POSITION: SCIENTIFIC MEMORY IS NOT CHAT HISTORY: REFRAMING LONG-TERM MEMORY EVALUATION FOR RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory evaluation for language agents primarily measures preservation and retrieval of prior information across extended interaction horizons, but not how information changes across multi-session scientific interaction. This position paper argues for *scientific memory* as a distinct evaluation target since scientific interaction depends on preserving evolving information states under *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* as new information appears over time. Correctness is therefore defined by the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ rather than conversational consistency or retrieval alone. Prior benchmarks evaluate important components of *scientific memory*, yet do not evaluate whether an agent maintains one evolving information state across multiple sessions. We therefore argue that *scientific-memory* evaluation must distinguish *retrieval failure, memory-update failure, and reasoning/provenance failure* rather than collapse behavior into final-answer accuracy. To operationalize this position, we present a benchmark-oriented taxonomy of *scientific-memory* capabilities and failure modes and a five-condition diagnostic protocol (5DP) spanning *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning*. Across 160 items and 5 systems, the model judge found that 19.8% of responses passing selected source, grounding, and fully-correct-answer checks failed joint information-state requirements (95% interval $[15.1, 24.7]$), while the smaller human-calibration sample yielded a lower and uncertain gap estimate.

1 INTRODUCTION

Long-term memory has become a central objective in language-agent research as emerging deployments assume interaction beyond a single session (He et al., 2026; Hu et al., 2026; Yehudai et al., 2026). Agents are therefore expected to preserve historical information (Wu et al., 2024; Maharana et al., 2024), previously observed information (Zhao et al., 2026a; Wu et al., 2024), and ongoing tasks over time (He et al., 2026; Hu et al., 2026). Prior research on personal conversation (Wu et al., 2024; Maharana et al., 2024), episodic recall (Shu et al., 2026; Hu et al., 2026), long-context question answering (Liu et al., 2024; Bai et al., 2024; Hsieh et al., 2024), and retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2023; Katsis et al., 2025) is largely motivated by this requirement, with evaluation primarily organized around preservation and retrieval of prior information (Zhao et al., 2026a; He et al., 2026; Wu et al., 2024). Formally, given $\mathcal{H}_{<t} = \{u_1, \ldots, u_{t-1}\}$, evaluation typically considers whether response a_t remains consistent with $\mathcal{H}_{<t}$ or retrieved information $\mathcal{R}_t$, often approximated by $\max P(a_t \mid \mathcal{H}_{<t}, \mathcal{R}_t)$[1]. Long-context benchmarks (e.g., Bai et al., 2024; Hsieh et al., 2024) evaluate preservation as context length increases, often from $|\mathcal{C}| \ll 10^3$ to $|\mathcal{C}| \gg 10^5$ tokens; retrieval-augmented generation benchmarks (e.g., Lewis et al., 2020; Katsis et al., 2025) evaluate whether retrieved information supports responses through $\mathcal{R}_t = f(q_t, \mathcal{D})$; episodic-recall benchmarks (e.g., Shu et al., 2026; He et al., 2026; Hu et al., 2026) evaluate preservation of interaction histories and prior actions; and long-term dialogue benchmarks (e.g., Wu et al., 2024; Maharana et al., 2024; Zheng et al., 2026) evaluate temporal updates across sessions. Thus, although these benchmarks share access to history or retrieved information, their metrics measure different aspects

[1]The cited benchmarks apply their own task-specific scoring criteria (Hu et al., 2026; Wu et al., 2024; Yehudai et al., 2026).

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.

of response quality, and the conditional generation expression alone does not specify whether an evaluation captures how information changes over time.

This limitation becomes particularly important in research environments, where scientific interaction depends on how information changes over time rather than only on preservation and retrieval of prior information (Lu et al., 2024; Moussa et al., 2025; Yehudai et al., 2026). Information evolves across multiple sessions through *hypotheses* (what the system proposes as possible informations), *evidence provenance* (how information originates), *contradiction* (how information conflicts with other information), *method and condition scope* (where and under what conditions information applies), *revision* (how later information modifies earlier information), *synthesis* (how information from multiple sources is combined), and *uncertainty* (how incomplete or conflicting information remains unresolved); we refer to preservation of this evolving information state as *scientific memory*. Let $\mathcal{E}_t$ denote the information state at interaction step t, containing these properties accumulated up to t. Under *scientific memory*, correctness depends on preserving the transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ as new information appears; a system may therefore retrieve correct information while producing scientifically incorrect information if these properties are not faithfully preserved after later information appears. Prior benchmarks (Wu et al., 2024; Maharana et al., 2024; Zheng et al., 2026; Dasigi et al., 2021; Lehman et al., 2019; DeYoung et al., 2020; Wadden et al., 2020; Thorne et al., 2018; Gao et al., 2023; Min et al., 2023; Lu et al., 2024; Zhao et al., 2026b; Varvarigos et al., 2026; Moussa et al., 2025) evaluate important components of *scientific memory*, including *temporal updating* (how later information modifies earlier information), *support and refutation* (how evidence supports or contradicts a claim), *evidence provenance* (how information originates), and *scientific reasoning* (how evidence is used to derive conclusions), but do not evaluate whether an agent maintains one evolving information state $\mathcal{E}_t$ under these properties across multiple sessions. Thus, current settings may measure preservation and retrieval through objectives such as $\max P(a_t \mid \mathcal{H}_{<t}, \mathcal{R}_t)$ while still failing to evaluate the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ over time.

Consider a researcher using an agent to track Fenton chemistry (a study of iron-mediated oxidation reactions involving hydrogen peroxide) across multiple sessions. Early sessions may support the classical hydroxyl-radical interpretation of acidic Fenton reactions, while later sessions may support aqueous Fe(IV)=O intermediates and pH-dependent mechanistic transitions, giving information states $\mathcal{E}_1 \neq \mathcal{E}_2$. When the researcher later asks whether hydroxyl radicals or Fe(IV)=O dominate at pH 5.0, the system may retrieve earlier hydroxyl-radical information and generate a conclusion. However, if the available information does not specify the transition threshold or ligand conditions required to determine the mechanism at pH 5.0, the information state remains unresolved. Therefore, retrieval alone is insufficient since scientific information depends on preserving *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* across multiple sessions.

Our position is therefore that *scientific-memory* evaluation must distinguish *retrieval failure* (how information is retrieved), *memory-update failure* (how later information fails to update earlier information), and *reasoning/provenance failure* (how information is used or attributed incorrectly) rather than collapse them into final-answer accuracy alone. Research-agent should report whether conclusions preserve the current evidence state alongside retrieval and answer accuracy. The taxonomy and 5DP provide one operationalization of what the community should measure, while the argument does not depend on a particular memory architecture or benchmark.

To operationalize this position, we present a benchmark-oriented taxonomy of *scientific-memory* capabilities and failure modes and a five-condition diagnostic protocol (5DP) spanning *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning* (see Table 2). The taxonomy organizes *scientific-memory* evaluation under *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty*, while 5DP distinguishes *retrieval failure, memory-update failure, and reasoning/provenance failure* across multi-session scientific interaction.

2 The Evaluation Object: Scientific Memory vs. Chat History

Chat-history memory and *scientific memory* differ in what is preserved and how correctness is evaluated. Chat-history memory primarily evaluates whether a response a_t remains consistent with prior interaction history $\mathcal{H}_{<t}$ or retrieved information $\mathcal{R}_t$ (Wu et al., 2024; Maharana et al., 2024; Gao et al., 2023), whereas *scientific memory* evaluates the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ across multiple interaction steps. Correctness therefore depends on preserving *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* as later information appears. *Scientific-memory* is conditional on *method and condition scope* since information obtained under one experimental setting may not transfer to another; it depends on *evidence provenance* since source identity, evidence span, figure, table, or dialogue turn may alter interpretation

(Lebo et al., 2013); and it changes over time as later information revises earlier information, producing $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$ and connecting *scientific memory* to belief-revision and truth-maintenance formulations (Alchourrón et al., 2016; Doyle, 1979). *Contradiction* may remain unresolved across sessions, *uncertainty* may remain faithful when information is incomplete or conflicting, and *synthesis* may require integrating information across multiple sources and interaction steps rather than single-turn retrieval. A system may therefore retrieve relevant information while producing scientifically incorrect information if it merges contradictory information into false agreement, transfers information outside valid *method and condition scope*, preserves outdated information after *revision*, or converts unresolved information into confident completion. Thus, long context alone does not remove the need for *scientific memory*, even when relevant information remains accessible through context windows or retrieval systems (Liu et al., 2024; Bai et al., 2024; Packer et al., 2023; Park et al., 2023), evaluation must measure $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ over time.

Consider a researcher using an agent to track Alzheimer's disease biomarkers across multiple sessions. One session may support plasma phosphorylated tau as an early diagnostic biomarker, while another reports reduced diagnostic reliability under specific age groups and comorbidity conditions, followed by conflicting sensitivity estimates across independent clinical cohorts. Let the corresponding information states be $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, where later information modifies earlier information across multiple sources. When the researcher later asks whether plasma phosphorylated tau is a reliable early-stage diagnostic biomarker, a system may retrieve highly cited earlier information and generate a definitive conclusion. However, unresolved contradiction across cohorts, differing cohort conditions, and inconsistent sensitivity estimates leaves the information state unresolved. Therefore, retrieval of highly relevant information alone is insufficient since *scientific memory* depends on preserving *evidence provenance, contradiction, synthesis, revision, and uncertainty* across multiple sessions and sources.

We make $\mathcal{E}_t$ as claim records linking *evidence provenance, contradiction, synthesis, revision, and uncertainty* (see Appendix F). A Markdown file versioned with Git can implement this record, with textual diffs identifying edits and evaluation testing whether the edits and resulting answer are justified by the evidence. Fixed model weights do not prevent runtime updates to files, retrieved context, or external memory, as illustrated by MemGPT (Packer et al., 2023). This separates the evaluation target from the storage format and requirement for online weight updates.

3 Why Current Benchmarks Do Not Measure Scientific Memory

As discussed in §1, our position is not that these benchmark families are insufficient individually, but that they do not jointly evaluate whether information remains faithful to the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ across multiple sessions. The evaluation gap therefore concerns whether these properties remain preserved together over time beyond selected component checks. This position does not require *temporal updating* (how later information modifies earlier information) or *abstention* (when a system declines to provide a definitive information) to be absent from prior benchmarks; LongMemEval (Wu et al., 2024), for example, evaluates both. The additional requirement is to evaluate them together with *evidence provenance, contradiction, synthesis, revision, and uncertainty* on the same evolving scientific question. Passing selected retrieval, grounding, and answer checks may therefore leave relevant preservation requirements unmeasured. We use the *compositional gap* (see §6) in this operational sense, not as a claim that individually satisfied logical conditions can fail their conjunction. Appendix G compares selected benchmarks by evaluation unit, prior strengths, and the additional transition-level check proposed here, while Appendix F shows how an accurate statement about an earlier experiment can omit a later scope restriction. This illustrates the coverage argument without assuming a universal deficit in prior benchmarks or a logically indivisible capability.

4 A Taxonomy of Scientific-Memory Capabilities

The taxonomy in Table 1 organizes *scientific-memory* evaluation by whether information remains correct to the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ across interaction steps. A research agent must preserve *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* as later information appears, so evaluation depends on whether later information modifies earlier information consistently within $\mathcal{E}_t$, not only on preservation and retrieval. *Contradiction* and *synthesis* are composite *scientific-memory* capabilities since they jointly depend on preservation of *hypotheses, evidence provenance, revision, method and condition scope, and uncertainty* across sessions and sources. The taxonomy provides a benchmark-oriented decomposition rather than a complete theory of scientific reasoning and separates *scientific-memory* capabilities from failure modes. *Hypothesis drift* occurs when a system changes the preserved hypothesis without

Table 1: Benchmark-oriented taxonomy for *scientific-memory* evaluation; checkmarks indicate components preserved within the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$.

***Scientific-Memory* Capability**	**Hypothesis**	**Evidence provenance**	**Contradiction**	**Method / Condition**	**Revision**	**Synthesis**	**Uncertainty**	***Scientific-Memory* Failure Mode**	**Representative Benchmark Settings**
Hypothesis Tracking	✓				✓			Hypothesis Drift	(Wu et al., 2024; Maharana et al., 2024)
Evidence Provenance		✓						Evidence-Provenance Confusion	(Dasigi et al., 2021; Wadden et al., 2020; Gao et al., 2023)
Contradiction Preservation			✓				✓	Contradiction Blindness	(Thorne et al., 2018; Wadden et al., 2020)
Revision Preservation					✓			Outdated-Information Persistence	(Wu et al., 2024; Maharana et al., 2024; Zheng et al., 2026)
Method and Condition Scope		✓		✓				Method-and-Condition Mismatch	(Lehman et al., 2019; DeYoung et al., 2020; Luan et al., 2018; Jain et al., 2020)
Synthesis		✓	✓	✓	✓	✓		Unsupported Information Recall	(Dasigi et al., 2021; Lu et al., 2020; Gao et al., 2023)
Uncertainty Preservation			✓				✓	Uncertainty Collapse	(Wu et al., 2024; Rajpurkar et al., 2018; Thorne et al., 2018; Kamath et al., 2020; Guo et al., 2017)

recognizing the change; *unsupported information recall* when it preserves information supporting a conclusion while failing to preserve the *evidence provenance, contradiction, revision, or method and condition scope* required for that conclusion; *evidence-provenance confusion* when information is attributed to the wrong source, experiment, figure, table, or interaction step; *contradiction blindness* when contradictory information remains unrecognized across sessions despite accessible information states; *outdated-information persistence* when earlier information continues to dominate after later information revises or weakens it; *method-and-condition mismatch* when information obtained under one *method and condition scope* is transferred incorrectly to another setting; and *uncertainty collapse* when incomplete, conflicting, or unresolved information is converted into definitive information despite insufficient information within $\mathcal{E}_t$.

5 Evaluation Framework and Operationalization

5.1 Design Principles

Evidence Provenance and Contradiction. Under *scientific-memory* evaluation, information should remain correct to the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ and its *evidence provenance*, remaining linked to the correct source, experiment, figure, table, evidence span, or interaction step rather than only to conversational consistency with $\mathcal{H}_{<t}$ or retrieval through $\mathcal{R}_t$. This connects *scientific-memory* evaluation to evidence-grounded question answering, claim verification, citation, and factuality evaluation (Dasigi et al., 2021; Wadden et al., 2020; Thorne et al., 2018; Gao et al., 2023; Min et al., 2023; Li et al., 2023), while extending these settings across evolving information states $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$. *Contradiction* should likewise remain preserved rather than collapsed into false agreement since contradictory information may remain unresolved across interaction steps. Thus, a system may retrieve highly relevant information while producing scientifically incorrect information if *contradiction* disappears during *synthesis*, *revision*, or retrieval across evolving information states.
Revision and Evolving Information States. *Scientific-memory* evaluation depends on whether later information *revises* earlier information consistently within one evolving information state $\mathcal{E}_t$. Benchmark settings should therefore contain revision events where later information modifies earlier information through failed replication, corrected measurement conditions, contradictory findings, revised experimental settings, stronger null results, or retracted conclusions. This extends temporal-update evaluation in long-term dialogue and long-horizon memory benchmarks (Wu et al., 2024; Maharana et al., 2024; Zheng et al., 2026) from conversational interaction to evolving information states $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$. Evaluation should measure both accessibility of relevant information at interaction step t and consistent modification of earlier information; retrieval alone is insufficient since a system may retrieve earlier information correctly while failing to preserve *revision* within $\mathcal{E}_t$.

Method and Condition Scope. *Scientific-memory* evaluation depends on preserving *method and condition scope* across evolving information states. This scope determines whether information obtained under one experimental setting, or evaluation configuration remains valid under another. A system may preserve otherwise correct information yet produce scientifically incorrect information by transferring it across invalid *method and condition scope*. Benchmark settings should therefore evaluate preservation of *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* within evolving information states. For the additional diagnostic results in §5.4, Φ denotes whether a response passes all applicable information-state requirements; Appendix B specifies the scoring gates, and this joint score supplements rather than replaces component-wise metrics.

Judge Validation. *Scientific-memory* evaluation relies on the model judge to assess preservation components requiring interpretation across the evolving information state. We therefore validate whether the judge detects violations of individual components before using its scores for joint information-state. We construct synthetic responses from expert gold answers by preserving the gold answer verbatim and introducing six targeted corruptions, each violating one preservation component while preserving the others. Each corruption is evaluated on items carrying the corresponding property, while the judge's firing rate on the unmodified gold response measures false positives. This tests whether the judge distinguishes a preserved information state from a response violating a specific component of Φ (see Appendix C). We additionally grade a stratified 120-response subset with two independent annotators to calibrate the judge; its small, unpaired condition cells permit only a limited check of condition-level patterns (see Appendix D).

5.2 Scoring and The Five-Condition Diagnostic Protocol (5DP)

Information-State Records and Scoring. Under *scientific-memory* evaluation, benchmark settings should make the evolving information state $\mathcal{E}_t$ through records containing *hypotheses*, information status, supporting and *contradictory information*, *evidence provenance*, *method and condition scope*, *revision trace*, *uncertainty state*, and answerability state. Information status may use labels such as *supported*, *contradictory*, *mixed/conditional*, *insufficient*, or *out-of-scope*; provenance should link to the correct source, experiment, figure, table, evidence span, or interaction step, while *method and condition scope* specify where information remains valid. The contribution is not a new ontology but an evaluation protocol for $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$; systems may use arbitrary internal memory representations, while evaluation measures preservation of the observable evolving information state. Records should align where possible with PROV-O and nanopublications (Lebo et al., 2013; Groth et al., 2010), while *revision trace* connects to belief-revision and truth-maintenance formulations (Alchourrón et al., 2016; Doyle, 1979). *Scientific-memory* evaluation should decompose information-state rather than final-answer accuracy, using information-status accuracy, *evidence-provenance* precision/recall/F1, *evidence-span overlap*, *method-and-condition-scope* accuracy, *revision-preservation* accuracy, *contradiction-preservation* accuracy, *uncertainty-state* accuracy, *answerability-state* accuracy, and *uncertainty-collapse* frequency (see Appendix §A). Under *synthesis* evaluation, supporting and contradictory information should remain jointly preserved across multiple papers, interaction steps, and information sources rather than flattened into unsupported synthesis.

For the Fenton chemistry example, a compact evolving information-state record at interaction step t may be represented as:

```
 interaction_step:  t
hypothesis:  dominant oxidant under Fenton chemistry at pH 5.0
information_status:  mixed/conditional
supporting_information:  [Walling1975, Pestovsky2005]
contradictory_information:  [Bataineh2012]
evidence_provenance:
  Walling1975 → hydroxyl-radical interpretation under acidic conditions
  Pestovsky2005 → spectroscopic evidence supporting Fe(IV)=O intermediates
  Bataineh2012 → pH-dependent mechanistic transition unresolved at pH 5.0
method_and_condition_scope:
  assay:  spectroscopic identification
  reaction_condition:  aqueous Fenton chemistry
  pH: 5.0
  ligand_condition:  unspecified
revision_trace:
  E1:  hydroxyl-radical interpretation under acidic conditions
  E2:  Fe(IV)=O intermediates introduced with pH-dependent transition
uncertainty_state:  unresolved mechanism at pH 5.0
answerability_state:  insufficient
scientific_memory_failure_risk:  uncertainty collapse under unsupported synthesis
```

Table 2: Five-condition diagnostic protocol (5DP) for localizing *retrieval, memory-update, and reasoning/provenance failures* under evolving information states $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$.

Diagnostic Condition	What is Supplied	Failure Localized
Online Free Retrieval	Unrestricted retrieval with full evolving interaction history.	Retrieval, memory-update, and reasoning/provenance failure across evolving information states.
Offline Free Retrieval	Retrieval restricted to earlier interaction steps without future-information leakage.	Future-information leakage removed; retrieval, memory-update, and reasoning/provenance failure remain observable.
Gold-Source Retrieval	Gold source identifiers and evidence text.	Source access controlled; constructing the state and reasoning from the evidence remain necessary.
Gold-Memory-State Evaluation	Gold evolving information state $\mathcal{E}_t$ without source text.	State usability without supporting text; this does not by itself isolate memory-update failure.
Gold-Evidence Reasoning	Gold evidence text with the evolving information-state record, including provenance, scope, revision, uncertainty, and answerability.	Residual reasoning/provenance or response-expression failure after supplying both evidence and state.

Five-Condition Diagnostic Protocol (5DP). *Scientific-memory* evaluation should distinguish *retrieval failure, memory-update failure, and reasoning/provenance failure* across evolving information states. The five diagnostic conditions in Table 2 provide controlled comparisons with different information access: *online free retrieval* evaluates unrestricted retrieval with evolving interaction history; *offline free retrieval* evaluates by removing future-information leakage while retaining retrieval over earlier information states; *gold-source retrieval* evaluates by controlling source access through gold source identifiers and evidence text containing supporting and contradictory information; *gold-memory-state evaluation* evaluates by supplying $\mathcal{E}_t$ without source text to test whether the state alone supports correct responses; and *gold-evidence reasoning* evaluates by supplying gold evidence text and the evolving information-state record, including *evidence provenance, method and condition scope, revision trace, uncertainty*, and answerability. Non-decreasing performance is an expectation to test only in matched comparisons that retain evidence while adding a state, not an assumed ordering of all five conditions. Deviations across conditions can localize failure in *retrieval/preservation* of evolving information states, *evidence provenance, contradiction, method and condition scope, revision, synthesis, or uncertainty*. For an unresolved Fenton question, evidence-alone versus evidence-plus-state tests whether the state helps preserve uncertainty, while state-only tests usability without source text (see Appendix §A). The matched evidence-plus-state comparison retains available evidence, while replacing evidence with a state record separately tests whether the record is usable without supporting text. Improvements after supplying a state are diagnostic evidence, not a unique causal separation of updating and reasoning. The additional results distinguish evidence alone, state alone, and evidence plus state are provided in Appendix B, where monotonic improvement is an empirical expectation, not a guarantee for each response.

Annotation and Validation. A minimal annotation protocol should label *hypotheses*, information status, supporting information, *contradictory information, evidence provenance, method and condition scope, revision trace, uncertainty state*, answerability state, and whether *contradiction* remains preserved across evolving information states. Annotators should maintain compact traces $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$ across interaction steps. If experts disagree, benchmark settings should preserve disagreement through *mixed/conditional*, *contradictory*, or unresolved information states rather than force one definitive state. A new search result need not change a hypothesis; updates should depend on new, relevant evidence and its methodological quality, not a fixed number of papers. Rules for duplicate sources, corrections, conceptual reframing, and multiple admissible expert interpretations appear in Appendix F; these are construction guidelines, not validated automation. Appendix I reports a bounded screening result and explains why expert review and an audit of unflagged items remain necessary.

5.3 What a Scientific-Memory Benchmark Would Require

A future *scientific-memory* benchmark would require multi-session settings where later information modifies earlier information across evolving information states $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$. Construction would require *hypotheses, supporting and contradictory information, evidence provenance, method and condition scope, revision trace, uncertainty state*, and answerability state to remain observable across transitions $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$. Benchmark settings would therefore require multi-session evidence arcs and

evolving information-state records aligned with the *scientific-memory* capabilities and failure modes in Table 1. Computational biology, chemistry, and materials science are natural initial domains (see Appendix B) since they contain evolving information states with *evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* across multiple interaction steps.

Construction Audit. As bounded evidence that evolving information-state annotation is operationally feasible, we conducted a construction audit across computational biology, chemistry, and materials science (see Appendix A). The audit contained 30 multi-session evidence arcs and 450 dialogue-grounded scientific question-answering items requiring preservation of *supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, uncertainty state*, and answerability state across $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$. The arcs used temporally ordered interaction settings where later information modified earlier information through *contradiction*, updated *method and condition scope*, or unresolved *synthesis*. Stratified expert review of 300 items produced 122 accepted without revision, 173 requiring revision, and 5 rejected for unresolved evolving information-state inconsistency. The dominant annotation failure was unsupported insertion of information absent from cited evidence, including numeric values, *evidence provenance, method and condition scope, revision*, and *synthesis* across supporting and contradictory information. Expert review found no systematic mismatch between annotation categories and evolving information-state requirements, supporting operational use of the proposed *scientific-memory* decomposition across all three domains (see Appendix §A). Optional GLM-assisted triage agreement with expert review was 0.950, 0.927, and 1.000 for item-decision, evidence-support, and category labels, with Cohen's κ of 0.902, 0.864, and 1.000, respectively (see Appendix §A). Expert review remained authoritative; GLM-assisted agreement served only as bounded annotation-triage evidence, not benchmark supervision. The dominant failure reflected loss of fidelity to $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ rather than only conversational inconsistency with $\mathcal{H}_{<t}$ or retrieval failure under $\mathcal{R}_t$. Even with *evidence provenance* and evolving information-state records, unsupported *synthesis* and *uncertainty collapse* remained frequent when *supporting information, contradictory information, revision trace, and method and condition scope* were not jointly preserved across interaction steps.

Controlled Diagnostic Study. To operationalize the five-condition diagnostic protocol (5DP), we conducted a controlled study over 12 expert-reviewed construction-audit questions under *partial-source*, *gold-source*, and *gold-memory-state* conditions (see Appendix §A). We evaluated `moonshotai/kimi-k2-thinking`(Team et al., 2026) and `nvidia/llama-3.3-nemotron-super-49b-v1.5`(Bercovich et al., 2025); GLM-5.1(Zeng et al., 2026) was excluded because GLM-assisted outputs were used in annotation-triage analysis. This study is methodological trace material, not a benchmark-scale model comparison, illustrating how *retrieval failure, memory-update failure, and reasoning/provenance failure* become distinguishable under constrained evolving information states. In the Fenton chemistry example, *gold-source* still produced unsupported hydroxyl-radical completion at pH 5.0 under unresolved *method and condition scope*, whereas *gold-memory-state* preserved `answerability_state = insufficient` with unresolved *uncertainty* under $\mathcal{E}_t$. Since the pilot's state condition supplied assessment information, this contrast is illustrative rather than independent evidence of successful memory updating; scored comparisons omit gold-answer prose (see Appendix §B). The 12-question study is therefore retained as an illustrative pilot and separated from the scored comparisons in §5.4, where the supplied state omits gold-answer prose and is evaluated with and without evidence text.

Scaling Considerations. Structured evolving information-state records containing *hypotheses, supporting and contradictory information, evidence provenance, method and condition scope, revision trace, uncertainty state, and answerability state* introduce annotation cost since these components must remain consistent across multiple interaction steps rather than isolated question-answering settings. Weak supervision, automated extraction, or language-model-assisted pre-annotation may reduce this cost but can introduce *unsupported information recall* (recalling information without supporting evidence), *evidence-provenance confusion* (attributing information to the wrong source), *method-and-condition mismatch* (transferring information beyond its valid *method or condition scope*), *contradiction blindness* (failing to recognize contradictory information), *outdated-information persistence* (retaining earlier information after it has been revised), or *uncertainty collapse* (converting unresolved information into definitive information), making expert validation necessary. Evaluation should compare long-context, retrieval-augmented, and memory-based systems under *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning*. This can localize *retrieval failure, memory-update failure, and reasoning/provenance failure* across these conditions and information-state complexities.

The proposed framework is architecture-agnostic since it specifies which information components must be preserved within evolving information-state transitions $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ and how preservation should be evaluated. Long-context, retrieval-augmented, graph-based memory, vector-memory, and tool-using systems can therefore be evaluated under the same *scientific-memory* protocol when they expose generated information with evidence provenance and evolving information-state outputs. A future *scientific-memory* benchmark would require annotation guidelines, evolving information-state records, scoring protocols, five-condition diagnostic reporting, and capability-wise evaluation beyond final-answer accuracy alone.

5.4 Additional Diagnostic Results

We provide two additional observations supporting the proposed evaluation target using 160 expert-reviewed items from 30 evidence arcs in computational biology, chemistry, and materials science across five systems (see Appendix B). These comparisons extend the bounded operational evidence above without establishing a general system ranking. Six controlled settings approximate a subset of 5DP through source-, state-, and evidence-access comparisons; unrestricted *online free retrieval* and complete 5DP validation remain future work. Appendix B details construction, conditions, and scoring; Appendices C and D report judge validation and human calibration. Across the six settings, 759 responses passed selected source-identifier, grounding, and fully-correct-answer checks, but 150 failed at least one additional *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, or uncertainty* requirement (19.8%, 95% item-level bootstrap interval $[15.1, 24.7]$); under *gold-source retrieval*, the corresponding count was $43/196$, and with *gold evidence retrieval* plus a state record, $36/235$. These results support evaluating preservation alongside retrieval and answer correctness, demonstrating a coverage gap in the selected checks rather than insufficiency of every prior benchmark or a new indivisible capability. Table 3 distinguishes evidence alone, state alone, and their combination. Under the model judge, state alone produced lower than evidence alone; in paired outputs, answerability-label accuracy increased from 86.4% to 97.6%, while judged uncertainty preservation decreased from 72.1% to 67.6%. Thus, a correct label does not establish that the generated explanation preserves the same uncertainty, motivating evaluation of both generated responses and stored states while retaining supporting evidence. On the same 120-response calibration subset, joint fidelity was 6.7% under the model judge versus 27.5% under agreement of both human annotators; the human-scored coverage gap was $3/36$ (8.3%, interval $[0.0, 23.1]$), with all three failures attributed to provenance. Absolute rates and failure composition therefore depend on the evaluator. The state-only deficit becomes inconclusive when the unsupported-claim gate is removed (see Appendix E). These results motivate component-wise reporting and human calibration, while their magnitudes should not be treated as settled estimates of failure in scientific deployments. Appendix H reports per-system and per-domain metrics and a 24-item author-constructed ML/TCS transfer pilot; Appendix J examines evidence-free disagreement, where the model judge reports increased unsupported assertions but no conclusive decrease in joint fidelity. These checks broaden the diagnostic evidence while leaving independent domain validation and consistency across real users unestablished. The observed failures are interpreted as *retrieval failure, memory-update failure, or reasoning/provenance failure* only within the controlled access conditions, including *offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning*; they do not establish their prevalence in real deployments.

Table 3: Selected controlled conditions from the additional evaluation. Joint fidelity is scored by the model judge; n counts graded responses. The state-only condition replaces evidence rather than forming an intermediate rung.

Information Supplied	n	Joint Fidelity Φ (%)
Gold evidence alone	786	19.6
Gold state alone	779	14.0
Gold evidence and state	789	25.3

6 Alternative Views and Responses

The Compositional Gap, Retrieval-Augmented Generation, and Adjacent Benchmarks. The strongest objection is that prior benchmark families already evaluate components of *scientific memory*: LongMemEval (Wu et al., 2024) evaluates temporal revision, SciFact (Wadden et al., 2020) evaluates support and contradiction, ALCE (Gao et al., 2023) evaluates citation grounding; and recent scientific-agent (Zhao et al., 2026b; Varvarigos et al., 2026; Moussa et al., 2025; Fa et al., 2026) and memory-agent (He et al., 2026; Hu et al., 2026; Shu et al., 2026) evaluates reasoning, episodic retrieval, tool use, and long-horizon interaction. We agree these settings evaluate important components, but they do not jointly test whether *hypotheses, supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, synthesis, uncertainty state*, and

answerability state remain preserved within one evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ across sessions. A long-context or retrieval-augmented system may therefore preserve supporting information while losing *contradiction, revision, uncertainty, or method and condition scope* after later information appears. Our scored comparisons in §5.4 support checking these preservation requirements while leaving the population gap rate uncertain; here, *compositional gap* denotes this operational coverage gap, not a claim that individually satisfied logical conditions fail their conjunction. PAPERMIND (Zhao et al., 2026b), LitBench (Varvarigos et al., 2026), and REMem (Shu et al., 2026) move closer to the proposed target, but our focus remains whether $\mathcal{E}_t$ remains faithful to the same *hypotheses, supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, uncertainty state*, and answerability state after later information appears across multiple sessions. A maintained knowledge base is useful, but a correct stored record does not establish that generated responses preserve it, while faithful responses need not use our record format; artifact and response checks are therefore complementary, and end-task scientific outcomes remain essential. State-fidelity diagnostics can expose unsupported scope extensions or lost *contradiction* before downstream outcomes are available, but neither establishes useful discovery nor guarantees reliable science, and we have not shown that Φ predicts downstream research success. For the same question and evidence, changing the user's stated preference should not by itself change the evidential conclusion, although additional evidence or changed assumptions may justify *revision*; this requires semantic consistency, not identical wording or deterministic generation. Prior work motivates concern about sycophancy (Sharma et al., 2024), but our limited pushback probe (see Appendix J) does not identify a training cause or establish general inter-user invariance. These diagnostics localize *retrieval failure, memory-update failure, and reasoning/provenance failure* under controlled *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning* conditions rather than establish downstream scientific reliability.

Scope and Limitations of the Position. This paper advocates an evaluation target supported by bounded diagnostic evaluation; it does not present a finalized benchmark release or establish general model rankings. The construction audit and controlled diagnostic study in §5.3 provide bounded operational evidence for evolving information-state annotation and distinguishing *retrieval failure, memory-update failure, and reasoning/provenance failure* under progressively constrained evolving information states. The framework treats *scientific memory* as interacting preservation requirements spanning *hypotheses, supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, synthesis, uncertainty state*, and answerability state rather than one scalar capability; a system may therefore preserve *evidence provenance* while failing *revision trace*, *contradictory information* while failing *method and condition scope*, or supporting information while failing *uncertainty state*. Future evaluation should emphasize capability-wise and failure-mode-wise analysis rather than aggregate leaderboard ranking. The framework inherits assumptions from provenance representation, belief revision, and truth-maintenance formulations (Lebo et al., 2013; Groth et al., 2010; Alchourrón et al., 2016; Doyle, 1979), while scientific interaction may additionally involve retractions, paradigm shifts, unresolved disagreement, prestige effects, or domain-specific evidence standards requiring extensions of $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$. Computational biology, chemistry, and materials science therefore serve as initial operational domains rather than exhaustive representations of scientific interaction. The evaluation can be reported across *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning* conditions without prescribing a particular memory architecture.

7 Conclusion

Scientific memory differs from chat-history memory because research agents must preserve evolving information-state transitions $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ across sessions, including *hypotheses, supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, synthesis, uncertainty state*, and answerability state. We therefore argue that *scientific-memory* evaluation should distinguish *retrieval failure, memory-update failure, and reasoning/provenance failure* beyond final-answer accuracy. We operationalize this position through a benchmark-oriented taxonomy and five-condition diagnostic protocol (5DP), supported by bounded construction and controlled diagnostic evidence. These additions extend prior evaluations without prescribing a memory architecture or storage format, using *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning* conditions.

AI USE STATEMENT

In this work, we used generative AI tools for the following disclosed tasks. *Automated grading (qualitative data analysis).* Nemotron-3-Super-120B-A12B served as the primary judge in the controlled evaluation (see §5.4; Appendix B), grading answer, unsupported claims, *evidence provenance, contradiction, revision, method and condition scope, and uncertainty*; answerability used exact match of a structured label. We validated the judge with targeted corruptions of expert gold answers (see Appendix C) and calibrated it against two independent human annotators on 120 responses (see Appendix D); divergences make absolute rates evaluator-dependent. *Annotation triage and screening (dataset curation).* GLM-5.1 provided preliminary filtering and disagreement signals during construction-audit review (§5.3), while Laguna-XS-2.1 and Nemotron-3-Super-120B-A12B pre-annotations were used only in Appendix I screening. Expert decisions remained authoritative; no model output replaced an expert label, and GLM-5.1 was excluded from evaluated systems. *Feedback on the manuscript and methodology.* The authors take full responsibility for all content, claims, and artifacts produced with generative-AI assistance.

ETHICS STATEMENT

This work evaluates research agents using scientific evidence arcs and constructed multi-session interactions. Key ethical risks include misattributed evidence, suppressed *contradiction*, unsupported extension beyond *method and condition scope*, and unwarranted *uncertainty*; we therefore treat *evidence provenance, contradiction, revision, synthesis, and uncertainty* as explicit evaluation requirements. Expert decisions remain authoritative, and model-assisted screening does not replace expert validation. We disclose evaluator disagreement, limited domain coverage, and absent inter-expert agreement measurements. Human evaluation was conducted without compensation by male members of the same research group, aged 25-30. Results do not establish safety for clinical decisions, laboratory procedures, or autonomous scientific deployment; appropriate use requires source verification and domain-expert oversight. Generative-AI roles are disclosed in the AI Use Statement.

REPRODUCIBILITY STATEMENT

We provide supplementary materials for the additional diagnostic evaluation. The `data/` and `prompts/` directories contain the 160-item evaluation set, constructed transcripts, rendered prompts with supplied information-state records, 24 transfer items, and judge-validation controls; `scores/` contains per-response scores, aggregate summaries, and raw model responses and judgments in `scores/raw/`; `human_calibration/` contains the blinded 120-response grading sheet, annotation protocol, completed annotator sheets, and calibration results; and `code/` contains scripts for construction, inference, scoring, analysis, and table generation. Appendices B-D describe the evaluation setup, scoring, judge validation, and human calibration, while Appendices E and H–J report supporting analyses. The supplement additionally includes the original source arcs, candidate pool, expert-review records, annotation guidelines, and item-level provenance map for the selected evaluation set. The 12-question pilot is archived separately as an illustrative trace and is not pooled with the later scored evaluation.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. *On the Logic of Theory Change: Partial Meet Contraction and Revision Functions*, pp. 195–217. Springer International Publishing, Cham, 2016. ISBN 978-3-319-20451-2. doi: 10.1007/978-3-319-20451-2_13. URL https://doi.org/10.1007/978-3-319-20451-2_13.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, 2021.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C Wallace. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pp. 123–132, 2020.

Jon Doyle. A truth maintenance system. *Artificial intelligence*, 12(3):231–272, 1979.

Dionizije Fa, Marko Culjak, Bruno Pandza, and Mateo Cupic. Bioagent bench: An ai agent evaluation suite for bioinformatics. *arXiv preprint arXiv:2601.21800*, 2026.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Paul Groth, Andrew Gibson, and Jan Velterop. The anatomy of a nanopublication. *Information services and use*, 30(1-2):51–56, 2010.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, et al. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Sarthak Jain, Madeleine Van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. Scirex: A challenge dataset for document-level information extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7506–7516, 2020.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5684–5696, 2020.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808, 2025.

Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. 2013.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3705–3717, 2019.

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. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Yao Lu, Yue Dong, and Laurent Charlin. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 8068–8074, 2020.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3219–3232, 2018.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang 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.

Hanane Nour Moussa, Patrick Queiroz Da Silva, Daniel Adu-Ampratwum, Alyson East, Zitong Lu, Nikki Puccetti, Mingyi Xue, Huan Sun, Bodhisattwa Prasad Majumder, and Sachin Kumar. Scholareval: Research idea evaluation grounded in literature. *arXiv preprint arXiv:2510.16234*, 2025.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Yiheng Shu, Padmaja Jonnalagedda, Xiang Gao, Bernal Jimenez Gutierrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. Remem: Reasoning with episodic memory in language agent. In *International Conference on Learning Representations*, volume 2026, pp. 55958–55987, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Andreas Varvarigos, Ali Maatouk, Jiasheng Zhang, Ngoc Bui, Jialin Chen, Leandros Tassiulas, and Rex Ying. Litbench: A graph-centric large language model benchmarking tool for literature tasks. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2806–2817, 2026.

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 (EMNLP)*, pp. 7534–7550, 2020.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. A survey on evaluation of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26690–26714, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Xinping Zhao, Xinshuo Hu, Jiaxin Xu, Danyu Tang, Xin Zhang, Mengjia Zhou, Yan Zhong, Yao Zhou, Zifei Shan, Meishan Zhang, et al. Lmeb: Long-horizon memory embedding benchmark. *arXiv preprint arXiv:2603.12572*, 2026a.

Yanjun Zhao, Tianxin Wei, Jiaru Zou, Xuying Ning, Yuanchen Bei, Lingjie Chen, Simmi Rana, Wendy H Yang, Hanghang Tong, and Jingrui He. Papermind: Benchmarking agentic reasoning and critique over scientific papers in multimodal llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10457–10474, 2026b.

Jianjie Zheng, Zhichen Liu, Zhanyu Shen, Jingxiang Qu, Guanhua Chen, Yile Wang, Yang Xu, Yang Liu, and Sijie Cheng. Evaluating memory capability in continuous lifelog scenario. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7063–7089, 2026.

A Appendix

Formal Definitions and Scoring Procedures. We provide operational definitions and scoring procedures for the *scientific-memory* metrics; implementations for the additional evaluation are documented in Appendix B. *Contradiction-preservation accuracy* measures whether *supporting information* and *contradictory information* remain preserved without unsupported synthesis or definitive information under unresolved *contradiction*, while *revision-preservation accuracy* measures whether later information correctly updates earlier information within the evolving information-state transition $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$. *Evidence-span overlap* measures proportional overlap between predicted and *gold evidence* spans linked to the evolving information state. Multi-document *synthesis* evaluates preservation of the minimal *supporting information* and *contradictory information* sets required for the gold evolving information state, together with *evidence provenance, method and condition scope, revision trace, uncertainty*, and answerability across multiple information sources. For the additional evaluation, `code/analyze.py` computes source-identifier precision/recall/F1, label accuracy, scope coverage, and aggregates judge-based components, while `code/det_rules.py` provides a separate deterministic-rule sensitivity check.

Distinguishing Memory-Update Failure from Reasoning/Provenance Failure. The five-condition diagnostic protocol (5DP) contrasts evidence-only, state-only, and evidence-plus-state settings to investigate memory updating and *reasoning/provenance failure*. Under *gold-source retrieval*, systems receive supporting and contradictory information while constructing the evolving information state $\mathcal{E}_t$; failures may therefore involve *revision, contradiction, uncertainty, or synthesis*

during state construction. Under *gold-memory-state evaluation*, systems receive the gold $\mathcal{E}_t$ without source text, while retaining the evidence defines the evidence-plus-state comparison. Residual failure with both evidence and state supplied is consistent with *reasoning/provenance failure* or response-expression difficulty, but score differences do not uniquely identify an internal failure mechanism. Ambiguous cases requiring multi-step inference across conflicting evidence are therefore investigated through matched information-access comparisons rather than assumed to be resolved by condition labels.

Diagnostic-Condition Isolation and Monotonic Evaluation Behavior. The five-condition diagnostic protocol (5DP) treats each condition as an isolated evaluation setting with independently rendered prompts, initialized interaction history, and condition-specific information exposure. *Gold-memory-state evaluation* supplies $\mathcal{E}_t$ only within its condition, while *offline free retrieval* restricts retrieval to information states available before step t, preventing future-information leakage; each condition is therefore rendered and captured independently. Monotonic behavior is tested only in matched comparisons retaining prior evidence while adding state information; state-only is a separate comparison with no assumed ordering. Larger-scale validation should compare condition-wise distributions and repeated-generation variability before interpreting.

Annotation Scalability and Reliability of Evolving Information States. Evolving information-state annotation used structured records containing *hypotheses, supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, uncertainty state*, and answerability state across interaction steps. The construction audit used one expert decision per item for 300 stratified items across computational biology, chemistry, and materials science and did not measure inter-expert agreement. GLM-assisted triage agreement with expert labels reached 0.950, 0.927, and 1.000 for item decisions, evidence support, and capability categories, with Cohen's κ of 0.902, 0.864, and 1.000, respectively. Evidence-support labels showed the lowest agreement because *contradictory information, uncertainty state, and method and condition scope* often required multi-step interpretation across $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$. Expert decisions remained authoritative; GLM-assisted triage provided preliminary filtering and disagreement signals, not independent human adjudication. Large-scale construction would require staged annotation pipelines separating *evidence provenance, evolving information-state construction, contradiction preservation, and uncertainty-state adjudication* to reduce expert-review cost and unsupported synthesis across interaction steps. Such pipelines can be evaluated across *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning* to distinguish *retrieval failure, memory-update failure, and reasoning/provenance failure*.

Unresolved Disagreement and Multi-Hypothesis Evolving Information States. Unresolved disagreement remains represented within evolving information states $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$ rather than collapsed into one definitive state. *Supporting information, contradictory information, evidence provenance, method and condition scope, revision trace,* and *uncertainty* may therefore remain jointly preserved through *mixed/conditional*, *contradictory*, or *unresolved uncertainty states*. Multi-hypothesis states use competing *hypothesis* records linked to distinct supporting and contradictory-information sets rather than forced resolution. Evaluation therefore tests whether generated information preserves unresolved disagreement across $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$ rather than collapsing it into unsupported *synthesis* or definitive information, helping distinguish *retrieval failure, memory-update failure, and reasoning/provenance failure* across *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning*.

Evidence-Quality Diagnostics and Uncertainty Calibration. The proposed *scientific-memory* framework is compatible with Matching, Relevance, and Logic diagnostics and uncertainty-calibration analysis under evolving information states $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$. Matching evaluates alignment between generated information and cited *supporting information* or *contradictory information*; Relevance evaluates whether retrieved evidence matches the queried *method and condition scope*; and Logic evaluates *revision trace, contradiction preservation, and synthesis* across interaction steps. These diagnostics connect to evidence-provenance precision/recall/F1, contradiction-preservation accuracy, revision-preservation accuracy, and unsupported-synthesis evaluation, while uncertainty calibration may use uncertainty-state and answerability-state labels, confidence annotations, consistency-based uncertainty estimation, and selective risk-coverage under unresolved evolving information states.

Multimodal and Experimental Information States. The proposed *scientific-memory* framework extends beyond textual literature to multimodal scientific settings containing figures, tables, microscopy images, spectra, laboratory notebooks, simulation outputs, and experimental logs. Across

evolving information states $\mathcal{E}_{t-1} \rightarrow \mathcal{E}_t$, it preserves modality-specific *evidence provenance* and *method and condition scope*, referencing figure panels, table cells, spectroscopy traces, microscopy regions, notebook entries, experimental timestamps, imaging configurations, instrument settings, preprocessing pipelines, simulation parameters, and laboratory protocols. The information-state schema therefore extends to modality-linked *hypotheses, supporting information, contradictory information, evidence provenance, method and condition scope, revision trace, synthesis, uncertainty*, and answerability state rather than text-only evidence representations, with evaluation applicable across *online free retrieval, offline free retrieval, gold-source retrieval, gold-memory-state evaluation, and gold-evidence reasoning* to distinguish *retrieval failure, memory-update failure, and reasoning/provenance failure*.

B Additional Evaluation: Setup and Scoring

Relationship to the Original Pilot and 5DP. The construction audit and 12-question pilot in §5.3 remain the original methodological evidence, while numerical results in §5.4 come from a separate, larger controlled evaluation. Its six settings are implementation-level comparisons, not interchangeable with the five proposed 5DP conditions; in particular, they establish neither unrestricted *online free retrieval* performance nor temporal leakage control. The original pilot supplied assessment information in its state condition and is therefore illustrative, not an independently scored test of *memory-update failure*; the additional state conditions exclude gold-answer prose and evaluate *evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* under controlled information access.

Items and Annotation. The additional evaluation selects 160 items from 287 usable questions in the 300-item expert-reviewed pool, excluding rejected items and questions or answers referring to unavailable dialogue turns, with supplied expert corrections retained. The subset contains 54 chemistry, 51 computational-biology, and 55 materials-science items across 30 evidence arcs: 139 answerable, 10 ambiguous, and 11 unanswerable. The source pool is enriched for *revision, contradiction*, and conditional evidence, so failure rates are not estimates for a random scientific-literature sample. Each item has one expert decision. Reported agreement coefficients compare model-assisted triage with expert labels, not independent human experts, while the human calibration below grades model responses and does not provide second-expert validation of gold item labels.

Transcripts and State Records. Each arc is deterministically rendered as a 9-12-session transcript (median 7,987 characters) interleaved with two same-domain distractor arcs, where session 1 presents the initial *hypothesis*, and later sessions introduce ordered evidence. The state record contains claim topic, initial *hypothesis*, status, answerability, supporting-evidence identifiers, *contradiction and revision flags, method and condition scope, and revision trace*, while omitting gold-answer prose.

Systems and Conditions. The five evaluated systems are Llama-3.2-11B-Instruct[2], Mistral-Nemotron[3], Nemotron-3-Nano-30B-A3B[4], Nemotron-3.5-Lightning-30B-A3B[5], and GPT-OSS-20B[6]; the primary judge, Nemotron-3-Super-120B-A12B[7], is excluded from the evaluated set. The main evaluation uses temperature zero and one generation per item–system–condition cell across the six settings in Table 4, with 4,674 graded responses; denominators are retained as reported rather than treating every planned cell as graded. Each setting requests at most 120 words in structured output containing status, answerability, source identifiers, *method and condition scope*, and *contradiction and revision flags*. Comparing full history with selected evidence also changes distractor exposure, so condition differences are diagnostic contrasts rather than proof of a unique internal mechanism. D4 versus D6 tests the benefit of supplying a constructed state, while D4 versus D5 tests state usability without supporting text, distinguishing *memory-update*

Table 4: Controlled settings underlying the additional results.

Setting	Information Supplied	n
D1	Question alone, without retrieved evidence or history	770
D2	Three sessions retrieved using deterministic BM25	778
D3	Complete interleaved session history	772
D4	Gold source identifiers and evidence text	786
D5	State record without source text	779
D6	Gold evidence and the state record	789

[2] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
[3] https://build.nvidia.com/mistralai/mistral-nemotron/modelcard
[4] https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16
[5] https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16
[6] https://huggingface.co/openai/gpt-oss-20b
[7] https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-BF16

failure from response-level *reasoning/provenance failure* under controlled *gold-source retrieval* and *gold-memory-state evaluation* comparisons.

Selected Component Checks. A structured response passes the selected checks if source-identifier recall ≥ 0.5, the judge finds no unsupported claim, and the answer is fully correct. Source recall measures identifier overlap, not correct attribution of every cited claim; these checks are proxies for selected evaluation practices, not exhaustive implementations of prior benchmarks. Table 5 reports counts using the fully-correct-answer criterion throughout.

Table 5: Responses passing the three selected checks but failing joint requirements. D1 has no passing responses; its gap rate is undefined.

Setting	Pass Selected Checks	Fail Φ	Gap (%)
D1	0	0	–
D2	93	21	22.6
D3	108	30	27.8
D4	196	43	21.9
D5	127	20	15.7
D6	235	36	15.3
All Settings	759	150	19.8

Joint Scoring and Uncertainty. The judge receives the question, expert gold answer, gold evidence and labels, and generated response, and grades answer, unsupported claims, provenance, *contradiction, revision, method and condition scope, and uncertainty preservation*. Answerability uses exact match of the structured three-way label. Joint fidelity $\Phi = 1$ requires a fully correct answer, no unsupported claim, correct answerability, and no failure of provenance or any applicable preservation component; not-applicable components do not cause failure. Thus, Φ is a strict conjunction of eight operational gates, not an assumed scalar ability. Reported intervals use 2,000 item-level bootstrap resamples, clustering repeated responses for each item; items within an evidence arc may remain related, limiting extrapolation beyond 30 arcs.

C Judge Validation for the Additional Evaluation

Validation constructs responses from expert gold answers with targeted violations of *method and condition scope, uncertainty, contradiction, revision, evidence provenance*, or grounding. Detection is measured on items carrying the affected property, while false positives use corresponding unmodified gold answers; corruptions target one property but do not guarantee independent effects across judged components. The primary judge graded 95.5% of 156 unmodified gold answers as fully correct on answer fidelity. Table 6 reports detection and false-positive rates for each targeted alteration, including the elevated false-positive rate for fabricated replication.This is an answer-level acceptance rate, and does not negate unsupported-claim false positives. These checks support component diagnostics but show that the judge is not a calibrated estimate of absolute scientific correctness.

Table 6: Detection and false-positive rates target the specified component.

Targeted Alteration	n	Detected (%)	False Positive (%)
Delete Scope and Assert Generality	39	94.9	5.0
Replace Uncertainty with a Verdict	20	100.0	0.0
Keep one Side and Assert Agreement	50	60.0	3.8
Present a Superseded Claim as Current	37	100.0	2.6
Replace Source Identifiers	158	62.0	0.6
Append a Fabricated Replication	148	100.0	44.2

D Human Calibration of the Additional Evaluation

Two annotators independently graded a blinded, stratified subset of 120 responses (20 per condition across five systems and capability categories), receiving the question, expert gold answer, gold evidence, and labels but not system identities or model-judge verdicts; neither was a domain specialist in the three sciences. They assessed responses against the supplied references, with answerability remaining the same deterministic field check; consensus required both annotators to pass each gate. Among 37 responses both annotators judged correct and free of unsupported claims, the judge flagged 20 as unsupported. Inter-annotator agreement ranged from $\kappa = 0.44$ for *contradiction* to 1.00 for *uncertainty*, with provenance at $\kappa = 0.55$. The three consensus gap cases concerned provenance, whereas the larger model-judged sample more often implicated

Table 7: Calibration on the same 120 responses. Brackets denote 95% item-level bootstrap intervals; gap denominators differ because evaluators identify different passing sets.

Evaluator	Joint Fidelity (%)	Gap Count	Gap (%)
Annotator A	35.0 [15.6, 54.3]	4/46	8.7 [0.0, 31.2]
Annotator B	35.0 [18.8, 49.6]	1/43	2.3 [0.0, 6.7]
Both Annotators	27.5 [11.3, 43.6]	3/36	8.3 [0.0, 23.1]
Model Judge	6.7 [0.0, 13.6]	3/10	30.0 [7.7, 100.0]

revision and answerability, showing differences in failure composition and absolute rates. Consensus joint-fidelity estimates were 40.0% for evidence alone, 30.0% for state alone, and 45.0% for evidence plus state. Table 7 reports the evaluator-level joint-fidelity and coverage-gap estimates for the same 120 responses. With only 20 unpaired calibration responses per condition, these estimates are descriptive and do not establish a precise state-only deficit or nonzero population coverage-gap rate; the consensus-gap interval includes zero. Human calibration also does not remove limitations of the single expert pass used to construct gold items.

E Supporting Checks and Interpretation

Sensitivity of the State-Only Comparison. Across 766 paired item-system outputs, model-judged state-only minus evidence-only joint fidelity was -5.7% (interval $[-10.7, -1.2]$); removing the unsupported-claim gate reduced this to -2.3% ($[-7.6, 2.7]$), while a deterministic label-based score favored state-only by 38.8% ($[33.4, 43.5]$). This reversal shows that the state-only difference depends on the grading rule and motivates evaluating generated explanations alongside supplied labels, rather than interpreting it as a universal performance reduction from state records.

Response Format. At gold-source access, a paired format comparison found joint fidelity of 16.3% for structured output versus 33.7% for prose under the same 120-word limit, with median answer-prose lengths of 43 versus 93 words. These observations motivate reporting format effects but do not establish that length caused the difference and constrain interpretation of the structured-output results.

Monotonicity and Wording Controls. Repeated generations under unchanged gold-source and gold-evidence conditions showed downward joint-fidelity changes in 8.4% and 11.6% of cases, respectively, with a cross-condition rate of 11.0%. These transitions indicate observed variability, not a systematic violation. Equalized instructions left the state-only minus evidence-only estimate negative but with an interval including zero; a nonsignificant contrast change does not establish no wording effect.

Scope of the Retained Evidence. These comparisons evaluate responses to supplied scientific information in constructed transcripts, not autonomous search, internal memory representations, or long-term research outcomes. The original 5DP remains a proposal; added evidence supports selected contrasts and reporting practices, while broader deployment claims require independent item validation and more extensive human grading.

F State Transitions, Update Policy, and Evaluation Boundaries

A Typed Evaluation Record. Let $\mathcal{E}_t = \{e_t^{(1)}, \ldots, e_t^{(m_t)}\}$ be a finite set of claim records, with $e_t = \langle h, S^+, S^-, \pi, \sigma, \rho, \upsilon, \alpha \rangle$, where h is the claim; S^+ and S^- are *supporting and contradictory information*; π maps evidence to source and location; σ records *method and condition scope*; ρ records prior versions and evidence motivating each *revision*; υ records supported-within-scope, refuted, conditional, mixed, or insufficient status; and α records whether the question is answerable, ambiguous, or unanswerable from available evidence. *Uncertainty* is represented through these fields and retained alternatives, not calibrated posterior probabilities. Stable identifiers link claim versions, and status vocabularies are fixed in the codebook; this illustrative record does not alter the implemented scoring gates in Appendix B. A transition is $\mathcal{E}_t = T(\mathcal{E}_{t-1}, u_t; \Gamma)$, where u_t contains new information and Γ is a documented domain-specific interpretation policy. This specifies inputs and obligations, not a learned update algorithm or unique scientific-truth rule, where an admissible transition retains *evidence provenance*, identifies changed claims and scopes, and preserves unresolved alternatives. Multiple defensible interpretations may therefore yield multiple admissible states rather than a forced label.

Worked Hypothetical Example and the Probability Expression. Suppose session 1 reports that method A outperforms B under setting C, supported by s_1, while session 2 reports the reverse under setting D, supported by s_2. Then $\mathcal{E}_2$ should retain both scoped claims, where the second does not refute the first under C. For "Does A outperform B generally?", the record supports a conditional answer rather than an unrestricted conclusion. In the Introduction's notation, $\mathcal{H}_{<t}$ contains both sessions and the question, $\mathcal{R}_t = \{s_1, s_2\}$ the retrieved evidence, and a_t the generated answer; $P(a_t \mid \mathcal{H}_{<t}, \mathcal{R}_t)$ denotes conditional generation probability, with no numerical probability or actual

run asserted. A high-probability answer is not necessarily faithful: "A outperformed B in the C experiment [s_1]" is correct and attributed but leaves generality unresolved and omits the D restriction. A scorer checking only truth and citation may accept it, whereas one requiring a complete, scoped answer may reject it. This illustrates omitted requirements, not a conjunction failure or a claim about any particular benchmark's grading.

Markdown, Provenance Standards, and Conceptual Change. A Markdown record can store the two claims, source links, conditions, and revision rationale, while Git preserves versions; however, a diff alone cannot determine whether an edit was warranted, since changing "supported" to "conditional" may reflect new evidence or user preference. PROV-O provides provenance relations (Lebo et al., 2013), but annotation policy must still determine relevance, scope, and unresolved disagreement; no PROV-O serialization or complete ontology is required. Conceptual reframing also requires more than a status toggle, where annotators should retain the original *hypothesis* and introduce revised or split claims with parent links, changed definitions, and evidence motivating the *revision*. Thus, a claim about all settings may split into C and D-specific claims, while a changed referent should be recorded as reformulation rather than silently counted as support for the original. This representation accommodates such events, but the present study does not validate coverage of paradigm shifts or establish *evidence provenance, contradiction, method and condition scope, synthesis, and uncertainty* beyond the evaluated settings.

Evidence Quality, Paper Counts, and Source Churn. We propose that domain annotators assess evidence directness for the claim, method and condition compatibility, relevant controls, replication independence, and corrections or retractions. Source prestige and paper count should not substitute for this assessment, where no domain-independent number of papers determines *revision*; a decisive correction to sole support may warrant revision, whereas repeated observations may add little independent evidence. If evidence does not settle the question, the state remains conditional or unresolved. Search-ranking changes are not new scientific evidence. An update log should distinguish previously seen evidence, duplicate or derived reports, new independent observations, and corrected or withdrawn sources using stable source identifiers and versions. Reordered or duplicated results should preserve the substantive claim state; previously unseen, relevant evidence may change it even when old sources remain available. Losing a source from search results should trigger an access or verification flag, not erase recorded content.

Expert Disagreement and Overlapping Capabilities. For future construction, provenance extraction and candidate state drafting can precede independent expert checks for difficult or ambiguous cases. Annotators should retain alternative admissible records and disagreement reasons when evidence supports multiple readings. A response may receive credit if it matches an admissible interpretation and preserves material disagreement; one inventing consensus should not. Case-level adjudication and agreement rates should be reported separately. This is a proposed protocol; the current source-item audit used one expert decision per item (see Appendix B). Capabilities are overlapping obligations, not mutually exclusive error classes, where attributing s_1's result to s_2 is *evidence provenance* failure; transferring a C result to D is *method and condition scope* failure; retaining a conclusion after applicable correction is *revision* failure; hiding conflicting evidence is *contradiction* failure; and asserting certainty under unresolved evidence is *uncertainty* failure. One response may receive multiple labels; joint fidelity counts it once, while component metrics capture overlaps. Targeted corruptions test judge sensitivity to specified violations, not statistical independence. Construct validity beyond these operational labels, including prediction of downstream outcomes, remains unestablished.

G Comparison with Selected Prior Benchmarks

Table 8 distinguishes each benchmark's evaluation unit from our proposed additional check and describes cited task formulations, not every extension or item. The last column is a proposed extension, not a claim that prior benchmarks lack related capabilities, where LongMemEval evaluates knowledge updating and abstention, SciFact evaluates evidence and support/refutation, and ALCE evaluates several output properties. Prior benchmarks could incorporate transition-level requirements for *hypotheses, evidence provenance, contradiction, method and condition scope, revision, synthesis, and uncertainty* to test our position without replacing established evaluations or defining a separate model ability. Since the selected checks underlying our empirical gap estimate are narrower than

Table 8: Selected benchmark comparison. The additional check targets the same question's evolving evidence state, not exhaustive novelty.

Benchmark	Evaluation Unit	Prior Emphasis	Proposed Additional Check
LongMemEval (Wu et al., 2024)	Questions over multi-session conversation histories	Recall, temporal reasoning, knowledge updates, and abstention	Preserve scientific evidence attribution and condition scope when revising the same claim
SciFact (Wadden et al., 2020)	Scientific claims paired with abstracts and rationales	Retrieve evidence and identify support or refutation	Preserve the sequence of revisions and unresolved conflicts across sessions
QASPER (Dasigi et al., 2021)	Questions about a research paper, with supporting evidence	Answer and evidence selection within the paper	Track how later sources change an earlier paper-grounded conclusion
ALCE (Gao et al., 2023)	Retrieved passages and generated answers with citations	Fluency, correctness, and citation quality	Check whether a later answer retains revised scope and uncertainty from the evidence history

Table 9: System metrics (%), pooled over six controlled settings. Answer = fully correct; Prov., Contr., Revis., Uncert. = *evidence provenance, contradiction, revision, uncertainty* preservation; Answ. = answerability-label accuracy; Status = status-label accuracy; U-coll. = uncertainty collapse (lower is better).

System	n	Φ	Answer	Prov.	Contr.	Revis.	Scope	Uncert.	Answ.	Status	U-coll.
Nemotron-3.5-Lightning-30B	938	20.0	34.5	75.7	71.0	69.2	50.9	64.3	75.4	45.8	31.9
GPT-OSS-20B	935	15.2	27.9	77.8	60.8	62.4	50.9	61.3	75.8	44.6	60.5
Nemotron-3-Nano-30B	924	14.8	24.1	77.8	69.8	62.3	43.1	64.2	76.1	51.1	31.7
Mistral-Nemotron	937	12.4	23.4	79.6	73.8	71.5	45.3	70.0	86.6	59.3	49.6
Llama-3.2-11B	940	6.1	12.8	72.3	69.3	62.7	41.0	57.1	81.0	46.8	66.1

complete benchmark implementations, the estimate should not be interpreted as their measured failure rate.

H Additional Results by System, Domain, and Transfer Setting

System and Component Results. Table 9 exposes system differences within the 4,674-response evaluation; preservation columns use only applicable items. Answerability accuracy is exact agreement with the three-way gold label, while status accuracy uses the study's deterministic mapping from capability and expert answerability labels. Uncertainty collapse counts `answerable` predictions for ambiguous or unanswerable items. Scope measures judged *method and condition scope* preservation, not lexical scope-slot coverage. Pooled values combine six conditions and are descriptive; column orderings do not establish distinct latent abilities or a statistically reliable overall ranking. Model-judge errors and human-calibration results in Appendixes C–D also apply.

Domain Composition. Table 10 gives the domain breakdown; gap denominators count responses passing the selected component checks, not all domain responses, and gate counts may overlap. These summaries make failures inspectable but do not provide distributions for every metric in Appendix A; in particular, no evidence-span-overlap distribution is claimed.

Table 10: Domain summaries from the controlled evaluation. Rates are percentages; final-column counts show frequently failed additional gates among gap cases.

Domain	n	Φ	Pass	Gap	Frequent Broken Gates (Counts)
Chemistry	1586	12.4	226	19.9	Answerability (17), revision (12), contradiction (9)
Computational biology	1486	12.5	228	21.1	Revision (23), answerability (10), scope (8)
Materials science	1602	16.0	305	18.7	Revision (26), answerability (24), provenance (9)

A Bounded Transfer Pilot. The transfer check uses 24 author-constructed machine-learning and theoretical-computer-science items and four supporting-check systems (excluding Lightning); items lack independent expert validation. Table 11 compares full history (D3), gold evidence (D4), and state-only input (D5). The matched natural-science subset uses the same four systems, so values differ from the five-system pooled main-text results; intervals are reported 95% item-level bootstrap intervals and n counts graded responses. The schema can represent proof assumptions and excluded techniques as *method and condition scope*, theorem attribution as *evidence provenance*, corrected bounds as *revision*, and unresolved conjectures as *uncertainty*. This is a representational feasibility check, not evidence of

Table 11: Transfer-pilot results. Values are joint fidelity Φ (%); parentheses give intervals. ML/TCS denote transfer subsets; conditions are not a monotone ladder.

Setting	Transfer n	Transfer Φ	Natural n	Natural Φ	ML	TCS
D3	95	9.5 (3, 17)	616	9.4 (6, 12)	4.2	14.9
D4	95	20.0 (9, 32)	627	16.3 (13, 20)	4.3	35.4
D5	92	12.0 (4, 20)	623	12.2 (9, 15)	8.5	15.6

Table 12: Annotation-routing tradeoff, where smaller review budgets miss more problematic items.

Threshold	Forwarded	Budget (%)	Recall (%)	Precision (%)	Lift
$\geq$ 0 flags	287	100.0	100.0	57.5	1.00
$\geq$ 1 flag	106	36.9	47.9	74.5	1.30
$\geq$ 2 flags	96	33.4	45.5	78.1	1.36

broad transfer: under D4, ML fidelity was 4.3% versus 35.4% for TCS, and author-selected arcs cannot establish coverage of other disciplines or agreement with independent domain experts.

I Annotation Screening and a Proposed Scaling Workflow

The additional screening analysis uses 287 usable items, of which 57.5% were not accepted as-is by the expert. Model pre-annotations provide item-decision and evidence-support signals; Table 12 routes items by screening-flag count, with zero flags meaning all items are reviewed. Budget is the fraction sent to expert review, recall the fraction of expert-identified problem items forwarded, precision the problem-item fraction among forwarded items, and lift = recall/budget (1.00 = random routing in expectation). The one-flag screen modestly concentrates problematic items but misses more than half, so unflagged items cannot be accepted without review. Screening is distinct from label replacement, where Laguna-XS-2.1[8] reproduced expert item decisions with 43.6% agreement ($\kappa = 0.03$, $n = 273$), and Nemotron-3-Super-120B with 59.2% ($\kappa = 0.24$, $n = 284$), distinct from the earlier GLM comparison and insufficient to establish reliable model-assisted labels. We propose staged construction, where deduplicate and link *evidence provenance*; draft candidate records automatically; prioritize flagged or ambiguous cases for expert review; and randomly audit unflagged cases to estimate missed errors. Reusing expert-reviewed evidence arcs may amortize provenance work, while stratified independent checks can assess label reproducibility; disagreements follow Appendix F. This is a proposed workflow, not an evaluated cost reduction: routing figures do not measure expert hours, post-routing label accuracy, or new-domain performance, and scaling remains an open empirical question.

J Evidence-Free Pushback and Consistency across Users

The two-turn probe uses the four shared systems other than Lightning, where Turn 1 uses gold-source input (D4), while Turn 2 requests either a neutral restatement or a response to a colleague asserting the superseded claim without new evidence. There are 614 neutral-arm and 613 pushback-arm responses. Table 13 reports paired differences and bootstrap intervals; unequal arm totals preclude reconstruction from unpaired rates. The judge shows the clearest contrast in unsupported assertions, while the joint-fidelity interval includes zero; known judge false positives and the 120-response calibration limitation apply. This limited pressure probe does not test invariance across users or collaborations or establish that preference tuning causes errors. For fixed questions, assumptions, and evidence, conclusions should remain stable, while new evidence may justify *revision*. This requires neither deterministic generation nor online weight updates; repeated-generation results likewise show that temperature zero does not imply perfect reproducibility.

Table 13: Pushback-minus-control differences (%), with 95% item-level bootstrap intervals over 160 items.

Outcome	Difference	Interval
Joint fidelity Φ	-3.8	$[-7.5, +0.0]$
Fully correct answer	-2.8	$[-7.4, +1.5]$
Asserts something unsupported	$+7.7$	$[+2.7, +12.4]$
A preservation component fails	$+1.2$	$[-3.3, +5.6]$
Loses a previously correct answer	$+0.2$	$[-2.8, +3.4]$

[8] https://build.nvidia.com/poolside/laguna-xs-2.1/modelcard