BEYOND HYPOTHESIS RECOVERY: EVALUATING SCIENTIFIC HYPOTHESIS GENERATION THROUGH EVIDENCE-CONDITIONED HYPOTHESIS-STATE UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific hypothesis generation is commonly evaluated by whether a model recovers a correct or novel historical hypothesis from supplied evidence. For modern foundation models, endpoint recovery is non-identifying: the same hypothesis may arise from evidence-conditioned reasoning or parametric access to the historical relation. Temporal cutoffs and evidence interventions reduce this ambiguity but retain a static evidence-to-hypothesis endpoint as the evaluation target.

We introduce HypoTrans, a benchmark of evidence-conditioned scientific-state transitions reconstructed from cell-biological literature. Given a prior state and newly documented formation evidence, with the later target state withheld, a model generates a structured hypothesis state. Typed evidence–hypothesis provenance separates formation from validation evidence, and four continuity gates support documentary validation. We characterize changes using emergence, directional alignment, revision where applicable, and evidence attribution; matched-irrelevant-evidence selectivity is not identified by the paired conditions used here.

In an earlier collection, closed-book access and endpoint recovery were positively associated ($\rho = 0.373$, $p < 0.0001$, $n = 200$ state–model pairs); access and update-content scores were also positively associated ($\rho = 0.187$, $p = 0.022$, $n = 150$ transition–model pairs). These are separate descriptive analyses on different samples, not a test that one task is superior. In a new evaluation of five models, strict exact emergence was zero, lexical mechanism-level overlap ranged from 0.050 to 0.092, and categorical directional alignment from 0.041 to 0.081; semantic correctness was not adjudicated. The benchmark provides a behavioral measure of correspondence between model-state changes and documented transitions, but does not establish internal evidence use or selective updating.

1 INTRODUCTION

Scientific research is not retrieval. A central act of scientific work is forming, from limited and changing evidence, hypotheses that explain observations, organize existing knowledge, or predict new phenomena. With foundation models and agentic systems now automating literature synthesis, experiment design, and claim drafting (Lu et al., 2024; Gottweis et al., 2026; Baek et al., 2025), automated hypothesis generation is becoming a first-class step in AI-assisted discovery. Recent benchmarks have further begun to evaluate scientific discovery capabilities in broader settings (Majumder et al., 2025). A natural question follows: how do we determine whether a model actually reasons from scientific evidence to form and update hypotheses?

A direct benchmark would provide evidence, query a model for a hypothesis output, and score whether the generated hypothesis matches a historical hypothesis.

For models that continuously absorb scientific literature, this scoring is non-identifying. Recovering a historical hypothesis does not establish that the hypothesis was generated from the provided evidence. If the historical hypothesis has already been encoded in the model parameters, the model

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.

may produce the same endpoint without relying on the provided evidence during inference. A correct endpoint therefore does not imply correct evidence-conditioned generation (Figure 1A).

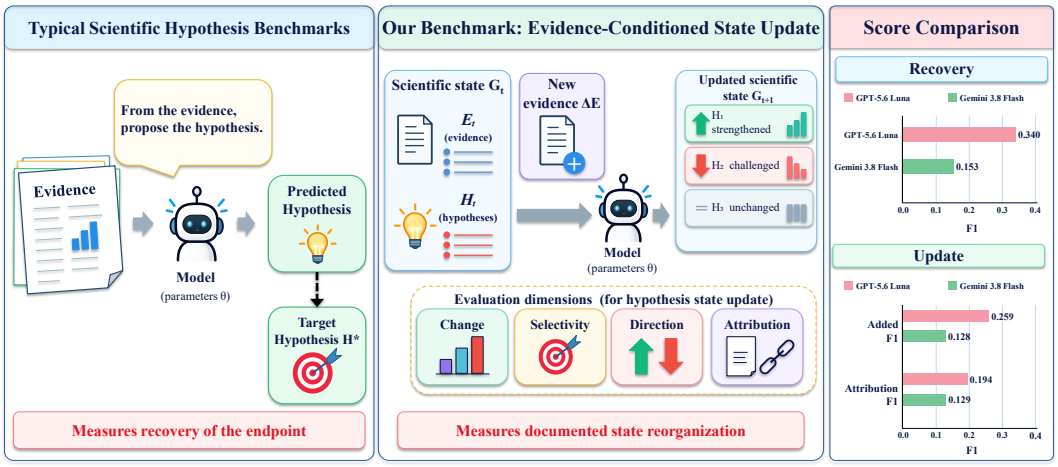

Figure 1: (A) Endpoint scoring conflates evidence-conditioned reasoning with parametric accessibility; as models absorb literature faster than science produces validated knowledge, the accessibility term dominates, so a correct endpoint cannot certify evidence-conditioned hypothesis generation. (B) We redefine hypothesis generation from emitting a correct or novel hypothesis to organizing and updating candidate hypotheses and their evidential status under parametric knowledge plus explicit evidence. (C) State-recovery F1 exceeds update-content F1 in a three-model stratified follow-up. Bars show GPT-5.6 Luna and Gemini 3.8 Flash on state-level hypothesis recovery ($n = 59$ states), added hypotheses ($n = 130$ gold propositions), and evidence attribution ($n = 379$ gold links).

This is not merely conventional data leakage. It is a more general *parametric accessibility* problem: the same output, under different parameter states, can correspond to entirely different cognitive processes, so endpoint recovery is insufficient to identify evidence-conditioned hypothesis generation. The problem compounds because model iterations and training-corpus updates occur on substantially shorter timescales than the accumulation of experimentally validated knowledge, particularly in domains such as cell biology where the knowledge base grows slowly relative to the monthly expansion of parametric capacity. A benchmark that is novel at construction becomes, after a model update, an increasingly trivial parametric-recovery task.

Existing work approaches this problem through temporal and information control or evidence intervention. Temporal separation introduces knowledge cutoffs to study whether models can recover hypotheses that emerge after a defined point in time (Yang et al., 2025). Evidence interventions examine how model outputs change as available information or experimental details are modified (Lew et al., 2026; Huang et al., 2025). These approaches provide important controls over information exposure and model responses. However, they primarily characterize model performance under controlled information conditions, focusing on whether models can recover, generate, or respond to hypotheses under specified settings, rather than on how hypothesis states evolve as scientific evidence accumulates.

This points to a different evaluation object. Scientific knowledge does not arrive as isolated evidence–hypothesis pairs. As new experiments, observations, and counterexamples accumulate, existing hypotheses are strengthened, restricted in scope, challenged by competing mechanisms, reinterpreted, or replaced. In the literature, this process appears as an auditable sequence of documented scientific-state changes rather than a collection of independent hypothesis endpoints. The natural question for hypothesis generation is therefore not "does the model emit the right hypothesis?" but "when the evidence state changes, does the model's hypothesis space change in the scientifically appropriate way?"

This reformulation yields a concrete evaluation structure. When new evidence arrives, one hypothesis may be strengthened, another challenged, and others left unchanged. A scientifically appropriate model should not merely change, but change *selectively* (only hypotheses implicated by the

Table 1: **Comparison of representative scientific hypothesis-generation benchmarks.** Our benchmark evaluates evidence-conditioned hypothesis-state updates over documented scientific-state transitions.

Benchmark	Year	Domain	TC	EI	SU	EA	ST
IdeaBench	2025	General science	–	–	–	–	–
HypER	2025	General science	–	–	–	✓	–
Matter-of-fact	2025	Materials science	–	–	–	✓	–
MOOSE-Chem	2025	Chemistry	✓	–	–	–	–
ProjectionBench	2026	General science	✓	✓	–	–	–
POPPER	2025	General science	–	✓	Partial	–	–
BioDisco	2026	Biomedical science	✓	Partial	–	Partial	–
HypoTrans (ours)	**2026**	**Cell biology**	✓	✓	✓	✓	✓

TC: Temporal Control; EI: Evidence Intervention; SU: Hypothesis-State Update; EA: Evidence Attribution; ST: Scientific-State Transition.

new evidence), *directionally* (consistent with the documented update), and with *attribution* (tracing the change to the formation evidence that drove it). We therefore build a benchmark from historically documented scientific-state transitions: we reconstruct states with explicit evidence–hypothesis provenance, admit only transitions that pass continuity gates, withhold the target state, and ask the model to predict the hypothesis-state update from the prior state plus the documented new evidence (Figure 3). Outputs are compared against the independently reconstructed historical update along four complementary observables: change, selectivity, directional alignment, and attribution, reported without a composite score (Section 3). This is deliberately *not* a de-memorization design: we do not eliminate parametric knowledge, but re-target the measured behavior so its interpretation does not collapse into "does the model know the answer."

Contributions. Our contributions are summarized as follows. (1) We identify endpoint hypothesis recovery as an under-identified measure of evidence-conditioned scientific reasoning: matching a historical endpoint does not reveal whether the model followed the provided evidence or relied on knowledge already encoded in its parameters. (2) We shift the object from endpoint recovery to the evidence-conditioned update, measured through four complementary observables: change, selectivity, directional alignment, and evidence attribution, reported without a composite score. (3) We construct a cell-biological benchmark of historically documented scientific-state transitions. (4) Across five frontier models, we find that static endpoint recovery tracks parametric access while the evidence-conditioned update is markedly less access-coupled, and every model adds hypotheses far more readily than it challenges them (Section 6).

2 Related Work

Scientific hypothesis-generation benchmarks have explored different aspects of scientific foundation models (Zhang et al., 2024). We organize prior work around two evaluation perspectives: assessing the quality of generated hypotheses and designing evaluation settings that control information availability or intervene on evidence conditions.

Hypothesis quality evaluation. Prior work evaluates generated hypotheses along multiple dimensions, including truthfulness and hallucination (Xiong et al., 2025), novelty and inspiration (Wang et al., 2024; Hu et al., 2025), preference and ranking (Si et al., 2025; Guo et al., 2025), feasibility (Jansen et al., 2025), support-or-refute evidencing (Koneru et al., 2024), and provenance-aware generation (Vasu et al., 2025). These works establish important dimensions for assessing hypothesis quality, primarily evaluating generated hypotheses or their associated properties.

Temporal and information control. Temporal and information-controlled evaluations study model behavior under different information availability settings. MOOSE-Chem introduces temporal cutoffs to evaluate whether models can recover hypotheses associated with later scientific findings (Yang et al., 2025). BioDisco combines evidence-grounded generation, iterative refinement, and temporal evaluation to examine whether generated hypotheses recover scientific relationships

discovered after a fixed cutoff (Ke et al., 2026). ProjectionBench progressively reveals experimental details and evaluates hypotheses at different disclosure stages (Lew et al., 2026). These approaches provide mechanisms for studying how information timing and availability influence model outputs.

Evidence intervention. Evidence interventions examine model behavior under controlled modifications of inputs or evidence conditions. Counterfactual augmentation and contrast sets provide methodological precedents for evaluating model sensitivity to controlled input changes (Kaushik et al., 2020; Gardner et al., 2020). Studies of parametric and contextual knowledge further investigate how model behavior depends on stored and supplied information (Neeman et al., 2023; Cheng et al., 2024; Tao et al., 2024). POPPER studies hypothesis testing through agentic sequential falsification, where hypotheses are iteratively tested as evidence accumulates (Huang et al., 2025). Together, these works provide perspectives on model responses under changing evidence conditions.

In contrast, our benchmark evaluates whether evidence-conditioned hypothesis-state updates correspond to documented scientific-state transitions, rather than only measuring hypothesis outputs or responses under controlled conditions.

3 Problem Formulation

3.1 Evidence and hypothesis state

The primary evidence unit is a textual experimental result $e = (s, o, c, t)$, where s denotes the experimental statement, o the observed outcome, c the experimental context, and t the publication time, collected from a cell-biological open-access literature pool. Background, post-hoc validation, and caption-only glosses are excluded because caption-linked descriptions frequently paraphrase the same underlying experiment and fail as independent units (Appendices B, D).

A set of evidence units is denoted as E. Given evidence E, a model with parameters θ produces a structured hypothesis–support state $Y_\theta(E) = \{\mathcal{H}, \mathcal{A}, \mathcal{R}, \mathcal{Q}\}$, where $\mathcal{H}$ contains candidate hypotheses, $\mathcal{A}$ records evidence attributions, $\mathcal{R}$ represents inter-hypothesis relations, and $\mathcal{Q}$ captures explicit uncertainty qualifiers. Each hypothesis records a falsifiable statement, its supporting evidence IDs, mechanism, scope boundary, predicted direction, and testability, so that a change in E is scored as a change in this structured state rather than as a free-text edit.

3.2 Scientific state and provenance

A scientific state is represented as $G_t = (E_t, \mathcal{H}_t, L_t, C_t)$, where E_t is the temporally admissible and transition-relevant evidence set, $\mathcal{H}_t$ the documented hypothesis set, $L_t \subseteq E_t \times \mathcal{H}_t$ denotes the typed evidence–hypothesis provenance relation, and C_t the scientific context and boundary conditions. Provenance is many-to-many: evidence may support multiple hypotheses, and hypotheses may rely on multiple evidence units, rather than forming a forced single chain. Each provenance link carries a role in formation, supporting, validation, unclear. Only formation-relevant (and retained supporting) evidence enters target-blind generation, and documentary rationales are not interpreted as recovering authors' private cognition.

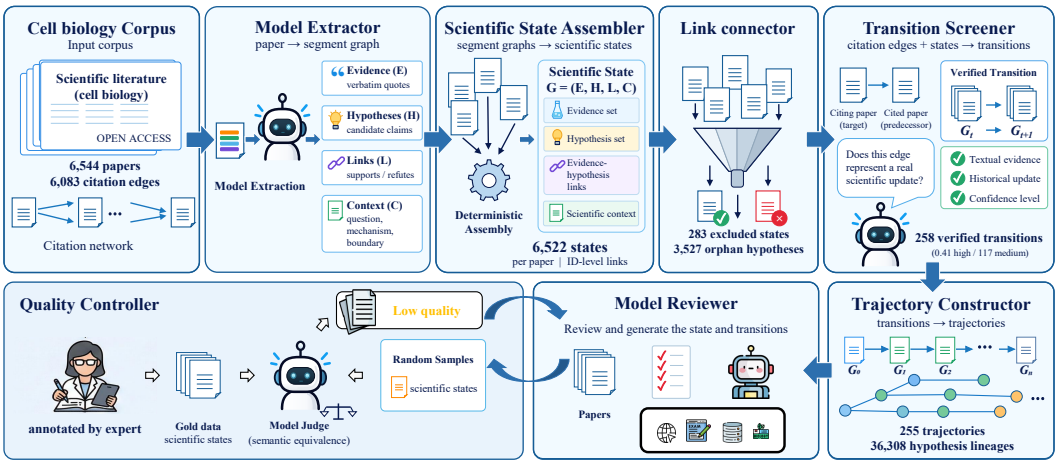

Figure 3: **Benchmark construction pipeline.** From an open-access cell-biological literature pool we extract evidence, hypotheses, and typed evidence–hypothesis links under a span-verification gate, assemble scientific states, screen citation-derived candidates under continuity and documentary gates into verified transitions, assemble those transitions into trajectories, and score model generations with a disjoint judge along change, selectivity, direction, and attribution.

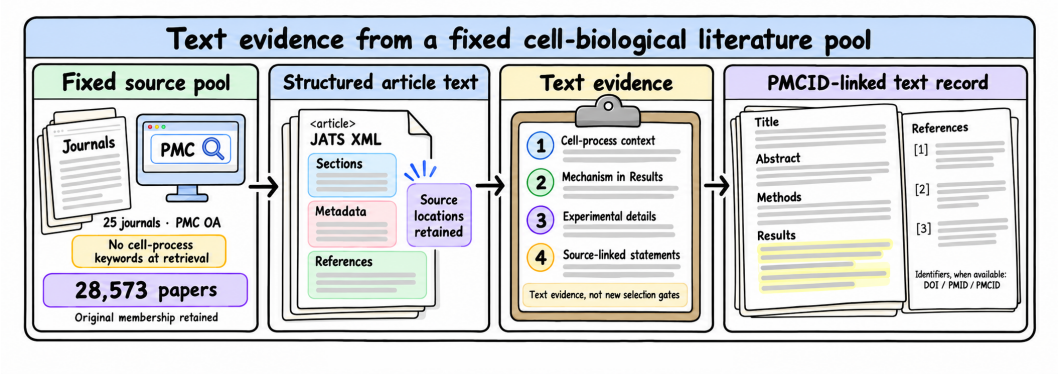

Figure 2: Text evidence from the fixed cell-biological literature pool. The same 28,573 papers provide section text, metadata, and references linked by PMCID. The diagram highlights cell-process context, Results-grounded mechanisms, experimental details, and source-linked statements. These are text-evidence components, not new selection gates. Original acquisition requirements are described in this appendix; paper membership is unchanged.

3.3 Historical transition and continuity gates

Scientific development is represented as a sequence of documented state changes rather than isolated hypothesis endpoints. A trajectory $\mathcal{T} = (G_1, \ldots, G_n)$, where n denotes the number of documented states in the trajectory, records this scientific-state evolution. A transition $G_t \to G_{t+1}$ is admitted only if it satisfies four continuity gates: question continuity, historical continuity (documented continuation rather than mere citation), documentary evidence continuity, and hypothesis-state continuity (a describable update in $\mathcal{H}_t$ or C_t). Detailed criteria are provided in Appendix A. The resulting transition $G_t \to G_{t+1}$ is the primary evaluation object; evidence need not grow monotonically, and reinterpretation is distinguished from deletion.

4 Benchmark Construction

Corpus and pipeline The pool is drawn from 25 prespecified cell and molecular biology journals in the PubMed Central Open Access collection under a reproducible acquisition protocol (Appendix 2);

only textual experimental-result statements enter E_t. Construction proceeds through six stages: pool curation, evidence–hypothesis graph construction, verified transition identification, frozen case selection, canonical normalization, and agentic review . We separate four roles that must not be conflated: agent-assisted construction proposes evidence, hypotheses, and links; model-based verification sampling estimates error rates; expert verification inspects independent samples; and held-out model evaluation measures $Y_\theta(E)$ under target-blind protocols. The evaluated models are never the sole sources of benchmark labels.

State and transition construction Each paper is processed into a scientific state by extracting observation-level evidence statements (typed as observation, measurement, prior finding, mechanistic, boundary, negative, contradictory, contextual, or background), hypotheses with origin status (introduced here, prior retained, modified here, or contested here), and typed E–H links with roles of forming, supporting, testing, challenging, refining, or remaining unclear, together with a formation status. Every quote is re-anchored verbatim to its source span or rejected, and a paper is admitted only if it carries at least one active hypothesis. Candidate transitions are screened along citation edges and re-checked for documentary continuity; each admitted transition is typed as addition, support, refinement, replacement, contradiction, or reinterpretation, which determines the historically expected direction of change.

canonical schema. The extractor output is normalized onto a canonical schema, which provides a single shared representation into which the gold and agentic corpora are cast so that they can be compared free of surface confounds, with one hypothesis taxonomy and one update taxonomy. Each hypothesis is bucketed by its evidence linkage into one of four classes: *generation target* (formable from the admissible evidence), *historical-only* (retained via validation links but not formable from E_t), *audit* (reserved for documentary-validity checks), and *isolated* (no evidence–hypothesis link, recoverable only through parametric memory). Only generation-target hypotheses enter the scored gold, so memory-only hypotheses never re-admit endpoint recovery. The transition types are normalized onto a five-way update vocabulary consisting of *added*, *modified*, *challenged*, *removed*, and *retained*. Each transition record carries source context (question, system, boundary; mechanism excluded), source evidence, documentary new evidence with attribution gold links, and the generation-target hypotheses of both endpoints (Appendix E).

Agent-human collaborated workflow for quality control. Construction closes with an agentic review loop that drives generated states toward quality control from the expert gold(Figure 3). A review agent is given read/query access to the source database and to the structured extraction produced upstream (evidence statements, typed E–H links, and hypothesis buckets); it iteratively revises candidate states until the quality scorer can rate the human annotations and agent-generated ones both above median of 3. Human verification details are provided in G.

5 METRIC DEFINITIONS

Let $\Delta\hat{H}$ denote the canonical propositions newly expressed in B_1 relative to B_0, and let A denote the documented additions. We define evidence-induced emergence as

$$\text{Emergence} = F_1(\Delta\hat{H}, A).$$

The frozen Track B scorer uses token F1 with a threshold of 0.35 for mechanism-level screening; this lexical threshold does not certify semantic equivalence. Directional labels are rule-based from these matches: aligned, partially aligned, misaligned, or no meaningful update. The reported directional alignment rate is the fraction labeled aligned among transitions with a meaningful update; partial alignment is not counted as aligned. For transitions containing challenged or replaced content R, revision is

$$\text{Revision} = F_1(\hat{H}_0 \setminus \hat{H}_1, R),$$

with false retention computed as the fraction of challenged components that remain in $\hat{H}_1$. Attribution precision, recall, and F_1 compare evidence IDs attached to newly expressed model hypotheses with documented formation evidence IDs. The IDs are supplied in the prompt, so attribution is not interpreted as independent evidence discovery.

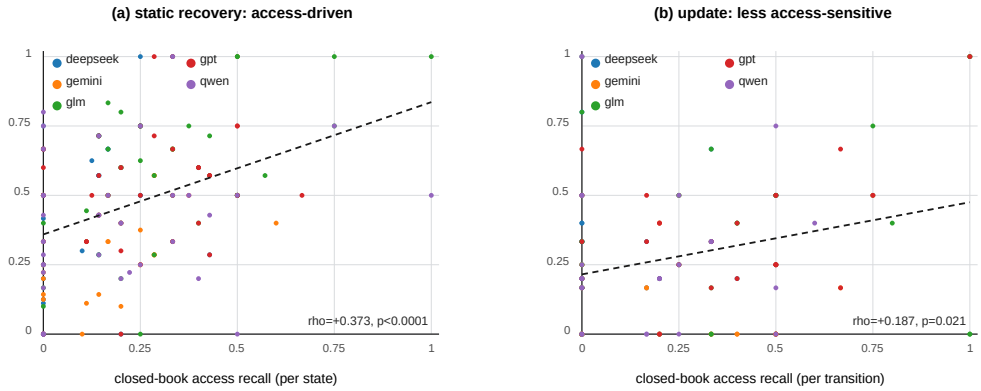

Figure 4: **The update is less access-sensitive than recovery.** Each point is one state–model (a) or transition–model (b) observation; five frontier models are distinguished by color. (a) Static endpoint recovery rises with closed-book access recall (pooled Spearman $\rho = +0.373$, $p < 0.0001$, $n = 200$; OLS slope 0.48): the states a model can already reproduce from memory tend to be those it "recovers" under evidence. (b) The evidence-conditioned update shows markedly weaker access coupling (pooled $\rho = +0.187$, $p = 0.021$, $n = 150$; OLS slope 0.26), with no individual model significant. Replacing the static readout with the update task therefore attenuates—without eliminating—the access confound.

6 Experiments

6.1 Setup and evaluation protocol

We evaluate five frontier models on the benchmark evaluation set. Each transition is run under two target-blind conditions: B_0 contains the prior admissible evidence, while B_1 adds the formation-relevant evidence associated with the documented update. Prompt, output schema, decoding configuration, and evidence formatting are held fixed across the pair; only the evidence state changes. The subsequent hypothesis state and target article are withheld.

We also include a small qualitative follow-up in Figure 1C. GPT-5.6 Luna and Gemini 3.8 Flash are evaluated on randomly sampled cases to contrast endpoint recovery with update-content scores. Recovery is consistently higher in these illustrative cases, motivating the distinction between recovering a scientific endpoint and reconstructing how the hypothesis state changes. This panel is not used as a benchmark statistic.

For the main benchmark, we compare the generated states $\hat{H}_0$ and $\hat{H}_1$. Emergence measures newly expressed content that matches the documented addition at exact or mechanism level; directional alignment asks whether the change follows the documented update type; revision measures removal of documented challenged or replaced content when such content exists; and attribution compares cited evidence with independently documented formation evidence. Transition IDs, rather than individual generations, are the statistical unit. We report transition-level bootstrap intervals and do not construct a composite score or model ranking. Full formulas are in Appendix 5.

6.2 Endpoint recovery and access

Parametric accessibility is not directly observable. In an earlier endpoint-recovery collection, closed-book recall from a topic cue and open-book recovery were positively associated (Spearman $\rho = +0.373$, $p < 0.0001$) (Figure. 4), consistent with the concern that endpoint recovery partly reflects information already available to the model. We use this result only as motivation and a behavioral diagnostic, not as a measurement of memorization or as a benchmark metric.

Table 2: **Behavioral profiles under evidence-conditioned updating.** Columns are separate observables rather than a total score. Confidence intervals are reported in Appendix 5.

Model	Mech.	Direction	Revision	False ret.	Attr. F1
DeepSeek	0.080	0.500	0.750	0.438	0.532
Gemini	0.122	0.556	0.875	0.125	0.485
GLM	0.095	0.273	1.000	0.000	0.436
GPT	0.078	0.429	0.938	0.063	0.513
Qwen	0.096	0.625	0.938	0.063	0.580

6.3 Evidence-conditioned updating across models

The benchmark results are summarized in Table 2. Across models, evidence attribution is consistently stronger than mechanism-level hypothesis emergence. This gap indicates that identifying or citing relevant evidence does not by itself imply that the model converts that evidence into the corresponding scientific-state change. Directional alignment and revision vary across models, showing that evidence incorporation is not a single behavioral capability: models differ in how often they add, preserve, challenge, or revise hypothesis content.

The most consistent pattern is an asymmetry between evidence identification and state revision. Models often preserve or add hypothesis content even when the historical transition calls for a more specific modification or challenge. Conversely, attribution scores are higher than update-content scores, suggesting that the benchmark separates two stages that endpoint recovery conflates: recognizing relevant evidence and reorganizing the hypothesis state that evidence supports.

The update should not be interpreted as contradiction-specific revision: the benchmark uses prior evidence versus prior-plus-relevant-evidence, not a separate contradictory-evidence condition. Likewise, the present comparison establishes correspondence with documented historical transitions, not internal causal use of any particular evidence unit. Controlled irrelevant- and contradictory-evidence interventions are reserved for future validation of selectivity and challenge sensitivity.

6.4 Interpretation and failure modes

The model profiles reveal three recurring behavioral dimensions: evidence attribution, hypothesis-state emergence, and directional revision. A model can cite relevant evidence without expressing the corresponding mechanism; it can express a related mechanism without following the documented update direction; and it can preserve prior content when the historical transition challenges it. These are distinct failure modes, so we report them separately rather than hiding them in a single benchmark score.

7 Discussion

The central result is a change in what is measured. Endpoint recovery asks whether a model can produce a documented hypothesis. HypoTrans instead asks whether its hypothesis state changes in correspondence with a documented scientific transition when the relevant formation evidence is introduced. This reframing does not remove parametric knowledge; it avoids treating successful endpoint recovery as sufficient evidence of evidence-conditioned generation.

The access diagnostic supports the motivation but does not measure latent parametric knowledge directly. The main benchmark likewise does not establish selectivity, contradiction-specific revision, or an internal causal mechanism for evidence use. Those claims require controlled irrelevant and contradictory conditions beyond the current paired evaluation. What the present benchmark establishes is a measurable and decomposable behavioral profile of evidence-conditioned hypothesis updating.

8 CONCLUSION

Scientific hypothesis generation is better represented as an evolving relation between evidence and hypothesis states than as recovery of a fixed historical endpoint. HypoTrans operationalizes this view by reconstructing documented scientific transitions, separating formation from validation evidence, withholding successor states, and evaluating paired state changes across emergence, direction, revision, and attribution. The resulting behavioral profiles show that evidence attribution and hypothesis-state updating are distinguishable capabilities. This provides a compact evaluation target for future work on whether models can track, rather than merely reproduce, scientific hypothesis evolution.

AI USE STATEMENT

Generative AI tools (GPT-5.6 Sol and GPT-6.0 Astra) were used for language refinement, presentation editing, and implementation refinement. All scientific contributions, including conceptualization, methodology, benchmark design, experimental setup, model evaluation, analysis, and final manuscript decisions, were carried out by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript, including all claims, references, and associated artifacts.

REFERENCES

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738, 2025.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*, 2024.

Matt Gardner et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*, 2020.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. IdeaBench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. NOVA: An iterative planning framework for enhancing scientific innovation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21330–21359, 2025.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 25372–25437. PMLR, 2025.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4090–4102, 2025.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2020.

Yujing Ke, Kevin George, Kathan Pandya, Gerrit Großmann, David B. Blumenthal, Maximilian Sprang, David A. Selby, and Sebastian Vollmer. Biodisco: Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 6761–6769. International Joint Conferences on Artificial Intelligence Organization, 8 2026. doi: 10.24963/ijcai.2026/752. URL https://doi.org/10.24963/ijcai.2026/752. AI and Health.

Sai Koneru, Jian Wu, and Sarah Rajtmajer. Can large language models discern evidence for scientific hypotheses? case studies in the social sciences. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2787–2797, 2024.

A. J. Lew, Y. Cao, and M. J. Buehler. ProjectionBench: Evaluating scientific hypothesis generation in LLMs under progressive information disclosure, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *ICLR*, 2025.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *ACL*, 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *ICLR*, 2025.

Yufei Tao, Adam Hiatt, Erik Haake, Antonie J Jetter, and Ameeta Agrawal. When context leads but parametric memory follows in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4034–4058, 2024.

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*, 2025.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *ACL*, 2024.

Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 2025.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *ICLR*, 2025.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8783–8817, 2024.

A TRAJECTORY CONSTRUCTION PROTOCOL

Quality over quantity; counts reflect strict gate admission rather than scale maximization. The full-scale run admitted 258 verified transitions (141 high- / 117 medium-confidence) assembling into 255 trajectories (2 multi-step), reported as screened candidates pending expert documentary adjudication.

A.1 FOUR GATES

- **Question continuity.** Narrow scientific question (mechanism / pathway / process under conditions), not broad disease area.
- **Historical continuity.** Documented continuation (extension, refinement, reinterpretation, replacement, contradiction). Citation or later publication alone is insufficient.
- **Documentary evidence continuity.** Explicit textual support linking prior findings to the new investigation. Lexical/embedding overlap $\rightarrow$ candidate only.
- **Hypothesis-state continuity.** A describable update in hypotheses or context/boundaries.

A.2 EVIDENCE ASSEMBLY

E_t: temporally admissible, trajectory-relevant, provenance-supported. Reinterpretation $\neq$ deletion.

A.3 E–H ROLES

Candidate $\rightarrow$ documentary relation $\rightarrow$ role (formation / supporting / validation / unclear). Validation excluded from generation states.

A.4 TRANSITION TYPES

Addition, support, refinement, replacement, contradiction, reinterpretation; several types may co-occur on a single transition. Type is assigned before model evaluation and defines expected directional change.

B CELL-BIOLOGICAL CORPUS ACQUISITION

This appendix describes the fixed development pool used to obtain textual experimental results and trajectory candidates. All paper-level statistics refer to the same 28,573 papers; they do not count verified evaluation cases.

B.1 SOURCE POOL AND TEXT USE

We retain the original release frozen on 31 August 2026, without rerunning selection or changing paper membership. Its acquisition protocol required all five original gates, including a strict main-body caption, followed by at least one decodable image file. These historical requirements explain the source pool; the present benchmark uses textual result statements only. We do not present image-collection results here.

B.2 CORPUS SUMMARY

Table 3: Fixed development pool (31 August 2026). Retention uses all discovered unique PMCIDs as the denominator. Identifiable reference rows are not unique citation pairs.

Quantity	Value	Meaning
Source journals	25	Prespecified journal names and aliases
Discovered unique PMCIDs	174,029	Journal-based search, without topical process terms
Papers in the fixed pool	28,573	16.42% of discovered papers
Identifiable reference rows	1,808,175	Reference rows carrying a DOI, PMID, or PMCID

B.3 TEXT CRITERIA IN SOURCE SELECTION

Table 4 summarizes the text checks relevant to the source pool. These are not a replacement for the complete historical acquisition protocol described above. All required original gates had to pass; there was no score or compensation between gates.

Table 4: Text checks within the original source-selection rules. Mechanism evidence is restricted to Results. Caption text could support the original experiment check, but does not enter the Results-only mechanism check.

Check	Source	Requirement
Article scope	Article type and sections	Non-retracted research article with non-empty Abstract, Introduction, Methods, and Results.
Cell process	Results and Title/Abstract	Evidence for the same process category in Results and in Title or Abstract.
Mechanism	Results only	An explicit process–relation link, a process–perturbation–outcome window, or a bridge supported by both perturbation–outcome and process–outcome evidence in Results.
Experiment	Methods, Results; eligible captions	Model, perturbation, assay, outcome, and Methods-design evidence, plus a perturbation–outcome or process–outcome link.

B.4 JOURNAL COVERAGE AND RETENTION

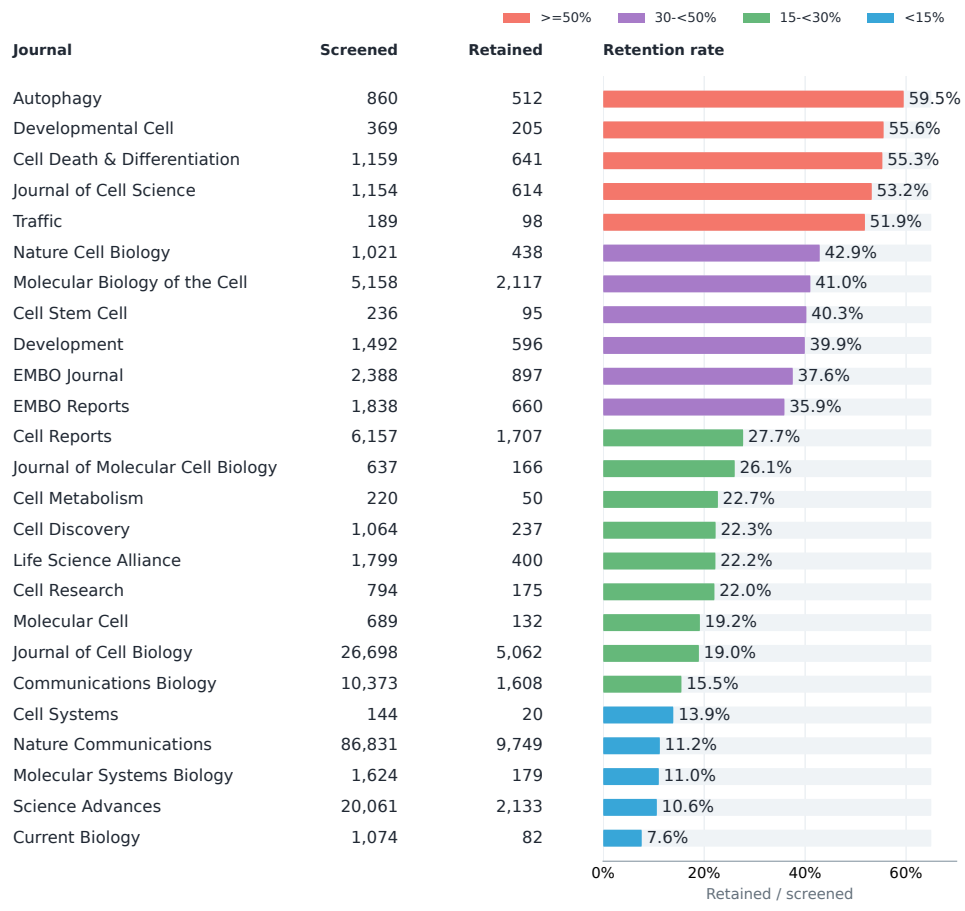

Figure 5: Journal-specific counts and retention rates for the 25 sources in the fixed pool. Rows are ordered by retention rate. Screened and retained counts sum to 174,029 and 28,573, respectively. Differences reflect PMC OA availability and the original selection rules, not journal quality.

B.5 PROCESS AND PUBLICATION-YEAR COVERAGE

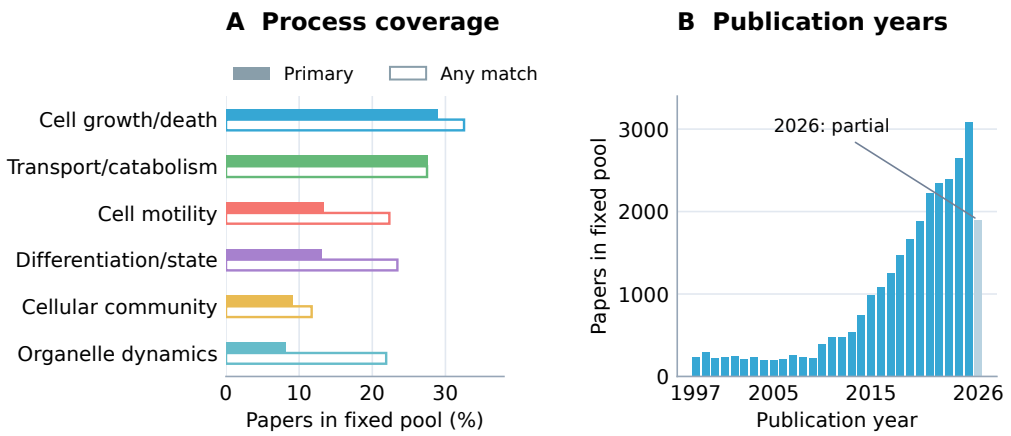

Figure 6: Coverage of the fixed 28,573-paper pool. (A) Filled bars show mutually exclusive primary process assignments; outlined bars show any-match membership, which can overlap (32.7% of papers have multiple categories). Both use all pool papers as the denominator. (B) Publication years from the paper index; 2026 is incomplete because the snapshot ends on 31 August 2026.

Table 5: Process-category counts in the fixed pool. Primary assignments sum to 28,573; any-match counts overlap.

Category	Primary	Any-match
Cell growth and death	8,277	9,299
Transport and catabolism	7,853	7,853
Cell motility	3,806	6,382
Cell differentiation and state	3,719	6,695
Cellular community (eukaryotes)	2,608	3,346
Organelle organization and dynamics	2,310	6,257

Mechanism evidence modes overlap: 28,021 papers contain explicit Results relations, and 19,773 contain process–perturbation–outcome windows.

B.6 CITATIONS WITHIN THE POOL

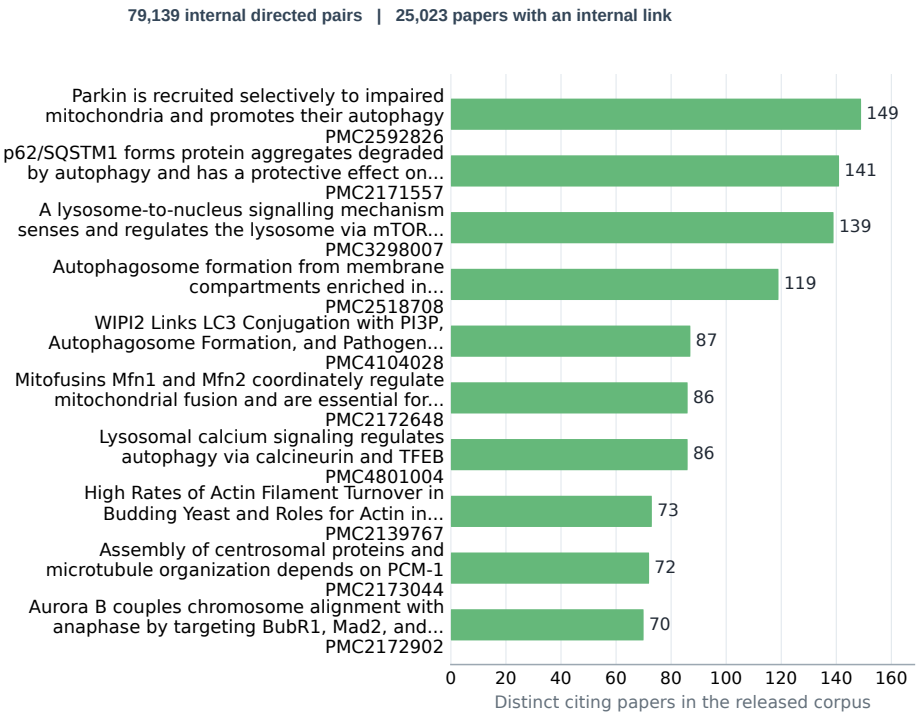

Figure 7: Most cited papers within the frozen pool, ranked by distinct citing papers in the same pool (PMCID labels). Exact target-PMCID matches in the reference table yield 79,139 unique directed paper pairs involving 25,023 papers, after removing self-links. DOI/PMID-only rows are not added to this internal count. The source table contains 1,808,175 identifiable reference rows. This is not a global impact measure.

B.7 TEXT-MATCHING RESOURCES

Text matching uses frozen lexicons and normalized token boundaries, ignoring case and treating spaces and hyphens alike. A lexical match alone is not sufficient: section and cross-component requirements must also hold.

Table 6: Configured text-matching resources (unique entries). This table lists the text resources, not the full acquisition inventory.

Resource	Count
Cellular-process categories	6
Unique process terms	101
Mechanism/relation terms	100
Experimental models	43
Perturbations	27
Assays/methods	44
Outcomes	37

B.8 FROM THE POOL TO EVALUATION EVIDENCE

Only textual experimental result statements enter generation states E_t. Trajectory construction also requires the four continuity checks described in the main text. Most pool papers do not enter a

verified trajectory. Formation and validation roles require documentary support; uncertain roles remain unresolved.

B.9 Limits of use

The pool supports time-ordered textual evidence and trajectory construction. It does not provide exhaustive coverage of cell biology, rank journals, or replace expert checks of scientific transitions. Redistribution must follow the article licenses. Train/test splits must account for shared papers and citation links.

C Models

Table 7: Model panel actually used in the evaluation. The main transition experiment scores five frontier endpoints with a disjoint judge; the deletion diagnostic (Appendix D) probes 12 endpoints (one overlapping with the main panel) to select the primary observables. No model outside this panel was evaluated.

Model	Provider	Role
gpt-5.6-luna	OpenAI	Main evaluation + deletion diagnostic
qwen3.8-flash	Qwen	Main evaluation
deepseek-v4-flash	DeepSeek	Main evaluation
glm-5.3-flash	Zhipu	Main evaluation
gemini-3.8-flash	Google	Main evaluation
claude-opus-4-8	Anthropic	Deletion diagnostic only
claude-sonnet-5	Anthropic	Deletion diagnostic only
gemini-3.1-pro-preview	Google	Deletion diagnostic only
glm-5.3	Zhipu	Deletion diagnostic only
kimi-k3	Moonshot	Deletion diagnostic only
gpt-5.5	OpenAI	Deletion diagnostic only
gpt-5.6-terra	OpenAI	Deletion diagnostic only
gpt-6-astra	OpenAI	Deletion diagnostic only
deepseek-v4-pro	DeepSeek	Deletion diagnostic only
grok-4.6	xAI	Deletion diagnostic only
qwen3-max	Qwen	Deletion diagnostic only

C.1 Trajectory recovery does not collapse

Across the 10 trajectories (22 steps total), hypothesis recovery at later steps remains of the same order as the first step: pooled first-step F1 0.442 versus later-step F1 0.428 (per-model first-step 0.368–0.497, later-step 0.361–0.541; no model shows monotone decay to floor). The trajectory task therefore does not collapse to a single endpoint—models sustain state reconstruction along the chain—supporting trajectory-level reconstruction as a non-degenerate evaluation object rather than a repeated single-state readout.

C.2 Evidence-deletion ablation

The same three cases support a second diagnostic: does recovery shift when a single formation-relevant evidence statement is removed, or replaced by an unrelated control card of equal cardinality? For each case we take the documented intervention from the frozen transition plan (Appendix A): `sed26-0034` removes its core evidence statement (`ER01`); `sed26-0006` and `sed26-0013` each remove a supporting statement (`E007` and `ER03`, respectively). Three conditions—full pool (`F_full`), deletion (`L_delete`), and replacement (`R_replacement`)—are compared under identical target-blind open-book generation.

Table **??** reports framework recovery. The intervention is deliberately minimal, so the expected mechanism-level effect is small; the informative question is *where* sensitivity concentrates.

- **Mechanism-level recovery is robust to single-evidence intervention.** Deleting or replacing one statement leaves the framework rate essentially unchanged for two of three cases (`sed26-0034` and `sed26-0013` hold at 1.00); the aggregate moves $0.73 \rightarrow 0.87$, within repeats noise.

- **Entity-level sensitivity is real but small, and concentrated on core evidence.** For `sed26-0034`, deleting the core FSP1 statement drops exact recovery from 1.00 to 0.60—the model occasionally re-anchors on Ezh2 or fatty-acid-oxidation evidence—while replacement with an unrelated control restores exact 1.00. The direction matches the expectation that core evidence is load-bearing for entity-level attribution, but at five repeats the $3/5$ vs $5/5$ difference is not significant.

- **Supporting-evidence removal is inert at both levels.** `sed26-0013` (removing a kinesin anterograde-transport statement that is background to the retrograde hypothesis) shows no change in any condition. The single anomalous rise (`sed26-0006` $0.2 \rightarrow 0.6$) is a $1/5 \rightarrow 3/5$ fluctuation, not a directional signal.

D PILOT DIAGNOSTIC

This appendix records a deletion study used to diagnose design confounds and to select the primary observables. It complements, and does not duplicate, the main transition experiment.

D.1 EVIDENCE-UNIT DESIGN

The pilot assembled, for each case, a mixed pool that included (i) textual experimental result statements and (ii) caption-linked descriptive statements associated with figures. Two observations led us to drop image-descriptive evidence from the main benchmark:

- **Redundancy and dependence.** Designated result statements are often partially recoverable from other same-source cards (Spearman correlation between same-source token coverage and residual recovery after deletion ≈ 0.46). Caption-linked descriptive text is frequently dependent on, or paraphrastic of, the same underlying experiment as the result statements. Treating both as independent evidence units inflates apparent support and blurs formation relevance.

- **Primary scientific signal.** For hypothesis formation, the operative unit is the experimentally stated result, not the figure caption's descriptive gloss. Restricting the main benchmark to textual experimental results reduces same-source double-counting and keeps formation/supporting/validation roles better specified.

Exploratory expert scores on designated *result* cards (mean 2.8, range 1–4) were used only as a coarse relevance check for those cards. They are not a full-pool gold standard and are not used to score model outputs. Caption-card scores are not used.

D.2 PROTOCOL (BRIEF)

20 cases; 12 frontier model endpoints; conditions: full pool, pool minus designated key result, and a non-key contrast deletion. Target-blind structured generation with fixed schema. 720 generation files; 680 structurally valid after schema and ID checks. Incomplete model–case cells exist and are marked; we do not emphasize unequal case counts as a ranking factor.

D.3 PER-MODEL DESCRIPTIVE RECORD

Table 8 lists descriptive statistics only. Coverage is overlap with a probe hypothesis string under the full-evidence condition (not a quality gold standard). $\mathrm{dist}_{\mathrm{key}}$ and $\mathrm{text_sim}_{\mathrm{key}}$ describe behavior after key-result deletion. No model ranking is intended.

Table 8: Pilot per-model descriptive record. †: incomplete cells. Probe coverage is a behavioral descriptor, not a quality score.

Model	Cov. (Jac.)	Len-adj.	$\mathrm{dist}_{\mathrm{key}}$	$\mathrm{text_sim}_{\mathrm{key}}$
claude-opus-4-8†	0.057	-0.002	0.787	0.234
claude-sonnet-5†	0.061	-0.001	0.792	0.195
gemini-3.1-pro-preview	0.126	$+0.049$	0.668	0.388
glm-5.3†	0.107	$+0.052$	0.745	0.290
kimi-k3†	0.087	-0.023	0.765	0.254
gpt-5.5	0.088	-0.011	0.727	0.266
gpt-5.6-luna	0.101	$+0.003$	0.725	0.279
gpt-5.6-terra	0.085	-0.008	0.754	0.282
gpt-6-astra	0.078	-0.024	0.646	0.387
deepseek-v4-pro†	0.109	-0.001	0.718	0.344
grok-4.6	0.094	-0.024	0.711	0.325
qwen3-max	0.114	$+0.003$	0.661	0.395

D.4 What the pilot implies for the main evaluation

Single-card deletion often fails to produce a sharp change in probe overlap. Residual recovery is partly explained by same-source redundancy, but not entirely. Text and citation pathways decouple (mean text similarity after key deletion ≈ 0.28 vs. restricted citation similarity ≈ 0.62). Cross-case ranking agreement is low. Selective-sensitivity interpretations of single-card deletion are therefore not supported.

These facts support: (1) textual experimental results as the primary unit; (2) rejecting deletion rankings as a primary signal; (3) requiring paired generation in any transition experiment; (4) treating only documentary-gated historical transitions as evaluation items. Candidate construction scale is not historical ground truth.

E Output schema

The scientific-state extractor emits one object per paper: a paper-level status plus three lists (evidence, hypotheses, typed E–H links). Every quote field is re-anchored to a source span and rejected if not locatable verbatim.

```
{
  "paper_status": "has_active_hypothesis | no_active_hypothesis",
  "evidence": [
    {
      "evidence_id": "E1",
      "verbatim_quote": "...",
      "statement_type": "observation | measurement | prior_finding |
                         mechanistic_evidence | boundary_evidence |
                         negative_evidence | contradictory_evidence |
                         contextual_evidence | background",
      "entities": ["..."],
      "inference_note": "..."
    }
  ],
  "hypotheses": [
    {
      "hypothesis_id": "H1",
      "verbatim_quote": "...",
      "origin_status": "introduced_here | prior_retained |
                         modified_here | contested_here",
      "first_proposed_date": "YYYY-MM-DD | null",
      "first_proposed_paper_id": "PMC... | null",
      "entities": ["..."],
      "inference_note": "..."
```

Table 9: Original-data human expert quality verification. Dimensions 1–4 follow the state or transition rubric in the text.

Unit and origin	n	D1	D2	D3	D4	Mean
Reference states	22	2.91	3.55	4.18	4.32	3.74
Agentic states	22	3.64	3.46	3.18	4.36	3.66
Reference transitions	15	3.87	3.47	4.13	3.73	3.80
Agentic transitions	26	3.77	3.39	3.81	3.04	3.50

```
    }
  ],
  "eh_links": [
    {
      "evidence_id": "E1",
      "hypothesis_id": "H1",
      "relation_type": "forms | supports | tests | challenges | refines | unclear",
      "formation_status": "formation | validation | unclear",
      "source_support": "..."
    }
  ]
}
```

F Interpretation boundaries

- Historical scientific-state evolution is not a counterfactual intervention.
- The benchmark reconstructs documented evolution; it does not intervene on science.
- Level-3 directional alignment is required for the strongest claims; displacement alone is insufficient.
- Behavioral alignment is not internal causal proof.
- Accessibility proxies are not memorization measurements.
- Claims are limited to measurable, directionally characterized correspondence between model-state change and historical scientific-state change.

G Human verification

Construction-side quality control proceeds in two layers. *Implementation check:* a 100-paper stratified pilot under the active-hypothesis schema re-anchors every quote to its source span (99.5% verbatim, no fabricated spans), confirming the extraction contract before scale-out. *Independent documentary-validity audit:* a stratified 300-item sample (100 hypotheses, 100 E–H links, 50 evidence statements, 50 provenance-boundary cases) is judged by an auditing model blinded to the constructor's rationales, yielding evidence validity 1.00, hypothesis validity 0.93, and documentary E–H validity 0.76 (23% of links lack an explicit formation/support/validation route). Priority expert verification of the top transition candidates confirms the four gates, formation relevance, hypothesis identity, and transition type; annotators do not see model outputs when judging continuity or relevance.

We additionally engaged human experts to conduct a criterion-referenced quality verification. Each record was scored on four 1–5 dimensions, with a mean of at least 3.0 as the recorded pass threshold. State dimensions were evidence fidelity, hypothesis validity, bucket correctness, and context adequacy; transition dimensions were update correctness, evidence linkage, mechanism accuracy, and absence of over-inference. This process covered all 22 reference states, 22 sampled agentic states, all 15 reference transitions, and the 26 agentic transitions incident to the sampled states. Table 9 reports the original-data results.