EVIDENCEWEAVE: COMPILING TIME-BOUNDED EVIDENCE INTO INSPECTABLE MECHANISMS AND DISCRIMINATIVE TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

One omitted contradiction can reverse the rescue experiment chosen from an otherwise well-formed mechanism graph. EVIDENCEWEAVE keeps that failure visible by compiling the path from each causal edge to its source span, accepted or uncertain state, unresolved prerequisite, rival mechanism, and outcome-separated test. The accompanying benchmark contains 312 CRISPR perturbation cases and 4,210 expert-identified support or contradiction opportunities, including 715 contradictions, from literature published before January 2022. Chronology is enforced before retrieval ranking: fixed-volume future-document injection yields 0.089 leakage and 8.3% support inflation at 20%, whereas the date-filtered pipeline has zero detected leakage. Against a typed Llama-3.1-70B control matched on training objects, retrieval, schema, and decoding, EVIDENCEWEAVE lowers unsupported-edge rate from 0.246 to 0.142 while raising contradiction coverage from 0.302 to 0.478 and evidence-opportunity recall from 0.601 to 0.672; 299 additional annotated opportunities reach reviewers with similar accepted-edge mass. A separately blinded 72-case annotation panel preserves the edge ordering, while a 48-case human target yields written-outcome-separation AUC of 0.706 versus 0.584 for the control. Counterbalanced review also falls from 11.1 to 8.7 minutes, with faster review in 49 of 60 pairs. The resulting model-selection criterion rewards evidence exposure and adverse-evidence recovery only when asserted edges remain supportable and the linked object reduces expert reconstruction.

1 INTRODUCTION

The decision-bearing unit in a mechanism hypothesis is a path, not a sentence. In the schematic knockout of Figure 5, the claim that X represses STAT1 targets is connected to an accepted support span, a conflicting rescue result, an indirect candidate, and the assumption that the knockout does not trigger generic stress. The proposed rescue becomes informative only after the graph records two outcomes: collapse of the interferon signature under the main mechanism and persistence under a DNA-damage rival. Removing any link leaves a fluent hypothesis, but changes what a molecular biologist can verify before selecting the experiment.

Existing evaluation units cut across this path. Hypothesis benchmarks judge the final proposal, evidence-retrieval benchmarks recover relevant text, fact checkers score supplied claims, and experiment benchmarks assess a proposed action. A typed mechanism graph can pass all field-level validity checks while still asserting an unsupported edge, omitting a known contradiction, or pairing a test with outcomes that do not separate its rival. The missing technical object is the linked state that lets a reviewer locate such failures without reconstructing the generator's reasoning.

Evidence-obligation compilation constructs that state. The compiler receives observations and a retrieval corpus bounded at time t, then emits claims whose evidence edges carry source-exact spans, stance, confidence, and an accepted or uncertain state. Assumption nodes expose unresolved prerequisites; rival nodes anchor alternatives; experiment nodes record the different outcomes predicted by the main mechanism and each rival. The resulting obligation graph supports local inspection from an edge to the experiment it licenses.

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.

This graph requires coupled measurements. Unsupported-edge rate (UER) evaluates assertions among accepted edges. Evidence-opportunity recall (EOR) measures how many expert-identified support or contradiction opportunities are shown as accepted or uncertain under a shared ledger capacity. Contradiction coverage (CC) isolates adverse evidence. Whole-object review measures the corrections and time required to inspect the graph, while a separately adjudicated target measures whether ranked tests actually separate written outcomes. A system is preferred only when assertion quality, evidence exposure, and contradiction coverage improve together at comparable accepted-edge mass.

The benchmark instantiates the graph on 312 CRISPR perturbation cases and 4,210 expert-enumerated opportunities from PubMed and PMC records published before January 2022. EVIDENCEWEAVE enforces chronology before retrieval ranking, generates the typed ledger, gates each claim–span edge into accepted or uncertain state, and ranks experiments against named rivals. Its central control shares the Llama-3.1-70B family, 847 training objects, 3,200 augmentation examples, retrieval pool and budget, output schema, and decoding budget, but omits the verifier transition and rival-conditioned compilation.

The matched result moves all three automatic axes in the preferred direction: UER falls from 0.246 to 0.142, CC rises from 0.302 to 0.478, and EOR rises from 0.601 to 0.672. A new annotation panel preserves the edge ordering, human adjudicators preserve the written-outcome ordering, and paired expert review falls from 11.1 to 8.7 minutes. The automatic frontier is aligned with the observed reduction in reconstruction time under the paired-review protocol.

Contributions. The principal contribution is the edge-to-experiment obligation graph and its joint scoring rule: assertions, annotated evidence exposure, counterevidence, and written-outcome separation become observable without collapsing uncertainty into correctness. EVIDENCEWEAVE makes this object executable through an ordered chronology boundary, constrained records, verifier state transitions, and rival-conditioned outcome fields. The matched comparison establishes the practical consequence of those linked semantics; the remaining analyses test whether the ordering survives changes in annotators, biological subfield, decoding seed, representation, and evidence chronology.

2 RELATED WORK

Evidence retrieval and scientific synthesis. EvidenceBench directly evaluates hypothesis-conditioned sentence retrieval from biomedical papers (Wang et al., 2025). MedRAGChecker scores support and contradiction for atomic biomedical claims (Ji et al., 2026), while PaperQA2 and ContraCrow evaluate scientific synthesis and literature-scale contradiction discovery (Skarlinski et al., 2024). These systems establish strong local evidence operations. Our object links those operations to a generated mechanism: the scorer asks which source-exact opportunities were surfaced at each causal edge, whether the edge was asserted or deferred, and what downstream test that evidential state supports.

Hypotheses and falsifiable proposals. BioDisco and DBench-Bio evaluate biomedical discovery under temporal evidence (Ke et al., 2025; Yang et al., 2026); HypER generates literature-grounded hypotheses through provenance-bearing reasoning chains (Vasu et al., 2025); HypoArena ranks prospective hypotheses (Zhong et al., 2026); and ProjectionBench discloses scientific information progressively (Lew et al., 2026). Concurrent Lit2Test requires a proposal-level contract with supporting and falsifying outcomes (Wang et al., 2026). The closest overlap is the insistence that a proposal encode how it can be rejected. Evidence-obligation evaluation adds edge-local source states and annotated omission denominators, then measures whether the linked representation changes expert inspection time.

From proposal to execution. POPPER executes sequential falsification under statistical control (Huang et al., 2025), SciGym evaluates experiment-selection trajectories in systems-biology simulators (Duan et al., 2025), and the Hypothesis Evolution Protocol externalizes hypotheses, tests, attached evidence, and belief-state transitions in a longitudinal registry (Takahara & Mizoguchi, 2026). Our benchmark occupies the pre-execution boundary between literature synthesis and action:

the evaluated unit contains the evidence needed to challenge a mechanism and the rival-specific outcomes needed to reject a non-discriminative experiment.

Retrieval, verification, and extraction supply the compiler's local operations (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020; Asai et al., 2024; Edge et al., 2024). MedGraphRAG organizes evidence for medical answers (Wu et al., 2025); scientific fact-checking evaluates claims and rationales (Wadden et al., 2020; Wright et al., 2022; DeYoung et al., 2020; Min et al., 2023); and biomedical systems recover relations and spans (Gyori et al., 2017; Lee et al., 2020; Beltagy et al., 2019; Gu et al., 2021; Wei et al., 2019). The novelty here is operational composition: chronology acts before retrieval, verification changes an edge's assertion state without hiding it, and rival-specific outcomes connect the resulting graph to a test. Table 10 compares the nearest evaluation objects directly.

3 Benchmark Objects and Obligation Scoring

Each instance contains an observation bundle $\mathcal{O} = \{o_1, \ldots, o_m\}$ and a retrieval corpus $\mathcal{C}^{<t}$ whose records precede cutoff t. A system emits $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ with claim nodes for the main mechanism and dependencies; evidence edges with stance, source identifier, span, confidence, and state; assumption nodes for unresolved prerequisites; rival nodes; and experiment nodes pairing the predicted outcome under the main mechanism with the outcome under a named rival. Claims create provenance obligations, evidence states mark assertions and deferrals, assumptions expose conditions requiring judgment, and paired outcomes identify tests whose observations separate mechanisms.

Evidence edges produce the three automatic full-benchmark metrics. Assumptions remain visible during full-object expert review, whose corrections, review time, and final confidence summarize the representation as a whole. Experiment obligations are assessed against a separately adjudicated pairwise target. This measurement map keeps one linked object while assigning each quantitative claim an explicit denominator: accepted edges for assertion quality, annotated opportunities for exposure, and adjudicated outcome pairs for test separation.

Each emitted evidence edge has a system state and an orthogonal expert label. An *accepted* edge asserts that its span entails or contradicts the linked claim; an *uncertain* edge presents a below-threshold candidate for inspection without making that assertion. Expert adjudication labels accepted edges as supported or *unsupported*. An annotated opportunity is *surfaced* when an accepted or uncertain edge matches it and *omitted* from the system object when no emitted edge does. The case specification fixes a system-independent capacity K_i before any output is scored; accepted and uncertain edges both consume one slot, blank slots receive no credit, and duplicates collapse under normalized head, relation, tail, stance, and source identifier. A partial span matches when endpoints and stance agree, the source record is identical, and token intersection-over-union is at least 0.5.

For case i, let A_i be unique accepted edges, $B_i \subseteq A_i$ the unsupported subset, G_i the annotated support-or-contradiction opportunities, G_i^- its contradiction subset, and S_i the opportunities matched by accepted or uncertain edges. The three primary metrics are

$$\mathrm{UER} = \frac{\sum_i |B_i|}{\sum_i |A_i|}, \qquad \mathrm{CC} = \frac{\sum_i |S_i \cap G_i^-|}{\sum_i |G_i^-|}, \qquad \mathrm{EOR} = \frac{\sum_i |S_i|}{\sum_i |G_i|}. \tag{1}$$

UER measures accepted-edge precision loss; EOR measures recovery of the frozen expert opportunity set; and CC isolates its adverse-evidence subset. Credit for an uncertain match is credit for exposure, never a correctness judgment. A system with no accepted edges has undefined UER. The shared K_i prevents unbounded candidate lists, while the accepted-edge denominator reveals whether a lower UER accompanies fewer assertions. The benchmark therefore uses a joint preference rule: lower UER must coincide with higher EOR and CC at comparable accepted-edge mass. EOR and CC describe the annotated reference population rather than claiming an exhaustive census of all literature.

All full-benchmark intervals use 10,000 case-clustered bootstrap resamples. System differences use paired case resampling; timing inference clusters repeated judgments by case and expert. The four biological subfields are resampled within stratum. Five fixed decoding seeds test sensitivity with retrieval held constant. Captions state the evaluated population and uncertainty unit.

4 EVIDENCEWEAVE AS AN OBLIGATION COMPILER

EVIDENCEWEAVE realizes the graph as an ordered five-stage compilation. First, normalization binds genes, phenotypes, directions, and significance values to claim nodes. Second, hard date filtering restricts candidates to $\mathcal{C}^{<t}$ before a SciBERT bi-encoder selects its top 20 source-preserving passages. Third, a LoRA-tuned Llama-3.1-70B (Grattafiori et al., 2024) emits ledger records under constrained decoding: claim references must resolve, evidence records must contain source, span, stance, and confidence fields, and every experiment must name its main and rival outcomes. Fourth, PubMedBERT assigns $q_{\text{ver}}(e)$ to every claim–span pair. Finally, the test compiler ranks $\tau = (a, y_{\text{main}}, y_{\text{rival}})$, where a is the intervention and the two y fields are its predicted outcomes, by a fixed scalar divergence d whose order increases with outcome separation.

For an edge e, the decisive state transition is

$$z(e) = \begin{cases} \text{ACCEPTED}, & q_{\text{ver}}(e) \geq 0.7, \\ \text{UNCERTAIN}, & q_{\text{ver}}(e) < 0.7, \end{cases}$$

and the experiment queue is ordered by decreasing d. The 0.7 operating point is fixed for all benchmark systems that use the verifier. The transition order enforces four invariants: accepted edges retain retrievable pre-cutoff provenance; low-confidence candidates remain visible but non-asserted; ranked tests explicitly compare outcomes for a main mechanism and rival; and corpus dates act before retrieval scores can influence generation. These are state semantics rather than field names: post-hoc citation filtering cannot undo information already consumed by generation, and a schema alone cannot decide whether an edge is asserted or deferred.

The primary control preserves the Llama-3.1-70B family, the same 847-object LoRA set and 3,200 augmentation examples, retrieval pool and top-20 budget, decoding budget, and output schema. It bypasses verifier state transitions and rival-conditioned test compilation while retaining typed output. The final two rows of Table 1 therefore identify the joint package effect of these transitions. The single-component interventions in Table 5 are diagnostic: verifier removal bypasses $z(e)$; rival removal excludes rival-conditioned records, outcomes, and test ranking from the end-to-end path; schema removal relaxes typed serialization; and constrained-decoding removal retains the requested fields without grammar enforcement. Their downstream metric changes are descriptive end-to-end associations rather than separately identified module effects.

Training used PyTorch 2.3.1 and CUDA 12.4 for 32 hours on 8×A100 80GB GPUs. A provenance comparison finds 0 shared objects between the 847 training objects and 312 evaluation cases and 0 evaluation source-paper overlaps among the 3,200 augmentation examples. The 60-case timing and 48-case falsification subsets are deterministic proportional subfield samples obtained by sorting stable case hashes, making their membership independent of system output.

5 EXPERIMENTS

5.1 FROM PERTURBATION SCREENS TO SCORED OBLIGATION GRAPHS

The benchmark begins with 312 molecular-mechanism cases derived from eight CRISPR perturbation screens (Dixit et al., 2016; Adamson et al., 2016; Jaitin et al., 2016; Datlinger et al., 2017; Norman et al., 2019; Shifrut et al., 2018; Frangieh et al., 2021; Replogle et al., 2022). Each observation bundle fixes one knockout and 5–20 differentially expressed genes. Reference construction then searches only PubMed abstracts and PMC full text published before January 2022. Across 480 expert hours, three molecular biologists recorded claim-linked support and contradiction spans, assumptions, rivals, and preferred tests; a separate edge-adjudication pass normalized endpoints and resolved stance labels. Duplicate collapse yields 4,210 scored opportunities, including 715 contradictions. This is the frozen reference population for EOR and CC. A second cohort later constructs a panel-specific opportunity set on 72 cases with original labels hidden, directly testing whether the principal edge ordering depends on the first reference set.

We compare direct RAG (GPT-4o) (OpenAI, 2024), GraphRAG, an extraction-then-planning pipeline, Schema-Constrained GPT-4o, the matched typed Llama-3.1-70B control, and EVIDENCEWEAVE. The first three systems produce free-form mechanisms that are mapped once into the common obligation schema; the schema-constrained system emits those fields directly. A blinded audit over

Table 1: Evidence-obligation results on all 312 cases. Brackets are 95% intervals from 10,000 case-clustered bootstrap resamples. UER uses asserted accepted edges; CC uses 715 contradictions; EOR uses 4,210 evidence opportunities. The upper rows position broader systems after common-schema mapping; the final pair is the primary controlled contrast and shares model family, training, retrieval, decoding budget, and schema.

Method	UER ↓	CC ↑	EOR ↑
RAG (GPT-4o)	0.342 [0.301,0.384]	0.231 [0.194,0.271]	0.529 [0.493,0.565]
GraphRAG	0.287 [0.248,0.328]	0.312 [0.274,0.352]	0.557 [0.521,0.592]
Extraction+Planner	0.298 [0.258,0.339]	0.267 [0.229,0.307]	0.548 [0.512,0.583]
Schema-Constrained GPT-4o	0.261 [0.224,0.301]	0.284 [0.247,0.323]	0.584 [0.548,0.620]
Matched Typed Llama-3.1-70B	0.246 [0.212,0.283]	0.302 [0.267,0.339]	0.601 [0.564,0.638]
EVIDENCEWEAVE	0.142 [0.118,0.168]	0.478 [0.438,0.519]	0.672 [0.636,0.706]

72 held-out cases and 216 outputs gives the free-form conversion 0.882 macro F1; adjudicated corrections shift every primary metric by at most 0.017 and preserve the ordering. The broader four rows establish cross-family position under one evaluator. Only the final typed pair is used to identify the obligation-compilation package because it holds model family, training, retrieval, decoding budget, and schema fixed.

5.2 RQ1: IS SCHEMA COMPLIANCE SUFFICIENT?

No. All three typed rows in Table 1 satisfy the field contract, yet EVIDENCEWEAVE is the only system in the preferred region on UER, CC, and EOR. Schema-Constrained GPT-4o provides a cross-family check that valid fields alone do not recover the same annotated evidence. The matched typed pair then isolates the comparison from model family, training, retrieval, decoding budget, and schema. What changes is the behavior attached to each record—assert, defer, retain adverse evidence, and rank a rival-separating test.

5.3 RQ2: WHAT IS THE MATCHED PACKAGE EFFECT?

Relative to Matched Typed Llama-3.1-70B, EVIDENCEWEAVE changes UER by -0.104 [-0.126, -0.082] and CC by +0.176 [+0.142, +0.210]. Raw decisions move from 597/2427 to 337/2373 unsupported accepted edges; surfaced contradictions move from 216/715 to 342/715. Accepted-edge mass differs by 2.2%, while unsupported accepted edges fall by 43.6%. This controlled contrast identifies the compiler package. Descriptive component interventions localize the largest UER degradation to verifier removal and the largest CC and test-separation degradation to rival removal (Table 5).

Human adjudication of the fixed 48-case target preserves the RAG, matched-control, and EVIDENCEWEAVE ordering in written-outcome separation (Appendix Table 7), providing a second endpoint for the full compiler package.

5.4 RQ3: IS LOWER UER PURCHASED BY WITHHOLDING EVIDENCE?

The joint frontier rejects wholesale withholding at the benchmark's fixed ledger capacity. EOR rises from 0.601 to 0.672, a paired gain of +0.071 [+0.050, +0.092], while UER falls. The systems surface 2530 and 2829 of 4210 annotated opportunities, respectively; Schema-Constrained GPT-4o surfaces 2458. Thus 299 additional reference opportunities reach the reviewer despite 54 fewer accepted

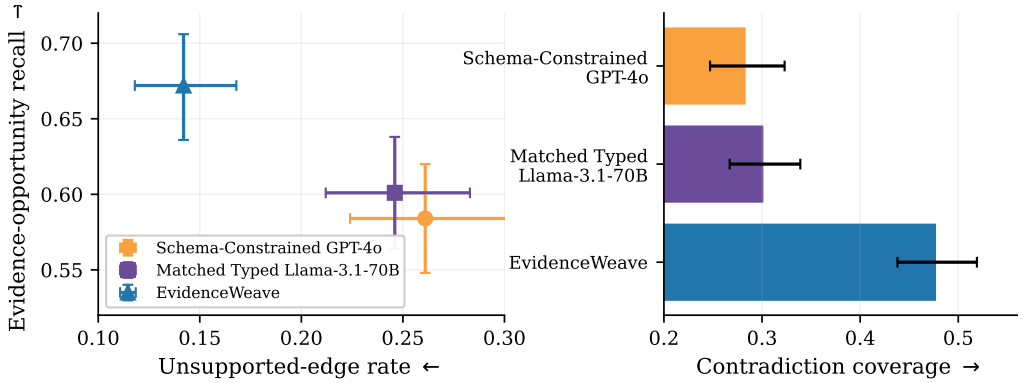


Figure 1: Model-selection frontier for the three typed systems on 312 cases. Left: unsupported-edge rate versus evidence-opportunity recall, where upper-left is preferable; both whiskers are case-bootstrap intervals. Right: contradiction coverage. The matched pair changes compiler transitions while preserving model family, training data, retrieval, decoding budget, and schema.

assertions. The gain is evidence exposure through accepted or explicitly uncertain matches, evaluated separately from correctness among accepted edges.

5.5 RQ4: Does a new annotation panel preserve the edge ordering?

A separate expert cohort re-annotated a fixed proportional panel of 72 cases using the frozen corpus and common rubric, with system identities and original labels hidden. It recovers the matched-control versus EvidenceWeave difference in UER (0.247 versus 0.158) and case-macro CC (0.309 versus 0.452). Cross-panel edge-label agreement is 0.70 [0.65, 0.75], with favorable changes in every biological subfield (Appendix Table 9). The panel's CC is averaged over paired case-level coverage because re-annotation creates a panel-specific opportunity set; replication evidence is the paired direction and interval, while full-benchmark CC remains the pooled estimand in Equation 1.

5.6 RQ5: Does the ledger reduce expert reconstruction?

The counterbalanced 60-case study compares the matched typed control directly with EvidenceWeave. Three experts each reviewed 20 paired cases with blinded labels, randomized within-batch order, and balanced system order. Mean review time falls from 11.1 to 8.7 minutes; the case-clustered paired change is -2.4 minutes [-3.1, -1.7], with EvidenceWeave faster on 49/60 cases. Corrections per object fall from 2.8 to 1.6, unsupported-edge discoveries during review from 1.9 to 0.8, and final confidence rises from 3.4 to 4.1 on a five-point scale (Appendix A.1). The same representation that moves the automatic frontier therefore reduces reconstruction work in paired review.

Table 2: Expert-stratified counterbalanced audit study. Each row contains the same 20 paired cases under both systems. Means and SDs are minutes per case.

Expert	Matched typed mean	Matched typed SD	EvidenceWeave mean	EvidenceWeave SD
1	10.9	2.8	8.2	2.0
2	11.7	3.3	9.1	2.5
3	10.8	2.9	8.8	2.3
Pooled	11.1	3.0	8.7	2.3

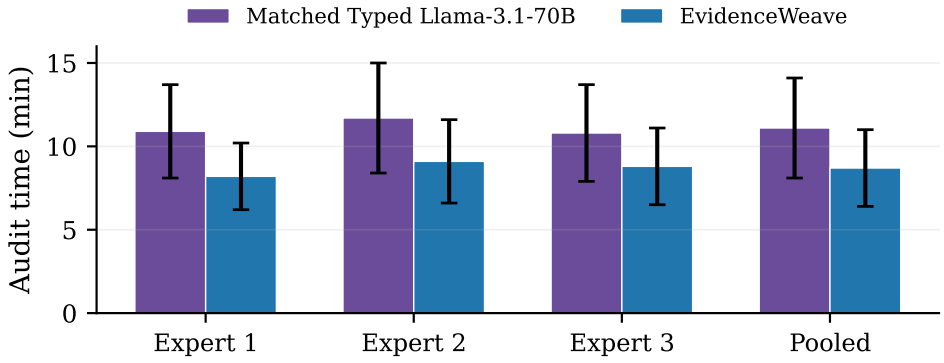


Figure 2: Counterbalanced time to discharge the audit obligations in Matched Typed Llama-3.1-70B and EVIDENCEWEAVE objects. Each expert reviews the same 20 cases under both systems; bars show means and whiskers show SD. The pooled effect uses paired case- and expert-aware inference over all 60 cases.

5.7 WHERE CHRONOLOGY ENTERS THE COMPILATION PATH

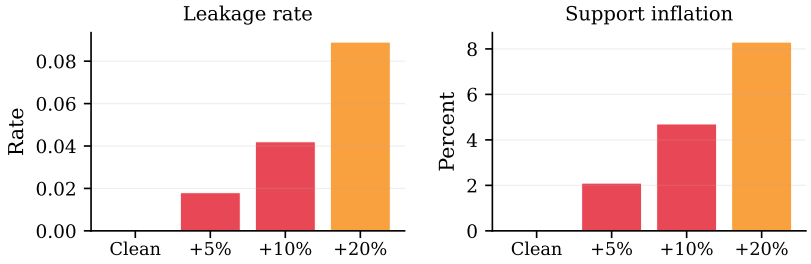


Figure 3: Retrieval-chronology stress test on all 312 cases. Future-document injection of 5%, 10%, and 20% yields leakage rates 0.018, 0.042, and 0.089 and support inflation of 2.1%, 4.7%, and 8.3%; the unmodified pre-January-2022 retrieval pipeline yields zero detected leakage.

Automated PubMed-ID and PMC-deposition scans find zero post-cutoff records in the date-filtered retrieval pool. Controlled injection produces monotone detector response and edge-support inflation (Figure 3), making retrieval-corpus chronology an experimental variable rather than a metadata check. Pre-retrieval filtering is therefore a compiler boundary: once a future passage influences generation, removing its citation afterward cannot restore the earlier information state.

At each intervention level, post-cutoff papers are sampled within biological subfield and retrieval-score decile and replace an equal number of pre-cutoff candidates, holding retrieval-pool size and generation budget fixed. Duplicate PubMed identifiers collapse before sampling. Electronic PubMed publication date defines eligibility, with PMC deposition date used as a consistency check. Leakage rate is the fraction of retrieved evidence edges attached to a post-cutoff record; support inflation is the relative increase in expert-accepted support edges over the clean condition. The injected sets and decoding seed are fixed across all 312 paired cases, so the curve isolates retrieval-corpus chronology from retrieval volume and generation stochasticity.

Appendix Table 4 maps every obligation breach to its protected state, enforcement boundary, strongest probe, measured property, and expert consequence.

6 DECISION IMPLICATIONS

The matched pair changes system selection for a reason that schema checks miss. With backbone, training objects, retrieval, decoding, and fields held fixed, EVIDENCEWEAVE accepts nearly the

same number of edges but produces 260 fewer unsupported accepted edges, surfaces 126 additional contradictions, and exposes 299 additional annotated opportunities. The independently annotated edge set, human test target, and paired review study then recover the same preferred system through different endpoints.

This yields an operational decision rule. Compare systems on paired cases under one ledger capacity; require lower UER together with higher EOR and CC at comparable accepted-edge mass; then use review time and adjudicated outcome separation to test whether the frontier matters to experts. The matched pair clears each criterion, with paired intervals excluding equality on all three automatic metrics. Cross-family rows remain positioning evidence, while the matched pair carries the package claim.

The component pattern explains why the graph matters. Verification removal produces the largest UER increase (0.142 to 0.238); rival removal produces the largest CC and outcome-separation decreases (0.478 to 0.312 and 0.724 to 0.582); and removing typed structure raises review time from 8.7 to 11.8 minutes. These are descriptive localizations within a package identified by the matched contrast, not separate causal claims for every module.

The benchmark can be instantiated in another domain through an interface rather than a domain-specific score: claims must have inspectable endpoints, the retrieval corpus must have a meaningful temporal boundary, experts must be able to enumerate support and contradiction opportunities, and actions must admit rival-specific outcomes. Domain experts determine ledger content; the distinction among assertion quality, evidence exposure, and test separation determines evaluation.

Interpretation envelope. The evidence identifies a benchmark-local compiler-package effect in 312 cases from eight CRISPR screens. UER measures claim–span support, EOR and CC recover curated expert-enumerated opportunities, and test AUC measures separation between written outcome pairs; these endpoints evaluate the inspectable representation rather than biological correctness of the entire mechanism, and assumptions remain visible in whole-object review without a separate automatic score. The 72-, 60-, and 48-case studies use new labels or review procedures but remain subsets of this population; paired review measures reconstruction burden rather than downstream decision accuracy; and chronology constrains retrieved provenance rather than knowledge latent in model parameters. Two annotation cohorts, four biological subfields, five decoding seeds, paired human review, and separately adjudicated tests provide complementary checks within that operating regime and define the evidence required for a transfer claim.

7 Conclusion

EvidenceWeave connects the evidence behind a causal edge to the experiment that can distinguish its mechanism from a rival. On 312 perturbation cases, its matched compiler lowers unsupported accepted edges while surfacing more annotated evidence and contradictions; a new annotation panel preserves that ordering, human-adjudicated written-outcome-separation AUC rises from 0.584 to 0.706, and paired review saves 2.4 minutes. Fixed-volume future-document injection separately shows how retrieval chronology changes apparent support. The resulting design principle is concrete: preserve source and uncertainty state at every edge, evaluate assertion quality and evidence exposure together, and require proposed tests to encode rival-specific outcomes.

AI use statement

Generative AI tools assisted in drafting and revising portions of this manuscript, including wording, presentation, consistency checking, citation verification, and formatting. The final manuscript remains the authors' responsibility.

References

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein

response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048. URL https://doi.org/10.1016/j.cell.2016.11.048.

Akari Asai, Zeqi Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, 2019. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donát Alpár, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017. doi: 10.1038/nmeth.4177. URL https://doi.org/10.1038/nmeth.4177.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038. URL https://doi.org/10.1016/j.cell.2016.11.038.

Haonan Duan, Stephen Zhewen Lu, Caitlin F. Harrigan, Nishkrit Desai, Jiarui Lu, Michał Koziarski, Leonardo Cotta, and Chris J. Maddison. Measuring scientific capabilities of language models with a systems biology dry lab. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3733. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a23760ba51036d530a2656c3835f826c-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi: 10.48550/arXiv.2404.16130. URL https://arxiv.org/abs/2404.16130.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, Maryann Zhao, Casey R Ager, Meri Rogava, Lila Hovey, Asaf Rotem, Chantale Bernatchez, Kai W Wucherpfennig, Bruce E Johnson, Orit Rozenblatt-Rosen, Dirk Schadendorf, Aviv Regev, and Benjamin Izar. Multimodal pooled perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3):332–341, 2021. doi: 10.1038/s41588-021-00779-1. URL https://doi.org/10.1038/s41588-021-00779-1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929–3938, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11):954, 2017. doi: 10.15252/msb.20177651. URL https://doi.org/10.15252/msb.20177651.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437, 2025. URL https://proceedings.mlr.press/v267/huang25n.html.

Diego Adhemar Jaitin, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 167(7):1883–1896.e15, 2016. doi: 10.1016/j.cell.2016.11.039. URL https://doi.org/10.1016/j.cell.2016.11.039.

Yuelyu Ji, Min Gu Kwak, Hang Zhang, Xizhi Wu, Chenyu Li, and Yanshan Wang. MedRAGChecker: Claim-level verification for biomedical retrieval-augmented generation. *arXiv preprint arXiv:2601.06519*, 2026. URL https://arxiv.org/abs/2601.06519.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Yujing Ke, Kevin George, Kathan Pandya, David Blumenthal, Maximilian Sprang, Gerrit Großmann, Sebastian Vollmer, and David Antony Selby. BioDisco: Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation. *arXiv preprint arXiv:2508.01285*, 2025. URL https://arxiv.org/abs/2508.01285.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Andrew J. Lew, Yuan Cao, and Markus J. Buehler. ProjectionBench: Evaluating scientific hypothesis generation in LLMs under progressive information disclosure. *arXiv preprint arXiv:2605.30284*, 2026. URL https://arxiv.org/abs/2605.30284.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438. URL https://doi.org/10.1126/science.aax4438.

OpenAI. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L Bonnar, Marco Jost, Thomas M Norman, and Jonathan S Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013. URL https://doi.org/10.1016/j.cell.2022.05.013.

Eric Shifrut, Julia Carnevale, Victoria Tobin, Theodore L Roth, Jonathan M Woo, Christina T Bui, P. Jonathan Li, Morgan E Diolaiti, Alan Ashworth, and Alexander Marson. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971.e15, 2018. doi: 10.1016/j.cell.2018.10.024. URL https://doi.org/10.1016/j.cell.2018.10.024.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024. doi: 10.48550/arXiv.2409.13740. URL https://arxiv.org/abs/2409.13740.

Izumi Takahara and Teruyasu Mizoguchi. Toward auditable AI scientists: A hypothesis evolution protocol for LLM agents. *arXiv preprint arXiv:2607.09195*, 2026. URL https://arxiv.org/abs/2607.09195.

Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra, Cristina Sarasua, Peter Clark, and Abraham Bernstein. HypER: Literature-grounded hypothesis generation and distillation with provenance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25413–25438, 2025. doi: 10.18653/v1/2025.emnlp-main.1292. URL https://aclanthology.org/2025.emnlp-main.1292/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Jianyou Wang, Weili Cao, Kaicheng Wang, Xiaoyue Wang, Ashish Dalvi, Gino Prasad, Qishan Liang, Hsuan-lin Her, Ming Wang, Qin Yang, Gene W. Yeo, David E. Neal, Maxim Khan, Christopher D. Rosin, Ramamohan Paturi, and Leon Bergen. EvidenceBench: A benchmark for extracting evidence from biomedical papers. In *Proceedings of the 2025 Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lEQnUI5lEA.

Ziyue Wang, Aomufei Yuan, Yiran Yao, Linli Yao, Hongyao Zuo, Ziwen Gong, Yuanxin Liu, Shicheng Li, Yishuo Cai, Tong Yang, Xu Sun, Xiaohui Li, and Haoli Bai. What proves you wrong: Benchmarking language models on falsifiable research ideation. *arXiv preprint arXiv:2608.22948*, 2026. URL https://arxiv.org/abs/2608.22948.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593, 2019. doi: 10.1093/nar/gkz389. URL https://doi.org/10.1093/nar/gkz389.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2448–2460, 2022. doi: 10.18653/v1/2022.acl-long.175. URL https://aclanthology.org/2022.acl-long.175/.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 28443–28467, 2025. doi: 10.18653/v1/2025.acl-long.1381. URL https://aclanthology.org/2025.acl-long.1381/.

Chaoqun Yang, Xinyu Lin, Shulin Li, Wenjie Wang, Ruihan Guo, Fuli Feng, and Tat-Seng Chua. Can large language models derive new knowledge? a dynamic benchmark for biological knowledge discovery. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817545. URL https://doi.org/10.1145/3770855.3817545.

Tianyun Zhong, Wangyi Jiang, Wei Wang, Xuanang Chen, Yaojie Lu, Shiwei Ye, Yuzhen Shi, Boyu Yang, Jinghang Wang, Han Li, Weiqi Zhai, Bing Zhao, Hu Wei, Haiyang Yu, Yongbin Li, Hongyu Lin, Le Sun, and Xianpei Han. Before the action: Benchmarking LLMs on prospective hypothesis discovery. *arXiv preprint arXiv:2607.15766*, 2026. URL https://arxiv.org/abs/2607.15766.

A Annotation, splits, and inference

The eight source screens supply perturbation observations rather than evidence labels. A retained case specifies one knockout and 5–20 differentially expressed genes, after which experts search only the date-filtered PubMed/PMC pool to construct its reference graph. The four subfields partition all 312 cases: kinase signaling (89), transcription regulation (78), metabolic (71), and immune (74).

Reference construction records source identifiers, exact spans, stance, linked claim endpoints, assumptions, rivals, and preferred tests. A separate edge-level pass normalizes endpoints and adjudicates support versus contradiction labels. Stable source identifiers and normalized relation endpoints drive duplicate collapse; adjudicators inspect all partial-span matches at the 0.5 token-overlap threshold. The resulting 4,210 opportunities are the frozen EOR reference population, and their 715 contradiction members are the CC reference population. The independent 72-case panel repeats opportunity construction with system identities and original labels hidden and produces its own denominator. Assumptions remain visible during whole-object review rather than entering either automatic denominator.

Full-benchmark and subfield intervals use 10,000 case-clustered resamples. Paired system effects resample case-aligned outputs together. The timing model resamples cases within expert and preserves each two-system pair. The independent panel is sampled before system scoring with the same hash-based proportional rule. Its $\kappa = 0.70$ agreement and same-direction UER and CC effects test whether the principal ordering survives an independently constructed label set; panel CC remains a case-macro estimand because the second panel creates its own opportunity set.

A.1 Counterbalanced timing details

Matched typed review has median 10.7 minutes and IQR [9.2, 12.8]; EvidenceWeave has median 8.4 and IQR [7.1, 10.0]. Mean corrections per object decline from 2.8 to 1.6, and unsupported-edge discoveries during review decline from 1.9 to 0.8. Final confidence on a five-point scale rises from 3.4 to 4.1.

Table 3: Obligation-audit results for EvidenceWeave by biological subfield. Counts partition all 312 cases. Brackets are 95% intervals from 10,000 within-subfield case resamples; audit minutes summarize the timed cases in each subfield, and proxy AUC retains the full-benchmark model-predicted target for heterogeneity analysis.

Subfield	n	UER	CC	Audit min	Proxy AUC
Kinase signaling	89	0.151 [0.119,0.186]	0.462 [0.421,0.504]	9.1	0.708 [0.671,0.744]
Transcription regulation	78	0.138 [0.108,0.171]	0.491 [0.448,0.534]	8.2	0.731 [0.695,0.766]
Metabolic	71	0.149 [0.116,0.184]	0.455 [0.411,0.499]	8.9	0.712 [0.674,0.749]
Immune	74	0.128 [0.098,0.161]	0.503 [0.461,0.545]	8.1	0.735 [0.701,0.768]

B OBLIGATION STRESS-TEST MATRIX

Table 4: Obligation boundaries, probes, and evidence traces. Each row connects one breach to its protected state, trust assumption, enforcement point, strongest control, measured property, and expert consequence.

Obligation breach	Protected state, trust, boundary	Threat capability; implemented control/probe	Property, utility, outcome, trace
Pre-cutoff provenance	Trusted dates; pre-generation filter	Threat: future retrieval; probe: 5–20% injection against clean control	Chronology integrity at fixed generation budget; support inflation when disabled; Figure 3
Supported assertion	Accepted edge; trusted expert spans; verifier gate	Threat: non-supporting spans; control: matched verifier-off ablation	Edge precision and audit burden; higher UER without gating; Tables 1, 5
Evidence coverage	Accepted/uncertain opportunity state; trusted opportunity labels; scorer	Threat: withhold difficult edges; control: joint UER–EOR scoring	Precision paired with surfaced evidence; omission remains visible; Figure 1
Counter-evidence exposure	Contradiction set; trusted expert labels; rival compiler	Threat: ignore counterevidence; control: rival-handling ablation	Contradiction discovery at fixed schema; coverage collapse; Table 5
Assumption exposure	Prerequisite nodes; trusted expert judgment; object schema	Threat: implicit conditions; control: whole-object expert audit	Expert-visible prerequisite; no separate assumption score; Appendix A.1
Discriminative test	Paired outcomes; trusted human target; test compiler	Threat: non-separating tests; probes: proxy ablation and separate adjudication	Component-level proxy AUC and human-adjudicated system AUC; Tables 5, 7
Representation fidelity	Six mapped field types; trusted blinded adjudication; evaluation mapper	Threat: dropped/merged fields; probe: held-out 72-case audit	Baseline comparability; metric shift at most 0.017; Table 8
Package identification	Compiler transitions; matched budgets and training; system boundary	Threat: model confounding; control: bypassed gates/rivals in same family	Package effect at equal retrieval and decoding budgets; Table 1
Panel robustness	Expert edge labels; trusted common rubric; annotation boundary	Threat: annotation-culture shift; probe: independent 72-case panel	Ranking stability without extra inference cost; direction retained; Table 9

C COMPONENT ANALYSIS AND ROBUSTNESS

Model-proxy AUC is a within-case diagnostic over all 312 cases. A frozen outcome-pair classifier receives the observation, intervention, and two predicted outcomes with system identity and rank hidden, then labels whether the outcomes separate the main mechanism from its rival. The compiler's test score is evaluated against these binary labels; ties receive half credit, AUC is computed within case, and the table reports the macro average. The classifier and labels are fixed before ablation scoring. This proxy traces component effects at full-benchmark scale, while the separately adjudicated 48-case target in Table 7 supplies the human system comparison.

The largest UER increase follows verification removal (0.142 to 0.238), whereas rival removal produces the largest CC and model-proxy-AUC decreases (0.478 to 0.312 and 0.724 to 0.582). Removing typed structure increases review time from 8.7 to 11.8 minutes, and free-form output is

Table 5: Descriptive obligation ablations on all 312 cases; audit time uses the 60-case subset. Model-proxy AUC measures predicted pairwise outcome separation on all cases and supplies component diagnosis; Table 7 separately reports the controlled matched comparison against a human-adjudicated target.

Variant	UER	CC	Audit min	Model-proxy AUC
Full EVIDENCEWEAVE	0.142	0.478	8.7	0.724
− Verification module	0.238	0.451	10.2	0.698
− Typed schema	0.197	0.367	11.8	0.641
− Rival explanations	0.151	0.312	9.1	0.582
− Constrained decoding	0.213	0.423	10.8	0.687
Free-form output	0.324	0.248	13.9	0.534

worst on all four reported endpoints. The interventions diagnose where each measured degradation enters the compilation path; the matched pair and independently adjudicated targets supply the inferential comparisons.

Table 6: Five-seed decoding sensitivity on all 312 cases with retrieval fixed. Worst denotes the least favorable seed. The paired intervals below the table resample cases while preserving seed blocks.

Metric	System	Mean	Between-seed SD	Worst
UER	Matched typed	0.247	0.006	0.255
UER	EVIDENCEWEAVE	0.143	0.004	0.149
CC	Matched typed	0.301	0.008	0.289
CC	EVIDENCEWEAVE	0.477	0.009	0.463
EOR	Matched typed	0.600	0.006	0.591
EOR	EVIDENCEWEAVE	0.671	0.005	0.663

Across 5 seeds, every EVIDENCEWEAVE seed outperforms every matched-control seed. Full-minus-control paired intervals are [-0.118, -0.090] for UER, [+0.151, +0.201] for CC, and [+0.055, +0.086] for EOR.

D SEPARATELY ADJUDICATED TARGETS AND REPRESENTATION AUDITS

Table 7: Human-adjudicated written-outcome-separation target on the fixed 48-case subset. AUC scores blinded system tests against pairwise labels from adjudicators who did not contribute scored rankings; ties receive half credit, and intervals use 10,000 case-bootstrap resamples with case as the uncertainty unit.

System	AUC	95% interval
RAG (GPT-4o)	0.529	[0.486, 0.571]
Matched Typed Llama-3.1-70B	0.584	[0.543, 0.623]
EVIDENCEWEAVE	0.706	[0.668, 0.742]

The human target contains 1128 adjudicated pairwise decisions and 96 ties. AUC is computed within case from the ranked system tests against pairwise labels and then aggregated over cases; inter-adjudicator agreement is $\kappa = 0.69$. Human adjudication shows that the full compiler package ranks written outcome pairs better than the matched typed control; this comparison does not separately identify the rival-conditioned stage.

Two blinded molecular-biology annotators and an adjudicator compare 3157 extracted field instances against 3228 reference instances, with 2781 adjudicated matches. Applying corrections changes any primary metric by no more than 0.017 and leaves all six systems in the same order.

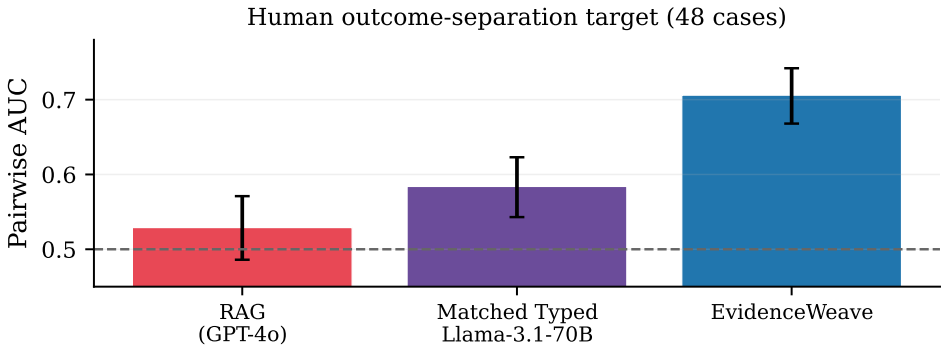


Figure 4: Human-adjudicated written-outcome-separation AUC on 48 cases for RAG, Matched Typed Llama-3.1-70B, and EVIDENCEWEAVE. Whiskers are 95% intervals from 10,000 case-bootstrap resamples; the dashed line marks chance.

Table 8: Blinded free-form canonicalizer audit over 216 outputs from RAG, GraphRAG, and Extraction+Planner on 72 proportional held-out cases. Precision, recall, and F1 use adjudicated field instances; intervals for macro F1 are case-bootstrap intervals.

Field	Precision	Recall	F1
Claim nodes	0.949	0.933	0.941
Provenance spans	0.926	0.910	0.918
Contradiction edges	0.899	0.876	0.887
Assumptions	0.866	0.838	0.852
Rival explanations	0.875	0.854	0.864
Falsification tests	0.841	0.817	0.829
Macro	—	—	0.882 [0.866,0.897]

Table 9: Independent-panel replication of supported-assertion and counterevidence obligations on 72 cases. CC is the paired case-macro estimand for the panel-specific opportunity set; aggregate brackets are 95% paired case-bootstrap intervals. Subfield rows report EVIDENCEWEAVE minus matched-control changes; negative UER and positive CC are favorable.

Comparison	UER	Case-macro CC
Matched typed	0.247 [0.208,0.289]	0.309 [0.262,0.358]
EVIDENCEWEAVE	0.158 [0.126,0.194]	0.452 [0.402,0.501]
Kinase signaling change	-0.090	+0.139
Transcription regulation change	-0.092	+0.147
Metabolic change	-0.089	+0.137
Immune change	-0.084	+0.152

Pooled unsupported decisions are 142/575 for the matched control and 89/563 for EVIDENCEWEAVE. The panel surfaces 51 and 75 matched contradictions; case-macro CC averages each case's covered share before paired inference and is therefore interpreted by within-panel system difference.

E CLOSEST EVALUATION OBJECTS

The nearest benchmarks expose different projections of the edge-to-experiment path. The target here is the linked pre-execution graph: source-exact evidence and adverse evidence at claim edges, explicit assertion state and prerequisites, rival-specific outcomes, retrieval dates, and measured review burden.

Table 10: Closest evaluation objects and their primary scoring units. The final column states the additional unit measured here rather than implying that adjacent systems cannot represent it.

Work	Primary evaluation object	Scored evidence/test unit	Additional unit scored here
EvidenceBench	Hypothesis–paper pair	Relevant biomedical sentences	Edge state, contradiction denominator, downstream test
PaperQA2	Literature answer or extracted claim	Synthesis accuracy and contradictions	Mechanism linkage and paired audit cost
HypER	Generated hypothesis and reasoning chain	Provenance-grounded chain quality	Accepted/deferred edges and annotated omissions
Hypothesis benchmarks	Hypothesis, future claim, or hypothesis set	Discovery quality and temporal targets	Edge-local adverse evidence and review cost
Lit2Test	Six-field research proposal	Supporting/falsifying outcomes and pairwise preference	Source-exact edge obligations and expert inspection
Execution protocols	Executed test, trajectory, or evolving state	Actions, outcomes, and belief transitions	Pre-execution ledger under bounded retrieval
This work	Evidence-obligation graph	Accepted/uncertain evidence, rivals, paired outcomes	Joint frontier and human reconstruction cost

The comparison places the contribution at the linkage boundary. The benchmark evaluates the state presented to a reviewer, while the matched typed pair identifies the effect of compiling that state.

F WORKED EVIDENCE-OBLIGATION OBJECT

The schematic object below follows the same local review sequence used by the benchmark. A reviewer can inspect the accepted support, challenge a contradiction edge, retain an indirect candidate without treating it as established, and compare the proposed rescue outcome with the named rival.

Main claim: KO(X) increases IFN signaling through STAT1.
Claim nodes: c1, X represses STAT1 targets; c2, differentially expressed genes enrich IFN pathways.
Support edges: c2 ← pre-cutoff span, accepted, confidence 0.93.
Contradiction edges: c1 ← pre-cutoff rescue span, accepted, confidence 0.88.
Uncertain edges: c1 ← indirect binding span, uncertain, confidence 0.61.
Assumption nodes: KO(X) does not trigger generic stress.
Rival explanations: DNA-damage response explains the IFN signal.
Experiment nodes: rescue X while leaving DNA-damage signaling unperturbed; the main mechanism predicts collapse of the IFN signature, while the DNA-damage rival predicts persistence.

Figure 5: A mechanism object as an obligation ledger: support, contradiction, uncertainty, assumptions, rivals, and outcome-separated tests remain inspectable in one local state.

The key property is linkage rather than field count: every span has a stance and claim endpoint, every uncertain edge remains visibly non-asserted, and the proposed experiment records outcomes whose disagreement can change the next scientific action.