VERIFIED MEMORY FOR KNOWLEDGE REUSE IN VARIANT-TO-GENE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agents that prioritise causal genes at genome-wide association (GWAS) loci are typically stateless: every locus is re-analysed from scratch, even when the same gene–trait relationship was already established in an earlier study. We present EviMem+CAVE, a memory layer for such agents that stores verified gene–trait claims in a four-state contract (CAVE) and retrieves prior loci by a numeric Evidence Fingerprint (EviMem). On a frozen benchmark of 844 loci (296 held-out test loci), the memory short-circuits 31.7% of within-ancestry test loci and 79.1% of cross-ancestry replication loci, answering them without an LLM call (top-1 precision 0.925 and 0.977), and on the within-ancestry test loci raises GPT-5.2 top-1 accuracy over the same agent without memory by +0.118 (95% CI +0.077 to +0.159). A mechanistic decomposition shows that the gain comes from verified reuse of recurrent gene–trait knowledge through the CAVE store rather than from fingerprint similarity, and that on novel gene–trait pairs we detect no improvement. We also show that a supervised baseline's apparent lead is an artefact of how gold genes are inserted into candidate sets, and we report the evaluation protocol under which reuse can be measured separately from such artefacts.

1 INTRODUCTION

Locus-to-gene (L2G) prioritisation asks which gene at a GWAS locus is causal. It is a high-stakes step between genetic association and drug-target selection: targets with human genetic support succeed more often in the clinic (Nelson et al., 2015). Supervised scores (Mountjoy et al., 2021; Ghoussaini et al., 2021), polygenic enrichment (Weeks et al., 2023) and combined SNP-to-gene strategies (Gazal et al., 2022) rank candidate genes from functional evidence, and LLM agents are increasingly used to synthesise that evidence into a call with a rationale.

Such agents are *stateless*. A pleiotropic gene that has been validated as causal for a trait in one study is re-derived from raw evidence the next time a locus for the same trait is analysed, at the full cost of an LLM call and with the full variance of a fresh synthesis. Human analysts do the opposite: they accumulate validated knowledge and consult it first. The question we study is therefore not whether memory makes an agent reason better about unseen biology, but whether an agent can *reuse verified knowledge* safely and cheaply, and how to measure that reuse honestly.

We make four contributions.

- **C1 — A verified memory layer for L2G agents.** EviMem+CAVE combines a four-state verification contract over (gene, trait) claims, populated from training data only, with evidence-fingerprint retrieval over previously analysed loci. We describe the system exactly as it was run (§3), including the two short-circuit paths and their measured contributions.
- **C2 — Reuse delivered.** Verified reuse short-circuits 31.7% of within-ancestry and 79.1% of cross-ancestry loci without an LLM call at high precision, and improves top-1 accuracy over the memory-free agent on both backbones (§5.1).
- **C3 — What memory reuses.** Decomposing test loci by overlap with training shows that the gain is concentrated on recurrent gene–trait pairs; no improvement is detected on novel pairs, unseen traits or zero-overlap loci. Fingerprint similarity alone, evaluated without an LLM, reaches 0.436 top-1 accuracy, and nearly all short-circuits come from the verified store (§5.2).

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

- **C4 — Evaluation that separates reuse from artefacts.** Gold genes that the candidate generator misses are force-added to candidate sets; a classifier that sees only which evidence fields are missing exploits this and explains a supervised model's apparent lead. We report matched-locus comparisons with published and classical methods and a protocol that keeps these effects apart (§5.3, §5.4).

2 Related work

L2G methods. Open Targets L2G trains a gradient-boosted classifier on curated gold-standard assignments using distance, molecular QTL colocalisation, chromatin and coding-consequence features (Mountjoy et al., 2021; Ghoussaini et al., 2021). PoPS prioritises genes by polygenic enrichment of gene features (Weeks et al., 2023); cS2G combines SNP-to-gene linking strategies (Gazal et al., 2022). Gene-level association (MAGMA) (de Leeuw et al., 2015) and transcriptome-wide association (Barbeira et al., 2018) provide classical, non-learned alternatives. All are evaluated by ranking candidate genes against curated gold genes.

Memory and retrieval for LLM agents. MemGPT manages long-term context for LLM agents (Packer et al., 2023); Hindsight-style memories retrieve prior experience by text (Latimer et al., 2025); reward-driven approaches learn what to store (Xiong et al., 2026); retrieval-augmented generation grounds outputs in retrieved documents (Lewis et al., 2020). In genomics, GeneGPT augments an LLM with database tools (Jin et al., 2024), and LLMs are known to be unreliable at precise numeric reasoning (Kambhampati et al., 2024). EviMem+CAVE differs in storing *verified claims* with an explicit lifecycle and retrieving by numeric evidence rather than text.

Shortcuts and leakage in evaluation. Apparent competence can reflect shortcut learning (Geirhos et al., 2020), "Clever Hans" predictors (Lapuschkin et al., 2019), annotation artefacts (Gururangan et al., 2018; Kaushik & Lipton, 2018), leakage (Kaufman et al., 2012) or train–test contamination (Carlini et al., 2021; Magar & Schwartz, 2022); sampled-candidate evaluation in recommendation is a known source of bias (Cremonesi et al., 2010; Rendle et al., 2019; Krichene & Rendle, 2020). We use these lenses to measure what a memory actually reuses.

3 EviMem+CAVE

Pipeline. For each GWAS locus, LD expansion and fine-mapping define candidate genes. Six deterministic stages then attach evidence to each candidate: variant annotation, colocalisation, cis-eQTL, Mendelian randomisation, disease–gene linkage and composite ranking. A memory-assisted LLM step produces the final ranking unless memory resolves the locus first (Figure 1). In the benchmark used here, the per-candidate evidence is pre-computed from Open Targets, so all methods rank the same candidate sets from the same evidence.

CAVE: verified claims. CAVE stores (gene, trait) claims in four states: PROVISIONAL, SUPPORTED, VALIDATED and REJECTED. Only VALIDATED claims can trigger a short-circuit; REJECTED claims suppress one. SUPPORTED records that a top-ranked gene satisfied at least one independent evidence path; it is metadata and does not alter rankings. Runtime write-back never creates VALIDATED claims. For all reported runs the store was seeded with 432 VALIDATED claims taken only from training gold genes; no claim exists that is specific to a test or cross-ancestry locus (0 test-only claims; Appendix A).

EviMem: evidence retrieval. Each candidate is encoded as a 16-dimensional Evidence Fingerprint (EF) whose dimensions are fixed, hand-engineered transforms of the evidence (Appendix Table 2). EviMem retrieves stored anchors by cosine similarity and proposes the gene of the most similar anchor. The encoder is not learned or optimised; we make no claim about its optimality or portability.

Short-circuit logic as run. The implementation checks the verified store first. If any candidate has a VALIDATED claim for the locus trait, that gene is returned without an LLM call (the *direct* path). Otherwise the EF-retrieved gene is approved (VALIDATED) or suppressed (REJECTED) by CAVE, and a hard short-circuit requires approval, cosine similarity at or above an adaptive threshold that depends on how many EF dimensions depart from their defaults (Appendix Table 3; the schedule is not monotone), and the gene being a candidate at the current locus (the *EF* path). All other loci are answered by the LLM with retrieved hints injected.

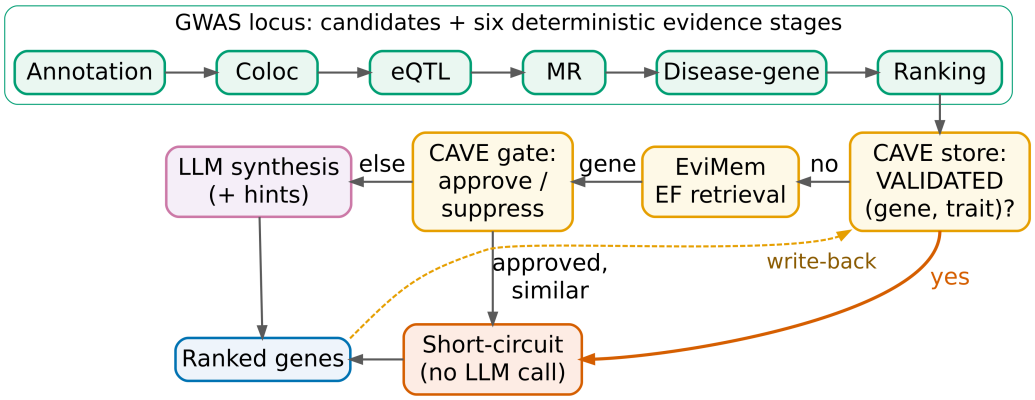


Figure 1: EviMem+CAVE as run. After candidate definition (LD expansion and fine-mapping) and the six deterministic evidence stages (variant annotation, colocalisation, cis-eQTL, Mendelian randomisation, disease–gene linkage, composite ranking), the verified store (CAVE) is consulted first: if any candidate has a VALIDATED (gene, trait) claim, the locus is short-circuited without an LLM call. Otherwise EviMem retrieves a gene from similar prior loci by Evidence Fingerprint (EF); CAVE approves (VALIDATED) or suppresses (REJECTED) it, and a short-circuit requires approval, similarity at or above an adaptive threshold and the gene being a current candidate. Remaining loci go to LLM synthesis with retrieved hints. The dashed arrow is gated write-back; the SUPPORTED state it records is metadata only and never changes rankings.

Two disclosures about the fingerprint. First, on this benchmark the fingerprint varies on only 5 of 16 dimensions; the others sit at their defaults because the benchmark supplies no MR, GWAS p-value, druggability or allele-frequency inputs. Second, the anchor store was populated by a bootstrap routine that omits eQTL tissues, while runtime queries include them; stored anchors are reproduced exactly by the bootstrap encoding (575 of 575 compared anchors) but agree with the runtime encoding for only a fraction 0.383 of them. We quantify the effect of this mismatch in §5.2.

4 Benchmark and Evaluation Protocol

Benchmark. We use a frozen benchmark of 844 loci built from the Open Targets curated gold standard (548 training and 296 test loci spanning 174 EFO traits) and an 88-locus cross-ancestry replication panel. Every population and denominator used in the paper is listed in Appendix Table 5.

Labels. Gold genes are curated assignments rather than model scores. They were, however, established by curators with access to the same kinds of evidence that the models receive (colocalisation, eQTL, coding consequence and distance), so independence of labels from model inputs is not established. We therefore make no claim that the benchmark is free of circularity; independently derived labels remain future work.

Candidate construction and populations. When the generator does not retrieve the gold gene, the gold gene is force-added so that every test locus is rankable. We call a test locus *natural-gold* when its gold candidate has non-missing eQTL and distance evidence (59 loci), a proxy for natural retrieval. This population differs from the public-gold statistic that the gold gene is naturally in the Open Targets pool for 66.1% of 1078 public loci, which uses a true retrieval flag; the two are not comparable. Natural-gold loci are better annotated by construction (mean non-missing evidence fields 6.220 versus 5.068, Mann–Whitney $p = 5.1 \times 10^{-15}$).

Overlap with training. A test locus has a *recurrent* pair if one of its gold genes is a training gold gene for the same trait name, the key used by CAVE (87 loci); otherwise the pair is *novel* (209). We also report unseen traits (EFO absent from training, 80), unseen gold genes (142) and zero-overlap loci with both (51). Keying pairs on EFO instead gives 99 recurrent loci; results are reported for both keys in the appendix.

Methods compared. EviMem+CAVE and the same agent without memory on GPT-5.2 and Qwen3-32B (five seeds each; runs audited and fixed before this analysis); XGBoost on the evidence features with and without median imputation and on the missingness mask alone; ranking by V2G or L2G score; and two parameter-free recurrence lookups fitted on training loci. The *gene lookup* scores a candidate by whether it is any training gold gene; the *trait lookup* by whether it is a training gold gene for the locus EFO. Identity-only versions use gene identity alone, and ties are broken by candidate order; the evidence-assisted versions use the V2G score as a tie-break. The Qwen3-32B agent without memory returns the first-listed candidate in 99.7% of runs, so we treat it as a pass-through of candidate order and do not use it as a headline baseline; the GPT-5.2 agent without memory does so in 85.3% of runs.

Metrics and statistics. Top-1 precision (P@1) is the fraction of loci whose top-ranked gene is a gold gene; R@3, MRR and NDCG@5 use binary relevance with all gold genes and are reported in the appendix. For agents, each locus value is the mean over five seeds. Paired differences use a bootstrap over loci (10,000 resamples, percentile 95% intervals) and an exact McNemar test on per-locus hits; families of comparisons are Holm-corrected. We write "no difference detected" whenever an interval includes zero; we did not test equivalence.

5 RESULTS

5.1 REUSE DELIVERED

Figure 2 summarises reuse. On the within-ancestry test set, memory short-circuits 31.7% of loci (seed SD 0.002) with P@1 0.925 on those loci (95% CI 0.898–0.947), saving 31.7% of LLM calls. On the cross-ancestry panel it short-circuits 79.1% of loci with precision 0.977 (0.955–0.990). Against the same GPT-5.2 agent without memory, P@1 rises from 0.493 to 0.610 (difference +0.118, 95% CI +0.077 to +0.159; McNemar $p = 1.0 \times 10^{-7}$) within ancestry and from 0.568 to 0.773 (+0.205, +0.125 to +0.295) across ancestry, where loci replicate training associations and the gain is reuse rather than transfer (§5.2). On natural-gold loci the difference is +0.068 (-0.031 to +0.173; no difference detected).

The reuse is carried by the verified store. Within ancestry, 460 of 469 short-circuit decisions across seeds (92 loci) come from the direct CAVE path, with precision 0.935; the EF path fires on 2 loci with precision 0.444. Across ancestry the direct path covers 69 loci at precision 0.986, and the EF path 1 locus at precision 0.000. Where memory does not short-circuit, retrieved hints change little: P@1 is 0.464 with hints versus 0.459 for the memory-free agent on the same loci and seeds, and both agents return the first-listed candidate in most such runs (87.9% and 89.0%).

5.2 WHAT MEMORY REUSES

Table 1 and Figure 3 decompose accuracy by overlap with training. On the 87 loci with a recurrent gene–trait pair, EviMem+CAVE reaches 1.000 against 0.575 without memory (difference +0.425, +0.329 to +0.524). On the 209 novel-pair loci the difference is -0.011 (-0.038 to +0.015), on unseen traits +0.020 (-0.005 to +0.052) and on zero-overlap loci +0.016 (-0.008 to +0.051); no difference is detected in any of these. Memory therefore reuses knowledge where knowledge exists, and adds nothing detectable where it does not. A change of ancestry does not by itself test transfer to new biology: cross-ancestry loci replicate training associations, and their gain reflects the same recall.

Parameter-free lookups make the same point from the other side. The identity-only gene lookup reaches 0.611 on all test loci, and EviMem+CAVE differs from it by -0.001 (-0.044 to +0.041; no difference detected). Against the trait lookup the bootstrap interval excludes zero (+0.036, +0.004 to +0.068) but the paired McNemar test does not (Holm-corrected $p = 0.256$), so we do not claim a difference. EviMem+CAVE's accuracy is thus at the level of verified recall; its contribution is to deliver that recall inside an LLM agent, with a train-only verification store, suppression of rejected claims, provenance for every short-circuit and fewer LLM calls, not to exceed recall baselines.

Retrieval without an LLM. To ask whether fingerprint retrieval alone explains the agent, we rank candidates by EF nearest neighbours using the same encodings, training anchors and candidate sets, with k and voting chosen by leave-one-trait-out cross-validation on training loci only. Transferring

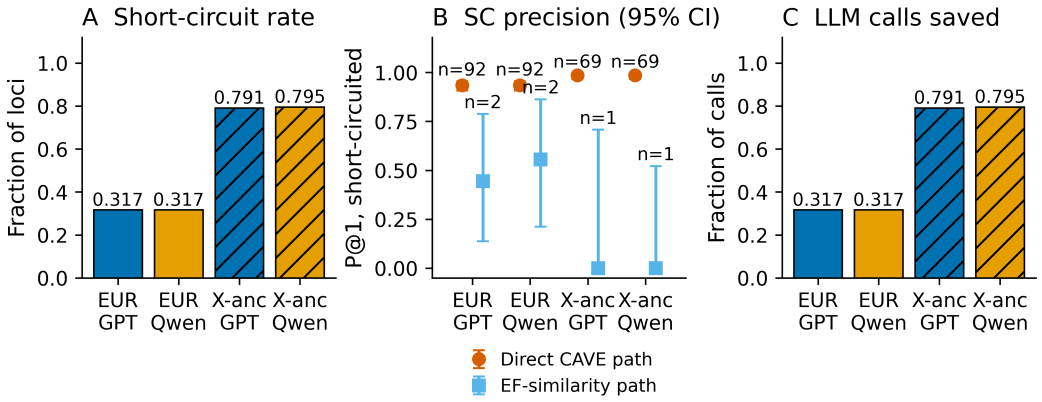


Figure 2: Reuse delivered by EviMem+CAVE. (A) Fraction of loci short-circuited without an LLM call. (B) Top-1 precision on short-circuited loci with 95% Clopper–Pearson intervals, separately for the direct CAVE path and the EF-similarity path (n = loci). (C) Fraction of LLM calls saved relative to the agent without memory. EUR: within-ancestry test set; X-anc: cross-ancestry replication panel.

Table 1: P@1 by method and test population (within-ancestry test set). Agent rows are five-seed means. Recurrent and novel pairs are keyed on trait name. Natural-gold is the leak-free population of §5.3.

Method	All	Natural-gold	Recurrent pair	Novel pair	Novel EFO	Zero overlap
EviMem+CAVE (GPT-5.2)	0.610	0.715	1.000	0.448	0.503	0.494
EviMem+CAVE (Qwen)	0.567	0.661	1.000	0.387	0.412	0.412
No-memory LLM (GPT-5.2)	0.493	0.647	0.575	0.458	0.483	0.478
No-memory LLM (Qwen)	0.395	0.407	0.414	0.388	0.412	0.412
Gene lookup (identity)	0.611	0.712	0.920	0.483	0.487	0.373
Gene lookup (+V2G tie-break)	0.618	0.712	0.943	0.483	0.487	0.373
Trait lookup (identity)	0.574	0.661	0.977	0.407	0.412	0.412
Trait lookup (+V2G tie-break)	0.574	0.661	0.977	0.407	0.412	0.412
XGBoost (full features)	0.782	0.437	0.726	0.806	0.807	0.780
XGBoost (median-imputed)	0.775	0.431	0.715	0.800	0.792	0.761
XGBoost (missingness only)	0.733	0.441	0.724	0.737	0.738	0.725
Rank by V2G	0.395	0.407	0.414	0.388	0.412	0.412
Rank by L2G	0.395	0.407	0.414	0.388	0.412	0.412

the gene of similar training gold anchors reaches 0.436 (encoding as run) and 0.443 (consistent encoding), below EviMem+CAVE by +0.174 (+0.125 to +0.225). The encoding mismatch therefore costs little at the ranking level; in a retrieval-level simulation of the EF path with the adaptive thresholds it lowers precision from 0.686 to 0.568 (Appendix B). A k-NN *classifier* over all training candidates reaches 0.645 but falls to 0.390 on natural-gold loci; removing the eQTL dimension drops it to 0.250. Its apparent strength is the same force-add artefact analysed next: the EF gives every evidence-free candidate the same default vector, so such candidates cluster together.

5.3 An evaluation artefact: force-added gold genes

Force-added gold genes carry little evidence, and that absence is detectable. XGBoost reaches 0.782 on all test loci and 0.775 after median imputation, but a classifier trained only on which evidence fields are missing reaches 0.733 (Figure 4A). On natural-gold loci all three fall to 0.437, 0.431 and 0.441, while EviMem+CAVE reaches 0.715 (Figure 4B). On these loci EviMem+CAVE exceeds XGBoost with full features by +0.278 (+0.125 to +0.427; Holm p = 0.024) and median-imputed XGBoost by +0.285 (+0.142 to +0.427). Stratifying test loci by how many quantitative evidence fields the gold candidate has makes the mechanism explicit: on the 108 loci whose gold gene has none, the missingness classifier scores 1.000, ranking by V2G scores 0.000, and EviMem+CAVE scores 0.263 (Appendix E).

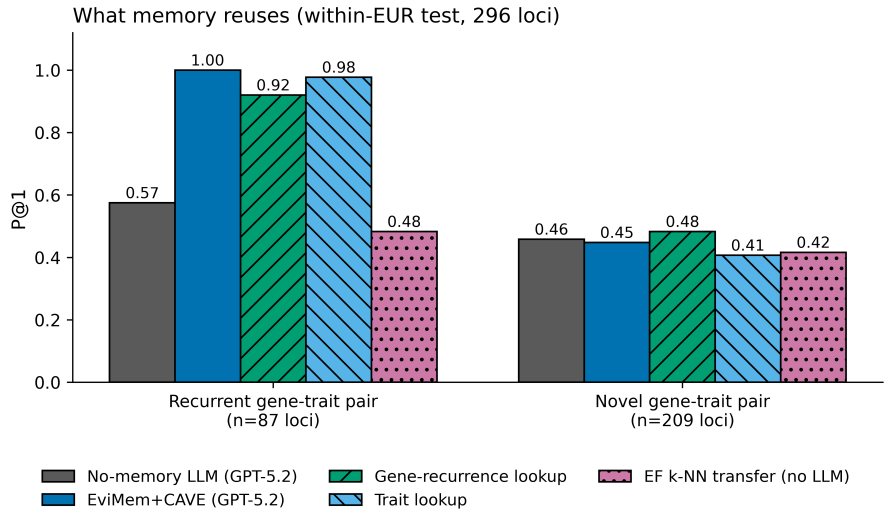


Figure 3: P@1 on test loci with a recurrent versus a novel gene–trait pair. EF k-NN transfer ranks candidates by similarity to training gold anchors of the same gene, without an LLM.

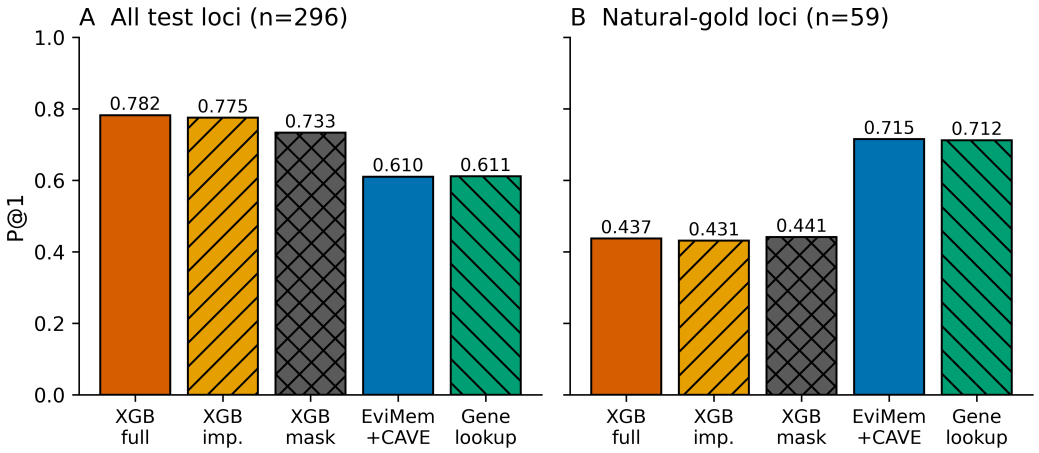


Figure 4: Force-added gold genes inflate supervised accuracy. (A) All 296 test loci. (B) The 59 natural-gold loci. XGB full: all evidence features; XGB imp.: median imputation; XGB mask: only indicators of which features are missing; Gene lookup: identity-only recurrence lookup.

5.4 Comparison with published and classical methods

Each comparator is evaluated on the test loci it covers, against the other methods on exactly those loci (Appendix Table 8). EviMem+CAVE exceeds cS2G (236 loci, cS2G P@1 0.432; difference −0.159, −0.224 to −0.095, expressed as comparator minus EviMem+CAVE), Open Targets L2G (211 loci, 0.460; −0.157), MAGMA (110 loci, 0.445; −0.209) and S-PrediXcan TWAS (82 loci, 0.427; −0.207); all intervals exclude zero. For PoPS (62 loci, 0.548) no difference is detected either against EviMem+CAVE (−0.081, −0.226 to +0.065) or against the gene lookup (−0.113, −0.242 to +0.016).

6 Discussion and limitations

Scope of the claim. EviMem+CAVE reuses verified knowledge; it does not improve prioritisation where no relevant knowledge has been accumulated. Novel gene–trait pairs, unseen traits and zero-

overlap loci show no detectable gain, and on them the agent's accuracy is that of the underlying LLM. The benefit therefore grows with the recurrence structure of the analyses a group performs, which is high for pleiotropic genes and repeated traits, and vanishes for genuinely new biology.

The fingerprint. The EF is a fixed, hand-engineered encoder; on this benchmark it varies on few dimensions, its anchors were encoded differently from its queries, and similarity-based short-circuits were rare and imprecise. Learned or ablated encoders, and inputs such as MR and allele frequency, are untested here.

Previously reported failures. The earlier version of this work reported zero accuracy on three atrial fibrillation and coronary artery disease loci. None was short-circuited; in each, the agent chose the candidate with the highest V2G score over a weakly supported gold gene, and for one locus the Qwen3-32B agent was correct in every seed. Two of the three loci are training loci in the frozen benchmark; the third, whose gold gene has no quantitative evidence, is missed by every evidence- and memory-based method and found only by classifiers that detect missing evidence (Appendix E; those runs predate the verification-store fix and are used only for this diagnosis).

Difficult cases. Allele frequencies are not available for benchmark candidates, so rare variants are not evaluated. Poorly annotated loci are evaluated only through evidence completeness, which is a different property from rarity (Appendix E).

Evaluation. Labels may share evidence with model inputs, the leak-free population is small (59 loci across 42 traits) and selects better-annotated loci, and conclusions are limited to the benchmark, backbones and settings tested. We recommend, but have not ourselves completed on independent data, evaluations that report recurrence decompositions and recall baselines alongside every method, report natural-gold results separately, and use labels derived independently of model inputs.

7 Conclusion

A stateless L2G agent repeats work that has already been verified. A verified memory lets it reuse that work: in our benchmark it short-circuits a third of within-ancestry loci and most cross-ancestry loci without an LLM call at high precision and raises top-1 accuracy over the memory-free agent. The gain is recall of verified gene–trait knowledge, and it should be measured as such, against recall baselines and on populations free of candidate-construction artefacts.

Reproducibility statement

All analyses were computed locally from audited per-locus agent runs (five seeds per backbone) without new LLM calls. The supplement contains the analysis scripts, the frozen benchmark definition, the numbers ledger from which every number in the paper is generated, integrity-check reports and the independent audit record.

AI use statement

We used generative AI tools (Anthropic Claude, via Claude Code) to assist with the analysis. Large language models (GPT-5.2 and Qwen3-32B) are also components of the evaluated system; that use is part of the method rather than assistance with this work. We take responsibility for the final content of this work.

References

Alvaro N Barbeira et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, 9:1825, 2018. doi: 10.1038/s41467-018-03621-1.

Nicholas Carlini et al. Extracting training data from large language models. In *USENIX Security*, 2021.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *RecSys*, pp. 39–46, 2010. doi: 10.1145/1864708.1864721.

Christiaan A de Leeuw, Joris M Mooij, Tom Heskes, and Danielle Posthuma. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11(4):e1004219, 2015. doi: 10.1371/journal.pcbi.1004219.

Steven Gazal et al. Combining snp-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics*, 54:827–836, 2022. doi: 10.1038/s41588-022-01087-y.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Maya Ghoussaini et al. Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49:D1311–D1320, 2021. doi: 10.1093/nar/gkaa840.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT*, 2018.

Qiao Jin et al. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024. doi: 10.1093/bioinformatics/btae075.

Subbarao Kambhampati et al. LLMs can't plan, but can help planning in LLM-modulo frameworks. In *ICML*, volume 235 of *PMLR*, pp. 22895–22907, 2024.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6:1–21, 2012. doi: 10.1145/2382577.2382579.

Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *EMNLP*, 2018.

Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *KDD*, pp. 1748–1757, 2020. doi: 10.1145/3394486.3403226.

Sebastian Lapuschkin et al. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019. doi: 10.1038/s41467-019-08987-4.

Chris Latimer et al. Hindsight is 20/20: building agent memory that retains, recalls, and reflects. *arXiv preprint arXiv:2512.12818*, 2025.

Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *ACL*, 2022.

Ellen Mountjoy et al. An open approach to systematically prioritize causal variants and genes at all published human gwas trait-associated loci. *Nature Genetics*, 53:1527–1533, 2021. doi: 10.1038/s41588-021-00945-5.

Matthew R Nelson et al. The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47:856–860, 2015. doi: 10.1038/ng.3314.

Charles Packer et al. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems. *arXiv preprint arXiv:1905.01395*, 2019.

Elle M Weeks et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nature Genetics*, 55:1267–1276, 2023. doi: 10.1038/s41588-023-01443-6.

Yiming Xiong et al. Learning to continually learn via meta-learning agentic memory designs. *arXiv preprint arXiv:2602.07755*, 2026.

APPENDIX

A INTEGRITY CHECKS PERFORMED BEFORE ANALYSIS

Before any new analysis we checked the audited runs and snapshot copies of the stores. (i) All canonical results of the previous version were recomputed exactly from source files. (ii) The verification store holds 432 VALIDATED claims, all from training gold genes, with 0 claims specific to test or cross-ancestry loci; the anchor stores contain no anchors written from test loci. (iii) The Qwen3-32B agent without memory returns the first-listed candidate in 99.7% of runs, so it is not used as a headline baseline. (iv) Fingerprints recomputed with the bootstrap encoding reproduce the stored anchors (1.000 agreement), confirming the encoding difference described in §3.

B EVIDENCE FINGERPRINT AND SHORT-CIRCUIT THRESHOLDS

Table 2 defines the fingerprint and Table 3 gives the adaptive short-circuit thresholds used by the implementation.

Ablation without an LLM. Table 4 reports test P@1 when each feature group is removed. Removing groups that are constant on this benchmark changes little. The k-NN classifier depends on the eQTL dimension because a zero value marks candidates without eQTL evidence, which include force-added gold genes; on the five varying dimensions alone, evidence-free candidates become zero vectors with undefined cosine similarity and the classifier collapses. A retrieval-level simulation of the EF path with the implemented thresholds (not a replay of the runs, in which the direct CAVE path is consulted first) short-circuits 37 loci at precision 0.568 (0.395–0.729) with the encoding as run and 86 loci at 0.686 (0.577–0.782) with a consistent encoding.

Table 2: The 16 Evidence Fingerprint dimensions (fixed, hand-engineered). "Missing" is the value used when the input is absent; "Varies here" indicates whether the dimension takes more than one value across benchmark test candidates.

#	Dimension	Group	Definition	Missing	Varies here
0	cir	priority_druggability	$\min(1,\ \mathrm{cir}/4)$	0	no
1	druggability	priority_druggability	$\min(1,\ \mathrm{druggability}/3)$	0	no
2	best_pp_h4	coloc	coloc PP.H4	0	yes
3	mr_pval_log	mr	$\min(1,\ -\log_{10} p_{\mathrm{MR}}/10)$	0	no
4	mr_estimate_sig	mr	$\sigma(\hat{\beta}_{\mathrm{MR}})$	1/2	no
5	tdoe_norm	priority_druggability	$(\mathrm{tdoe}+1)/2$	1/2	no
6	has_coloc	coloc	has colocalisation	0	yes
7	has_mr	mr	MR significant flag, or p_{MR} below 0.002	0	no
8	has_vep	coding_vep	coding consequence (VEP)	0	yes
9	has_v2g	v2g	V2G score above 3/10	0	yes
10	gwas_pval_log	gwas_signal	$\min(1,\ -\log_{10} p_{\mathrm{GWAS}}/20)$	0	no
11	idoe_norm	mr	MR direction (activate 1, inhibit 0)	1/2	no
12	n_eqtl	eqtl	$\min(n_{\mathrm{eQTL\ tissues}}, 10)/10$	0	yes
13	druggable_surface	priority_druggability	druggable surface	0	no
14	mr_precision	mr	$1-\min(1,\ 2\,\mathrm{se}_{\mathrm{MR}})$	1/2	no
15	gnomad_af_log	population_af	$\min(1,\ 10\,\mathrm{AF}_{\mathrm{gnomAD}})$	1/2	no

Table 3: Adaptive cosine threshold for an EF-path short-circuit, by the number of EF dimensions that depart from their defaults, as implemented (the schedule is not monotone).

Informative EF dimensions	Cosine threshold
2	0.93
3	0.98
4	0.97
5	0.96
6	0.95
7	0.94
8 or more	0.93
fewer than 2	no short-circuit

Table 4: EF feature-group ablation: test P@1 on all test loci for the three k-NN variants (no LLM).

EF configuration	Dims	k-NN classifier	Transfer (as run)	Transfer (consistent)
full EF (16 dims)	16	0.645	0.436	0.443
varying dims only	5	0.294	0.409	0.412
drop coloc	14	0.649	0.429	0.432
drop eqtl	15	0.250	0.436	0.436
drop coding_vep	15	0.642	0.432	0.439
drop v2g	15	0.574	0.439	0.446
drop mr	11	0.645	0.436	0.446
drop gwas_signal	15	0.645	0.436	0.443
drop priority_druggability	12	0.645	0.436	0.446
drop population_af	15	0.645	0.436	0.446
z-scored varying dims (unweighted baseline)	5	0.645	0.426	0.443

C POPULATIONS, SHORT-CIRCUIT PATHS AND ADDITIONAL METRICS

Table 5 lists every population used in the paper; public-gold rows refer to the full Open Targets curated gold standard and are not subsets of the frozen benchmark. Keying gene–trait pairs on EFO rather than trait name gives 99 recurrent loci, on which EviMem+CAVE (GPT-5.2) reaches 0.919 (novel pairs 0.455). On the 142 loci whose gold gene never appears in training, EviMem+CAVE reaches 0.461 against 0.470 without memory. Table 6 decomposes short-circuits by path, and Table 7 reports R@3, MRR and NDCG@5; method orderings are not identical across metrics (ranking by V2G and by L2G have equal P@1 but R@3 of 0.727 and 0.663).

Table 5: Every population used in the paper.

Population	n
Public OT curated gold standard: loci with natural candidates	1078
Public gold: gold gene naturally in OT candidate pool	66.1%
Public gold: held-out natural-candidate loci	226
NeurIPS-era panel (loci)	170
Frozen benchmark (loci, train+test)	844
Train	548
Test (all-296)	296
Test natural-gold (leak-free)	59
Test gold in candidate pool	296
Test novel-EFO (unseen trait)	80
Test seen-EFO	216
Test recurrent gene-trait pair	87
Test novel gene-trait pair	209
Test recurrent gene-EFO pair (sensitivity)	99
Test unseen gold gene	142
Test zero-overlap (unseen gene and trait)	51
Natural-gold and novel-EFO	12
Natural-gold and novel gene-trait pair	31
Cross-ancestry replication	88
Test traits (EFO) represented	174
Natural-gold traits (EFO) represented	42

Table 6: Hard short-circuits by path, pooled over five seeds (rows), with distinct loci and 95% Clopper–Pearson intervals on precision.

Panel	Backbone	SC path	Loci	Rows	Precision	CI lo	CI hi
Within-ancestry	GPT-5.2	direct CAVE	92	460	0.935	0.908	0.956
Within-ancestry	GPT-5.2	EF similarity	2	9	0.444	0.137	0.788
Within-ancestry	Qwen	direct CAVE	92	460	0.935	0.908	0.956
Within-ancestry	Qwen	EF similarity	2	9	0.556	0.212	0.863
Cross-ancestry	GPT-5.2	direct CAVE	69	345	0.986	0.967	0.995
Cross-ancestry	GPT-5.2	EF similarity	1	3	0.000	0.000	0.708
Cross-ancestry	Qwen	direct CAVE	69	345	0.986	0.967	0.995
Cross-ancestry	Qwen	EF similarity	1	5	0.000	0.000	0.522

Table 7: Ranking metrics on all test loci. Multiple gold genes: R@3 is the fraction of gold genes in the top three; MRR uses the first gold gene; NDCG@5 uses binary relevance with an ideal ranking over all gold genes.

Method	P@1	R@3	MRR	NDCG@5
EviMem+CAVE (GPT-5.2)	0.610	0.843	0.739	0.778
EviMem+CAVE (Qwen)	0.567	0.775	0.696	0.727
No-memory LLM (GPT-5.2)	0.493	0.801	0.659	0.709
No-memory LLM (Qwen)	0.395	0.707	0.579	0.626
Gene lookup (identity)	0.611	0.847	0.745	0.787
Gene lookup (+V2G tie-break)	0.618	0.879	0.760	0.806
Trait lookup (identity)	0.574	0.765	0.693	0.725
Trait lookup (+V2G tie-break)	0.574	0.810	0.709	0.747
XGBoost (full features)	0.782	0.926	0.864	0.882
XGBoost (median-imputed)	0.775	0.919	0.860	0.882
XGBoost (missingness only)	0.733	0.913	0.834	0.858
Rank by V2G	0.395	0.727	0.583	0.632
Rank by L2G	0.395	0.663	0.561	0.603

D COMPARATORS ON THEIR COVERED LOCI

cS2G: loci where cS2G makes a call within 2 kb. PoPS: loci whose trait maps to the public PoPS results with a score for at least one candidate (reproducing the published P@1 of 0.548 on 62 loci). Open Targets L2G: loci with an L2G score for at least one candidate. MAGMA and S-PrediXcan TWAS: loci covered by public European-ancestry summary statistics for the trait. Differences are comparator minus reference on identical loci, each with its own McNemar test; comparisons against the trait lookup and the Qwen3-32B agent are in the supplementary results.

Table 8: Comparators versus the gene lookup and EviMem+CAVE (GPT-5.2) on each comparator's covered loci.

Comparator	n	P@1	Reference	Ref P@1	Δ	lo	hi	McNemar p
cS2G	236	0.432	Gene lookup (identity)	0.593	−0.161	−0.233	−0.089	3.2×10^{-5}
cS2G	236	0.432	EviMem+CAVE (GPT-5.2)	0.592	−0.159	−0.224	−0.095	9.1×10^{-6}
PoPS	62	0.548	Gene lookup (identity)	0.661	−0.113	−0.242	0.016	0.143
PoPS	62	0.548	EviMem+CAVE (GPT-5.2)	0.629	−0.081	−0.226	0.065	0.383
OT L2G	211	0.460	Gene lookup (identity)	0.607	−0.147	−0.204	−0.090	7.8×10^{-7}
OT L2G	211	0.460	EviMem+CAVE (GPT-5.2)	0.617	−0.157	−0.209	−0.109	2.1×10^{-9}
MAGMA	110	0.445	Gene lookup (identity)	0.645	−0.200	−0.309	−0.100	4.7×10^{-4}
MAGMA	110	0.445	EviMem+CAVE (GPT-5.2)	0.655	−0.209	−0.309	−0.109	1.9×10^{-4}
TWAS (S-PrediXcan)	82	0.427	Gene lookup (identity)	0.610	−0.183	−0.293	−0.073	0.004
TWAS (S-PrediXcan)	82	0.427	EviMem+CAVE (GPT-5.2)	0.634	−0.207	−0.317	−0.110	4.9×10^{-4}

E PREVIOUSLY REPORTED AF/CAD FAILURES AND EVIDENCE STRATA

The earlier version reported P@1 of zero on three atrial fibrillation (AF) and coronary artery disease (CAD) test loci (Table 9; original runs, five seeds per backbone; these runs predate the train-only verification store, whose fix did not affect these loci). No run short-circuited. At the AF locus (gold GPD2), GPT-5.2 chose NR4A2, the highest-V2G candidate, whereas Qwen3-32B chose GPD2 in every seed. At the CAD locus with gold NDUFA12, both chose FGD6, again the highest-V2G candidate. At the CAD locus with gold TUBB, in the MHC region, the gold gene has no distance, eQTL, colocalisation or L2G evidence; in the frozen benchmark this is a test locus that every evidence- and memory-based method misses and only the XGBoost variants, including the missingness-only classifier, rank first. The other two loci are training loci in the frozen benchmark. Allele frequencies are absent from benchmark candidates, so strata by variant frequency cannot be formed. A pre-specified count over nine evidence fields was uninformative because presence flags are never missing, so Table 10 uses a post hoc count over the four quantitative fields that can be missing. On the 108 loci whose gold gene has none of them, ranking by V2G scores 0.000, EviMem+CAVE 0.263 and XGBoost 1.000; strata with fewer than ten loci should not be interpreted.

Table 9: AF/CAD loci in the original runs (P@1 = mean over five seeds; SC frac = fraction of seeds short-circuited).

Trait	Gold	Backbone	P@1	Predicted top-1	SC frac	Frozen split
coronary artery disease	NDUFA12	GPT-5.2	0.000	FGD6	0.000	train
coronary artery disease	NDUFA12	Qwen	0.000	FGD6	0.000	train
atrial fibrillation	GPD2	GPT-5.2	0.000	NR4A2	0.000	train
atrial fibrillation	GPD2	Qwen	1.000	GPD2	0.000	train
coronary artery disease	TUBB	GPT-5.2	0.000	IER3—SFTA2	0.000	test
coronary artery disease	TUBB	Qwen	0.000	SFTA2	0.000	test

Table 10: P@1 by number of quantitative evidence fields present for the gold candidate (post hoc).

Method	0 fields (n=108)	1 fields (n=8)	2 fields (n=165)	3 fields (n=14)	4 fields (n=1)
EviMem+CAVE (GPT-5.2)	0.263	1.000	0.801	0.786	1.000
EviMem+CAVE (Qwen)	0.263	1.000	0.724	0.786	1.000
XGBoost (missingness only)	1.000	0.875	0.564	0.571	1.000
XGBoost (full features)	1.000	0.825	0.647	0.657	1.000
Gene lookup (identity)	0.333	1.000	0.764	0.714	1.000
No-memory LLM (GPT-5.2)	0.043	1.000	0.731	0.829	1.000
Rank by V2G	0.000	0.875	0.612	0.571	1.000
Trait lookup (identity)	0.287	1.000	0.727	0.714	1.000

F THE FORCE-ADDED ARTEFACT BEYOND GENOMICS

In a synthetic rank-a-candidate task in which no method can reason from evidence, a recurrence lookup's P@1 rises from 0.081 to 0.815 as the recurrence rate increases while an evidence model stays near chance (0.123 to 0.122), and a missingness-only classifier rises from 0.125 to 0.998 as the force-add rate increases. In MovieLens-100k (943 users), when the held-out item is force-added to a retrieval-generated candidate set (48.7% of users), a detector using only insertion signatures identifies it with P@1 0.196 on force-added cases and 0.006 on naturally retrieved ones (chance 0.010).