each true positive occupies a top-k position, giving drift at least $\mu_{(k)}$. Combining Propositions A.7 and A.8 with Theorem A.3 gives

$$T_{\text{GP-BELIEF}} \;=\; \mathcal{O}\left(\frac{L^2}{\mu_{(k)}^2} \log \frac{N}{\delta}\right) \quad \text{with probability at least } \; 1 - k(N-k)\, e^{-\gamma \Delta_\tau}.$$

For GP-Entropy, the corresponding drift approaches $\mu_{(N-k+1)} = o(\delta_J)$, so the post-saturation argument supplies no fixed positive floor and no logarithmic guarantee. At finite horizons, beliefs do not fully saturate and empirical middle positions retain nonzero diagnosticity; GP-Entropy can therefore continue converging more slowly, as in Fig. 2. The comparison concerns asymptotic rates rather than categorical viability.

B Calibration and Detector Validation

B.1 Calibration Protocol

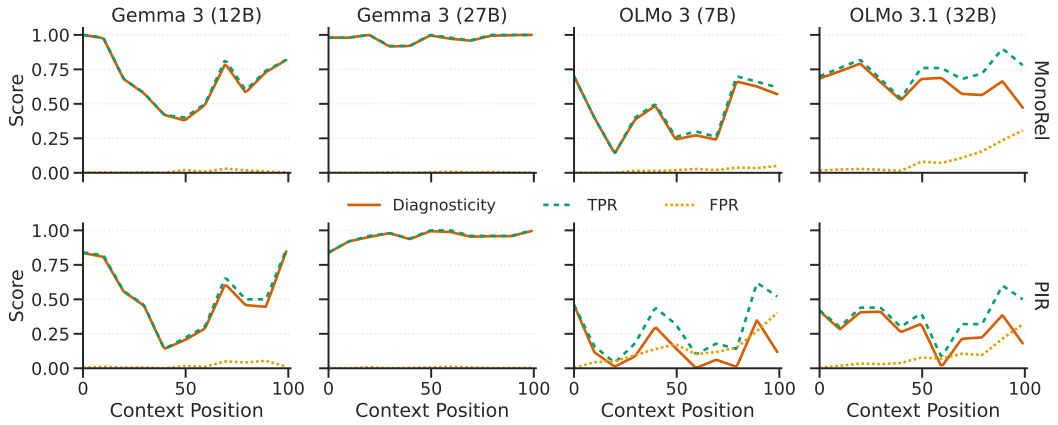

Figure 5: Position profiles across models and tasks ($N = 100$): diagnosticity $|\mathrm{TPR} - \mathrm{FPR}|$ (solid), TPR (dashed), and FPR (dotted). Gemma-3-27B has little positional variation.

For each dataset (MonoRel, PIR), place a known gold document at position j among $N-1$ random distractors and query for citation IDs. Gold citations estimate TPR_j; distractor citations estimate FPR_j. Repeat with 50 gold/distractor samples per position and average over 10 independent trials. Compute $d_{\text{diag}}(j) = |\mathrm{TPR}_j - \mathrm{FPR}_j|$.

The update adopts conditional independence given labels and assignment, while calibration estimates position-wise marginal rates under a fixed protocol. Citation budgets can induce within-read dependence; §B.5 reports the realized-layout checks and their limits.

B.2 Sparse-Grid Validation

We compare interpolation from $K = 11$ grid positions against calibration over all $N = 100$ positions. Table 3 reports mean absolute diagnosticity residuals of at most 0.071, supporting the sparse protocol used in §4.1.

Table 3: Mean absolute residual Δ (mean$_{\pm SE}$) between uniformly-sampled ($N = 100$) and fixed-grid interpolated diagnosticity ($K = 11$ positions). All reported mean residuals are at most 0.071.

Dataset	Model	Mean Δ Diagnosticity
MonoRel	Olmo-3-7B-Think	$0.054_{\pm 0.002}$
	Olmo-3.1-32B-Think	$0.051_{\pm 0.002}$
	Gemma-3-12B-IT	$0.027_{\pm 0.002}$
	Gemma-3-27B-IT	$0.008_{\pm 0.000}$
PIR	Olmo-3-7B-Think	$0.041_{\pm 0.002}$
	Olmo-3.1-32B-Think	$0.071_{\pm 0.003}$
	Gemma-3-12B-IT	$0.028_{\pm 0.002}$
	Gemma-3-27B-IT	$0.013_{\pm 0.001}$

B.3 Held-Out Validation

To assess sensitivity to the sampled facts, we split the raw 550-trial calibration data (11 positions $\times$ 50 trials) per-trial into independent train and hold-out halves of 25 trials per position, and recompute $(\mathrm{TPR}_j, \mathrm{FPR}_j)$ separately.

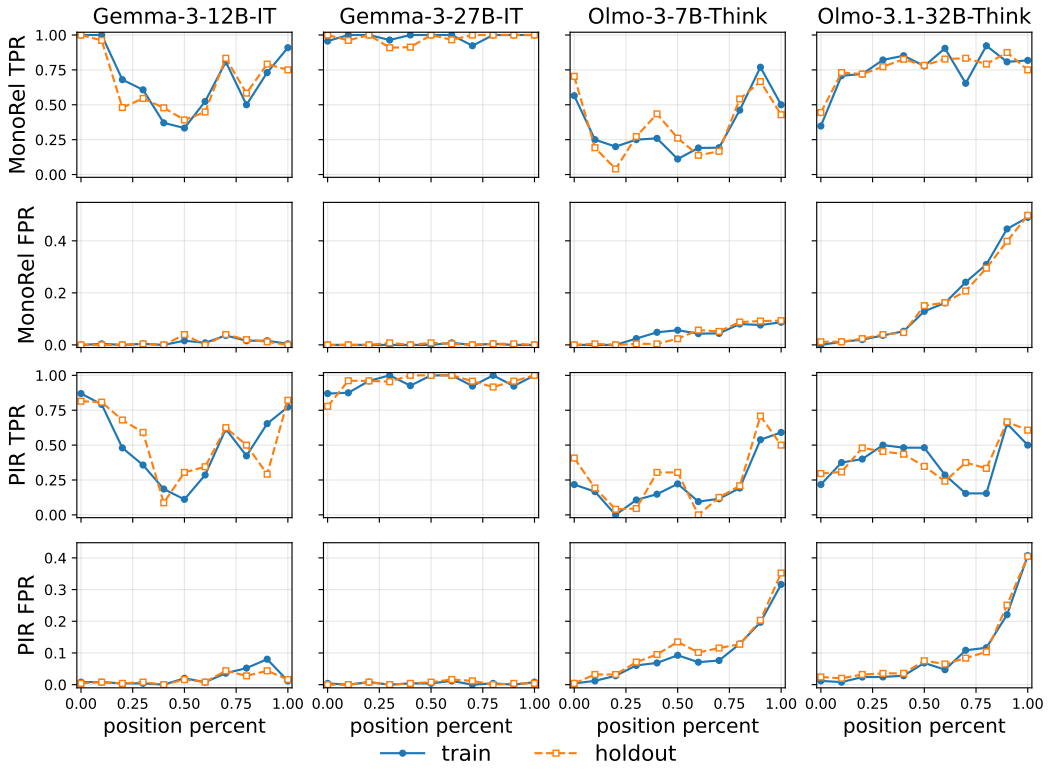

Figure 6: Held-out profile validation across four models. Train: solid, circles; hold-out: dashed, squares. Rows: MonoRel TPR, MonoRel FPR, PIR TPR, PIR FPR. Both halves contain 25 trials per grid position.

Across all 16 panels, both halves recover the same positional structure without systematic separation (Fig. 6), complementing the grid-density check.

B.4 Cross-Task Profile Stability

To quantify the cross-task stability suggested by Fig. 4, we correlate the MonoRel and PIR diagnosticity rankings at $N = 100$ (Tab. 4).

Table 4: Spearman's ρ and Kendall's τ between diagnosticity profiles calibrated on MonoRel vs. PIR. Red entries are not statistically significant ($p > 0.05$); these correspond to degenerate cases where the model exhibits minimal position bias.

Model	ρ	τ
OLMo-3-7B-Think	0.595	0.455
OLMo-3.1-32B-Think	0.488	0.315
Gemma-3-12B-IT	0.959	0.842
Gemma-3-27B-IT	0.102	0.112

Correlations are positive for models with exploitable position bias; Gemma-3-27B-IT's nearly flat profile has nonsignificant correlations.

B.5 Calibration and Citation Dependence Under Realized Layouts

Instrumented Gemma-3-12B reruns ($N = 100$, rounds ≥ 2) compare calibration against adaptive layouts. Realized and predicted gold TPR use the same visited-position mix (Tab. 5).

Table 5: Realized rates versus calibration on Gemma-3-12B, N=100. Each entry is realized / profile. Position 0 accounts for 91% / 84% of gold visits on MonoRel / PIR.

Rate	MonoRel	PIR
False-cite rate per position	0.0030 / 0.0129	0.0084 / 0.0213
Visitation-weighted TPR	0.977 / 0.983	0.910 / 0.908
TPR at position 0	1.000 / 1.000	0.964 / 0.950

False-citation rates fall below the profile mean, and visitation-weighted TPR matches calibration. Gold visits concentrate at position 0; sparse mid-context visits contain harder, not-yet-locked items and have lower TPR at several positions. These marginals do not establish uniform conditional calibration.

False citations per read when gold is cited versus not cited are 0.18/1.44 (MonoRel) and 0.65/1.51 (PIR) under GP, versus 0.63/1.44 and 1.22/2.33 under PSC. Table 6 adjusts for instance and round fixed effects with trial-cluster bootstrap 95% CIs. GPT-5.4-Nano random-layout calibration ($N = 800$, 550 trials) instead gives 0.310/0.308 ($\Delta = 0.002$, roughly 0.5% of marginal FPR per position); this different model/distribution does not establish independence under adaptation.

Table 6: Residual within-read dependence, $\Delta = (\text{gold not cited}) - (\text{gold cited})$, in false citations per read after instance and round fixed effects. Positive values indicate fewer false citations on reads that cite the gold; brackets are cluster-bootstrap 95% CIs.

Layout policy	MonoRel	PIR
GP-Belief	$+1.16\ [0.89, 1.45]$	$+0.17\ [-0.37, 0.64]$
PSC	$+0.78\ [0.56, 0.94]$	$+1.07\ [0.80, 1.35]$

Gold citation is associated with fewer false citations on MonoRel under both policies; the GP PIR interval includes zero. This supports a citation-budget association, not conservative or calibrated posterior updates in general.

B.6 Retrieval-Head Attention Probe

We probe positional and content effects on retrieval-head attention in Gemma-3-12B-IT, following Wu et al. [33].

Setup. For 100 deduplicated MonoRel facts, place the gold among $N = 100$ documents at 11 positions $(0\%, 10\%, \ldots, 100\%)$, with randomized distractors. A final-layer forward hook records attention from the last decoded token to gold-fact tokens for all 16 heads: $16 \times 11 \times 100 = 17{,}600$ observations. A deterministic 50/50 fact split separates head identification from evaluation. Rank heads by mean gold attention on the identification half; carry Head 8 into the held-out analysis, with top-K checks at $K \in \{3, 5\}$. Head 8 attends to gold approximately $12\times$ more than the least retrieval-active head.

Statistical model. Fact length and query overlap are constant within fact identity, making a model with both identity dummies and these covariates rank-deficient. We therefore fit two Type-II ANCOVAs:

- Model A: attention_mass $\sim C(\text{position}) + C(\text{content_id})$.
- Model B: attention_mass $\sim C(\text{position}) + \text{fact_length} + \text{jaccard_overlap}$.

Table 7: Type-II ANCOVA on Head 8 attention in the held-out 50-fact set. Model A adjusts for fact identity; Model B for length and lexical overlap.

Factor	$\boldsymbol{\eta^2}$	$\boldsymbol{p}$
Position	0.202	$< 10^{-20}$
Content ID	0.073	0.466
Fact Length	0.001	0.444
Jaccard Overlap	0.003	0.185

Results and interpretation. Position is the only significant systematic predictor (Tab. 7), accounting for about 20% of variance; content identity, fact length, and lexical overlap show null effects. The same pattern holds at $K = 3, 5$. The unexplained $\sim 72\%$ is attributed to trial-level stochasticity, consistent with the position-only noisy-detector model. The probe and robustness checks support the observation model; realized-layout citation checks are reported separately in §B.5.

C Additional Real-Model Experiments

C.1 Time to First Positive Belief Margin

For each of 24 open-weight model×task×N settings (600 trials), record the first checkpoint in $\{1, 2, 4, 8\}$ where gold log-odds exceed every distractor's. "Never" means no positive margin observed within eight rounds. This diagnostic does not certify permanent lock-in or bound discovery time in Proposition A.8.

Table 8 shows rapid crossings for Gemma-3-12B at $N = 100$ but substantial never-crossing mass in weaker regimes. Across settings, never-crossing trials have pooled round-8 $F_1 \approx 0.006$ (below 0.05 wherever they exceed 2%). On OLMo-3-7B PIR at $N = 100$, the approximately 46% that cross attain about 0.80 F_1, consistent with aggregate 0.37 versus PSC's 0.17. This localizes failures without excluding recovery by other policies. Gemma-3-27B crosses rapidly despite its flat profile: informative reads support discovery, while positional heterogeneity determines the opportunity for anchoring.

Table 8: Time to first positive belief margin: percentages of trials, MonoRel / PIR. "Never" is restricted to the eight-round budget.

Model	N	$\tau \leq 1$	$\tau \leq 2$	$\tau \leq 4$	Never
Gemma-3-12B	50	77 / 60	99 / 86	100 / 96	0.0 / 2.2
	100	60 / 47	85 / 72	98 / 90	0.7 / 2.0
	200	28 / 21	59 / 25	80 / 41	4 / 41
OLMo-3-7B	50	54 / 25	82 / 45	95 / 65	0.3 / 17
	100	28 / 17	48 / 26	65 / 32	18 / 54
	200	16 / 5	32 / 5	37 / 6	56 / 92
OLMo-3.1-32B	50	61 / 40	82 / 66	97 / 84	0.0 / 7
	100	48 / 15	76 / 34	94 / 58	1.2 / 20
	200	26 / 13	69 / 20	86 / 31	3 / 53
Gemma-3-27B	50	88 / 84	99 / 98	100 / 100	0.2 / 0.0
	100	78 / 83	95 / 97	98 / 99	0.8 / 0.3
	200	72 / 79	93 / 93	98 / 98	0.5 / 1.3

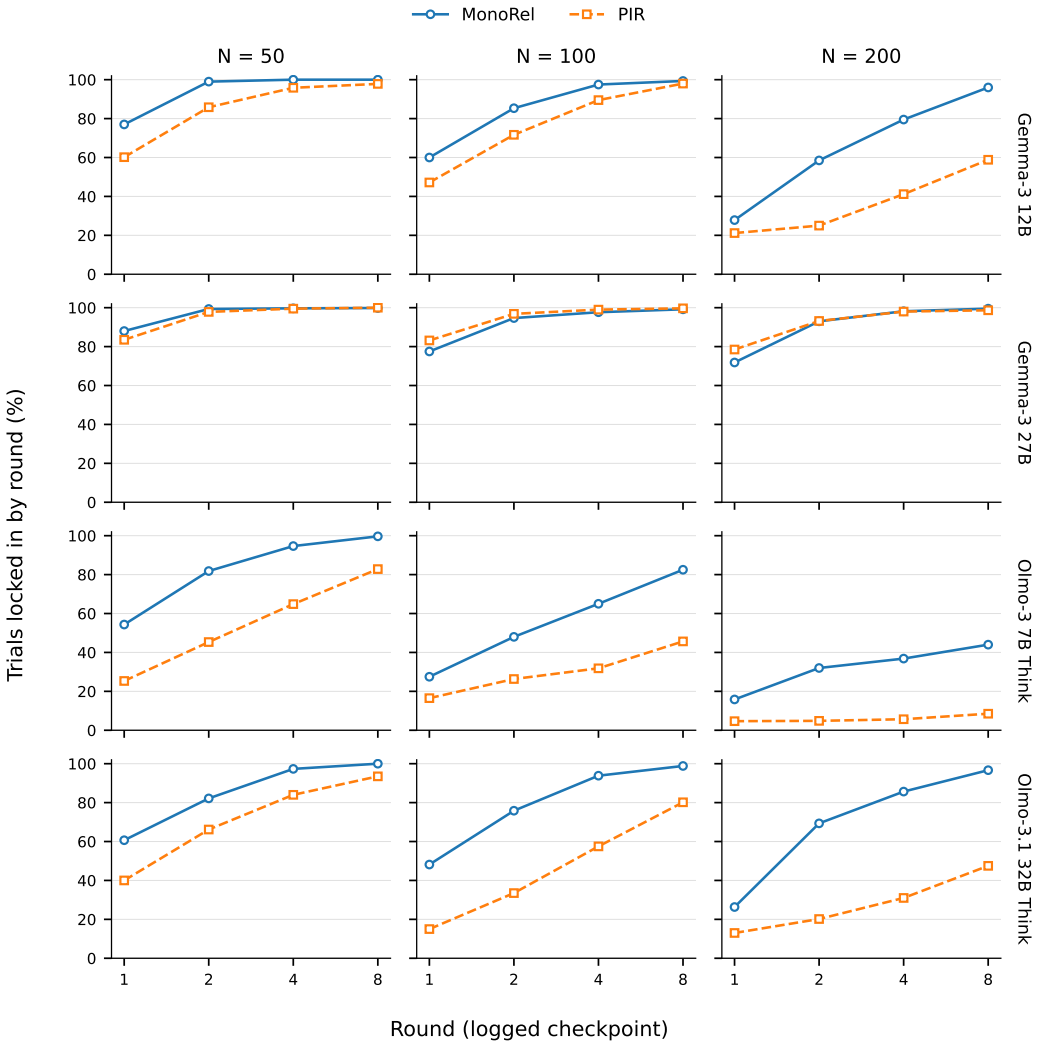

Figure 7: Time to first positive belief margin for the four open-weight models and three context sizes. Curves show the cumulative fraction with a positive margin at a logged checkpoint; they do not certify permanent lock-in.

C.2 Static Placement and Sequential Updating

All three static arms share the calibrated position ranking without Bayesian updates (Gemma-3-12B, $N = 100$, eight rounds, 600 trials per arm/task; Tab. 2). Static-Fixed repeats one layout and votes; Static-Rotate sweeps diagnosticity-ranked slots; Static-FilterAnchor pins round-1 citations to the best slots thereafter, with predictions starting at round 2.

Rotate finishes $+0.02/+0.04$ above PSC on MonoRel/PIR, versus GP-Belief's $+0.04/+0.16$; Rotate and PSC are within a few points through round 4. GP-Belief matches the best static arm's eight-round result by round 4 (50% fewer calls). It matches Fixed after 1/2 rounds on MonoRel/PIR, FilterAnchor after 2/2, PSC after 4/3, and Rotate after 4/4. On mid-context PIR starts, Fixed ends at $F_1 = 0.15$–0.32, versus GP-Belief's 0.95–1.00 at every start. This supports combined updating and adaptive placement over the tested static rules, without establishing uniqueness of Bayesian updating.

C.3 Multiple Needles and Semantic Distractors

C.3.1 Multiple Needles

Gemma-3-12B is evaluated on both tasks with $k = 5$, $N = 100$, eight rounds, and 300 trials per setting. Distinct needles pair five questions with disjoint, individually sufficient gold facts; redundant needles repeat one gold fact five times for a single question. Neither construction requires multi-hop evidence. Table 9 separates answers, citation exposure, and recovery of the full gold set.

Table 9: Gemma-3-12B real-model k=5, round 8, 300 trials per setting. Entries are GP-Belief / PSC; answer and citation metrics are distinct from recovery of the entire gold set.

Construction	Metric	MonoRel	PIR
Distinct needles	Answer accuracy	0.891 / 0.839	0.700 / 0.612
	Golds surfaced per read (of 5)	4.91 / 2.75	4.16 / 1.90
	Identify-all-five set recovery	0.883 / 0.823	0.350 / 0.397
Redundant copies	Answer accuracy	0.947 / 0.897	0.817 / 0.697
	Any copy cited per read	0.97 / 0.62	0.96 / 0.62

Answer and citation advantages persist in both constructions. On distinct-needle PIR, PSC answer accuracy declines from 0.732 at round 1 to 0.612 at round 8, versus GP's 0.700. However, PSC leads on identify-all-five PIR recovery. In redundant MonoRel, a co-cited distractor out-believes every gold copy in approximately 17% of GP trials without affecting answers. These metrics therefore measure different outcomes.

C.3.2 Semantic Distractors

With bge-large nearest-neighbor distractors (Gemma-3-12B, MonoRel, $N = 50$), mean diagnosticity is 0.852 versus 0.797 for random distractors, with similar FPR. Citations per read fall from 1.4 to 1.1; from round 3, GP cites gold in at least 97% of reads, versus about 77% for PSC. The random-distractor calibration retains its benefit.

Similarity-ordered pools expose tied-belief cold starts: worst-start round-1 F_1 is 0.15–0.37, recovering in rounds 2–3. Randomizing the initial permutation mitigates placement sensitivity. Saturation by round 4 limits this probe to profile robustness and transfer, rather than additional query savings or general adversarial robustness.

C.4 Model-Specific Efficiency

Across four open-weight models, two tasks, and three context sizes, matched-PSC-accuracy savings are 50–87.5% in 9 of 24 settings. Other unsaturated settings gain 4–12 percentage points at matched budget; flat or saturated settings offer little saving. Table 10 separates these quantities with model-specific examples.

Table 10: Model-specific efficiency examples from the reported results. Matched-accuracy savings and matched-budget accuracy gains are different quantities.

Model	Reported comparison
Gemma-3-12B	At N=100, four rounds match or exceed PSC's eight-round accuracy on both tasks (50% fewer calls).
OLMo-3-7B	PIR, N=100: round-8 F_1 is approximately 0.37 versus 0.17; substantial never-locked mass limits recovery.
OLMo-3.1-32B	PIR, N=100: +11.7 percentage points over PSC at the matched eight-round budget.
Gemma-3-27B	Little separation at N=50–100; at N=200 on MonoRel, mild heterogeneity permits four rounds to match PSC's eight.
GPT-5.4-Nano	At N=800, +10 percentage points over PSC at four rounds; see Tab. 1 for eight-round reranking comparisons.

Symmetric empirical comparison. Using the comparison in Corollary A.6, let $\hat{u}_{\mathrm{GP}}$ be gold's visitation-weighted per-round KL under GP in rounds ≥ 2, and $\bar{u}_{\mathrm{PSC}}$ the position-averaged KL. Across 24 settings, $\hat{u}_{\mathrm{GP}}/\bar{u}_{\mathrm{PSC}}$ spans $[0.86, 3.62]$ (median 1.26); its squared round-ratio proxy spans $[0.75, 13.1]$ (median $\approx$ 1.6). Flat profiles predict near-unit ratios and show little saving. These empirical means are not uniform drift floors, so the proxy is not an unconditional runtime guarantee.

D Simulation Ablations

The first two ablations vary calibration noise, detector heterogeneity, and context size ($N = 100$, 5,000 trials unless noted), reporting ΔF_1 against PSC. Section D.3 separately tests a policy modification on real calibrated profiles.

D.1 Single-Needle Ablations ($k = 1$)

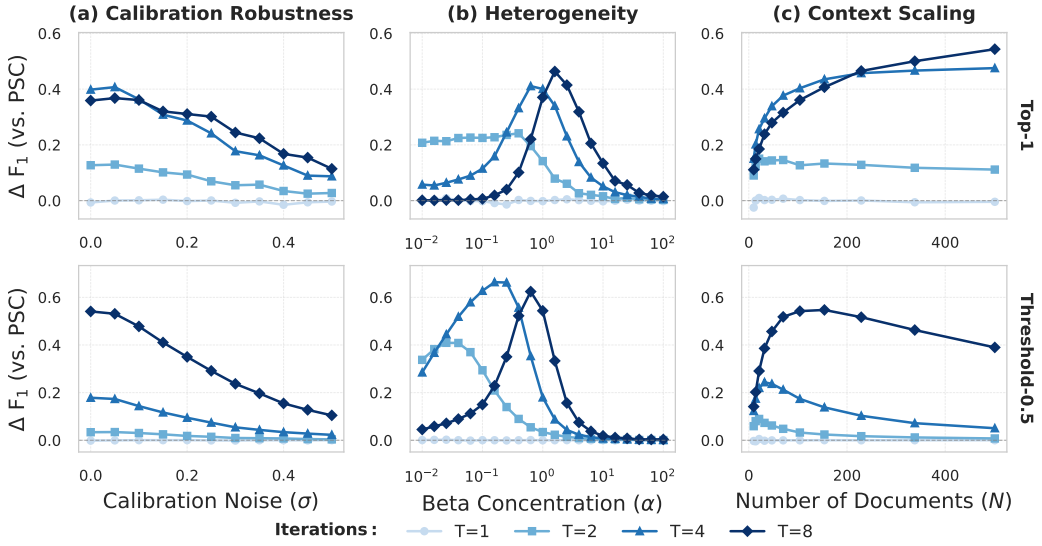

Figure 8: Single-needle simulation ablations ($k = 1$; $N = 100$, 5,000 trials unless noted). Top row: Top-1 selection; bottom row: Threshold-0.5. Panels vary (a) Gaussian noise in TPR/FPR estimates, (b) detector heterogeneity, and (c) context size.

With Gaussian calibration noise $\sigma \in [0, 0.5]$, GP-Belief retains $\Delta F_1 \approx 0.15$–0.25 at $\sigma = 0.4$ (Fig. 8a). For Beta-distributed detector quality, the advantage peaks near $\alpha \approx 1$ and vanishes as $\alpha \to \infty$ makes positions homogeneous (b). Increasing context size raises the advantage to about 0.5 at $N = 500$ (c), as random placement increasingly misses diagnostic positions.

D.2 Multiple-Needle Ablations ($k = 5$)

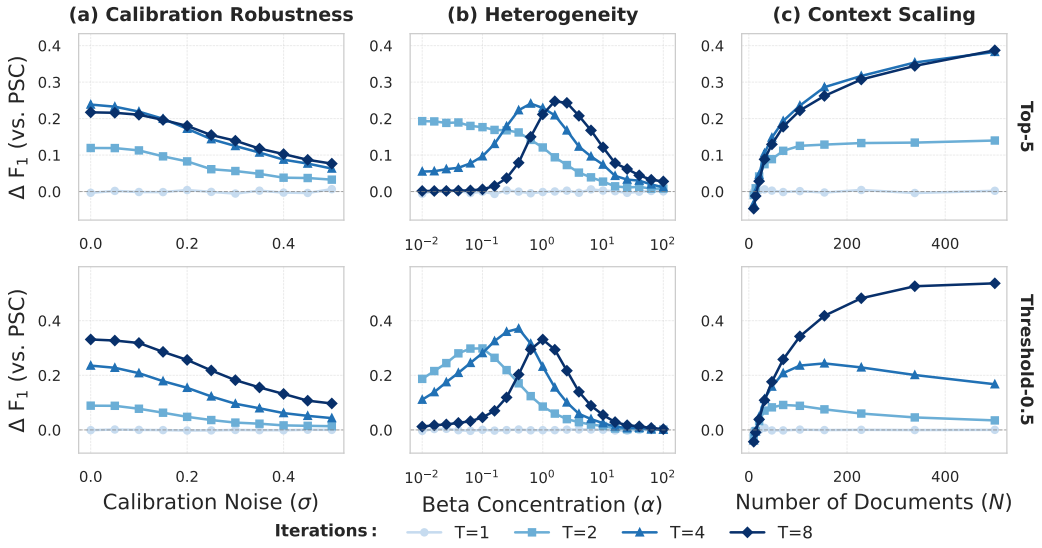

Figure 9: Ablation studies for k=5 relevant documents (simulation; N=100, 5,000 trials unless noted). Top row: Top-5 selection; bottom row: Threshold-0.5. GP-Belief maintains a positive ΔF_1 over PSC across all axes, though the magnitude is modestly attenuated relative to k=1 (Fig. 8) under the Top-5 rule.

Figure 9 repeats the three axes with $k = 5$. Top-5 gains at $T = 8$ decrease from about 0.22 at $\sigma = 0$ to 0.13 at $\sigma = 0.5$; Threshold-0.5 starts at about 0.34 (a). Heterogeneity gains peak at $\alpha \approx 0.5$–1: about 0.24 for Top-5, versus 0.45 for Top-1, and 0.30 for Threshold-0.5 (b). At $N = 500$, gains reach about 0.40/0.55 for Top-5/Threshold-0.5 at $T = 8$ (c). Threshold-0.5 does not require knowing k.

D.3 Protecting Confident Positives in an Entropy Policy

GP-ENTROPY-WITHHOLD retains entropy scoring but exempts documents with belief above 0.9 from demotion, keeping them in the positions they would receive under GP-BELIEF. Simulations on real calibrated profiles test this change separately from real-model inference. For Gemma-3-12B PIR, Top-1 performance at rounds 2 / 3 / 4 is 0.77 / 0.90 / 0.95 for GP-BELIEF, 0.77 / 0.88 / 0.95 for GP-ENTROPY-WITHHOLD, and 0.61 / 0.68 / 0.78 for plain GP-ENTROPY. Thus, protecting found signal closes most of this gap; the comparison does not establish that entropy-based policies necessarily fail.

E Implementation Details and Compute

E.1 Output Parsing and Reliability

Prompts request reasoning followed by boxed JSON with an answer and citation indices. The parser extracts the final parsable object and maps indices to document IDs. Calibration parse success is ≥ 0.95 except in one setting (0.948), with truncation at most 5.3%. Real-model $k = 5$ runs have parse success ≥ 0.98 and no truncation. Parsability does not establish faithfulness or relevance calibration.

E.2 Recalibration and Amortized Cost

Calibration costs 550 queries per model×task deployment configuration. Saving four calls per request breaks even after about 140 requests, assuming equal per-call accounting; token and latency costs depend on prompt/output lengths. Prompt, format, model-version, context-length, or domain changes may require recalibration. The transfer probes cover only the tested shifts.

E.3 Hardware and Compute Accounting

Local inference uses vLLM on H100/H200 GPUs. Each open-weight model×task×N setting has three policies, 600 trials, and eight rounds (14,400 calls). Table 11 gives single-GPU wall-clock hours, also GPU-hours. GPT-5.4-Nano uses the API; simulations use CPU.

Table 11: Wall-clock hours per open-weight model×task×N setting on one H100/H200 GPU, ranges over tasks and reruns.

Model	N=50	N=100	N=200
Gemma-3-12B	2.1–2.7	4.4–5.5	9.4–12.1
Gemma-3-27B	4.8–5.9	9.6–12.1	20.2–25.4
OLMo-3-7B	1.4–1.8	3.0–4.1	7.0–10.3
OLMo-3.1-32B	6.3–8.1	14.5–17.3	26.2–34.0

F Unequal Numbers of Documents and Positions

More documents ($N_{\mathrm{doc}} > N_{\mathrm{pos}}$). Add $N_{\mathrm{doc}} - N_{\mathrm{pos}}$ dummy positions with $\mathrm{TPR} = \mathrm{FPR} = 0$. Assigned documents are omitted, receive no observation, and retain their beliefs. GP-Belief omits the lowest-belief documents; GP-Entropy omits the lowest-uncertainty documents. Convergence requires eventual informative real-position coverage. With T rounds, the TN_{pos} available document–position measurements suffice when T is moderate.

More positions ($N_{\mathrm{pos}} > N_{\mathrm{doc}}$). Add $N_{\mathrm{pos}} - N_{\mathrm{doc}}$ known-irrelevant dummy documents with fixed $b = 0$. They occupy the lowest-diagnosticity positions and are excluded from belief updates.

Implementation. Experiments use equal document and position counts; asymmetric cases were not tested.

G Extended Limitations

Conditional independence. Citation budgets, attention saturation, and constrained output schemas can couple within-read citations. Realized-layout checks reveal residual dependence and do not establish generally calibrated posterior confidence (§B.5).

Semantic distractors. The nearest-neighbor probe supports calibration transfer but exposes initial-layout sensitivity; it does not establish arbitrary adversarial or open-domain robustness (§C.3.2).

Evaluation scope. Controlled MDQA uses deterministic labels, including the closed-source and multiple-needle experiments. Open-domain benchmarks such as NQ-Open add unevaluated partial relevance, answer ambiguity, and heterogeneous sources. Graded relevance requires richer beliefs; multi-hop retrieval requires jointly sufficient evidence sets.

Recalibration. Robustness to $\pm 15\%$ Gaussian calibration noise (Fig. 8) does not eliminate recalibration needs when shifts invert the positional profile. Section E.2 gives triggers and costs.

Parsing. High parse success does not establish citation faithfulness or exclude difficulty-dependent failures and harm under other protocols (§E.1).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction qualify the $O(\log N)$ rate as conditional on positive-margin lock-in and the detector assumptions, and describe 50–87.5% matched-accuracy query savings from exploiting positional heterogeneity, with comparable accuracy when profiles are nearly flat.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated limitations discussion is provided in Section 6 and expanded in §G, addressing the conditional independence approximation, reliance on deterministic labels, and the need for open-domain evaluation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix A provides full formal proofs and explicit assumptions (e.g., A1 and A2') for consistency (Theorem A.1), sample complexity (Theorem A.3), and drift behaviors (Propositions A.7 and A.8) of the framework.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 details the benchmark datasets (FLenQA), exact model scales, generation hyperparameters, and baseline configurations, while Appendix B formalizes the calibration procedure.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: A public code release is planned at `https://github.com/abyerly2/gold-panning`. A complete public reproduction package with instructions for the main experiments is not yet available. Algorithm 1 and Appendix E provide algorithmic and implementation details.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 explicitly outlines the evaluation setup, including the $N = 100$ default configuration, $T = 8$ iteration budgets, and decoding temperatures/top-p values for each model family.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All relevant empirical plots (Figures 2, 4, 8, and 9) display 95% confidence intervals as shaded regions to indicate statistical variance across evaluation trials.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section E.3 reports H100/H200 GPUs, vLLM serving, and per-setting GPU-hour ranges. GPU memory capacities and total compute across all experiments, including API and auxiliary runs, are not fully specified.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research involves standard benchmarks and model inference optimization, conforming fully to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work focuses on the mathematical and algorithmic fundamentals of inference-time retrieval and position bias mitigation; it does not directly enable new high-risk societal applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new generative models, web-scraped datasets, or artifacts with a high risk for dual-use or misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The open-weight models (Gemma-3, OLMo-3) and datasets (FLenQA) used are appropriately cited in the text and bibliography.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new datasets, foundational models, or similar deployable assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The research does not involve human subjects or crowdsourced data collection.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the core subject of the research; their use as evaluation targets and inference engines for the proposed algorithm is explicitly detailed throughout Section 4.