Does Attaching LBrain Lift LLM Benchmarks?

A Rigorous With/Without Lift-Study Design

Metavolve Labs • pre-registration draft • not legal or scientific advice until reviewed

This design was produced by a grounded survey of LBrain's actual code, a benchmark-to-mechanism mapping, a confound analysis, and an adversarial reviewer pass. It is a draft to pre-register and refine before running.

1. Executive Summary

The claim under test: **attaching LBrain to a leading LLM measurably improves grounded-memory and faithfulness benchmarks** — not only versus a bare model, but versus a strong retrieval-augmented baseline. The design's central, honest finding shapes everything: LBrain is a **retrieval + injection + guard layer, not a generation model**, and its differentiators fire **only when the corpus carries structural signals** (priority tiers, wikilinks, supersession edges, curated core-memory). On a flat passage corpus, LBrain ≈ a competent vanilla hybrid RAG. Therefore the credible, defensible thesis is **scoped**:

> *"Structure-aware, newest-wins memory (LBrain) beats strong RAG on temporal / multi-session grounding"*

— **not** the inflated "LBrain makes any LLM smarter," which a reviewer would dismantle. The single highest-value, lowest-cost first experiment (§ 7) tests the crux for roughly a day of compute and gates all expensive work.

2. What LBrain Actually Does (verified in code)

LBrain indexes a corpus of markdown "lairs" + a memory journal into SQLite (FTS5 + sqlite-vec) and, at query time, performs **hybrid retrieval** (vector + BM25 fused by Reciprocal Rank Fusion) with **always-on multiplicative boosts**: priority-folder 1.3×, inbound-wikilink graph $1+0.05 \cdot \min(\text{inlinks}, 5)$, and **supersession de-ranking** 0.25×. It adds an injection layer (token-budgeted context, an untrusted-data **fence** framing retrieved text as data-not-instructions, a provenance footer, and an optional always-on "core-memory" block) and a zero-LLM Lair Protocol (save-worthiness + anti-pattern/feedback-conflict checks). It performs **no answer generation**; the harness must add the generate step.

Scoping caveat (bounds the whole thesis). The boosts, fence, core-memory, and supersession/recency guards only affect results if the corpus carries those signals. A flat plain-passage benchmark exercises only the hybrid core, so LBrain's lift collapses toward "RRF-hybrid vs. single-retriever." The mechanism a flat baseline **cannot** replicate is **supersession (newest-wins) temporal memory** — which is why the study centers there.

3. Design: Four Arms, One Changed Variable

Base model, decoding, and injection-template shell are held **byte-identical** across arms; the only thing that varies is which passages are selected, their ordering, and LBrain's ranking machinery.

Arm	Name	Role
A	Bare (closed-book)	Parametric-knowledge floor. Never sees the corpus — not LBrain's fair contrast.
B	Vanilla-RAG (dense)	Same corpus, same embedder/chunker, plain cosine top-k. The "retrieval-helps" contrast.
C	**Strong-RAG (hybrid)**	BM25+dense RRF, **without** LBrain's boosts/core-memory/protocol. A genuinely tuned baseline.
D	**LBrain (full)**	All always-on boosts + fence + provenance + core-memory.

Primary estimand: paired Δ = LBrain − Strong-RAG (isolates LBrain's specific design from mere hybrid fusion). Secondary: LBrain − Vanilla, and RAG − Bare (the "retrieval helps" lift). Internal ablations (−structural-boosts, −fence, −supersession, −core-memory, structure-stripped corpus) decompose *which* component carries any lift. Reference bounds: BM25-only lower bound, long-context stuffing control, and an oracle-retrieval upper bound.

4. Benchmark Slate (mechanism-mapped)

Pri	Benchmark	Why it stresses LBrain's mechanism
P0	**Supersession / Knowledge-Update probe** (bespoke; TempLAMA/TemporalWiki/FreshQA edit-pairs)	Cleanest isolation of the unique newest-wins mechanism. *Generation-free.*
P0	**LongMemEval**	Flagship long-term conversational-memory setting.
P0	**Metavolve Confabulation Probe** (Paper 4 reuse)	Lowest harness cost; 4-way CORRECT/ABSTAIN/CONFABULATE scoring.
P0	**KILT provenance** (R-precision / Recall@k)	Generation-free head-to-head on the retrieval design. Contamination-immune.
P0	**GSM8K** (negative control)	Pure reasoning — **predicted flat**; movement = contamination alarm.
P1	HotpotQA + MuSiQue / 2WikiMultihop (supporting-fact F1)	Direct public test of the wikilink-graph structure claim.
P1	FreshQA / RealTimeQA	Post-cutoff answers — contamination-robust; tests newest-wins.
P1	RAGTruth	Span-level hallucination of generation vs. retrieved evidence.
P1	Distractor / irrelevant-corpus inertness (safety control)	**Predicted zero degradation** — injection must not add noise.
P1	Poison-relay guard (Paper 4 Condition P)	**Pre-registered NULL** (see § 5).
P2	ALCE, PopQA, SQuAD v2, FActScore	Attribution, long-tail QA, abstention, long-form atomic-fact precision.
P2	MMLU over private corpus	Negative control + contamination sentinel.

5. Pre-Registered Hypotheses and Honest Nulls

Directional hypotheses (frozen before scoring; publish regardless of direction):

- **H1 (crux):** LBrain − Strong-RAG > 0 on the supersession-probe current-answer rate *and* the LongMemEval knowledge-update subtask. *Falsified if either CI crosses 0 ⇒ the headline temporal differentiator is reported UNSUPPORTED.*
- **H2 (retrieval crux):** LBrain − Strong-RAG > 0 on KILT R-precision. Falsified if ≤ 0.
- **H3 (structure):** LBrain − Strong-RAG > 0 on HotpotQA/MuSiQue supporting-fact F1. Falsified ⇒ "structure-aware retrieval > flat RAG" unsupported.
- **H4:** modest, non-headline faithfulness lift on RAGTruth / Paper-4 correctness.

Pre-registered NULLS — these MUST be published, not buried:

- **Poison-relay:** LBrain does *not* reduce verbatim relay of a plausible-but-false planted fact (∼100% relay). **Grounding induces faithfulness, not truth**; the fence stops instruction-injection, not false-fact relay.
- **Honesty-axis ceiling:** on already-calibrated modern base models, the abstention/confabulation lift is near-null; only confab-prone bases have headroom. **Correctness lift is universal; honesty lift is generation-specific.**
- **Reasoning negative control (GSM8K/BBH):** predicted $\Delta \approx 0$. Any lift indicts leakage or a confounded harness.
- **Retrieval-uplift attribution:** a benchmark where vanilla-RAG already captures the full lift is reported as a null for LBrain's *differentiator*, even if bare→LBrain looks large.

6. Key Confound Controls

- **Fair baseline:** the mandatory Strong-RAG arm; the thesis-critical estimand is LBrain − Strong-RAG, never bare − LBrain.
- **Corpus-provisioning leakage:** convert only the benchmark's supporting *evidence* (never question or gold answer) to markdown via one frozen, answer-blind recipe; isolated `LBRAIN_HOME` per benchmark so the founder's real lairs never leak; anchor the retrieval thesis on generation-free metrics (KILT R-precision) that are contamination-immune.
- **Structural-signal bias (researcher DoF):** pre-register *one* deterministic, answer-blind structure-derivation rule and prove it does not correlate with the gold label.
- **Parity:** byte-identical prompt/fence/template; equalized k, per-chunk char cap, and total token budget; identical decoding (temp 0 or fixed seed set); identical embedder+dim+chunker across retrieval arms; matched candidate-pool depth; the AMP gate disabled or applied to all arms.
- **Statistics:** per-item paired analysis (McNemar / paired deltas), Wilson CIs, a priori power for a +3–5 point minimum effect, BH-FDR across the confirmatory family, model treated as a random factor with a small-k (Hartung–Knapp) correction; ≥ 3 model families and ≥ 2 sizes, requiring **sign-consistent** lift before any general claim.
- **Reproducibility:** disable the running instance's heartbeat/autosync/consolidation crons during runs; pin and publish the exact LBrain git SHA (not just `config.toml`).

7. The First Experiment — run this week

A generation-free supersession / knowledge-update probe. Hand-author ∼50–100 update-pairs from public temporal data (TempLAMA / TemporalWiki / FreshQA edit-pairs) — each pair a *stale* note and its *superseding* note carrying a `Supersedes:` edge — into a fresh isolated `LBRAIN_HOME`; `lbrain import && lbrain embed`. Over the library `search()` path, compare **Arm C strong-RAG (supersession OFF)** vs. **Arm D LBrain (supersession ON)** on three automatic metrics: **current-over-stale rank-win rate**, **stale-relay@k**, and **MRR-of-the-current-answer**; add the oracle upper bound. **No generation, no judge, ≈ embedding cost, runnable in a day.**

This directly tests H1/H2. **If LBrain does not rank the superseding note above the superseded one relative to a genuine (timestamp-aware) strong-RAG baseline, the unique-mechanism claim is in trouble** — and the expensive multi-session/LongMemEval build should be gated on fixing it. One cheap experiment, a decisive go/no-go.

8. Adversarial Review — verdict and required fixes

Reviewer verdict: *"Conditionally credible — survives as a SCOPED result, not the universal thesis."* Four potentially-fatal issues, each with a concrete non-redesign fix:

1. **Baseline strawmanning (critical).** The supersession crux is near-tautological against a recency-*blind* baseline. **Fix:** give Strong-RAG a genuine timestamp/metadata-aware temporal baseline, and generate supersession edges via an automated pipeline (not hand-authored oracle edges).

2. **Prompt/compute non-parity (critical).** The nominated primary (full injected-context) bundles ranking with the fence, core-memory, and provenance footer. **Fix:** make the *ranking-only* treatment (Hits injected through the identical wrapper both RAG arms use) the primary; decompose fence/core-memory as additive ablations.

3. **Contamination / answer-blindness (critical).** The structures LBrain ranks on could correlate with answers. **Fix:** publish the exact structure-derivation rule and *prove* answer-blindness with a reported number; anchor on generation-free KILT R-precision.

4. **Generalization (critical).** Every differentiator is tuned to the founder's private corpus. **Fix:** elevate the **structure-stripped-corpus** arm to a co-primary generalization endpoint and add ≥ 1 genuinely external structured corpus.

Plus high-severity fixes: match the cross-encoder reranker (or bar it from the primary), pin candidate-pool depth, disable crons, report total (write+read) compute on memory benchmarks, and re-establish judge–human κ per generation benchmark.

LBrain Benchmark Study — lift-study design — pre-register, apply the red-team fixes, run the first experiment, publish every direction.