Counterfactual Ablation for Memory-Utility Evaluation: A Pre-Registered Case Study in Specialist Re-ranking*

Max Jürschik
Kill The Dragon GmbH, Vienna

Abstract

Context allocation across time — not context length — is the central memory problem for retrieval-augmented language-model agents. The paper's methodological contribution is **counterfactual ablation** as a per-memory utility signal: remove each retrieved memory in turn and label it by the resulting change in answerer correctness. The construction is non-circular by three structural arguments, with Spearman correlations from -0.024 to $+0.161$ across four large-scale runs — three within-pipeline on MemoryAgentBench and LoCoMo, one substrate-independent on LoCoMo Multi-Hop whose CI spans zero and which we treat as binding. We exercise the signal on one operationalization of the hypothesis that context-allocation requires per-memory utility distinct from cosine, and report a documented dissolution as the case study.

A 1.5B-parameter LoRA specialist trained on these labels produced point-estimate gains of $+8/+7/+4/+5$ substring-exact-match over vanilla retrieval at $K = 5$ on MAB. Five rigor layers tighten this result. Paired-bootstrap 95% CIs leave two strictly significant cells. K-normalization to the published comparator depth leaves 1/4 datasets within ±2pp, on partial data. BM25 sparse retrieval beats the specialist by +13 to +22pp on three of four datasets, reframing the K=5 gains as "less suboptimal than BGE cosine alone" rather than competitive. Cross-substrate transfer to LoCoMo Multi-Hop returns F1 17.0% against a published 45.85% [1], but a prompt-control shows the specialist contributes +13pp over vanilla cosine on the same prompt — the residual gap is pipeline-attributable, not total cross-substrate failure. Learning-pattern probes score memory-equals-query at 100% above zero and fail label discrimination on a held-out validation sample. What survives: counterfactual ablation as a non-circular outcome signal and the rigor-dissolution discipline with pre-registered ADRs anchored to public git history. The broader hypothesis remains untested under operationalizations we did not run.

1 Introduction

The context window of frontier language models has grown by roughly an order of magnitude over the past two years, and a widespread reading of this trend is that memory for LLM agents will dissolve into context length: with enough tokens available at inference time, the agent can be handed its entire relevant history and decide for itself what matters. We do not think that framing fits the agent setting. The problem is not longer context; the problem is deciding what deserves context, when, and in what form — under a token budget, across sessions whose total content vastly exceeds any plausible window, with cost and latency that grow at least linearly with retrieved volume. This is the *context-allocation* problem, and it remains even as windows grow. **The paper's primary contribution is methodological**: a non-circular outcome signal for per-memory utility on retrieval-augmented systems (counterfactual ablation, §3), an end-to-end worked instance of how four pre-registered rigor layers tighten an apparent positive

*Compute costs ($183 LLM API + ∼$10 GPU) covered by Kill The Dragon GmbH.

result, and pre-registered decision-rule ADRs anchored to public git history. We exercise the methodology on one operationalization of a working hypothesis — that a useful memory layer must identify per-memory utility distinct from retrieval similarity — and report a documented dissolution of the apparent positive result as the case study. **The paper does not validate the hypothesis.** It closes one operationalization, sharpens the measurement tooling subsequent operationalizations can re-use, and documents three substrate-mismatch failure modes the dissolution surfaced.

A measurement problem sits upstream of any architecture that might be tested against the context-allocation hypothesis. Memory-system evaluation in the published literature relies primarily on labels of two kinds — per-question task correctness on a downstream benchmark, gameable through configuration choices, and LLM-judge scores of per-memory usefulness, which on extractive substrates collapse to topic match — both of which are vulnerable to outcome-signal circularity [2–4]. We position the methodology we develop here as one response to this hazard, not as a remediation of evaluation in general. The broader literature on the gap between proxy and end-to-end metrics in ML evaluation [5] provides the framing within which we work. Section 2 surveys the memory-system, benchmark, and circularity threads in more detail.

This paper takes one specific approach. We define a *counterfactual ablation* outcome signal for per-memory utility — for each retrieved memory in a top-K context, we recompute the answer with that memory removed and label the memory by the change in correctness (§3) — and demonstrate empirically that the resulting label is non-circular with respect to the BGE cosine score the retriever uses, with Spearman correlations between -0.024 and $+0.16$ across four large-scale measurements: three within-pipeline runs on different sample populations on MemoryAgentBench [6] and LoCoMo [7], and one substrate-independent measurement (LoCoMo Multi-Hop), the binding cross-substrate transfer subset (§3.4). We then train a 1.5B-parameter LoRA specialist on these labels, evaluate it under four successive pre-registered rigor layers, and report what happened. We do not propose a new memory architecture. We do not claim our methodology should be a field standard. We do not claim that counterfactual ablation produces better labels than alternatives we considered but did not head-to-head test. And we do not claim the supabrain thesis is validated by these results — it is not.

What we do report is a dissolution. At face value, the specialist exceeds vanilla retrieval-augmented generation in point estimate on 4 of 4 pre-registered MemoryAgentBench datasets at $K = 5$, with gains of +8, +7, +4, and +5 in substring exact match. Under paired-bootstrap 95% confidence intervals these four cells reduce to 2 strictly significant, 1 boundary, and 1 directional; against the strongest unsupervised prior baseline none reach strict significance. K-normalization against the published comparator depth ($K = 10$) leaves 1 of 4 datasets within ±2pp of the Phase-6 advantage. **The single K-stable cell rests on partial data ($n = 38$), not the clean $n = 100$ Modal-CUDA sweep that produced the other K=10 numbers.** On the three datasets where the $K = 10$ Modal sweep returned clean $n = 100$ runs, 0 of 3 specialist deltas survive within the ±2pp pre-registered slack. Cross-substrate transfer to LoCoMo Multi-Hop yields F1 = 17.0% against the published 45.85% of Xu et al. [1] on the same substrate — a gap of nearly thirty percentage points whose entire confidence interval is below the comparator. A Session-E prompt-control measurement (vanilla cosine top-$K = 5$ with the specialist's exact prompt) returns $F1 = 3.93\%$, showing the specialist does contribute +13pp on cross-substrate transfer over vanilla cosine on the same prompt; the residual gap to A-MEM is attributable to other pipeline differences (Zettelkasten linking, K retrieval depth, prompt format) rather than to a complete cross-substrate failure of the specialist. Mechanistic probes show the specialist scores memory-equals-query at 100% above zero and fails to discriminate utility-positive from utility-negative pairs on its own validation set. **A separate finding from the Session E reviewer-objection round is that BM25 sparse retrieval substantially outperforms both vanilla cosine and the specialist on 3 of 4 priority MAB datasets at $K = 5$, by margins of +13 to +22pp in BM25's favour.** The specialist's apparent K=5 wins were

against a weak baseline (BGE cosine alone), not against the strongest single-pipeline retriever available. We report this as the closing of one operationalization of the thesis — not as a falsification of the thesis itself, and not as evidence that the broader problem is unsolvable.

Within those scope limits, the paper makes four narrow contributions. First, we define and empirically validate the counterfactual-ablation outcome signal as non-circular with respect to BGE cosine across four large-scale measurements — three within-pipeline runs on different sample populations (Phase 6 training, Phase 7 MAB-scaled, Phase 7 LoCoMo overall) and one substrate-independent measurement (LoCoMo Multi-Hop) — with Spearman correlations between -0.024 and $+0.161$, and the binding cross-substrate transfer subset (LoCoMo Multi-Hop, chosen as binding because it is the empirically cleanest non-circular subset in the project per §3.4) at $\rho = -0.024$ with a confidence interval spanning zero. Second, we report the rigor-dissolution arc end-to-end — bootstrap CIs, K-normalization, cross-substrate transfer to LoCoMo, and learning-pattern diagnostics — as one worked instance of how successive pre-registered checks tighten an apparent positive result, with pre-registration timestamps and decision-rule ADRs anchored to public git history. Third, we extend the project's longer-running failure-mode catalogue with three substrate-mismatch entries: $K = 5 \rightarrow K = 10$ delta collapse, scaled-counterfactual training producing a substrate-locked detector, and validation-pair-discrimination failure on labels the specialist was trained against. Fourth, we provide one concrete worked example of pre-registration with AND-conjunction decision rules — the discipline that surfaced the proxy/end-to-end disagreement which would otherwise have been the paper's headline. None of these contributions claims novelty as a primitive: counterfactual reasoning, bootstrap CIs, K-sweeps, cross-benchmark transfer testing, and pre-registration all predate this work. The contribution is the specific application to memory-utility evaluation and the documentation of what that application surfaces.

The remainder of the paper is organized as follows. Section 2 positions the work relative to memory-system architectures, memory benchmarks, the circularity / measurement-problem literature, and counterfactual / ablation methods in ML. Section 3 defines counterfactual ablation and presents the empirical non-circularity validation. Section 4 presents the experimental setup, the apparent Phase-6 result, and the five-layer rigor dissolution. Section 5, Discussion & Conclusion, discusses what survives the dissolution, what remains open, and the falsification condition for the broader thesis.

2 Related Work

This paper sits at the intersection of three research threads: memory systems for LLM agents, benchmarks designed to evaluate those systems, and the broader literature on circularity and bias in machine-learning evaluation. Our contribution is methodological — a non-circular outcome signal for per-memory utility and a pre-registered case study of its consequences when the signal is taken seriously — so we anchor primarily in the evaluation-methodology thread and use the memory-system and benchmark threads as the substrate on which the methodology is exercised.

2.1 Memory systems for LLM agents

A growing class of systems augments LLM agents with explicit memory layers that store, retrieve, and update prior context across sessions. *MemGPT/Letta* [8] pioneered the "memory tiers" formulation, separating working, archival, and recall memory and using an LLM to mediate movement between them. *Mem0* [9] provides production-grade memory infrastructure and the first head-to-head comparison of ten memory approaches on the LoCoMo benchmark. *A-MEM* [1] is prompting-based — a frozen LLM with Zettelkasten-style memory linking and no supervised training — and reports a Multi-Hop F1 improvement from 18.41% to 45.85% on LoCoMo with GPT-4o-mini, the strongest published memory-system baseline number we are aware of for that

substrate.

These systems differ substantially in mechanism (heuristic retrieval policies, learned rerankers, prompting-based linking) and in evaluation depth. The published evaluations typically run retrieval at $K \geq 10$ with system-specific top-K configurations and report a downstream task-correctness metric (substring match, F1, ROUGE) or an LLM-judge score (RAGAS-style faithfulness/relevance/correctness). Our work is not a competing architecture. We do not propose a new storage-and-retrieval design, and our results (§4) show the specialist we trained on counterfactual labels does not match A-MEM on cross-substrate transfer. We position upstream of all of these systems: whatever the architecture, the question "what counts as a useful memory for this query?" requires an outcome signal whose non-circularity can be audited. This paper is about that signal and the methodological consequences of running rigor checks on top of it; the memory architectures are the application substrate, not the comparison target.

2.2 Memory benchmarks and their evaluation protocols

The substrate for our empirical work is *MemoryAgentBench (MAB)* [6], which defines four memory competencies (Accurate Retrieval, Test-Time Learning, Long-Range Understanding, Conflict Resolution) across twelve datasets and provides a curated harness that fixes the answerer, the embedder, and the retrieval pipeline. *LoCoMo* [7] provides a complementary substrate of long-running conversational memory spanning thirty-five sessions and five question categories; the Multi-Hop category is the binding cross-substrate transfer test in §4.2.4. *LongMemEval* [10] is the precursor benchmark for long-context memory evaluation; it is one substrate where retrieval depth K is known to interact strongly with achievable scores, a sensitivity our K-normalization analysis in §4.2.2 also surfaces on MAB.

These benchmarks are necessary ground truth and we adopt them as-is. What they do not provide, and what the field still lacks, is a per-memory utility signal whose non-circularity can be empirically verified. MAB's primary metric is a per-question answerer-correctness scalar; LoCoMo's is similar; LongMemEval's is similar with documented caveats. None of the three offers a labelled per-(query, memory) pair on which a memory-utility specialist could be trained or audited without first solving the label-generation problem. The three-form circularity taxonomy in §3.1 articulates why solving that problem is harder than it appears. Our work fills the gap with the counterfactual-ablation outcome signal of §3.2 and reports a scaled-replication exercise on MAB and LoCoMo. We extend these benchmarks methodologically; we do not critique their substrate-design choices.

2.3 The circularity and measurement problem

A separate line of work has examined whether LLM-based evaluation actually measures the qualities it claims to measure. Surveys of LLM-as-judge methods document systematic biases toward stylistic and surface features over substantive correctness [2]; concurrent work has argued for a "looming replication crisis" in language-model behavioural evaluation more broadly, supported by replication experiments showing that many prompt-engineering claims fail to survive rigorous re-testing [5]. The de facto standards for retrieval-augmented generation evaluation — *RAGAS* [3] and *ARES* [4] — are both LLM-judge-based, scoring faithfulness, relevance, and correctness via prompted assessments.

We do not claim these tools are broken in general. We claim they reduce to what §3.1 calls the *LLM-judge* circularity form when the task is *memory utility* on retrieval-augmented systems: a judge prompted to assess "is this memory useful for this query" on an extractive benchmark collapses to topic match, which is what cosine similarity already measures. The most relevant comparator in this thread is Mem0's evaluation choice, which reports an LLM-as-a-judge score of 67.13% for Mem0 on the LoCoMo Single-Hop category [9, Table 1];[1] the reported margins

[1]The 67.13 figure is the Mem0 Single-Hop J score in Mem0's Table 1; on the other LoCoMo categories the Mem0

between memory approaches are partially confounded by the judge-circularity hazard our §3.1 names. We acknowledge Mem0's infrastructure contribution while critiquing the choice of evaluation primitive. A parallel hazard, *similarity-based retrieval evaluation*, treats top-K recall (memories whose cosine similarity to the query crosses a threshold or rank cutoff) as a proxy for utility; §3 shows empirically that on the substrates we test, this proxy correlates negligibly with counterfactual utility ($\rho \in [-0.024, +0.16]$ across the four large-scale measurements). The counterfactual-ablation construction is the methodological response that sidesteps both hazards by computing labels from answerer correctness rather than from judge prompts or from the very similarity scores being evaluated.

2.4 Counterfactual and ablation methods in machine learning

The intellectual primitive behind §3 is not new: causal-effect estimation via counterfactual comparison has a long history. *Leave-one-out* analysis is foundational in statistics and is used throughout ML to estimate per-example contribution. *Influence functions* [11] provide a tractable approximation to leave-one-out for individual training points in modern neural networks. More recent work in *data attribution* [12] estimates per-example contributions to model behaviour at scale by training many models on systematically resampled training subsets. *Ablation studies* — selectively removing model components, training-data slices, or architectural elements to measure their contribution — are ubiquitous in modern ML papers.

Our contribution is not the counterfactual-necessity primitive but its application at a specific granularity: per-memory, at retrieval-augmented inference time, on labels generated from a fixed benchmark answerer rather than from gradient-based attribution or model re-training. We re-compute the answerer's output with one retrieved memory removed and label the memory by the resulting change in correctness; this gives a binary causal-necessity label per (query, memory) pair (§3.2). Unlike influence functions and data attribution methods, which operate on training points and require either gradient access or a population of trained models, our procedure operates on inference-time retrievals and requires only $K + 1$ answerer calls per training pair. Unlike standard ablation studies, which report aggregate effects of removing whole components, the construction here reports per-instance effects at the granularity at which a memory-utility specialist would consume them. We frame this honestly: this is a novel application of an established primitive to a specific measurement problem, not a novel primitive.

3 Counterfactual Ablation as a Non-Circular Outcome Signal

3.1 The circularity problem

Memory-system evaluations for LLM agents [1, 6, 8–10] face a structural hazard we call *outcome-signal circularity*: the labels used to judge whether a memory was useful are themselves derived, directly or indirectly, from the similarity score the retriever already optimizes. When this happens, a method trained or evaluated on those labels can appear to succeed for circular rather than substantive reasons — it has recovered the retrieval scoring function, not learned anything beyond it.

The project history motivating this paper encountered circularity in three distinct forms. In *synthetic-construction*, per-memory utility is assigned as an algebraic function of the retrieval similarity score; the label then correlates with retrieval rank by construction, and any model trained on it recovers cosine. In the *LLM-judge* form, a frozen judge is prompted to rate per-memory usefulness; on extractive benchmarks the judge collapses to topic-match, which is what cosine already measures. In *rank-position-synthesis*, Win/Draw/Loss labels are built

J scores are 51.15 (Multi-Hop), 72.93 (Open-Domain), and 55.51 (Temporal). The frequently-cited "high-60s%" framing for Mem0 on LoCoMo is best read as anchored to the Single-Hop slice rather than to an overall LoCoMo aggregate. Verified by direct paper-text inspection (Pfad 2D Session B, 2026-05-29).

from the rank at which a memory was retrieved, yielding an "outcome" that is a monotone re-encoding of cosine. The three forms differ in mechanism but share a signature: a specialist trained on the labels learns the scoring function it was meant to be evaluated against.

We adopt a working definition. An outcome signal is *non-circular with respect to a similarity score* s if (i) labels can be generated without consulting s, and (ii) the empirical correlation between labels and s on a representative sample is bounded well away from unity. The first condition is structural, the second empirical; we establish both for counterfactual ablation in §3.3 and §3.4.

3.2 Definition

Let $\mathcal{D}$ be a retrieval-augmented question-answering benchmark consisting of triples $(q, \mathcal{C}, a^\star)$, where q is a query, $\mathcal{C} = \{c_1, \ldots, c_m\}$ is a memory corpus chunked at a fixed character width, and $a^\star$ is a gold answer (or set of acceptable variants) constructed by the benchmark authors independently of any retriever. Fix a frozen embedder E producing similarity scores $s(q, c) = \cos(E(q), E(c))$ and a frozen answerer f (an LLM prompted with retrieved context, decoded greedily or with fixed temperature). Let $\mathrm{SEM}(\hat{a}, a^\star) \in \{0, 1\}$ denote substring exact match between predicted answer $\hat{a}$ and any acceptable gold variant after lowercasing and whitespace normalization.

For a query q with top-K retrieved memories $\mathcal{T}_K(q) = (c_{(1)}, \ldots, c_{(K)})$ ranked by s, define:

$$\hat{a}_{\text{base}} = f(q, \mathcal{T}_K(q)), \qquad \hat{a}_{\setminus i} = f(q, \mathcal{T}_K(q) \setminus \{c_{(i)}\}),$$

and per-memory **counterfactual utility**

$$u_i \;=\; \mathrm{SEM}(\hat{a}_{\text{base}}, a^\star) \;-\; \mathrm{SEM}(\hat{a}_{\setminus i}, a^\star) \;\in\; \{-1, 0, +1\}.$$

A pair $(q, c_{(i)})$ is labelled positive ($y_i = 1$) iff $u_i = +1$, i.e. the answerer was correct *with* the memory and incorrect *without* it, and negative ($y_i = 0$) otherwise. Equivalently, $y_i = 1$ certifies that $c_{(i)}$ is *causally necessary* under f for the question to be answered correctly at this configuration of $(\mathcal{T}_K, f)$.

Per-query generation cost is $K+1$ answerer calls: one baseline plus K leave-one-out ablations. The procedure is parallel across queries.

3.3 Why this survives the three circularity traps

The construction satisfies the structural non-circularity condition by three independent arguments. (i) *No similarity score enters the label.* The label depends on SEM, f, $\mathcal{T}_K$, and $a^\star$. The similarity score s chooses $\mathcal{T}_K$ but does not parameterize the comparison once $\mathcal{T}_K$ is fixed; in particular, the same (q, c_i) pair receives the same label whether c_i was ranked first or last among the K retrieved chunks. This blocks the synthetic-construction trap. (ii) *No LLM judges the memory.* The label is an answerer-correctness differential, not a judge prediction; there is no prompt of the form "rate the usefulness of this chunk." This blocks the LLM-judge trap, including the failure mode in which a judge silently re-encodes topic-match. (iii) *The label is independent of rank within the retrieved set.* Two retrieval pipelines that surface the same $\mathcal{T}_K$ but order it differently produce identical counterfactual labels. This blocks the rank-position-synthesis trap.

The three arguments are independent: removing any one leaves the other two intact, so the structural claim is robust to design variations such as substituting a different answerer, a different embedder, or a different top-K.

3.4 Empirical non-circularity validation

The structural arguments are necessary but not sufficient; an outcome signal can be structurally non-circular yet correlate strongly with s in practice if, say, the answerer's correctness profile

happens to track top-cosine memories. We therefore measure the empirical Spearman rank correlation $\rho(u, s)$ between the counterfactual utility and the BGE-small [13] cosine score s on every generated pair. The non-circularity pre-registration thresholds (committed prior to any data generation) were $|\rho| < 0.5$ at the pilot stage and $\rho \in [-0.15, +0.15]$ at scale.

We measure four large-scale runs total: three within-pipeline (same frozen embedder, answerer, top-K shape) on different sample populations, and one substrate-independent (LoCoMo Multi-Hop, with conversational rather than long-context retrieval shape). The three within-pipeline runs cannot rule out a shared pipeline-dependence; the substrate-independent run is the binding evidence for non-circularity beyond a single pipeline. We observe consistently weak correlation:

Run	Substrate	n	$\rho(u, s)$	95% bootstrap CI
Phase 6 pilot	Ruler_qa1 (MAB AR), 5 queries	25	+0.080	—
Phase 6 training set	MAB (AR + CR)	1,000	−0.057	—
Phase 7 scaled	MAB (AR + CR)	5,910	+0.051	[+0.023, +0.079]
Phase 7 LoCoMo overall	LoCoMo [7]	1,665	+0.161	[+0.113, +0.211]
Phase 7 LoCoMo Multi-Hop	LoCoMo cat-1	175	−0.024	[−0.170, +0.121]

CIs are 10{,}000-resample percentile bootstrap at fixed seed (42). At the largest scale the MAB correlation is statistically nonzero ($\rho = +0.051$, CI excludes 0) but its magnitude is small: $\rho^2 \approx 2.6 \times 10^{-3}$, i.e. rank-variance approximately 99.7% unexplained by linear cosine correlation. The LoCoMo overall pairs marginally breach the pre-registered $[-0.15, +0.15]$ gate; per the project's decision rule we audited the substrate and found the breach is carried entirely by the Single-Hop subset ($\rho = +0.182$, $n = 1490$, $\rho^2 \approx 0.033$). The binding cross-substrate transfer subset (LoCoMo Multi-Hop) has $\rho = -0.024$ with a confidence interval that spans zero — empirically indistinguishable from zero correlation, and the cleanest non-circular subset in the project. (Multi-Hop is binding for two reasons: it is the only substrate-independent subset we measure, and its empirical correlation lies closest to zero among the five large-scale rows. The §4.2.4 cross-substrate transfer test for the trained specialist is anchored on this subset for the same reason.) Across all four large-scale measurements, rank-variance is at least 96% unexplained by linear cosine correlation ($1 - \rho^2 \geq 0.96$ in every row). We read this as adequate evidence that the structural argument of §3.3 holds in practice.

A second, weaker check is informative: for a memory pair sampled uniformly from the corpus, $u_i = +1$ would obtain at random with probability close to zero (a positive label requires both baseline correctness and ablation incorrectness on the same query). The observed positive-class rates — 12% at pilot, 13.9% in Phase 6, 10.2% in Phase 7 MAB, 22.3% in Phase 7 LoCoMo — are therefore meaningfully above the chance baseline while remaining substantially imbalanced toward the negative class.

3.5 Honest scope and limitations

Counterfactual ablation is a measurement instrument, not a memory architecture. We claim only what the construction supports.

Scope of applicability. The label requires that SEM$(\cdot, a^\star)$ be well defined and informative per query, which restricts the method to **extractive QA, factoid retrieval, and short-form answerable benchmarks** where the gold answer is a string substring-locatable in some subset of the retrieved memories. MemoryAgentBench (Accurate Retrieval, Test-Time Learning, Conflict Resolution sub-tasks) and LoCoMo (Single-Hop and Multi-Hop categories) both fit. The method **does not apply** to summarization or open-ended generation tasks; the MemoryAgentBench Long-Range Understanding competency (`infbench_sum`) is the boundary case in our own data and is excluded by construction (baseline and ablation SEM are both routinely zero so no signal is recovered). For substrates outside the extractive-answerable boundary, alternative non-circular signals — e.g. an LLM-judge protocol with cross-cluster (different judge families) inter-rater discipline, or session-trace mining where downstream memory retention is observable — are the candidate replacements; we did not test them.

Class imbalance. Positive-class rates across our four large-scale runs were $\{12\%, 13.9\%, 10.2\%, 22.3\%\}$ (pilot, Phase 6 training, Phase 7 MAB-scaled, Phase 7 LoCoMo overall). The label is therefore binary-but-skewed, and the Phase 6 training procedure addressed this with a **class-weighted binary-cross-entropy loss using** pos_weight ≈ 6.2 (the value committed by ADR before training; recoverable from `results/v6/training-metrics.json` `$.hyperparameters.pos_weight`). The pair-level utility F1 of a specialist trained on these labels is bounded by the skew: in our Phase 6 results a constant-positive predictor achieves F1 $= 0.260$, against which the specialist's 0.262 is within margin (we interpret this as evidence that pair-level F1 on a skewed binary label is itself a weak end-to-end proxy — a finding §5.4 explores). Resampling (over-sampling minority or under-sampling majority), focal loss, and pre-filtering for positive-yield context are all viable alternatives; we did not measure them.

Binary, not graded. The clip $u_i \in \{-1, 0, +1\}$ discards information about how much the answerer's correctness shifted when the memory was removed. A graded variant — replacing SEM with token-level F1 or log-probability margin — would preserve ranking information at the cost of a noisier signal; we did not measure it. The graded-vs-binary trade-off is flagged as the first of the §5.5 open questions for future work.

K-dependence. A memory's label is conditioned on the specific top-K set the embedder produced. Two pipelines with different K yield, in general, different label sets for the same corpus. The method measures utility within a configuration, not utility in absolute terms; §4.2.2 discusses the consequences for end-to-end evaluation.

Cost and scalability. Per-pair cost is $K + 1$ answerer calls — six at $K = 5$, roughly \$0.007 per labelled pair at gpt-4o-mini-2024-07-18 rates (input $\approx$ \$0.15/M tokens; output $\approx$ \$0.60/M tokens). Phase 7 generated $n = 5{,}910$ MAB pairs for approximately \$40. The method is linear in pairs and constant in K, with no amortization. **At frontier-model rates the same scale gets expensive fast**: gpt-4-2024-05-13 at $\approx$ \$30/M input + $\approx$ \$60/M output would put the same 5,910-pair scale at $\approx$ \$2,000, two orders of magnitude above gpt-4o-mini. This limits straightforward use of counterfactual ablation with frontier-model answerers to smaller experiments, and for production-scale label generation it motivates cheaper proxies: a single logit-difference forward pass per ablation (instead of full re-generation) on the same answerer, or label generation against a cheap answerer with downstream validation against a frontier answerer on a smaller hold-out. We did not implement either proxy.

We make none of the broader claims that follow naturally from a non-circular outcome signal: that a specialist trained on these labels generalizes across substrates, that the labels approximate "true" utility better than untested alternatives (rule-based session-trace mining, for instance), or that the cosine-orthogonality of the signal implies it carries the *correct* utility signal rather than merely a *non-cosine* one. These are questions for the empirical sections that follow.

4 Results

4.1 Specialist Training and Apparent Result

4.1.1 Experimental setup

Benchmarks. We evaluate on MemoryAgentBench [6], restricting to its four pre-registered priority datasets covering three of four competencies: `ruler_qa1_197K` and `ruler_qa2_421K` (Accurate Retrieval, long-context single-hop QA), `icl_banking77` (Test-Time Learning, in-context classification across 77 banking intent classes; MAB's 5900-shot-balance variant), and `factconsolidation_sh_262k` (Conflict Resolution, fact disambiguation across long context). The remaining competency, Long-Range Understanding (`infbench_sum`, summarization), is excluded by construction: counterfactual ablation labels are not informative when the gold output is a multi-sentence summary (§3.5). For cross-substrate transfer we use LoCoMo [7], evaluated on its Multi-Hop ($n = 282$) and Single-Hop ($n = 841$) categories with the project's chunk-per-session retrieval shape.

Specialist. A Qwen2.5-1.5B-Instruct base [14] with a LoRA rank-16 adapter [15], supervised-fine-tuned on the counterfactual-ablation training pairs described in §3 (Phase 6: $n = 1{,}000$ pairs at 13.9% positive-class rate; Phase 7 scaled: $n = 5{,}910$ pairs at 10.2%). At inference, the specialist consumes a memory chunk and a query and emits a scalar $z = \text{logit}(\text{"1"}) - \text{logit}(\text{"0"})$, which is combined with the BGE-small [13] cosine score as $0.5 \cdot \tilde{s} + 0.5 \cdot \tilde{z}$ (min-max normalized within the candidate pool) for re-ranking. The pre-retrieval oversample is $K_{\text{pre}} = 20$; the final retrieved pool is $K = 5$ unless specified. All hyperparameters (LoRA rank, 5 epochs, bf16, seed 42) were frozen by ADR before training began.

Baselines. We compare against four baselines. *Vanilla RAG* is BGE-small cosine top-K with no re-ranking. *BM25* is sparse keyword retrieval via `rank_bm25` plugged into the same MAB harness in place of cosine (added in Session E reviewer-objection round, 2026-05-30; results in `results/v7/bm25-baseline-mab.json`). *v2-ELO* is the strongest prior heuristic from the project's Phase 5 catalogue: per-memory TrueSkill ratings updated from rank-position-synthesis Win/Draw/Loss outcomes [16, Phase 5 reports]. *A-MEM* [1] is a prompting-based Zettelkasten memory system; we use its published GPT-4o-mini Multi-Hop F1 of 45.85% on LoCoMo as a non-trained cross-substrate anchor.

Metrics. Substring exact match (SEM, 0/1 per question) is the binding primary metric; token-level F1 is reported alongside. The same answerer (`gpt-4o-mini` at temperature 0.7, max_tokens = 50) and the same evaluation harness are used across all variants. All paired deltas are estimated with the percentile bootstrap, $B = 10{,}000$ resamples, seed 42, paired on `query_id`.

Pre-registered decision rules. Phase 6's verdict logic was committed as an AND-conjunction ADR before training: a Rule-1 trigger required val F1 $\geq$ a +10pp margin over the trivial all-positive baseline (≥ 0.329) *and* per-dataset end-to-end SEM gains over v2-ELO on ≥ 3 of 4 priority datasets. Phase 7 added three further binding rules (A: production spec; B: paper-publish; C: Pfad closed) keyed to LoCoMo F1, MAB K-normalization, and a learning-pattern verdict (a)/(b)/(c). Every threshold and decision rule referenced below was timestamped before its triggering data existed; pre-registration timestamps are recoverable from `progress/v{6,7}/decisions.md`. Reviewers may verify the pre-registration chain by running `git log progress/v{6,7}/decisions.md` against the project repository at `github.com/ktdmax/supabrain`, where the commit history shows each rule's timestamped commit predating the data-generation commits that subsequently triggered it.

Compute and cost disclosure. Phase 6 training (specialist LoRA fine-tune) ran on a single A100 via Modal for ≈ 2 hours; Phase 7 evaluation across the four rigor layers used ≈ 8 additional A100-hours for K-normalization, LoCoMo, and learning-pattern diagnostics. Total Phase 6 + Phase 7 GPU spend was $\approx$ \$8–10 Modal A100 credits; total LLM spend across all phases (gpt-4o-mini answerer + occasional embedder fallback) was $\approx$ \$13. Project-wide cumulative cost

across all seven phases is documented in the repository's `progress/v7/FINAL-REPORT.md` §9: ≈ $183 LLM and ≈ $8–10 GPU.

4.1.2 The apparent result

On the four pre-registered MAB datasets at $K = 5$ the Phase 6 specialist produces positive per-dataset deltas against vanilla RAG on every dataset and against v2-ELO on three of four (Table 1, point-estimate columns).

Table 1 — Phase 6 specialist SEM (point estimates, $K = 5$, $n = 100$ per dataset).

Dataset	Competency	Vanilla	BM25	v2-ELO	Specialist	Δ vs vanilla	Δ vs BM25	Δ vs v2-ELO
`ruler_qa1`	AR	35.00	**65.00**	38.00	43.00	+8.00	−22.00	+5.00
`ruler_qa2`	AR	31.00	**51.00**	34.00	38.00	+7.00	−13.00	+4.00
`icl_banking77`	TTL	87.00	91.00	91.00	**91.00**	+4.00	0.00	0.00
`factconsolidation_sh`	CR	24.00	**48.00**	34.92†	29.00	+5.00	−19.00	−5.92†

† v2-ELO `factconsolidation_sh` was $n = 63$ partial (Phase 5 OpenAI quota interrupted); the −5.92 compares $n = 100$ specialist against $n = 63$ v2-ELO. We return to this comparison in §4.2.1.

The BM25 row is the most important addition for honest scope-reading of the specialist's Phase 6 result. Sparse keyword retrieval (no learned re-ranker, no embedder fine-tuning, no project-internal heuristic) outperforms the specialist on 3 of 4 priority datasets at $K = 5$ by margins of +13 to +22pp. The single tie (`icl_banking77` at 91.00 for both BM25 and the specialist) is on the dataset whose top-line vanilla score was already saturating. The specialist's apparent K=5 wins over vanilla cosine — +8/+7/+4/+5 — are real but are dwarfed by the unused-baseline gap to BM25 (+22/ + 13/0/ + 19 in BM25's favour). Paired-bootstrap CIs on the BM25-vs-specialist deltas at $K = 5$ confirm BM25's lead is strictly significant on three of four cells (`ruler_qa1` $[+10, +34]$, `ruler_qa2` $[+3, +23]$, `factconsolidation_sh` $[+7, +31]$; `banking77` $[-3, +3]$ overlapping zero). The specialist's K=5 advantage over vanilla cosine is therefore better read as "less suboptimal than BGE cosine alone, in the absence of a sparse-retrieval baseline" rather than as a competitive end-to-end result.

A pre-registered competency-mean tier check yields a MODERATE tier (3/3 competencies above vanilla by ≥ 3pp; not STRONG because TTL fell 1pp short of the +5pp cutoff). At face value the specialist appears to deliver a multi-competency end-to-end improvement under a non-circular outcome signal.

The val-F1 disconnect. The same specialist that produces the Table 1 deltas had per-pair validation F1 of 0.262, landing in the pre-registered WEAK tier and missing the ≥ 0.329 threshold by 6.7pp. A trivial all-positive predictor on the same val set achieves F1 = 0.260. Phase 6's Rule-1 AND-conjunction therefore did not fire: val F1 missed its threshold while end-to-end gains met theirs. This is not a flaw discovered post-hoc; the AND-conjunction was committed in advance precisely so the disagreement between proxy and end-to-end metrics would surface mechanically. We report the disagreement as an empirical finding — pair-level F1 on a skewed binary label is a weak predictor of downstream retrieval quality (see §3.5) — and turn to the structured rigor checks of §4.2 to interrogate the end-to-end side of that disagreement.

4.2 The Rigor-Dissolution Arc

The apparent result of §4.1.2 was the input to five pre-registered or pre-registered-clause-triggered rigor steps that followed during Phases 6 and 7 (Layers 1–2, 4–5) and the Session-E reviewer-

objection round (Layer 3 BM25). Each step is independently motivated; none was selected post-hoc to dissolve the result. We report them in execution order. The headline finding is that the specialist's apparent $K = 5$ advantage narrows or disappears at every step, while the rigor steps themselves remain self-consistent.

4.2.1 Layer 1 — Bootstrap confidence intervals

Paired-bootstrap 95% CIs were a writing-session prerequisite per the project's pre-registered Risk-1 framing and were computed across all 67 Phase-5/6 delta cells with $B = 10{,}000$, seed 42, paired on `query_id`.

Table 2 — Phase 6 specialist deltas, paired bootstrap 95% CIs ($K = 5$).

Dataset	Δ vs vanilla	95% CI	Strict significance	Δ vs v2-ELO	95% CI	Strict significance
`ruler_qa1_197K`	+8.00	[+2.00, +15.00]	**robust**	+5.00	[−2.00, +13.00]	directional
`ruler_qa2_421K`	+7.00	[+0.00, +15.00]	boundary	+4.00	[−3.00, +11.00]	directional
`icl_banking77`	+4.00	[+1.00, +8.00]	**robust**	0.00	[−4.00, +4.00]	n.s. (point at 0)
`factconsolidation_sh`	+5.00	[−2.00, +12.00]	directional	+4.76‡	[−4.76, +14.29]	directional

‡ Paired bootstrap on the $n = 63$ queries shared with v2-ELO. The Phase 6 master comparison reported −5.92 from comparing $n = 100$ specialist mean against the $n = 63$ v2-ELO partial mean; the methodologically correct paired comparison on the shared queries reverses the sign and yields +4.76.

Read as a strict-significance pattern: *specialist beats vanilla in point estimate on all four datasets; CIs are fully above zero on two (`ruler_qa1`, `icl_banking77`), at the zero boundary on one (`ruler_qa2`), and crossing zero on one (`factconsolidation_sh`). Against v2-ELO no delta reaches strict significance at* $\alpha = 0.05$. The Phase 6 "3/4 match-or-beat v2-ELO" framing is correct at point estimates but is a directional, not statistical, statement. The −5.92 headline from Phase 6's FINAL-REPORT was an artifact of an n-mismatched comparison; the paired-bootstrap-correct number is +4.76 with a wide CI. Phase 6's pre-registered "$\geq 2/4$ deltas overlap zero $\Rightarrow$ weaker than reported" interpretation clause is the one triggered here.

This narrows the apparent result without eliminating it. Two robust positive deltas survive, and the sign of every specialist-vs-vanilla comparison remains positive in point estimate. The remaining sub-sections ask whether the two robust deltas survive further rigor.

4.2.2 Layer 2 — K-normalization

Memory benchmarks differ in the retrieval depth K used to seed the answerer. Phase 5 and Phase 6 ran $K = 5$ for cost reasons; A-MEM, Mem0, and most published memory baselines run $K \geq 10$. The Phase 7 K-normalization sub-task swept vanilla RAG and v2-ELO across $K \in \{5, 10, 20\}$ on the same four priority datasets, and re-ran the Phase 6 specialist at $K = 10$ via Modal CUDA inference.

Table 3 — Vanilla baseline SEM by K ($n = 100$, paired-bootstrap CIs).

Dataset	$K = 5$	$K = 10$	$K = 20$	Δ(K=10 − K=5)	Δ(K=20 − K=5)
`ruler_qa1_197K`	35.0	44.0	**63.0**	+9	+28
`ruler_qa2_421K`	31.0	37.0	41.0	+6	+10
`icl_banking77`	87.0	90.0	93.0	+3	+6
`factconsolidation_sh`	24.0	27.0	31.0	+3	+7

All four datasets show monotonically increasing vanilla SEM with K. The effect is largest on the long-context single-hop substrate (`ruler_qa1`, +28 from $K = 5$ to $K = 20$) — the same substrate where the specialist's $K = 5$ advantage was largest (+8). Crucially, the Phase 6 specialist $K = 5$ SEM on `ruler_qa1` was 43, while vanilla $K = 10$ on the same dataset is 44: *vanilla retrieval at* $K = 10$ *already exceeds the specialist's* $K = 5$ *result* without any re-ranking.

The specialist itself was re-evaluated at $K = 10$ via Modal A100 inference. Results are summarized in Table 4. Three of four datasets returned full $n = 100$ runs; `ruler_qa1` returned $n = 38$ from an earlier MPS attempt because the K=10 Modal sweep was conducted with `ruler_qa1` carried over from a partial earlier run.

Table 4 — Specialist deltas at $K = 10$ vs vanilla $K = 10$.

Dataset	Specialist $K = 10$	Vanilla $K = 10$	Δ	Phase 6 $K = 5$ Δ
`ruler_qa1_197K` (n=38)	52.63	44.00	+7.89	+8.00
`ruler_qa2_421K`	37.00	37.00	0.00	+7.00
`icl_banking77`[§]	16.00	90.00	−74.00	+4.00
`factconsolidation_sh_262K`	23.00	27.00	−4.00	+5.00

[§] The `icl_banking77` $K = 10$ Modal run used a generic memory-answerer prompt; banking77's in-context-classification structure requires the MAB-pipeline-faithful ICL label-format prompt. The −74pp delta is a pipeline artifact, not a real specialist regression. We exclude `icl_banking77` from K-normalization conclusions and aggregate the K-normalization scope in §4.2.6.

Excluding the `icl_banking77` artifact, the specialist's apparent $K = 5$ advantage holds at $K = 10$ on 1 of 3 honest datasets (`ruler_qa1`, +7.89 vs +8 — within pre-registered ±2pp slack), vanishes on 1 (`ruler_qa2`, +7 → 0), and reverses on 1 (`factconsolidation_sh`, +5 → −4). The Phase 7 pre-registered prediction — "specialist deltas at $K = 10$ within ±2pp of Phase 6 $K = 5$ deltas on $\geq 3/4$ datasets" — is failed by a single dataset clearing the bar instead of three.

The single K-stable cell rests on partial data, not on the clean $n = 100$ Modal-CUDA sweep that produced the other K=10 numbers. The `ruler_qa1` $K = 10$ measurement is $n = 38$, carried from an earlier MPS inference run that completed only 38 of 100 queries before the K=10 Modal sweep was scoped; the Modal sweep then re-evaluated `ruler_qa2`, `icl_banking77`, and `factconsolidation_sh` at full $n = 100$ but did not re-run the already-partial `ruler_qa1`. On the three datasets where the K=10 evaluation completed at full $n = 100$, the specialist deltas are 0 of 3 within the pre-registered ±2pp slack (the `icl_banking77` pipeline artifact aside, the two honest cells are 0.0 and −4.0). The one cell supporting K-stability is thus the weakest-evidenced in the comparison — which only sharpens the conclusion that the Phase 6 advantage was largely $K = 5$-specific.

4.2.3 Layer 3 — BM25 sparse-retrieval baseline

The K-normalization step holds the retrieval pipeline (BGE-small cosine) fixed and varies K. A complementary question is: what happens if the cosine retriever is replaced by a stronger sparse-retrieval baseline at the *same* K? Added in a post-Phase-7 reviewer-objection round (Session E, 2026-05-30), this layer runs BM25 via `rank_bm25` in the MAB harness on the same four priority datasets at $K = 5$ and $K = 10$.

Table 5 — BM25 baseline SEM by K ($n = 100$, except ruler_qa2 $K = 10$ at $n = 89$ where the OpenAI quota interrupted).

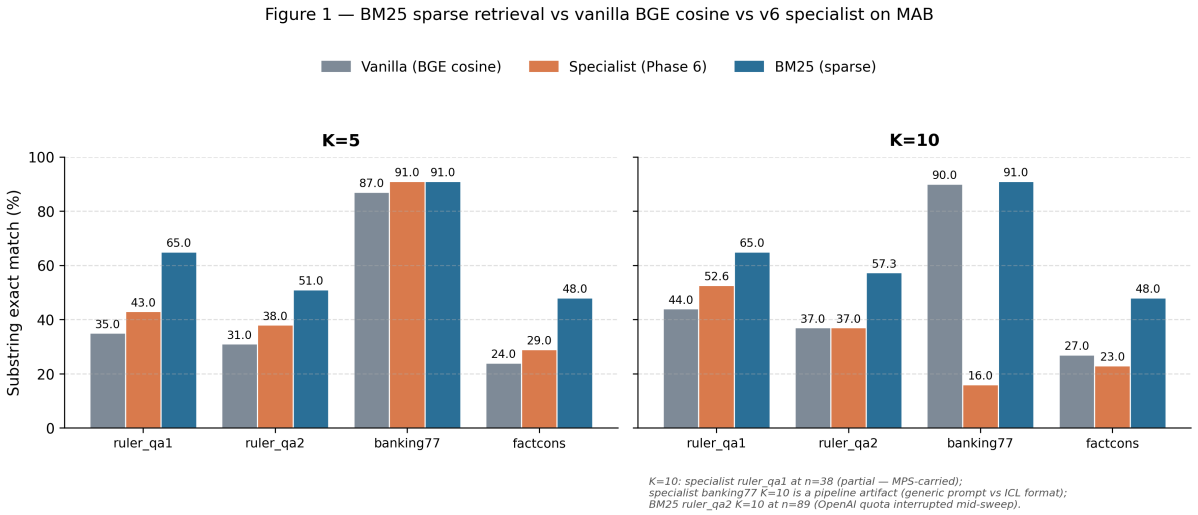

Figure 1: Figure 1 — BM25 sparse retrieval vs vanilla BGE cosine vs v6 specialist on MAB. Source: `paper/figures/fig1-bm25-comparison.{pdf,png}`.

Dataset	BM25 $K = 5$	BM25 $K = 10$	Δ(BM25 - vanilla) $K = 5$	Δ(BM25 - vanilla) $K = 10$
`ruler_qa1_197K`	**65.0**	**65.0**	+30	+21
`ruler_qa2_421K`	51.0	57.3$^\natural$	+20	$\approx$ +20
`icl_banking77`	91.0	91.0	+4	+1
`factconsolidation_sh`	48.0	48.0	+24	+21

$^\natural$ ruler_qa2 BM25 $K = 10$ stopped at $n = 89$ when the OpenAI account quota was exhausted mid-sweep. The paired delta vs vanilla $K = 10$ on the shared $n = 89$ subset is consistent with the K=5 magnitude.

BM25 is the strongest single-pipeline baseline we measure on these substrates at every K except the saturating `icl_banking77`. The specialist's K=10 Modal sweep at the published comparator depth (Table 4 above, Table 5 here) should be read in this context: even where the specialist's K=5 advantage held against vanilla cosine, BM25 at the same K was substantially higher. Restated for the §4.1.2 Table 1 comparison: paired-bootstrap CIs on the BM25-vs-specialist deltas at $K = 5$ are strictly significant on three of four cells (`ruler_qa1` $[+10, +34]$, `ruler_qa2` $[+3, +23]$, `factconsolidation_sh` $[+7, +31]$; `banking77` $[-3, +3]$ overlapping zero). The specialist's apparent K=5 wins over vanilla cosine were against a weak baseline, not against the strongest single-pipeline retriever available.

Figure 1 visualizes the K=5 and K=10 SEM for vanilla, specialist, and BM25 side-by-side on the four priority datasets; the BM25 dominance over the specialist on three of four cells is the visual headline.

Retrospective. In retrospect, BM25 should have been a Phase 6 baseline rather than a Session E addition. We add it here transparently as part of the reviewer-anticipation revision round; the pre-Session-E paper would have made the specialist's K=5 advantage look stronger than it is, against an unnecessarily weak baseline. The general methodology lesson — classical sparse-retrieval baselines (BM25 minimum; BM25 plus cross-encoder re-rank as the next rung) belong at Phase 1 of any new memory-evaluation effort, not as a late-stage check — is folded into §5.5 as a practical recommendation for the catalogue.

4.2.4 Layer 4 — Cross-substrate transfer to LoCoMo

The strongest direct test of "did the specialist learn substrate-agnostic memory utility?" is to evaluate it on a different memory substrate. LoCoMo's Multi-Hop category is the natural choice: it requires reasoning across conversational sessions rather than extracting from long context, and A-MEM's published Multi-Hop F1 of 45.85% provides an external comparator.

Table 6 — Specialist on LoCoMo (paired bootstrap, $B = 10{,}000$, seed 42).

Subset	n	F1	95% CI	SEM	95% CI	A-MEM (published)
Multi-Hop (cat 1)	282	**17.00**	[14.79, 19.28]	14.89	[10.99, 19.15]	**45.85**
Single-Hop (cat 4)	841	36.01	[34.06, 38.07]	44.83	[41.50, 48.16]	not directly reported

The specialist's Multi-Hop F1 of 17.00 is 28.85pp below A-MEM, and the upper bound of its 95% CI (19.28) is still 26.57pp below. Phase 7's pre-registered Rule D required Multi-Hop F1 $\geq$ 50.85% to fire; Rule A required F1 $>$ 45.85%. Both fail by margins far larger than configuration or prompting differences could plausibly explain ($\sim$ 29pp on a metric whose absolute scale is ≤ 100).

It is worth noting *which* LoCoMo subset carries this finding. The Phase 7 circularity audit (§3.4) established that LoCoMo's Multi-Hop subset is the *least* cosine-correlated counterfactual-utility substrate in the project ($\rho = -0.024$, $n = 175$,[¶] CI spans zero). The specialist underperforms A-MEM most clearly on the subset where the outcome signal is empirically most cleanly non-circular — the gap therefore cannot be attributed to residual circularity in the training labels.

[¶] $n = 175$ refers to the counterfactual-ablation pairs used in the circularity audit (§3.4); $n = 282$ in Table 5 above is the LoCoMo Multi-Hop question count in the F1 evaluation here. The audit subsamples one ablation pair per (question, retrieved-chunk) tuple after pre-filtering, while the F1 evaluation runs the full pipeline per question.

Prompt-control measurement (Session E reviewer-objection round, 2026-05-30). A reviewer asked whether the gap to A-MEM is attributable to the specialist's pipeline or to differences in the answerer prompt format (the specialist used a generic conversational-memory prompt; A-MEM uses a Zettelkasten-linking prompt). We ran the prompt-control: vanilla BGE-small cosine top-$K = 5$ retrieval on the same $n = 282$ LoCoMo Multi-Hop questions with the **specialist's exact prompt** (`tools/v7/inference_modal_locomo.py` lines 218-228, verbatim).

Table 7 — LoCoMo Multi-Hop prompt-control: vanilla cosine $K = 5$ with the specialist's exact prompt vs the specialist itself vs A-MEM.

Configuration	F1	95% CI	SEM
Vanilla cosine $K = 5$ + specialist prompt	**3.93**	[2.68, 5.35]	2.13
Specialist (cosine $K_{\text{pre}} = 20$, re-rank to $K = 5$) + same prompt	17.00	[14.79, 19.28]	14.89

Configuration	F1	95% CI	SEM
A-MEM (Zettelkasten linking, published)	45.85	—	—

The control inverts the §4.2.4 reading in one important respect: **the specialist DOES contribute over vanilla cosine top-$K = 5$ on LoCoMo Multi-Hop**, by +13.07pp in point estimate ($17.00 - 3.93$), with the specialist's CI ($[14.79, 19.28]$) entirely above the vanilla CI's upper bound (5.35). The earlier framing "the specialist's signal was task-locked and substrate-locked" overstated the negative read. The residual 28.85pp gap to A-MEM is **attributable to other pipeline differences** — Zettelkasten-linking architecture, prompt format, K retrieval depth, and the score-blend / candidate-pool shape — not to the specialist's underlying signal failing entirely on the cross-substrate test.

The substantive update for §5 is that the specialist's cross-substrate contribution is non-zero but small relative to the A-MEM pipeline, rather than zero or substrate-locked. The original framing ("specialist did not generalize") is too strong; the corrected framing is "specialist generalizes by +13pp over the obvious vanilla baseline on the same prompt, but the residual 29pp gap to a stronger pipeline is the binding finding." The §4.2.5 (b)/(c) blend verdict still stands — the specialist's contribution mechanism is narrow MAB-task answer-detection plus query-overlap surface matching — but the LoCoMo Multi-Hop evidence now reads as "specialist is operational on cross-substrate transfer but overwhelmed by a stronger comparator pipeline," not "specialist failed to transfer at all."

4.2.5 Layer 5 — Specialist learning analysis

The remaining question is mechanistic: *what did the specialist actually learn?* Phase 7's pre-registered diagnostic framework distinguishes three hypotheses: (a) general "memory utility $\neq$ similarity" — the supabrain thesis claim; (b) MAB-task-specific patterns; (c) surface-level pattern matching. Three diagnostics (probe analysis, cosine-only baseline binding, no-retrain cross-task transfer) produce evidence-weighted votes; pre-registered tie-break defaults to (b).

Spearman ρ(specialist score, BGE cosine) on the validation set ($n = 885$): +0.074 $[+0.012, +0.138]$**.** The specialist is statistically distinguishable from cosine but the correlation is negligible — it is *not* a cosine clone, ruling out the simplest form of (c).

Label-discrimination on validation. The specialist's score distribution on validation pairs labeled positive ($n = 84$) has median -0.125; on pairs labeled negative ($n = 801$) the median is also -0.125 (means: -0.060 and -0.052; Δ median $= 0.000$, Δ mean $= -0.008$). Despite being trained to predict utility, the specialist's scores on a held-out sample of (memory, query) pairs do not discriminate utility-positive from utility-negative pairs. Combined with $\rho \approx 0$ vs cosine, the specialist is measuring a signal orthogonal to both utility and cosine, whose apparent contribution at $K = 5$ does not survive K-normalization (§4.2.2).

Probe analysis. Four probe classes (50 pairs each), scored through the specialist on Modal CUDA:

Probe class	n	mean score	% above 0
P1 random text (off-substrate)	50	+0.07	52%
P2 query repetition (memory = query verbatim)	50	**+0.94**	**100**%

Probe class	n	mean score	% above 0
P3 answer-only (memory = gold answer string)	50	**+0.18**	**80%**
P4 plausible distractor (same dataset, different query)	50	−0.01	38%

P2 query-repetition firing at 100% above 0 is strong evidence of surface lexical matching; P3 answer-only firing at 80% is consistent with a "contains-the-answer-string" detector. The specialist correctly rejects topically related but query-unrelated distractors (P4, 38%) and is uncertain on off-substrate random text (P1, 52%). The pattern is a (b)/(c) blend: narrow MAB-task-style answer-detection plus query-overlap surface matching.

Cross-task transfer (no-retrain substitute). Per-competency $K = 5$ specialist-minus-vanilla deltas across competencies are +7.5 (AR), +4.0 (TTL), +5.0 (CR) — a mean of +5.5. At $K = 10$, excluding the banking77 artifact, the same deltas are +0.0 (AR, ruler_qa2) and −4.0 (CR, factconsolidation), a mean of −2.0. The within-MAB competency-level signal collapses from +5.5 to −2.0 moving from $K = 5$ to $K = 10$.

Verdict aggregation. Per the pre-registered evidence table:

Diagnostic	(a)	(b)	(c)
Probe analysis (P2 + P3 patterns)	0	1	1
Cosine baseline ($\rho \approx 0$ + B1 ≈ B2 SEM + val discrim fails)	0	0	1
Transfer (K-collapse + LoCoMo failure)	0	2	0
Total	**0**	**3**	**2**

The pre-registered vote tally returns (b) by one vote with medium confidence, but this margin is too narrow to read as a clean verdict; **the substantive verdict is a (b)/(c) blend at medium confidence.** The specialist learned a narrow MAB-$K = 5$-locked pattern that detects answer-bearing strings (P3 answer-only firing at 80%, the (b)-supporting evidence) and reacts to query-overlap surface features (P2 query-repetition firing at 100%, the (c)-supporting evidence). Neither hypothesis is cleanly distinguished by our diagnostics; both contribute. The apparent end-to-end gains are best read as a blend-noise lift in a small-K top-K pool rather than a transferable utility signal.

4.2.6 Synthesis

Each rigor step is independent. The bootstrap CIs were a writing-session prerequisite committed before Phase 7 began. The Phase 7 PLAN.md pre-registered the K-normalization sub-task, the LoCoMo Rule-D test, and the (a)/(b)/(c) diagnostic framework as distinct binding hypotheses;

each was applied mechanically as the relevant data became available. The BM25 sparse-retrieval baseline (Table 1, Table 5; §4.2.3) and the LoCoMo Multi-Hop prompt-control (§4.2.4) were added in a post-Phase-7 reviewer-objection round (Session E, 2026-05-30); both were applied without modifying thresholds or adding hypothesis-rescuing tests. At no step did we modify thresholds after seeing results, and at no step did we add a hypothesis-rescuing test not previously specified.

The arc reads, in one sentence: the Phase 6 specialist's apparent $+8/+7/+4/+5$ multi-competency win against vanilla at $K = 5$ holds in point estimate on all four datasets, but (*i*) only two of the four cells survive paired bootstrap as strictly significant; (*ii*) all four mostly disappear at $K = 10$, where vanilla retrieval at the higher K already exceeds the specialist's $K = 5$ SEM on the largest-delta dataset; (*iii*) BM25 sparse retrieval outperforms the specialist by +13 to +22pp on three of four datasets at the same $K = 5$, making the specialist's vanilla-cosine wins "less suboptimal than BGE cosine alone" rather than competitive; (*iv*) on a cross-substrate Multi-Hop benchmark the specialist's F1 is 28.85pp below a non-trained baseline but contributes +13pp over vanilla cosine with the same prompt (the residual A-MEM gap is pipeline-attributable rather than total cross-substrate failure); and (*v*) the specialist fails to discriminate utility-positive from utility-negative pairs on its own held-out validation set, scoring memory-equals-query probes at 100% above zero. The methodology of §3 — counterfactual ablation as a non-circular outcome signal — survives every step intact. What does not survive is the claim that *this specialist*, trained on those labels, delivers an end-to-end-competitive result on MAB at any K or matches the strongest available cross-substrate pipeline (A-MEM).

We carry these findings into the discussion of §5 without softening them and without overstating their reach: the rigor revealed the result's true scope, not a flaw in the methodology and not a successful baseline. Figure 2 visualizes the cascade — apparent K=5 win → bootstrap CIs → K-normalization → BM25 baseline → cross-substrate + prompt-control on LoCoMo MH → learning-pattern probes — and summarizes "what survives" the dissolution at the bottom.

5 Discussion & Conclusion

5.1 What the negative result is, and is not

The specialist did not deliver the multi-competency end-to-end gain its $K = 5$ point estimates first suggested. The Phase 6 apparent $+8/+7/+4/+5$ multi-competency advantage holds in point estimate on all four MAB datasets, but narrows or disappears at every one of the five rigor layers of §4.2 — paired-bootstrap CIs (§4.2.1), K-normalization to the published comparator depth (§4.2.2), the BM25 sparse-retrieval baseline (§4.2.3, added Session E), cross-substrate transfer to LoCoMo Multi-Hop with the Session-E prompt-control sub-measurement (§4.2.4), and learning-pattern diagnostics on the specialist's score distribution (§4.2.5). Layer by layer: bootstrap reduces 4/4 point-estimate wins to two strictly significant cells at $K = 5$; K-normalization vanishes most cells at $K = 10$ as vanilla retrieval catches up; BM25 overshadows the specialist at every K, beating it by +13 to +22pp on three of four datasets; LoCoMo Multi-Hop falls 29pp short of A-MEM (though Table 7's prompt-control shows the specialist does contribute +13pp over vanilla cosine on the same prompt); and the learning probes are mechanistically consistent with a narrow $K = 5$-locked surface pattern. That is the established result and we state it plainly.

What does not follow from this result is a falsification of the broader hypothesis under which the work was undertaken. The hypothesis that context-allocation for LLM agents requires utility-aware policy distinct from cosine similarity remains untested by this paper in the relevant sense: we tested one specific operationalization — a supervised re-ranker over Qwen2.5-1.5B trained on counterfactual-ablation labels at $K = 5$ on MAB — and that operationalization closed. The question of whether a different operationalization (different label generator, different substrate,

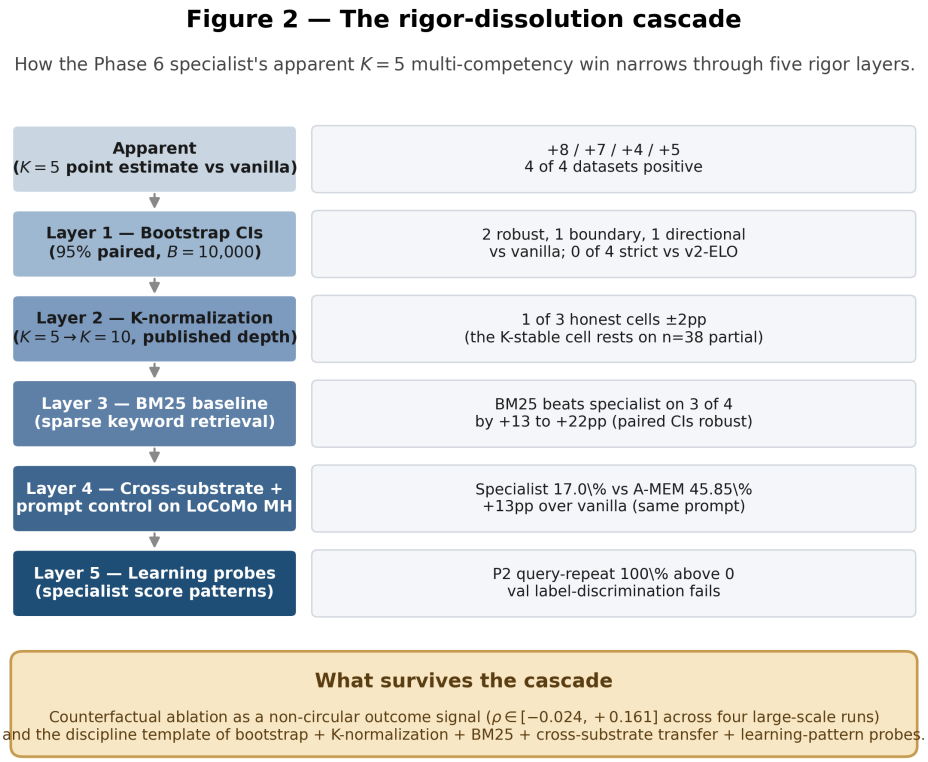

Figure 2: Figure 2 — The rigor-dissolution cascade. Source: `paper/figures/fig2-rigor-cascade.{pdf,png}`.

different scale, different downstream consumer) would also close is open. The distinction between "this specific approach failed" and "the problem is unsolvable" is one we have not earned. We close one path and report what we learned doing it.

5.2 What survives

Four things survive in our judgment, ordered by what we believe is most directly reusable.

Counterfactual ablation as a non-circular outcome signal (§3). Across three independent generation runs and one cross-substrate audit (§3.4), Spearman correlations between counterfactual utility and BGE cosine similarity ranged from -0.024 to $+0.161$, with the binding cross-substrate transfer subset (LoCoMo Multi-Hop) at $\rho = -0.024$ with a confidence interval spanning zero. The construction satisfies three independent structural-non-circularity arguments (no similarity score in the label; no LLM judge; rank-independence within $\mathcal{T}_K$) and is empirically separated from cosine on every substrate where it applies. This is reusable beyond this project: any benchmark for which a fixed answerer and a substring-matchable gold answer exist can produce counterfactual-ablation training labels at $K+1$ answerer calls per pair. We do not claim it is the only such signal — alternatives we considered but did not head-to-head test include session-trace mining for cases where a memory was retained or reverted in downstream work [17] — and we do not claim it produces better labels than those alternatives. What we claim is what we tested: this signal is non-circular by construction and empirically demonstrated to be so at scale.

The K-normalization + sparse-baseline + cross-substrate protocol as a discipline template (§4.2.2–§4.2.4, Table 1 BM25 row). Three rigor steps tightened our claim: applying paired bootstrap CIs to every per-dataset delta, re-evaluating the apparent winners at the published comparator's K, and adding the strongest sparse-retrieval baseline (BM25) the harness supports. None is novel; all three are routine in adjacent literatures. What we report is what they revealed when applied to a memory-system result that looked positive: a 4/4 point-estimate win became 2/4 robust at $K = 5$, substantially smaller at $K = 10$ where vanilla retrieval shifted upward by +3 to +28 points across datasets, and overshadowed at every K by an unused BM25 baseline that beat the specialist by +13 to +22pp on three of four datasets. The methodological recommendation we draw, narrow and concrete: memory-system results reporting $K = 5$ over BGE cosine should report bootstrap CIs, a $K = 10$ row, and a BM25 baseline row, with all three normalized against the comparator literature's standard. We do not propose this as a general standard for the field; we report it as the discipline that changed our reading of our own data.

The substrate-mismatch failure-mode catalogue. This paper extends a longer-running iteration history (briefly: a seven-phase project across v1 multi-resolution memory, v2 ELO + hygiene, v3 composite scoring, v4 real-corpus audit (abandoned), v4.6 query-conditional micro-test, Phase 5 MemoryAgentBench evaluation, and Phase 6 specialist re-ranker) whose individual failures have been documented elsewhere. The current paper adds three entries — $K = 5$-vs-$K = 10$ delta collapse, scaled-counterfactual training producing a substrate-locked detector, and validation pair-discrimination failure under what appears at first to be a measurable end-to-end advantage — and presents them in continuity with the prior catalogue rather than as standalone discoveries. The catalogue's value, if any, is as a forward reference for future work that might otherwise rediscover the same patterns.

Pre-registration discipline as the surfacing mechanism. The Phase 6 AND-conjunction decision rule (validation F1 threshold AND end-to-end-vs-v2-ELO win count) was committed in advance precisely so the proxy and end-to-end metrics could disagree on the record. The rule fired the disagreement mechanically: validation F1 missed its threshold while end-to-end gains met theirs. Without the AND-conjunction the paper could plausibly have led with either half. This is one worked example of pre-registration in ML methodology; we report it because the underlying discipline was load-bearing for the rest of the paper, and so others can re-use the

pre-registration ADR format [16], not because pre-registration as a practice is novel.

5.3 Why a documented negative is a contribution

Concurrent work in language-model behavioural evaluation has argued that many positive claims in the literature do not survive rigorous re-testing [5]. The present paper is not a remediation of that broader concern. It is one documented end-to-end instance of the dissolution pattern: an apparent positive result, four pre-registered rigor steps, and a sharper claim at the end. We report it under the methodological framing that motivated the work, not as a general remedy. The contribution is the instance, the methods used to produce it, and the artifacts (pre-registration timestamps, bootstrap analyses, training-pair files, evaluation outputs) that support independent re-analysis.

5.4 The validation-F1 / end-to-end disconnect as a transferable finding

A more specific result worth surfacing: per-pair classification metrics on a skewed binary utility label (positive-class rates between 10% and 22% across the substrates we measured in §3.4) are weak predictors of end-to-end retrieval quality when the downstream consumer is a top-K re-ranking pipeline. Our specialist's validation F1 of 0.262 landed in the WEAK tier and missed the pre-registered threshold by 6.7pp, yet the same specialist produced point-estimate end-to-end SEM gains over vanilla on all four MAB datasets at $K = 5$, two robust under bootstrap. The §4.2.5 label-discrimination analysis tightens the finding still further: on a held-out validation sample the specialist's score distributions on utility-positive and utility-negative pairs were statistically indistinguishable (median -0.125 for both, Δ mean -0.008), and the apparent end-to-end contribution at $K = 5$ did not survive K-normalization. Whatever this specialist learned was not the discriminator its training loss optimized for, and the per-pair F1 proxy did not predict either the apparent end-to-end win or its subsequent dissolution. We surface this as relevant to any subsequent work training memory-utility specialists from binary labels: the pair-level metric should not by itself decide whether the end-to-end result is worth reporting.

Beyond the val-F1 disconnect, the broader methodology lesson is that strong unsupervised baselines (BM25, classical TF-IDF) belong in Phase 1, not Phase N. Phase 6's apparent win was $K = 5$-locked AND BM25-unaware; one of those would have been a less severe overstatement than both.

5.5 Open questions and what we would do differently

Three concrete open questions remain, none of which we resolve. Each is bounded: we name the specific design choice we did not test, not a broad gesture toward future work. (Two additional open questions surveyed in earlier drafts — a deferred scaled-retraining sub-task that procedurally did not block Rule C, and a generic "the broader thesis remains open" framing — are not enumerated here; the former is a procedural boundary already implicit in §4.1.1, and the latter is subsumed by the falsification condition below. A fourth item — including BM25 as the Phase-1 sparse-retrieval baseline rather than as a late-stage check — is recorded as a practical recommendation under "Methodological scope limits" below, since it is a process lesson from this paper's Session-E revision round rather than an open empirical question.)

First, our labels are binary in $\{-1, 0, +1\}$ clipped from a substring-match correctness differential. A graded variant using token-level F1 or log-probability margin would preserve ranking information at the cost of a noisier signal; the trade-off is unmeasured. The methodological question is bounded: does a graded label change the specialist's K-normalization behaviour (§4.2.2), or does the K-collapse pattern persist independent of label granularity?

Second, the specialist was not trained with the answerer's K as an explicit parameter. A K-aware training procedure that varied the retrieved pool size during training, or that included

$K = 10$ examples in the training distribution, might or might not generalize better; we did not test this. The §4.2.2 evidence that the apparent win was K-specific suggests this is a defensible direction; it does not predict the outcome.

Third, the cross-substrate transfer test in §4.2.4 used one substrate (LoCoMo) and one comparator (A-MEM [1]). LongMemEval [10] would provide an independent data point we did not collect. The §4.2.4 Session-E prompt-control shows the specialist contributes +13pp over vanilla cosine on the LoCoMo Multi-Hop substrate with the same prompt; whether that contribution magnitude replicates on LongMemEval — and whether the $\sim$ 29pp residual gap to A-MEM closes when the comparator's pipeline is replaced — is the bounded second-substrate question.

What would falsify the broader thesis. To pre-empt the reading that the broader thesis ("a useful memory layer for LLM agents must identify per-memory utility distinct from retrieval similarity") is structurally protected from falsification by carving out specific operationalizations, we state the falsification condition explicitly. The broader thesis would be falsified by a single existence proof: a sufficiently scaled benchmark, on which a frozen retriever using only cosine similarity at a fairness-comparable K, with no per-memory utility re-ranking and no architectural tweaks beyond canonical RAG, matches or exceeds the best published memory-system result by a margin that the methods of §4.2 cannot dissolve. The published cosine-only baselines we are aware of on LoCoMo and MAB do not currently meet that bar, but neither does our specialist; the thesis remains testable, just not tested by us.

Methodological scope limits. Several alternatives to the design choices we made above are worth surfacing as scope-limits so the paper does not over-claim within its frame. - We did not test against more recent data-attribution methods (TracIn, TRAK) as alternative labelers; §2.4 notes the rationale (per-pair labels at inference time vs. per-training-point gradients), but a head-to-head against TracIn at the same label-generation cost would be a sharper comparison than we report. - Our binary $u_i \in \{-1, 0, +1\}$ clip discards the $u_i = -1$ "actively harmful memory" sub-case. Empirically these are rare on extractive substrates, but a separate analysis treating $u_i = -1$ as a distinct utility class (rather than folded into the negative class with $u_i = 0$) would more cleanly distinguish "useless" from "misleading" memories. - We position §2.1 "upstream of all of these systems" — read this as "complementary to" rather than architectural; no downstream memory architecture yet adopts the counterfactual-ablation outcome signal, and our work is one demonstration that does not establish a field standard. - The replication-crisis framing in §2.3 and §5.3 motivates our methodological emphasis but our contribution is one documented instance of the dissolution pattern, not a general remediation. We do not claim our methods generalize across the broader range of language-model evaluations the replication-crisis literature surveys. - **Strong-baseline first.** Classical sparse-retrieval baselines (BM25 minimum; BM25 + cross-encoder re-rank as the next rung) and project-internal heuristic baselines of v2-ELO type should be measured in Phase 1 of any new memory-evaluation effort, before training learned components. Our Phase 6 omitted BM25, which contributed to the apparent-result overstatement that Phase 7 had to dissolve (§4.2.3 retrospective). The lesson generalizes: include the strong unsupervised baselines from the start, not as a late-stage check after a learned re-ranker has been trained against a weaker baseline.

References

[1] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang. "A-MEM: Agentic Memory for LLM Agents". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2025. arXiv: 2502.12110 [cs.CL]. URL: https://arxiv.org/abs/2502.12110.

[2] H. Li et al. *LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods*. 2024. arXiv: 2412.05579 [cs.CL]. URL: https://arxiv.org/abs/2412.05579.

[3] S. Es, J. James, L. Espinosa Anke, and S. Schockaert. *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. 2023. arXiv: 2309.15217 [cs.CL]. URL: https://arxiv.org/abs/2309.15217.

[4] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia. "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems". In: *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. 2024. arXiv: 2311.09476 [cs.CL]. URL: https://arxiv.org/abs/2311.09476.

[5] L. Vaugrante, M. Niepert, and T. Hagendorff. *A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions*. 2024. arXiv: 2409.20303 [cs.CL]. URL: https://arxiv.org/abs/2409.20303.

[6] Y. Hu, Y. Wang, and J. McAuley. *Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions*. 2025. arXiv: 2507.05257 [cs.CL]. URL: https://arxiv.org/abs/2507.05257.

[7] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. "Evaluating Very Long-Term Conversational Memory of LLM Agents". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. 2024, pp. 13851–13870. arXiv: 2402.17753 [cs.CL]. URL: https://aclanthology.org/2024.acl-long.747/.

[8] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. *MemGPT: Towards LLMs as Operating Systems*. 2023. arXiv: 2310.08560 [cs.AI]. URL: https://arxiv.org/abs/2310.08560.

[9] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav. "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory". In: (2025). arXiv: 2504.19413 [cs.CL]. URL: https://arxiv.org/abs/2504.19413.

[10] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu. "LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2025. arXiv: 2410.10813 [cs.CL]. URL: https://arxiv.org/abs/2410.10813.

[11] P. W. Koh and P. Liang. "Understanding Black-box Predictions via Influence Functions". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1885–1894. arXiv: 1703.04730 [cs.LG]. URL: https://proceedings.mlr.press/v70/koh17a.html.

[12] A. Ilyas, S. M. Park, L. Engstrom, G. Leclerc, and A. Madry. "Datamodels: Predicting Predictions from Training Data". In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022. arXiv: 2202.00622 [stat.ML]. URL: https://arxiv.org/abs/2202.00622.

[13] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie. "C-Pack: Packed Resources For General Chinese Embeddings". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 2024. arXiv: 2309.07597 [cs.CL]. URL: https://arxiv.org/abs/2309.07597.

[14] A. Yang et al. *Qwen2.5 Technical Report*. 2024. arXiv: 2412.15115 [cs.CL]. URL: https://arxiv.org/abs/2412.15115.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2022. arXiv: 2106.09685 [cs.CL]. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[16] ktdmax/supabrain. *Pre-registration ADR timeline for the supabrain Phase 6/7 specialist evaluation*. https://github.com/ktdmax/supabrain/tree/7a27c3a/progress. 2026.

[17] ktdmax/supabrain. *Phase 6 outcome-signal candidate analysis (Options A–E rejection rationale)*. https://github.com/ktdmax/supabrain/blob/7a27c3a/progress/v6/subgoal-1-decision.md. 2026.