Selection Bias Inflates Per-Query Rerank-Augmentation Headroom: A Shuffle-Axis Null Calibration with Cross-Platform Evidence

Anonymous authors
Paper under double-blind review

Abstract

A common way to motivate adding a per-query routing or re-weighting signal on top of a strong reranker is to measure the *oracle-direction headroom*: for each query, pick the auxiliary axis (and its sign) that most improves the ranking, and report the average gain. We show that this quantity is *systematically inflated by selection bias*: the "pick-the-best-axis-per-query" operator returns a positive gain even when the axes carry no query-conditional information, because it is an order statistic over noise. We calibrate the diagnostic with a *shuffle-axis null* that permutes each query's axis values to destroy any axis–gold relation while preserving the selection operator, and report the net (observed − null) with paired bootstrap confidence intervals. On the platforms we test, the surface headroom that would license a per-query router collapses to zero or below after calibration: the direction-level net is -0.0032 (95% CI $[-0.017, +0.011]$). On the primary Stage-B platform no single axis clears the pre-specified bar of 0.02; across the BGE/MonoT5 generalization cells (two rerankers, two benchmarks: LoCoMo, LongMemEval-S) the direction-level net remains ≤ 0 or statistically indistinguishable from zero. A bounded fixed-axis exception appears for entity overlap on LoCoMo, but the direction-level router remains null there, so this does not support per-query routing. A controlled lexical ablation *refutes* a natural mechanism explanation for the bounded fixed-axis exception (entity overlap): it is *not* that the base scorer's lexical features internalize entity overlap. The same ablation reproduces a *known* effect from the reranking literature, where a weaker base leaves more apparent headroom on every axis; we cite it and use it to delimit the scope of any per-axis result. This is a methodological reporting requirement with a precise scope: it concerns per-query axis augmentation on top of a strong shortlist, and it makes the classical winner's-curse concern operational for reranking feasibility checks. The contribution is a concrete null that turns a misleading "oracle headroom exists" into a testable "net headroom after selection correction."

1 Introduction

Suppose a practitioner has a strong reranker producing a shortlist of candidates, and asks whether a per-query auxiliary signal (recency, entity overlap, a conflict/validity score) would improve the ranking if routed adaptively. A natural feasibility check is the *oracle-direction headroom*: for each query, take the auxiliary axis a that (in the right sign) most increases a ranking metric, and average that best-per-query gain. If the oracle headroom is large, it appears to justify building the router.

The trap is that the oracle operator $\max_a$ over axes (and the $\pm$ over signs) is an *order statistic over noise*. Even when no axis carries query-conditional information, $\mathbb{E}[\max_a \text{gain}_a] > 0$, simply because picking the largest of several noisy gains is positive in expectation. The surface headroom is therefore an upper bound that is *always inviting*, and a CI on it that excludes zero does not establish a real signal.

Our diagnostic is deliberately minimal. We (i) define the oracle-direction gain on a frozen base ranker; (ii) construct a *shuffle-axis null* that permutes each query's axis values, destroying axis–gold correspondence while preserving the exact selection operator; and (iii) report the *net* (observed − null) with paired bootstrap CIs at both the query and conversation-cluster level. The net is the part of the headroom that survives selection correction.

Contributions.

1. A shuffle-axis null that isolates the selection-bias component of oracle-direction / best-of-N headroom diagnostics (Section 4).

2. Cross-platform evidence that the calibrated *direction-level* (per-query-router) net headroom is ≤ 0 or statistically indistinguishable from zero across two rerankers and two benchmarks, while the only calibrated fixed-axis exception is entity overlap on LoCoMo (Sections 5.2, 5.4, 5.6).

3. A controlled lexical ablation that *refutes* the natural mechanism explanation for the one positive axis (that the base scorer's lexical features internalize entity overlap), separating it from a previously-reported base-strength effect (Section 5.5).

Scope. Our evidence speaks to per-query rerank-augmentation on top of a strong shortlist, across the tested reranker × benchmark cells. Recall ceilings appear as shortlist-coverage context. The statistical core, correcting for selection over a max operator, is the classical winner's curse; the contribution here is the concrete null that makes it operational for reranking feasibility checks.

2 Related Work

Selective inference / winner's curse. Choosing the best of many candidate statistics and then reporting its value is the classical winner's-curse / selective inference problem (Taylor & Tibshirani, 2015; Berk et al., 2013). We build directly on it: the contribution here is a concrete null for the oracle-direction headroom pattern as it appears in reranking feasibility checks.

Optimistic bias in LLM evaluation and routing. Recent work studies optimistic bias when a best configuration is selected and then reported, in adaptive benchmarking (Xu et al., 2026) and in router evaluation (Li et al., 2026). These target a different object (evaluation protocols / model routers); we target per-query *axis* selection on top of a fixed reranker.

Headroom that exists but is not recoverable. A recent line reports settings where an oracle headroom is real but cannot be closed by a practical method: Entity-Collision (Deng, 2026) measures adaptive routing on a conversational-memory setting as a null with residual oracle headroom, and Beyond-Localization / RAG-APR (Zhao et al., 2026) finds within-pool headroom that is bounded and selector-sensitive. These read the headroom as a real quantity that existing selectors fail to recover. Our question is upstream of theirs: we ask whether the headroom *as measured* is real to begin with. The oracle-direction operator returns a positive gain even under a null with no query–axis relation (Section 4), so the headroom number is inflated by the selection operator before any selector is asked to recover it. Concretely, the surface $+0.0685$ direction headroom (Table 1) is -0.0032 after the shuffle-axis null. The distinction is not adversarial: a "real-but-unrecoverable" finding and a "partly-selection-inflated" finding can coexist, but they call for different reporting, and a headroom reported without a selection null cannot be assigned to either category.

Rerankers on conversational memory. Reranking and query rewriting are well-established as helpful for conversational-memory retrieval. Our claim is fully compatible with this: it concerns *additional per-query axis augmentation on top of an already-strong shortlist*. A strong reranker can matter while leaving no further per-query headroom for axis-style augmentation on top of it.

Backbone-strength dependence. TARG (Wang et al., 2026) reports that the usefulness of an auxiliary signal depends on backbone strength. Our base-strength confound (Section 5.5) is the same phenomenon surfacing in our ablation, and we cite it to delimit the scope of any per-axis positive result.

3 Diagnostic Setup

Two distinct diagnostics (not two metrics of one quantity). We run two methodologically different probes, and keep them separate throughout:

- **Stage A: additive separability (AUROC).** A logistic probe over gold-membership, asking whether an axis is *additively separable* from the content score. We measure the AUROC increment of adding the axis as a feature on top of content+query. Source: `gate1_diagnostic.json` (and the mixed-candidate retest `gate1_retest_mixed.json`).
- **Stage B: oracle-direction headroom (MRR).** The per-query axis-selection gain $\mathrm{gain}(s, g, a) = \max\big(\mathrm{MRR}(z_s{+}z_a, g), \mathrm{MRR}(z_s{-}z_a, g)\big) - \mathrm{MRR}(s, g)$, where z denotes z-scoring; we then take, per query, the best axis (direction-level) or a fixed axis (single-axis), and subtract the shuffle-axis null. Source: `gate1b_net_ci.json`, `gate1b_mag_vs_dir.json`, `gate1b_by_category.json`, `learnability.json`.

These are *different questions* (additive separability vs. per-query selectability) and *different quantities* (AUROC increment vs. MRR gain). We report them as such.

Platforms. The diagnostic is applied on three reranker backbones across two benchmarks:

- **CE-lite**: a lightweight cross-encoder-distilled scorer (encoder `all-mpnet-base-v2`) trained for conversational-memory reranking, used here as one experimental backbone on **LoCoMo** (Maharana et al., 2024), the primary Stage A/Stage B platform.
- **BGE-reranker** (`BAAI/bge-reranker-v2-m3` (Chen et al., 2024), `max_length`=512) and **MonoT5** (`castorini/monot5-base-msmarco` (Nogueira et al., 2020)) on **LoCoMo** and **LongMemEval-S** (Wu et al., 2025), the generalization cells (Section 5.4).

CE-lite is one experimental platform among three; it is not the subject of the paper.

Candidate-pool construction (stated explicitly). The candidate pools differ across the diagnostics and are *not* fully controlled:

- Stage A (`gate1_diagnostic.json`): v1 top-50 shortlist (cache `reach_cache.pkl`), $n_{\text{queries}} = 1977$.
- Stage A retest (`gate1_retest_mixed.json`): mixed candidates, union(v1 top-50, BM25 top-50), average pool size **82.06** (cache `reach_cache_v2.pkl`). This rules out a saturation-by-self-selection artifact of the v1-only pool.
- Stage B (`gate1b_*.json`): same mixed-candidate cache (`reach_cache_v2.pkl`); the oracle-direction gain is computed over a per-query top-K subset with $K = 10$ (field `topk`=10 in `gate1b_mag_vs_dir.json`; the subset is the reranker's own top-10 by score, `order = argsort(−s_match)[:10]` in `conflict_build.py`; cached as `top_local` in `conflict_dev.pkl`), $n_{\text{queries}} = 892$ (the subset with a defined conflict axis and at least one in-pool gold).
- Generalization (`gen_*.json`): dense (mpnet) top-100 $\rightarrow$ reranker top-50 (N_{DENSE}=100, N_{SHORT}=50); for LongMemEval-S the full candidate set ($\approx$ 48 sessions, $< N_{\text{DENSE}}$) is scored directly by the reranker (`reach_remote.py`).

Because Stage A (v1 top-50) and Stage B (mixed-candidate, per-query top-10) use different pools, the AUROC and MRR diagnostics are not on an identical candidate set; we treat them as two separate probes rather than as a within-design comparison.

Caches and provenance. All caches are Python pickle (`.pkl`) files (`reach_cache.pkl`, `reach_cache_v2.pkl`, `conflict_dev.pkl`), *not* npz. Stage A/B results carry `data_hash=03e2a21d4572`, `git_commit=a56497c`.

Benchmark statistics. LoCoMo: $n_{\text{queries}} = 1977$ (`gate1_diagnostic.json`: `n_queries`), $n_{\text{gold pairs}} = 2398$ (`n_gold`), evaluated over `n_pairs`=98850 query–candidate pairs. Per-category query counts (from `recall_ceiling.json` `n` fields): cat1 = 281, cat2 = 320, cat3 = 89, cat4 = 841, cat5 = 446 (total 1977). Mean gold per query = 1.42 (median 1, max 19); fraction with ≥ 2 gold = 21.4% (423/1977); every query has ≥ 1 gold (computed from `load_locomo_examples`). LongMemEval-S: $n = 500$ items (`gen_longmemeval_*.json`: `n`); question-type counts (from `by_qtype`): knowledge-update 78, multi-session 133, single-session-assistant 56, single-session-preference 30, single-session-user 70, temporal-reasoning 133 (total 500). Average candidate-pool size = 47.7 haystack sessions (min 38, max 62); evidence is session-level (`answer_session_ids`; mean 1.90 gold sessions per item, max 6); all 500 items carry `question_date` and `haystack_dates`, supporting a true query-reference time delta (computed from the LongMemEval-S loader).

4 Shuffle-Axis Null Calibration

The null. For each query, the axis values a (e.g. recency, entity overlap) are randomly permuted across that query's candidates, destroying any axis–gold correspondence while *preserving the selection operator* (we still take max over axes / $\pm$ over signs on the permuted values). The expected gain under this permutation is the pure selection-bias baseline. We use $B_{\text{perm}} = 1000$ permutations.

Net and confidence intervals. The reported quantity is net = observed − null, with a paired bootstrap ($B_{\text{boot}} = 2000$) resampling queries (and, in a second variant, conversation clusters), recomputing observed and null within each resample. We report the net point estimate and 95% bootstrap CI at both the `query` and `cluster` unit.

Null stability. The permutation null is stable: cross-permutation standard deviation $\sim$ 0.004–0.006 and convergence drift over the last 100 permutations $\sim 10^{-5}$ (`learnability.json`: `perm_stability`). Specifically null means are time 0.0753, entity 0.0422, conflict 0.0692, direction 0.0715 (`perm_stability.*.null_mean`).

4.1 Validation on synthetic controls

To establish that the null itself is correct (removing pure selection artifacts while preserving real signal), we run two synthetic controls with known ground truth. Both import the *same* null implementation used by the real diagnostic (a single module `reach_null.py`; the real Stage-B script imports it and reproduces its four result json byte-for-byte), so this validates the deployed null, not a look-alike. The synthetic data match the real scale: $Q = 998$ queries, $K = 10$ candidates, $A = 3$ axes, $B_{\text{perm}} = 1000$, $B_{\text{boot}} = 2000$ (`null_validation.json`).

Negative control (the null removes a pure artifact). With all axes drawn as i.i.d. noise (true signal = 0), the uncalibrated oracle-direction is strongly positive (observed $\approx +0.14$ across 5 seeds): the winner's-curse artifact is real and large. After subtracting the shuffle-axis null the net contains zero in every seed: query and cluster coverage are both $5/5 = 1.00$. The null removes the pure selection artifact cleanly.

Positive control (the null preserves real signal). We inject a query-conditional signal of strength δ on each query's true axis and sweep $\delta \in \{0, 0.01, 0.02, 0.04, 0.08, 0.2, 0.4\}$ (paired across δ via a shared seed). At $\delta = 0.4$ the net is significantly positive and clears the bar (net $+0.052$, 95% CI $[+0.034, +0.068]$), direct evidence the null preserves a genuine signal; an over-correcting null would push even a strong injected signal to zero (Figure 1).

Pre-specified criterion and its noise-robust replacement (both reported). Our pre-specified positive-control criterion required the net to be *strictly* monotone in δ (point estimates non-decreasing).

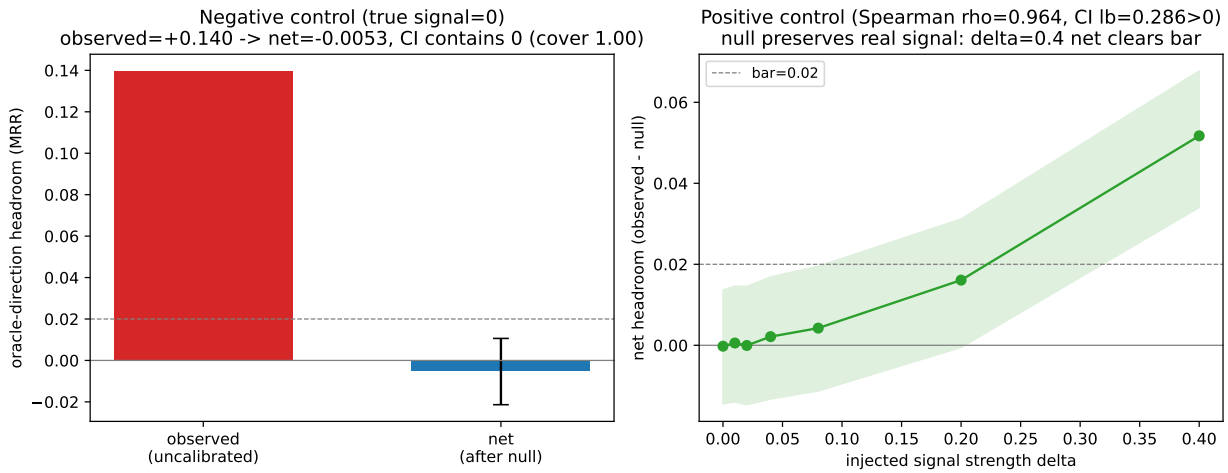

Figure 1: Null validation on synthetic controls. Left: negative control: the uncalibrated oracle-direction is large (observed $\approx +0.14$) while the net contains zero (coverage 5/5). Right: positive control: net vs. injected signal strength δ; the null preserves signal, with the net clearing the bar (0.02) by $\delta = 0.4$.

This criterion is triggered by a single non-monotone step, δ=0.01$\rightarrow$0.02, a drop of **0.0006**. That drop is two orders of magnitude below the bootstrap CI width ($\pm$0.015) and lies in the net ≈ 0 region where Monte-Carlo bootstrap noise dominates; point-estimate strict monotonicity is not a sound test for a quantity statistically indistinguishable from zero over the first five δ. We therefore adopt a *noise-robust* criterion, reported alongside the original: (i) the Spearman rank correlation $\rho(\text{net}, \delta)$ has a bootstrap 95% CI strictly above zero (point $\rho = 0.964$, CI $[0.29, 1.0]$), so the trend is significantly monotone; and (ii) the strong-δ net is significantly above the δ=0 net under a paired bootstrap ($\text{net}(0.4) - \text{net}(0) = +0.052$, CI $[+0.033, +0.070]$). Both hold, δ=0 contains zero (consistent with the negative control), and the net clears the bar by $\delta = 0.4$. Both verdicts are stored in `null_validation.json` (`verdict_prereg_strict`, `verdict_robust`; the `prereg` label is a legacy script name, and the criterion is pre-specified in our analysis plan, not externally preregistered).

Operating characteristic of the bar. The net crosses the bar$= 0.02$ between $\delta = 0.2$ and $\delta = 0.4$: a query-conditional signal must be roughly $\delta \approx 0.2$–0.4 in strength (the gold candidate raised $\sim$ 2–4 axis standard deviations on its true axis) before the diagnostic reports it as clearing the bar. Weaker signals (our $\delta \leq 0.08$) sit below the detection floor set by content and selection noise, and the null correctly reports them as ≈ 0, a detection floor rather than over-correction.

5 Results

The two probes answer one question from two directions. Stage A asks whether an auxiliary axis is *additively separable* from content, that is, whether, pooled across queries, the axis adds linearly-usable signal; it does not (the query-conditional time increment is -0.0034, Section 5.1). Stage B grants the axes the most generous possible form, *per-query* selection of the best axis and sign, and asks whether even that recovers signal once selection bias is removed; it does not (direction net -0.0032, Section 5.2). The two use different candidate pools and different metrics by design, so we do not combine their numbers; we read them as a weak and a maximally favorable diagnostic that point to the same conclusion. Taken together, these probes show no recoverable signal for *adaptive per-query axis selection* on top of the strong shortlist, from either the weakest or the strongest way of asking; the fixed-axis entity exception is treated separately in Sections 5.4–5.6.

Table 1: Stage B oracle-direction headroom (MRR), $n_{\text{queries}} = 892$, per-query top-K=10, mixed-candidate pool. Observed = `gate1b_mag_vs_dir.json` (`mean_gain` / `delta_direction`); null = `learnability.json` (`perm_stability.*.null_mean`); net + CI = `gate1b_net_ci.json` (`results.query`). Bar = 0.02.

axis / quantity	observed	null	net	95% CI (query)
time	0.0438	0.0753	−0.0314	[−0.0456, −0.0167]
entity	0.0497	0.0422	+0.0074	[−0.0041, +0.0191]
conflict	0.0658	0.0692	−0.0036	[−0.0204, +0.0142]
direction	0.0685	0.0717	−0.0032	[−0.0171, +0.0111]

5.1 Stage A: additive separability (AUROC)

On the v1 top-50 pool (`gate1_diagnostic.json`), content alone gives AUROC 0.8552 (`C0[s_match]`); adding the query raises it to 0.8876 (`C1[s_match,q]`). Adding raw recency on top of content+query changes AUROC by +0.0001 (`deltas.raw_recency_vs_C1`; recency is already absorbed by the base scorer's position feature), and adding a *query-conditional* time interaction changes it by **−0.0034** (`deltas.query_cond_time_vs_C1`), i.e. below the pre-specified bar of +0.02 and slightly negative. The raw time axis is near-orthogonal to the content score: Pearson −0.0192 (`corr_vs_smatch.time_raw.pearson`).

To rule out that this null result is an artifact of the v1-only pool saturating on its own selections, we retest on the mixed pool (`gate1_retest_mixed.json`): the query-conditional time increment is **−0.0025** (`delta_query_cond_time`), essentially unchanged, while content+query AUROC actually *rises* to 0.9171 (`C1[s,q]`). The BM25-only candidates (which v1 did not select) have a content-score spread comparable to the gold-bearing pool (`smatch_std`: both 2.918, bm25only 2.731, all 3.315), confirming that the pool has score variance and that saturation-by-self-selection is not the cause.

5.2 Stage B: oracle-direction headroom (MRR) and its null

Table 1 reports the Stage B observed gain, shuffle-axis null, and net for each single axis and for the direction-level (per-query best axis) quantity.

The headline is the **direction** row: the surface oracle-direction headroom is +0.0685 (`delta_direction`) with a bootstrap CI $[+0.0574, +0.0788]$ that excludes zero (`gate1b_mag_vs_dir.json`: `ci`); uncalibrated, this would license a per-query router. After subtracting the shuffle-axis null 0.0717 (`delta_direction_null`), the **net is −0.0032** with 95% CI $[-0.0171, +0.0111]$ (query) and $[-0.0152, +0.0094]$ (cluster, `gate1b_net_ci.json`: `results.cluster.direction`). The net is indistinguishable from zero; within uncertainty, the surface headroom is fully explained by the selection-preserving null.

Single axes. time net −0.0314 (CI all-negative), conflict net −0.0036 (CI contains 0), entity net +0.0074 (CI contains 0, upper bound +0.0191 < bar). No single axis clears the bar.

Per-query axis distribution. The per-query argmax axis is time 0.779 / entity 0.103 / conflict 0.118 (entropy 0.982 bits, `gate1b_mag_vs_dir.json`: `argmax_dist`, `argmax_entropy_bits`); the dominance of "time" is itself selection noise (the uniform direction-null is highest for time at 0.0753), not evidence that different queries genuinely need different axes.

Learnability (a second, independent negation of the router). Even granting the oracle, the per-query best axis is not predictable from the query: a leave-one-conversation-out classifier reaches accuracy 0.4776 vs. a majority-class chance of 0.7791 (`learnability.json`: `loco_cv_acc`, `chance`). A realizable selector using the predicted axis yields gain 0.0479 (`realize_gain`), below the best fixed axis 0.0658 (`best_fixed`); its null-subtracted net is **−0.0153** (`realize_net_minus_null`), i.e. −26% of the oracle delta (`realize_frac_of_oracle_delta`=−0.2610). The router fails twice: the headroom is selection bias *and* the choice is not learnable.

Per-category stratification. The pre-specified stratum (entity on multi-hop, cat1) does not clear the bar: entity net -0.0171, CI $[-0.0462, +0.0089]$ (`gate1b_by_category.json`: `category_1.entity`). One *post-hoc, exploratory* cell clears: entity on the adversarial stratum cat5, net $+0.0644$, CI $[+0.0326, +0.0980]$ (`category_5.entity`). We flag this as exploratory only: it is (i) not the pre-specified stratum (the pre-specified entity@multi-hop fails), and (ii) one of $\sim$15 axis$\times$stratum combinations, within multiple-comparison noise at $\alpha = 0.05$. It does not enter the claim and would require a pre-specified replication.

5.3 Top-K sensitivity of the direction-level net

The Stage B net in Table 1 is computed on a per-query top-K subset with $K = 10$. To rule out that the null direction-level result is an artifact of that particular truncation (for instance a top-10 pool so tight that every axis saturates), we recompute the direction-level net at $K \in \{10, 20, 50, \text{full}\}$ on the same mixed-candidate cache, holding the query cohort fixed across K (Table 2, Figure 2; `topk_sensitivity.json`).

Table 2: Top-K sensitivity of the direction-level net (MRR, shuffle-axis null, $B_{\text{boot}} = 2000$; 95% bootstrap CI, query unit). Same mixed-candidate cache as Stage B, a single query cohort ($n = 998$) held fixed across K. Direction oracle over {`time`, `entity`}; the conflict axis is excluded (see text). Source: `topk_sensitivity.json` (`results.*.query.direction`, `content_sat`, `mean_cand`). Bar = 0.02.

K	mean cand	content-sat	direction net	95% CI (query)
10	10.0	0.7785	+0.0034	[−0.0066, +0.0126]
20	20.0	0.8118	−0.0008	[−0.0121, +0.0095]
50	50.0	0.8496	+0.0010	[−0.0108, +0.0122]
full	81.9	0.8876	−0.0224	[−0.0351, −0.0079]

Three observations follow. First, the direction-level net stays at or below the bar for every K: its CI contains zero at $K = 10, 20, 50$ (net $+0.0034$, -0.0008, $+0.0010$) and is *significantly negative* at the full pool (net -0.0224, CI $[-0.0351, -0.0079]$), so the core negative does not depend on the truncation depth. Second, content-saturation AUROC rises monotonically from the tightest pool ($K = 10$, 0.7785) to the loosest (full, 0.8876), yet neither end shows a positive net; this rules out the "top-10 is too crowded / saturated" reading of the null, since loosening the pool surfaces no recoverable direction signal. Third, the full-pool net being significantly negative provides additional support for the selection-bias account: as the candidate pool grows, the shuffle-axis null baseline for an oracle max-over-axes rises faster than the observed gain, so the net becomes *more* negative, i.e. selection bias strengthens with pool size rather than revealing latent signal.

Scope and cohort differences (stated explicitly). Two differences from the Stage B main table are deliberate. First, this sweep holds a *single* query cohort fixed across all K so the columns are comparable, giving $n = 998$ queries (`n_queries`), whereas the Stage B main table uses $n = 892$ (the subset with a defined conflict axis and at least one in-pool gold at $K = 10$). The extra 106 queries are those whose gold falls outside the top-10 (gain = 0 at $K = 10$) but enters the candidate pool at larger K; the full pool averages 81.9 candidates (`mean_cand`), so those gold items sit around ranks 11–82. Second, the direction oracle here is taken over *two* axes (`time`, `entity`), not three: the conflict axis is excluded because its validity score was only computed on the top-10 cache, and recomputing it at larger K would violate the no-recompute / cache-reuse discipline (`topk_sensitivity.json`: `note`). The two-axis and three-axis direction nets are therefore not numerically identical to Table 1; the invariant we report is qualitative, namely that the direction-level net is at or below the bar at every K.

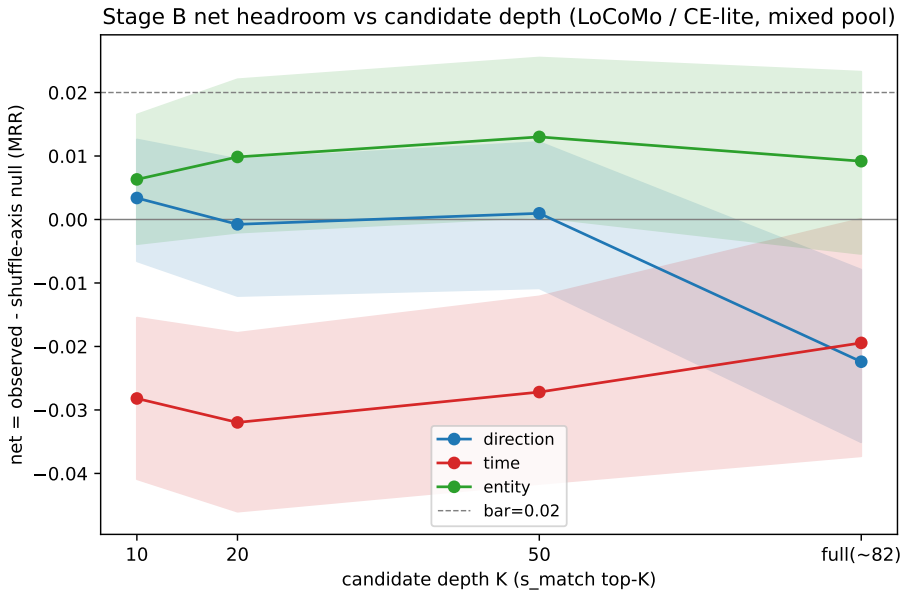

Figure 2: Top-K sensitivity of the direction-level net, with time and entity single-axis nets shown as references (MRR, shuffle-axis null; 95% bootstrap CI). The net stays at or below the bar (0.02) for all $K \in \{10, 20, 50, \text{full}\}$ and is significantly negative at the full pool, while content saturation rises from the tightest (K=10) to the loosest (full) pool; loosening the pool surfaces no recoverable signal, ruling out a top-10 crowding artifact.

5.4 Generalization across rerankers and benchmarks

Table 3 reports the four generalization cells ({LoCoMo, LongMemEval-S} × {BGE, MonoT5}); it gives direction and time nets, while Table 4 reports the entity single-axis nets separately because entity is the only bounded fixed-axis exception. All four cells have high content-saturation AUROC (0.868–0.947), and the direction net is ≤ 0 or its CI contains 0 in every cell.

The entity single-axis net is positive in all four cells and clears the bar on LoCoMo: BGE +0.0423 CI $[+0.0339, +0.0508]$, MonoT5 +0.0373 CI $[+0.0284, +0.0461]$; on LongMemEval-S its point estimates stay below the bar and do not robustly clear it (BGE +0.0186 CI $[+0.0099, +0.0282]$, MonoT5 +0.0166 CI $[+0.0085, +0.0260]$) (`gen_*.json`: `single_net.entity`). The time axis on LongMemEval-S, the only platform with a true query-reference timestamp delta and where temporal reasoning could plausibly help, is also null on the temporal-reasoning stratum: BGE time net -0.0059 CI $[-0.0199, +0.0064]$, MonoT5 -0.0043 CI $[-0.0148, +0.0039]$ ($n = 133$, `by_qtype.temporal-reasoning`). The entity exception is treated in Sections 5.5–5.6.

Table 3: Generalization, 4 cells. content-sat = `content_saturation_auroc`; direction and time net + CI from `gen_*.json`. MRR net, shuffle-axis null, $B_{\text{boot}} = 2000$. CIs are 95% bootstrap. The entity net is reported separately in Table 4.

benchmark	reranker	n	content-sat	direction net [95% CI]	time net [95% CI]
LoCoMo	BGE	1736	0.8836	$+0.0044$	-0.0177
				$[-.0011, +.0102]$	$[-.0247, -.0107]$
LoCoMo	MonoT5	1710	0.8684	-0.0219	-0.0086
				$[-.0295, -.0145]$	$[-.0202, +.0030]$
LME-S	BGE	500	0.9473	-0.0026	-0.0056
				$[-.0065, +.0013]$	$[-.0118, +.0003]$
LME-S	MonoT5	500	0.9102	-0.0038	-0.0032
				$[-.0081, +.0002]$	$[-.0091, +.0022]$

Table 4: Entity single-axis net (MRR, shuffle-axis null, $B_{\text{boot}} = 2000$; 95% bootstrap CI). Source: `gen_*.json` (`single_net.entity`). Entity robustly clears the bar (0.02) only on LoCoMo, where the lower CI bound is above the bar; on LongMemEval-S the point estimates stay below bar.

benchmark	reranker	entity net	95% CI
LoCoMo	BGE	+0.0423	[+0.0339, +0.0508]
LoCoMo	MonoT5	+0.0373	[+0.0284, +0.0461]
LME-S	BGE	+0.0186	[+0.0099, +0.0282]
LME-S	MonoT5	+0.0166	[+0.0085, +0.0260]

5.5 Lexical ablation: refuting the entity mechanism, and a cited base-strength caveat

The entity net clears the bar on LoCoMo under the BGE/MonoT5 backbones but not under CE-lite (Stage B entity net +0.0074, below bar). A natural hypothesis is that CE-lite's lexical features already internalize entity overlap, so removing them should make entity-reach re-appear. We test this with a controlled ablation: retrain the CE-lite scorer with vs. without its lexical features (the only changed variable), score a fixed union(full' top-50, BM25 top-50) pool, and recompute Stage B nets ($n = 1878$, `lexical_ablation_result.json`).

Removing lexical features weakens the scorer overall (MRR $0.6229 \rightarrow 0.5375$, a drop of 0.0853; `mrr_full`, `mrr_nolex`, `mrr_drop`), confirming lexical is a strong feature. The entity net rises from $+0.0114$ to $+0.0225$ ($\Delta = +0.0110$, CI $[+0.0019, +0.0208]$), *but the time net rises more*, from -0.0185 to $+0.0005$ ($\Delta = +0.0190$, CI $[+0.0042, +0.0345]$) (`by_unit.query.{time,entity}.delta_net`).

A known effect, confirmed. That weakening the base inflates the apparent rerank headroom on *every* axis (a weaker base leaves more oracle-rerank slack) is an established phenomenon in the embedding-robustness / reranking-depth literature (Magomere et al., 2025; Chen et al., 2025), and our ablation reproduces it cleanly (time, which has no lexical relationship, gains *more* than entity). The *diagnostic* implication we draw from it is precise: comparing one axis's headroom *across backbones* of different strength is confounded by base strength, so a per-axis positive net under a weaker backbone is not, by itself, a backbone-invariant signal.

A specific mechanism, refuted. The same ablation rules out the natural mechanism explanation for the entity exception, namely that CE-lite's lexical features *internalize* entity overlap, which would predict entity (and only entity) to re-appear when lexical is removed. Because time gains *more* than entity, the pre-specified specificity criterion ($\Delta_{\text{entity}} > \Delta_{\text{time}}$) is not met and the causal-confirmation flag is `CAUSAL_CONFIRMED=false`

(`lexical_ablation_result.json`); the lexical-internalization explanation is *refuted*. This excludes one plausible candidate for the entity mechanism, which remains open (Section 5.6).

5.6 The entity exception: bounded and mechanism-open

The entity axis is the one place a calibrated net clears the bar (LoCoMo × BGE/MonoT5, Table 4). We characterize it precisely as a bounded exception:

- It is *benchmark-specific*: it robustly clears on LoCoMo (turn-level evidence), while on LongMemEval-S (session-level evidence) the point estimates stay below the bar and do not robustly clear it, under the same backbones.
- Its obvious candidate mechanism, the base scorer's lexical features, is *excluded* by the ablation in Section 5.5, which refutes the lexical-internalization hypothesis.
- Its mechanism therefore remains *open* (a combination of backbone architecture, training, and candidate-pool differences), and we leave it open rather than assign one.
- It is a fixed global prior, not a per-query router: the direction-level net remains ≤ 0 in the cells where entity clears, so the gain (where used) applies globally, not adaptively.

6 Discussion

What this establishes. A reporting requirement: any feasibility check that selects the best-per-query axis / sign / model (oracle-direction, best-of-N) should subtract a selection-preserving null before its headroom is read as a real signal. The shuffle-axis null is one cheap, general instance.

Precise scope. The claim concerns per-query axis augmentation inside a strong shortlist. Recall ceilings (LoCoMo @50 = 0.910, @100 = 0.949; `recall_ceiling.json`) enter purely as shortlist-coverage context.

Base-strength caution (a known effect, cited). A per-axis positive net under a weak backbone can be an artifact of base weakening, an established phenomenon (Magomere et al., 2025; Chen et al., 2025); positive axis results should be checked for specificity against an axis with no plausible relationship (here, time), as our ablation does (Section 5.5).

Coverage. The evidence spans three reranker backbones (CE-lite, BGE, MonoT5) and two benchmarks (LoCoMo, LongMemEval-S), for per-query rerank-augmentation headroom inside a strong shortlist.

7 Reproducibility

Stage A/B results carry `data_hash=03e2a21d4572`, `git_commit=a56497c`, encoder `all-mpnet-base-v2`, base scorer as described in Section 3; $N = 50$ shortlist, bar = 0.02, $B_{\text{perm}} = 1000$, $B_{\text{boot}} = 2000$, null = shuffle-axis, direction = oracle-$\pm$. The lexical ablation uses `seed=23`. Caches are `.pkl`. Every number in this paper maps to its json field in `paper/NUMBERS_PROVENANCE.md`, and all citations resolve to `tmlr.bib`; all quantitative setup values are resolved from the loader and scripts.

References

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Liyang Chen, Yujun Cai, Jieqiong Dong, and Yiwei Wang. BRIGHT+: Upgrading the BRIGHT benchmark with MARCUS, a multi-agent RAG clean-up suite. *arXiv preprint arXiv:2506.07116*, 2025.

Youwang Deng. Entity-collision: A stratified protocol for attributing retrieval lift in agent memory. *arXiv preprint arXiv:2605.29630*, 2026.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. LLMRouterBench: A massive benchmark and unified framework for LLM routing. *arXiv preprint arXiv:2601.07206*, 2026.

Jabez Magomere, Emanuele La Malfa, Manuel Tonneau, Ashkan Kazemi, and Scott Hale. When claims evolve: Evaluating and enhancing the robustness of embedding models against misinformation edits. *arXiv preprint arXiv:2503.03417*, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fung. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.17753.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 708–718, 2020. doi: 10.18653/v1/2020.findings-emnlp.63. arXiv:2003.06713; model castorini/monot5-base-msmarco.

Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. doi: 10.1073/pnas.1507583112.

Yufeng Wang, Lu Wei, and Haibin Ling. Retrieval as a decision: Training-free adaptive retrieval gating for efficient RAG. *arXiv preprint arXiv:2511.09803*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10813.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking. *arXiv preprint arXiv:2605.05973*, 2026.

Pengtao Zhao, Boyang Yang, Bach Le, Feng Liu, and Haoye Tian. Beyond localization: Recoverable headroom and residual frontier in repository-level RAG-APR. *arXiv preprint arXiv:2603.29067*, 2026.