RESEARCH AGENTS GET LOST IN EVOLVING EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A research question may ask what was supported by the evidence available at a particular date. Scholarly corrections and revisions provide documented cases in which that answer changes. We study this requirement as temporal grounding and introduce Time-conditioned Scholarly Conflict Question Answering (TSC-QA), a benchmark of 200 question pairs across eight scientific fields. The questions in each pair differ only in their evidence cutoff and require different answers. Strict-pair accuracy requires both answers to be correct. To test this requirement across different ways of conducting research, we evaluate fixed answerers, planned workflows, and systems that control their own research loop. These include Qwen, Llama, Gemma, DeepSeek, and GPT models, as well as the complete Codex-Luna, Codex-Terra, and Codex-Sol systems. When systems must find or recall the evidence themselves, the highest pair accuracy is 35.50%, and the delegated research systems perform better at the later cutoff. Providing claims organized by source and public date improves strict-pair accuracy more than providing claim summaries alone. Models can recognize a supplied claim and answer the later question correctly while still missing the earlier answer. Response analysis connects these results to claim interpretation, source dating, and evidence use. TSC-QA tests whether research agents can recover the answer supported at each point in a changing evidence history.

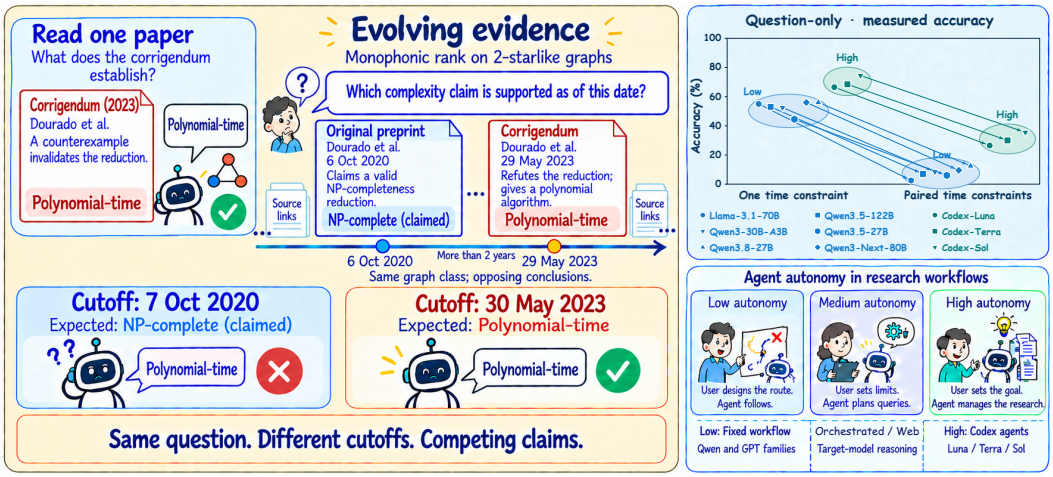


Figure 1: **Temporal grounding across a scholarly correction.** The monophonic-rank example contrasts a 2020 NP-completeness claim with a 2023 counterexample and polynomial-time algorithm for the same graph class. The two cutoffs select the original reported claim and the corrected position. Upper right: late and strict-pair accuracy without a supplied evidence packet. Lower right: representative workflows used to answer the same paired task, grouped by control over research decisions.

1 INTRODUCTION

Research agents search for sources, read documents, and combine evidence to answer a user's question (Asai et al., 2026; Zhu et al., 2025a; Agarwal et al., 2024; Shao et al., 2024; Li et al.,

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.

2026; Lu et al., 2024; Wang et al., 2023). Recent benchmarks assess difficult browsing tasks and extended research reports (Wei et al., 2025; Du et al., 2026). A user's question can also specify when the evidence must have been available. Corrections, revisions, and follow-up studies make this requirement concrete in scholarly research: a conclusion supported by an earlier source can change when new evidence appears. Answering at the requested date requires an agent to connect each claim to the source version in which it was publicly available.

Figure 1 illustrates this requirement with monophonic rank, a graph-convexity parameter. The original preprint, publicly available in 2020, claims that computing it on 2-starlike graphs is NP-complete. The corrigendum, publicly available in 2023, refutes the proposed reduction with a counterexample and provides a polynomial-time algorithm covering the same graph class (Dourado et al., 2022; 2024). For the displayed question, the early cutoff selects the original reported claim, and the late cutoff selects the correction. Answering both questions correctly requires recovering both positions and identifying when each applies.

We call this behavior **temporal grounding**: connecting an answer to evidence that matches the question's scientific scope, was publicly available by its cutoff, and satisfies its stated evidence rule. Temporal question answering studies changing facts, historical knowledge, and time-constrained retrieval (Jia et al., 2024; Zhu et al., 2025c; Cheng & Dou, 2025); TempoMed-Bench examines evolving medical knowledge and guideline recommendations (Guan et al., 2026). We study documented changes in the reported conclusion about the same scientific question under the same stated conditions. The source history establishes which answer applies at each cutoff and makes the change verifiable.

We introduce **Time-conditioned Scholarly Conflict Question Answering (TSC-QA)**, a benchmark containing 200 temporal question pairs. The two questions in each pair are answered independently and preserve the scientific scope, wording, answer options, and option order. Only the cutoff changes, and the reference answer changes with it. Strict-pair accuracy requires both answers to be correct, measuring whether the system succeeds across both evidence states. The construction protocol in Figure 2 links related publications to scoped claims, verifies their sources and dates, and builds paired questions around a documented change.

Recovering the earlier answer can require finding an original source, distinguishing it from a later correction, and deciding which evidence applies. Research systems organize these steps in different ways: some follow a fixed answering procedure, some plan within a configured workflow, and others control successive research actions and stopping. We evaluate representatives of all three to examine temporal grounding across these forms of research support. The evaluation includes Qwen, Llama, Gemma, DeepSeek, and GPT models in fixed or planned workflows, together with the complete Codex-Luna, Codex-Terra, and Codex-Sol research systems. Appendix Figure A2 details how research decisions are assigned in these workflows.

Research support improves accuracy, but answering correctly at both cutoffs remains difficult. When systems must find or recall the evidence themselves, the highest strict-pair accuracy is 35.50%, and the complete research systems answer later-cutoff questions more accurately than earlier ones. We then supply relevant claims directly to examine what helps recover both answers. Organizing these claims by source and public date improves accuracy more than supplying claim summaries alone. When a source supporting the reference answer is supplied, models can recognize its claim and answer the later question correctly while still missing the earlier answer. The response analysis in Figure 5 connects these results to patterns in saved answers and research records, including claim interpretation, source dating, and evidence use.

Our contributions are threefold. **First**, we introduce a paired task that measures temporal grounding across documented changes in scholarly evidence. **Second**, we provide a construction protocol linking shared scientific scope, source passages, public versions, and cutoff-specific answers. **Third**, we evaluate representative research workflows through paired outcomes, comparisons of supplied evidence, and response analysis, showing where historical answering remains difficult.

2 Related Work

Temporal knowledge and evolving evidence. Temporal benchmarks examine whether answers track changes in knowledge and the date specified by a question. EvolveBench evaluates temporal

awareness over evolving facts, FAITH performs temporal question answering over heterogeneous sources, and DailyQA evaluates web retrieval in a changing information environment (Zhu et al., 2025c; Jia et al., 2024; Cheng & Dou, 2025). TempoMed-Bench studies medical knowledge across historical and current guideline states, including retrieval-based answering (Guan et al., 2026). TSC-QA connects temporal answering to documented scholarly changes: each pair holds the scientific question and answer options fixed, and its source history establishes a different reference answer at each cutoff. Scoring the two answers jointly tests whether the system recovers both evidence states.

Scientific claims and knowledge conflicts. Scientific claim verification links assertions to passages that support or refute them, as in SciFact (Wadden et al., 2020). ConflictBank studies inconsistencies within retrieved knowledge and between retrieved and parametric knowledge, including temporal discrepancies (Su et al., 2024). Scholarly conflicts require both scope alignment and source history: competing claims must address compatible populations, conditions, and outcomes, and their qualifying source versions determine when each position became available. TSC-QA combines these checks in a passage-linked record and applies the question's evidence rule at each cutoff.

Research systems and workflow control. Research systems differ in how they find evidence and decide what to examine. ReAct interleaves reasoning with actions, while BrowseComp and DeepResearch Bench evaluate browsing and source-based report generation (Yao et al., 2022; Wei et al., 2025; Du et al., 2026). FinSearchComp includes time-sensitive retrieval and historical financial investigations, and DREAM evaluates the factual and temporal validity of research reports (Hu et al., 2026; Avraham et al., 2026). Data-agent taxonomies describe autonomy through the responsibilities assigned to humans and systems (Zhu et al., 2025b). We use this perspective to organize the systems in our evaluation into fixed answering, planning within a configured workflow, and delegated research. These groups cover different ways of carrying out the source finding and interpretation needed for the same temporal task.

Paired evaluation and response analysis. Contrast sets test whether predictions respond correctly to controlled input changes (Gardner et al., 2020). TSC-QA uses the evidence cutoff as the controlled change and scores the two answers jointly. The early and late outcomes distinguish success at both cutoffs from success at only one or neither. For analysis beyond final answers, recent work localizes errors in agent trajectories and tracks faithfulness and citation mistakes across calls (Wang et al., 2026; Hirsch et al., 2026). Our response analysis records temporal error categories and supporting passages in saved answers and source records.

3 Benchmark Construction

TSC-QA contains 200 question pairs across eight scientific fields. Each pair holds the scientific question, answer options and option order fixed while changing the evidence cutoff. The reference answer changes with the eligible evidence, following the selection rule stated in the question. Figure 2 shows the six construction steps; Appendix Figure A1 reports field coverage.

The monophonic-rank example in Figure 1 follows two positions on the same graph class: **A: NP-complete (claimed)** in the 2020 preprint and **B: polynomial-time** in the 2023 corrigendum (Dourado et al., 2022; 2024). The corrigendum refutes the proposed reduction and supplies the later algorithm. Applying the question's source-precedence rule selects A at 7 October 2020 and B at 30 May 2023.

1. Discover related literature. Starting from seed papers found through disciplinary searches, we follow citing papers, responses and linked versions. The *version-aware document graph* $G_D = (V_D, E_D)$ retains each document version as a node with a source link and public-date metadata. A typed edge $(u, r, v) \in E_D$ records that version u cites, responds to, or revises version v (Zhang et al., 2020). Following citations forward finds later sources; the stored citation edges point back to the cited versions.

2. Extract scoped claims. GPT-assisted extraction produces *source-linked claim records* $x = (p, s, e, d)$ containing an assertion p, scientific scope s, supporting passage e and source version $d \in V_D$. Scope preserves the object or population, conditions and outcome, including qualifiers such

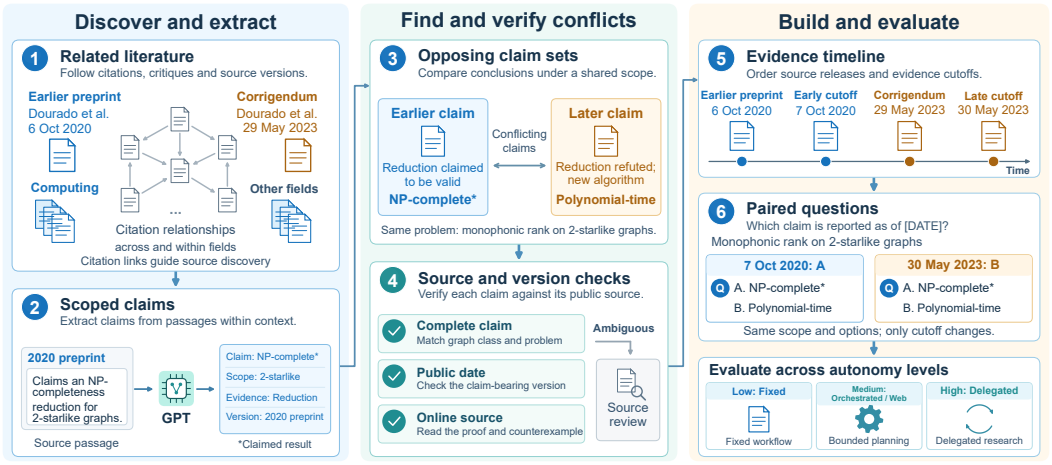


Figure 2: Constructing paired questions from a scholarly correction. Steps 1–2 link document versions to scoped claims; Steps 3–4 identify opposing positions and verify their passages and public dates; Steps 5–6 assign evidence cutoffs and render the pair. Positions A and B follow the monophonic-rank example. The discovery network is schematic; the bottom strip links the pairs to the workflows in Section 4.

as *on 2-starlike graphs* (Magnusson & Friedman, 2021). Each assertion retains its own passage, stated status and attribution; a document can yield several records (Appendix A.2).

3. Identify opposing claim sets. We compare claims for incompatible positions under compatible populations, conditions and outcomes (Thorne et al., 2018; Wadden et al., 2020). Each contrast forms a *local opposition bipartite graph* $B = (C_A, C_B, E_{\text{opp}})$: the disjoint claim sets support positions A and B, and an edge links scope-compatible records with proposed opposing assertions. Multiple records can support either position. In the example, the contrast is between the claimed validity of the reduction and its refutation by a counterexample.

4. Verify claims and source versions. Full-text review checks each claim and opposition edge against the complete passage, scope and qualifications. Scope mismatches lead to revised claims or removed edges. Version links and backward references establish the *claim origin*, the earliest verified public date of a source version containing the complete claim within the inspected history. Repeated claims retain this earlier origin. Each verified record links the position, passage, claim-bearing version and public date (Appendix A.1).

5. Build the evidence timeline. The early cutoff falls on or after the earlier claim origin and before the later origin; the late cutoff falls on or after the later origin. Evidence is eligible when the complete claim was publicly available by the cutoff. We apply the question's stated rule to the eligible positions. A *latest-origin* question selects the position whose complete claim first appeared most recently; other items specify correction, version precedence or source reporting. The example admits A early and both positions late, then selects B through source precedence. Appendix A.2 gives the assignment and unique-answer checks.

6. Render and check paired questions. We combine the shared scope, options, selection rule and answer format with the two cutoff dates. Mechanical checks require identical question text after date replacement, unchanged option order and distinct valid reference labels. Each pair retains both answers and their supporting sources and serves as the common test unit in Section 5.

4 Evaluation Setup

All configurations answer the same 200 question pairs. The workflows determine how the task is carried out; the evidence comparisons vary the source information supplied with each question.

Paired task and scoring. Each system receives the scientific question, cutoff, answer options and output format, with early and late questions answered in separate calls. *Strict-pair accuracy* requires both answers to be correct. Early and late accuracy use the same 200 pair IDs. Missing, failed and invalid responses count as incorrect. Appendix A defines scoring and matching.

Research workflows. Table 1 distinguishes how research decisions are organized. *Fixed* follows a prescribed answering procedure. In *Orchestrated*, a controller builds a question- and target-specific multi-agent task graph; the named target executes its nodes and produces the answer. *Web* lets the target plan queries within bounded retrieval. *Delegated* systems choose successive searches, reading, follow-up and stopping. The figure groups reflect this allocation of decisions (Zhu et al., 2025b).

Table 1: Workflow control and figure groups.

Workflow	Procedure	Group
Fixed	Prescribed	Low
Orchestrated	Task graph	Medium
Web	Bounded search	Medium
Delegated	Research loop	High

We evaluate 31 model–workflow configurations: twelve open targets from Qwen, Llama, Gemma and DeepSeek in Fixed and Orchestrated workflows, GPT-Luna and GPT-Sol in Fixed and Web workflows, and the delegated Codex-Luna, Codex-Terra and Codex-Sol systems. Appendix B and Figure A2 give implementations, tools and budgets; Section 7 discusses benchmark scope.

Question only supplies no prepared source packet; Fixed runs answer from model knowledge, and other workflows use their configured research procedures. The 28 open-target and GPT configurations cover six input/task conditions in Table 2. Each small evidence table averages the twelve open targets equally, preserving the same Fixed and Orchestrated comparison throughout.

5 Results and Analysis

The results address three questions: how well the workflows recover both answers, which supplied information helps them do so, and how they handle claims and dates in their saved responses.

5.1 Performance across Research Workflows

Figure 3 shows representative configurations on the Question-only task. The groups distinguish Fixed answering, planned research through Orchestrated or Web workflows, and complete delegated research systems. Table 2 gives all open-target and GPT scores.

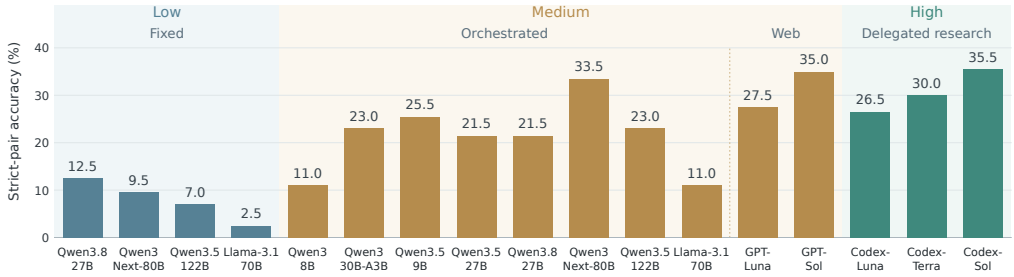


Figure 3: Question-only strict-pair accuracy for selected configurations. No prepared source packet is supplied; systems use the capabilities of their workflow. Low uses Fixed answering, Medium uses Orchestrated target models or GPT Web, and High uses complete Codex research systems. Each bar requires both cutoff answers to be correct. Table 2 gives the complete open-target and GPT Fixed/Web scores.

Across twelve open checkpoints, mean pair accuracy rises from 7.17% with Fixed answering to 20.38% with the Orchestrated workflow. All twelve targets improve; the respective model-level ranges are 2.00–12.50% and 11.00–33.50%. Bounded Web search also improves the same-target GPT comparisons: GPT-Luna rises from 10.50% to 27.50%, and GPT-Sol from 16.00% to 35.00%. The delegated Codex-Luna, Codex-Terra, and Codex-Sol systems reach 26.50%, 30.00%, and 35.50% paired accuracy.

Table 2: All model-level scores (%) on 200 pairs. **A:** Question only supplies no prepared packet; Claim summaries supplies claims without source identities or audited chronology; Claims + dates adds identities, public dates, ordering and cutoff guidance. **B:** Standard supplies a summary supporting the reference answer; Guided adds an instruction to use it under the cutoff. **C:** Evidence-only identifies the later source's position without chronological selection. Orch. denotes Orchestrated; GPT rows use Web. Pair accuracy requires both cutoff answers; the shaded columns score the later target alone.

Target model	**A. Source information** Pair accuracy						**B. Evidence use** Pair accuracy				**C. Recognition** Late accuracy	
	Question only		Claim summaries		Claims + dates		Standard		Guided		Evidence-only	
	Fixed	Orch.	Fixed	Orch.	Fixed	Orch.	Fixed	Orch.	Fixed	Orch.	Fixed	Orch.
Qwen3-8B	6.5	11	8	27	63.5	71.5	84	79.5	85.5	79.5	90	79.5
DeepSeek-R1-14B	7	18	7.5	27.5	79.5	66.5	71.5	80	88	78.5	94	92.5
Qwen3-30B-A3B	8.5	23	12.5	36	87.5	80	83.5	86	86	85	91.5	93
Llama-3.1-8B	2	20.5	4	28.5	32	38.5	71	81	78	80.5	89	92
Gemma-3-12B	9	17.5	5	23.5	27	36	44.5	52	42	51.5	79	78
Gemma-3-27B	8.5	18.5	10.5	46.5	72	73	85	94.5	90	94.5	96	98.5
Qwen3.5-9B	7	25.5	26	36.5	73	70.5	69.5	78.5	76	78	89.5	85
Qwen3.5-27B	6	21.5	21.5	44	72.5	70.5	67	78.5	85.5	83	82.5	86
Qwen3.8-27B	12.5	21.5	46	70	94	96.5	79.5	93	89	93	88	88
Qwen3-Next-80B	9.5	33.5	22.5	55.5	82.5	81	62.5	86.5	83	79	82.5	82.5
Llama-3.1-70B	2.5	11	11	19	19	32.5	22.5	35	24	26	51	63
Qwen3.5-122B	7	23	14	26.5	41	48	29.5	50.5	40.5	54	67.5	74
	Fixed	Web	Fixed	Web	Fixed	Web	Fixed	Web	Fixed	Web	Fixed	Web
GPT-Luna	10.5	27.5	34.5	40.5	64	68	42.5	49	42.5	53	71.5	68.5
GPT-Sol	16	35	32.5	46.5	72	77	39.5	47	49	55	71	71

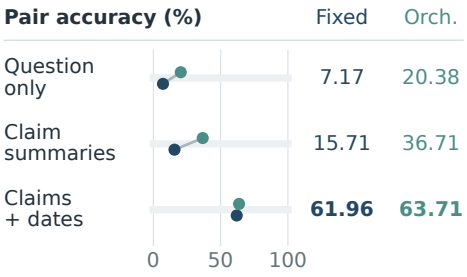
The delegated systems perform better at the later cutoff. Early accuracy trails late accuracy by 10.00, 10.00, and 22.00 percentage points for Codex-Luna, Codex-Terra, and Codex-Sol, respectively. Codex-Sol reaches 74.50% late accuracy, 52.50% early accuracy, and 35.50% paired accuracy. These results show why both dates matter: a workflow can answer the later question while missing the earlier position. The evidence comparisons examine how supplying relevant claims, their source histories, and instructions for using them changes recovery across the pair.

5.2 Answering with Supplied Evidence

Table 2 reports all model-level scores in three groups: source information, evidence-use instructions, and claim recognition. Its header marks the metric for each group. The three small tables summarize the comparisons below, averaging the same twelve open targets equally over all 200 pairs. GPT Fixed/Web results remain visible in the full table. Changes in percentage points use unrounded means.

Relevant claims and their source histories. All three conditions in Table 3 retain the original cutoff questions. *Question only* supplies no prepared source packet. *Claim summaries* provides relevant claims and their scientific qualifications, with explicit source identities and audited chronology withheld. *Claims + dates* provides claim content together with source identities, verified public dates, source ordering, and guidance on applying the cutoff. The comparison examines how these source-information packages help recover the two answers; Appendix B documents their preparation.

Table 3: Source information. Mean pair accuracy (%).

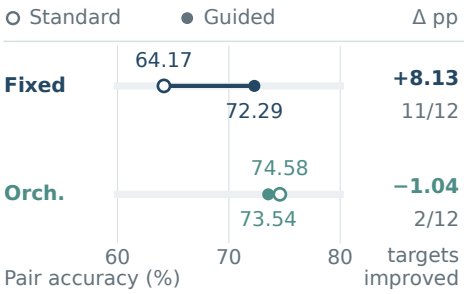


Providing claim summaries improves mean pair accuracy by 8.54 points for Fixed and 16.33 points for Orchestrated. Linking the claims to their sources and public dates produces a larger increase. Claims + dates reaches 61.96% for Fixed and 63.71% for Orchestrated, gains of 46.25 and 27.00 points over Claim summaries, with improvements in every open-model row under both workflows. The Fixed increase also holds under equal discipline weighting, from 14.90% to 61.79% (Table A6).

The mean Orchestrated–Fixed gap is 13.21 points with Question only, 21.00 with Claim summaries, and 1.75 with Claims + dates. With the organized source history supplied, both workflows recover the pair at similar average rates. Orchestrated scores higher for seven targets and Fixed for five; Group A of Table 2 gives the individual results.

Using a selected supporting source. Group B of Table 2 tests how the workflow uses a selected supporting source. Here, each cutoff question receives a summary of a source supporting its reference answer. *Standard Temporal QA* (Standard) appends this summary to the original question. *Evidence-Guided Temporal QA* (Guided) additionally directs the workflow to use the supplied summary while preserving every time restriction. Both conditions require answers at the early and late cutoffs, so Table 4 uses strict-pair accuracy. This comparison examines evidence-use instructions after the supporting source has been selected.

Table 4: Guidance with a selected source. Mean pair accuracy (%); changes use unrounded means.



Explicit guidance increases Fixed mean pair accuracy from 64.17% to 72.29%, while Orchestrated changes from 74.58% to 73.54%. The instruction helps eleven of twelve Fixed rows; Orchestrated has two improvements, seven declines, and three ties. Guidance therefore has its clearest benefit in the fixed answering procedure. These summaries also let us compare recognition of the later claim with answering under a cutoff.

Table 5: Late accuracy and matched early misses. The bottom block uses all model–pair cases.

Late accuracy (%)	Fixed	Orch.
Evidence-Only	83.38	84.33
Guided, late	84.33	86.25
Change (pp)	+0.96	+1.92

Both late correct; Guided early wrong

Share of all model–pair cases (%)

Fixed	**10.29**
Orch.	**10.67**

Recognizing the claim and answering under a cutoff. *Evidence-Only QA* (group C in Table 2) uses the later target's supporting summary and asks which offered position it supports, with chronological selection removed. We compare it with the *late answer* from Evidence-Guided Temporal QA. The upper block of Table 5 reports single-target late accuracy (%): the task changes from identifying the supplied claim to answering the original cutoff question. The lower block tracks pairs where both late tasks are correct but the Guided early answer is wrong, as a percentage of all pairs. Appendix B records source-summary matching and preparation details.

Mean late accuracy is high in both tasks, while the change varies across models. Restoring the temporal question improves six Fixed rows, decreases five, and leaves one unchanged; for Orchestrated, six improve and six decline. The matched outcomes show what high late accuracy leaves unresolved: 10.29% of all Fixed model–pair cases and 10.67% of Orchestrated cases have both late tasks correct but an incorrect Guided early answer. The pattern occurs in every open target, ranging from 2.0–27.5% for Fixed and 2.5–30.0% for Orchestrated (Table A5). Recognizing the supplied claim and using it at the later cutoff can coexist with an earlier-answer error.

5.3 Response Analysis

To examine how systems handle the original task, we review saved Question-only errors against the question, cutoff and reference answer. The review records how claim interpretation, source dating and evidence use appear in the saved answers and research records.

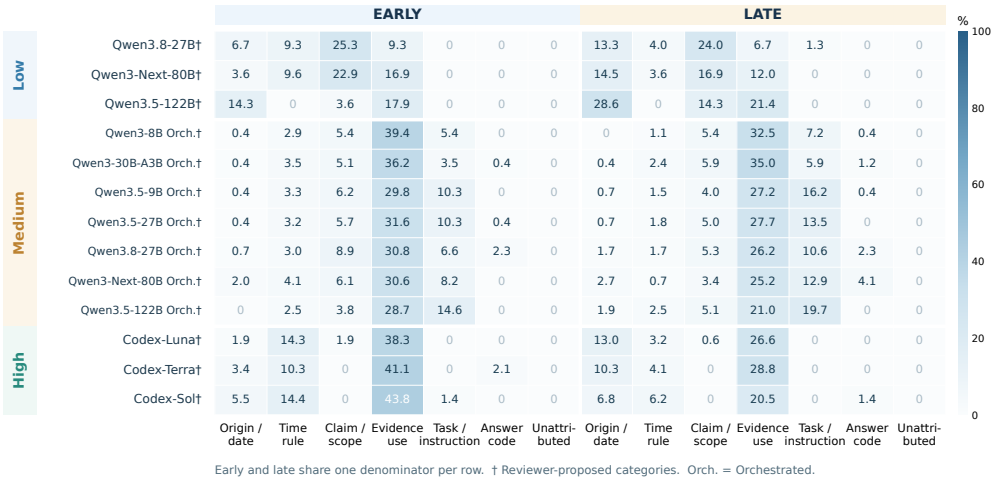


Figure 4: Assigned response patterns for reviewed Question-only errors. Low uses saved Fixed responses, Medium uses historical Orchestrated node records, and High uses retrospective Codex explanations. The † marks reviewer-proposed categories. EARLY and LATE counts share a denominator within each row and jointly sum to 100% before rounding. Colors use a common scale; Appendix D documents the audit populations.

Figure 5 follows the local input–output perspective of Hirsch et al. (2026). The review first checks the reference answer, scores both cutoff answers, and distinguishes observable outcomes such as using the later answer too early or retaining the earlier answer too late. A Luna-based reviewer then examines the saved response and available research records in two passes. The response categories are origin/date, claim/scope, temporal rule, evidence use, task instruction, and answer mapping. As shown in the figure, insufficient response evidence leaves the failure cause unattributed; a proposed category alone does not establish that cause.

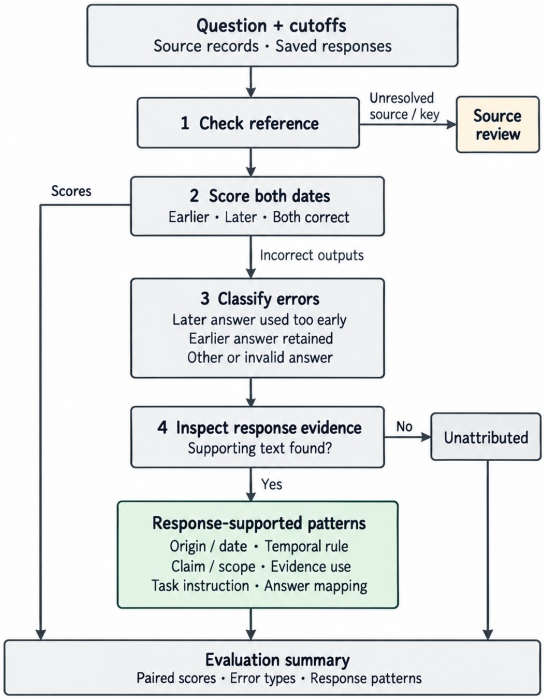


Figure 5: Response review links paired scores to answer patterns and supporting text.

The profiles cover three types of records: saved Fixed responses for Low, historical Orchestrated node records from before the context-repair rerun for Medium, and retrospective Codex explanations for High. Figure 4 summarizes reviewer-proposed response categories for these populations. Appendix D gives the population matching and passage checks.

In the Fixed profiles, claim/scope accounts for 49.3% of assignments for Qwen3.8-27B and 39.8% for Qwen3-Next-80B. Origin/date mapping is the largest category for Qwen3.5-122B, at 42.9%. These patterns concern interpreting the complete scientific position and connecting that position to its public origin, the information supplied together in Claims + dates.

Evidence use accounts for 61.1% of pooled assignments in Medium and 66.4% in High. In every High row, more evidence-use assignments occur at the early cutoff: their shares are 38.3–43.8% early

and 20.5–28.8% late, using the joint early–late denominator. Evidence-use assignments are especially common at the early cutoff in the reviewed High responses.

Supplementary matched cases across evidence conditions show revision-date substitution, selection by the order of supplied text, and mapping a correctly described claim to the wrong option (Table A7). Their source records let a reader check the scientific scope, public origin, cutoff rule and returned label.

6 Discussion and Conclusion

TSC-QA evaluates whether research systems can recover the evidence state requested by a question. Its paired questions hold the wording and options fixed across a documented change in the supported answer. On the Question-only task, research-enabled workflows attain higher paired scores than Fixed answering, but the delegated systems' stronger later-cutoff performance leaves many earlier answers unresolved. Pairing the two dates makes this gap visible in the same scientific question.

The evidence conditions separate two requirements for answering these questions: identifying the relevant claims and placing them in their source history. Supplying sources and public dates improves every open target under both workflows and brings their mean paired scores closer together. Explicit guidance helps most Fixed targets, while its effect varies under Orchestrated answering. The value of a workflow thus depends on the information available and the instructions for using it.

Accurate claim recognition and later answering can still coexist with an incorrect earlier answer. The construction records connect these paired outcomes to the scientific scope, public origin and cutoff rule behind each reference answer. Together with the reviewed responses, they support closer analysis of where a system loses track of changing evidence. TSC-QA makes recovery across evidence changes a measurable part of research-agent evaluation.

7 Limitations

TSC-QA contains 200 question pairs selected around documented scholarly changes in eight fields. The collection emphasizes cases with source-supported answer changes, following the challenge-set approach to evaluation (Kiela et al., 2021). Its scores describe performance on these pairs under their stated source-selection rules. Multiple-choice answers make the two dates directly comparable; open-ended research reports involve additional decisions about explanation and synthesis. The workflow comparison includes complete configurations with different models, tools and inference budgets. The Claims + dates condition combines source identities, public dates, ordering and cutoff guidance, with preparation details retained by source stream. Response analysis uses saved Fixed outputs, historical Orchestrated node records and retrospective explanations from the delegated systems. The profiles summarize reviewer-proposed response categories rather than established failure causes.

AI use statement

We used GPT models to assist with scoped claim extraction during benchmark construction (Section 3). Codex-Luna assisted with the qualitative coding of saved responses and the associated support records (Appendix D). We also used ChatGPT/Codex for feedback on the study's methodology, interpretation of experimental results, literature search, manuscript organization and editing, reference formatting, LaTeX preparation, figure design, and visualization code. The benchmark scores are computed from the stored target-model answers; the AI-assisted response annotations are described separately in the paper. The authors are responsible for the final text, claims, analyses and artifacts.

Reproducibility statement

Section 3 describes benchmark construction. Appendix A specifies the scoring population, metrics and matching procedure, and Appendix B records the input conditions, model configurations and execution protocols. Appendix D documents the response-analysis populations, coding procedure and normalization. These records distinguish benchmark scoring from the accompanying response annotations.

REFERENCES

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650(8103):857–863, 2026.

Elad Ben Avraham, Changhao Li, Ron Dorfman, Roy Ganz, Oren Nuriel, Amir Dudai, Aviad Aberdam, Noah Flynn, Elman Mansimov, Aditya Kalyanpur, et al. Dream: Deep research evaluation with agentic metrics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9879–9904, 2026.

Jiehan Cheng and Zhicheng Dou. Dailyqa: A benchmark to evaluate web retrieval augmented llms based on capturing real-world changes. *arXiv preprint arXiv:2505.17162*, 2025.

Mitre C Dourado, Vitor S Ponciano, and Rômulo LO da Silva. On the monophonic rank of a graph. *Discrete Mathematics & Theoretical Computer Science*, 24(Graph Theory), 2022.

Mitre C Dourado, Vitor S Ponciano, and Rômulo LO da Silva. Corrigendum to" on the monophonic rank of a graph"[discrete math. theor. comput. sci. 24: 2 (2022)# 3]. *Discrete Mathematics & Theoretical Computer Science*, 25(Graph Theory), 2024.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Zihan Guan, Qiao Jin, Guangzhi Xiong, Fangyuan Chen, Mengxuan Hu, Qingyu Chen, Yifan Peng, Zhiyong Lu, and Anil Vullikanti. Large language models lack temporal awareness of medical knowledge. *arXiv preprint arXiv:2605.13045*, 2026.

Eran Hirsch, David Wan, Han Wang, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. Who is the agent to blame? localizing faithfulness and citation mistakes in agentic deep research. *arXiv preprint arXiv:2608.24306*, 2026.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Dongyuan Mutu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, et al. Finsearchcomp: Towards a realistic, expert-level evaluation of financial search and reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 46402–46428, 2026.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020.

Zhen Jia, Philipp Christmann, and Gerhard Weikum. Faithful temporal question answering over heterogeneous sources. In *Proceedings of the ACM Web Conference 2024*, pp. 2052–2063, 2024.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124, 2021.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *Advances in Neural Information Processing Systems*, 38:120091–120131, 2026.

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.

Ian Magnusson and Scott Friedman. Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4651–4658, 2021.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International conference on machine learning*, pp. 31210–31227. PMLR, 2023.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*, 2024.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

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.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Jiaming Wang, Ziteng Feng, Jiangtao Wu, Ruihao Li, Qianqian Xie, Yuxiang Ren, He Zhu, Xueming Han, Fanyu Meng, Junlan Feng, et al. Where do deep-research agents go wrong? span-level error localization in agent trajectories. *arXiv preprint arXiv:2606.02060*, 2026.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Yi Zhang, Zachary Ives, and Dan Roth. "who said it, and why?" provenance for natural language claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4416–4426, 2020.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, 2025a.

Yizhang Zhu, Liangwei Wang, Chenyu Yang, Xiaotian Lin, Boyan Li, Wei Zhou, Xinyu Liu, Zhangyang Peng, Tianqi Luo, Yu Li, et al. A survey of data agents: Emerging paradigm or overstated hype? *arXiv preprint arXiv:2510.23587*, 2025b.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, and Yu Wang. Evolvebench: A comprehensive benchmark for assessing temporal awareness in llms on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16173–16188, 2025c.

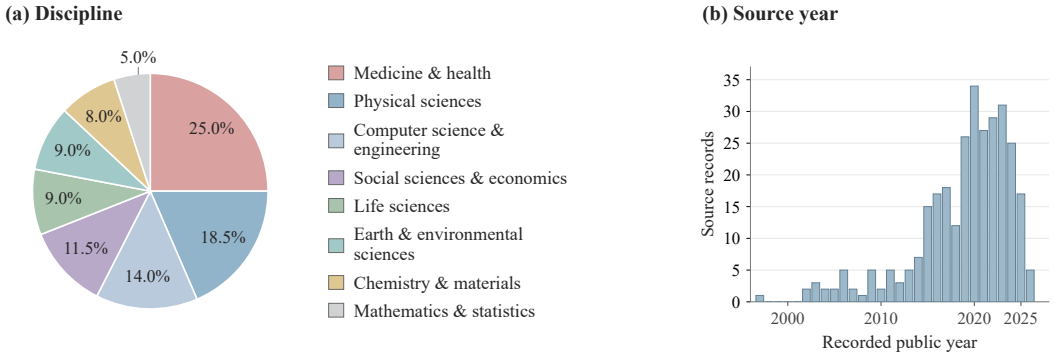


Figure A1: Disciplinary and temporal coverage. (a) Percentage of paired samples in each broad field. (b) Supporting source records by recorded public year. Previously resolved bibliographic aliases are merged; repository versions with the same stable identifier are counted once, regardless of question reuse. Cached publication years are used where available, otherwise the recorded public-source year. These inventory dates are not question cutoffs.

A Evaluation Coverage, Scoring and Reproduction

Metrics. Let S be the full benchmark set of question pairs, $|S|$ its size, and m the evaluated model configuration. For pair i, y_i^e and y_i^l are the reference options at the early and late cutoffs; equivalently, they are $y_i(c_i^{\text{early}})$ and $y_i(c_i^{\text{late}})$ under the applicable item rule. The predicted options are $\hat{y}_{im}^e$ and $\hat{y}_{im}^l$, where the hat marks a model prediction. The superscripts e and l abbreviate early and late. We compute

$$\text{PairAcc}(S, m) = \frac{1}{|S|} \sum_{i \in S} \mathbf{1}[\hat{y}_{im}^e = y_i^e] \, \mathbf{1}[\hat{y}_{im}^l = y_i^l]. \tag{1}$$

Here $\mathbf{1}[\cdot]$ is 1 when the enclosed statement is true and 0 otherwise. Multiplying the two indicators gives credit only when *both* answers in a pair are correct. The displayed formula is a proportion; multiplying by 100 gives the percentages reported in figures and tables. Marginal early and late accuracies use the same pair IDs as the paired score. Evidence-Only QA evaluates its single later target. For every model and workflow, missing targets, failed or incomplete requests, and completed responses without a valid final selection are scored as incorrect. For K_i options in pair i, independent uniform guessing has expected pair accuracy $|S|^{-1} \sum_{i \in S} K_i^{-2}$. Repeating one label fails a pair because its reference labels differ.

Let $n_{11}, n_{10}, n_{01}, n_{00}$ count both-correct, early-only, late-only and neither-correct pairs. In each subscript the first digit refers to the early answer and the second to the late answer; 1 means correct and 0 means incorrect. These counts refer to the same set S and model configuration m. Then $\text{PairAcc} = n_{11}/|S|$, whereas $\text{LateAcc} = (n_{11} + n_{01})/|S|$. Their difference is exactly the fraction of *late-correct, early-wrong* pairs. We also report $n_{11}/(n_{11} + n_{01})$: the probability of early success conditional on late success, defined when at least one late answer is correct. These complementary summaries quantify how often success at the later date extends to both dates using the same predictions.

Coverage, matching and snapshot. Every model and workflow is scored on the complete set of 200 pairs, including all GPT Fixed and Web conditions. Question identity, reference labels, valid-option flags and stored predictions are checked before scoring. Marginal and paired scores use identical IDs within each configuration. Codex-Sol's confirmed pair accuracy is 35.50%; Table A1 gives marginal and paired results for the open-target and GPT workflows; Appendix E reports the delegated systems' corresponding scores. Independent uniform guessing yields 23.33% expected pair accuracy on this benchmark.

The result snapshot combines the available workflow runs, supplied GPT archives and Fixed missing-only supplement. Imported outputs are matched to the benchmark question, temporal view and reference answer. GPT archives use the latest recorded attempt for each matching target. The cell-level manifest retains input versions and scored IDs. Response-category diagnostics use completed

Table A1: Orch. denotes the Orchestrated workflow. Early, late and strict-pair accuracy (%). Each triplet uses identical scored pair IDs. Unqualified open-model and GPT labels denote Fixed; Orchestrated and Web are separate workflows. High Codex-system question-only results appear in Figures 1 and 3. Every score uses the recorded predictions and the full-benchmark scoring rule defined above.

Model / workflow	Question only			Claim summaries			Claims + dates		
	Early	Late	Pair	Early	Late	Pair	Early	Late	Pair
Qwen3-8B	38.0	54.0	6.5	34.5	54.5	8.0	72.5	88.0	63.5
DeepSeek-R1-14B	36.0	51.5	7.0	30.5	52.0	7.5	84.5	93.5	79.5
Qwen3-30B-A3B	37.0	52.5	8.5	52.0	49.5	12.5	95.0	92.0	87.5
Llama-3.1-8B	19.0	26.0	2.0	37.5	52.0	4.0	57.0	67.5	32.0
Gemma-3-12B	45.0	51.5	9.0	28.5	51.5	5.0	48.5	70.5	27.0
Gemma-3-27B	43.0	54.0	8.5	41.0	63.5	10.5	79.0	91.0	72.0
Qwen3.5-9B	31.5	48.0	7.0	59.0	58.5	26.0	84.0	87.5	73.0
Qwen3.5-27B	27.5	44.5	6.0	55.5	54.0	21.5	86.5	85.5	72.5
Qwen3.8-27B	43.0	56.0	12.5	68.5	72.0	46.0	94.5	99.0	94.0
Qwen3-Next-80B	41.0	56.0	9.5	59.0	54.5	22.5	89.0	92.5	82.5
Llama-3.1-70B	45.5	55.0	2.5	41.5	58.5	11.0	37.5	54.0	19.0
Qwen3.5-122B	28.5	53.0	7.0	49.0	50.5	14.0	69.0	66.0	41.0
Qwen3-8B + Orch.	36.5	51.5	11.0	55.0	68.0	27.0	88.0	82.0	71.5
DeepSeek-R1-14B + Orch.	49.5	55.5	18.0	44.5	70.5	27.5	72.5	90.0	66.5
Qwen3-30B-A3B + Orch.	53.5	62.0	23.0	55.5	73.0	36.0	87.5	89.0	80.0
Llama-3.1-8B + Orch.	47.5	60.0	20.5	45.0	73.0	28.5	56.5	74.5	38.5
Gemma-3-12B + Orch.	51.5	55.0	17.5	51.0	65.0	23.5	58.5	73.0	36.0
Gemma-3-27B + Orch.	48.5	62.0	18.5	60.5	84.0	46.5	78.0	94.0	73.0
Qwen3.5-9B + Orch.	54.0	63.0	25.5	65.0	66.5	36.5	85.0	84.0	70.5
Qwen3.5-27B + Orch.	48.5	61.0	21.5	66.0	75.0	44.0	87.0	82.0	70.5
Qwen3.8-27B + Orch.	47.5	68.0	21.5	83.5	86.5	70.0	96.5	100.0	96.5
Qwen3-Next-80B + Orch.	54.5	72.5	33.5	72.0	80.5	55.5	90.5	89.0	81.0
Llama-3.1-70B + Orch.	42.0	62.5	11.0	44.5	59.0	19.0	61.0	61.0	32.5
Qwen3.5-122B + Orch.	50.5	60.5	23.0	58.5	59.0	26.5	72.0	73.0	48.0
GPT-Luna	32.0	60.5	10.5	60.5	60.0	34.5	71.0	69.5	64.0
GPT-Luna + Web	44.5	64.5	27.5	63.0	63.0	40.5	72.0	72.5	68.0
GPT-Sol	38.5	72.5	16.0	55.5	61.0	32.5	77.5	78.5	72.0
GPT-Sol + Web	51.5	77.5	35.0	62.5	65.5	46.5	78.0	82.0	77.0

matching outputs, with review membership recorded separately from the full-benchmark scoring population.

Matching prior runs to the benchmark. The scorer matches question text and temporal view, resolves duplicated text using the original pair ID, and recomputes correctness against the frozen reference answers. Ambiguous matches are excluded from the imported prediction records and scored as missing if no valid matching output remains. For Standard Temporal QA/Evidence-Only QA/Evidence-Guided Temporal QA, an oracle-source-selecting gold mismatch also excludes the imported record because the packet depends on the reference answer. New-literature outputs are joined through their frozen input dataset.

The released prediction records link each answer to its question, temporal view, gold label and input provenance. The accompanying coverage audit identifies unmatched questions, ambiguous duplicate prompts and oracle-gold mismatches. The archive preserves the frozen questions and individual workflow answers used to recompute the scores.

The reproducibility materials include executable scoring checks, the typeset tables and vector figures. Validation covers strict-pair and marginal scores, composition, matched transitions, table cells, figure

endpoints and joint heatmap normalization. The construction figure can be regenerated without changing the underlying data. Offline scoring and document compilation require neither credentials nor new model calls.

A.1 Source Scope and Timeline Details

Human verification of reference answers. One PhD-level reviewer checked the gold answers for 50 questions with GPT assistance. The reviewer confirmed all 50 gold answers, yielding a verification pass rate of 100% on the reviewed sample.

Running-example scope. The monophonic-rank example uses the 2-starlike graph class as its common scope. The original preprint claims NP-completeness for every fixed $k \geq 2$; the corrigendum refutes the cited reduction and gives an algorithm for the broader starlike class. Thus its algorithm covers the original subclass. The direct opposition concerns the validity of the reduction. The figure's complexity labels summarize the original claim and the corrected algorithm; the relationship between P and NP remains separate from this proof correction. Both records are by Dourado, Ponciano and da Silva (Dourado et al., 2022; 2024).

Formal timeline and date-event checks. For pair i, let τ_i^{old} and τ_i^{new} be the verified origins of the earlier and later opposing positions. They are the two corresponding values of τ_{ij} in Equation 3. Let c_i^{early} and c_i^{late} be the cutoffs assigned in Step 5 and used to render the questions in Step 6. The word superscripts identify the position or view. The intended temporal contrast requires

$$\tau_i^{\text{old}} \leq c_i^{\text{early}} < \tau_i^{\text{new}} \leq c_i^{\text{late}}. \tag{2}$$

The first cutoff admits the earlier position and precedes the later one; the second admits both. Both cutoffs are inclusive. All offered positions still require checking under the question's stated selection rule. The running example gives 6 October 2020 $\leq$ 7 October 2020 $<$ 29 May 2023 $\leq$ 30 May 2023.

The date audit separates the public appearance of the complete claim from submission, acceptance, metadata deposit and notice dates. It inspects versions and backward references for earlier in-scope support. For arXiv, this includes upload versus public announcement, historical release timing, weekends and delays. Publisher and repository events are compared against the claim-bearing text. Earlier complete support changes the assigned origin.

A discovery page and an eligible source can have different dates. For example, a 2024 review may link to a 2019 study. For a 2020 cutoff, the 2019 text is the candidate evidence; the later review helps locate it. The qualifying date belongs to the version containing the claim, rather than the page through which it was discovered. Whether an answering system retrieves that source is a separate evaluation question.

A.2 Construction Details and Answer Assignment

Document and claim records. Each document-version node stores its work identifier, version identifier, URL and public-date metadata. Identifiers merge repeated references to the same version while keeping distinct versions separate. Response edges include comments, replies and reanalyses; revision edges include corrections and version updates. Claim records retain scientific status, such as a conjecture versus an established result, and attribution when a passage reports another author's position. Different passages from one paper can support different positions; neither common authorship nor a document-level link alone establishes claim opposition.

Reference-answer assignment. For a latest-origin question, let i identify a pair, j an offered option, and K_i the number of options. Let τ_{ij} be the verified origin of option j's complete claim and c the evidence cutoff. The eligible option indices are $E_i(c)$, and the selected reference option is $y_i(c)$:

$$\begin{aligned} E_i(c) &= \{j \in \{1, \ldots, K_i\} : \tau_{ij} \text{ is verified and } \tau_{ij} \leq c\}, \\ y_i(c) &= \arg \max_{j \in E_i(c)} \tau_{ij}. \end{aligned} \tag{3}$$

The first line filters by date; the second returns an option index, mapped to its displayed letter. All offered positions are checked. An empty eligible set, tied origins or an unverified supporting source

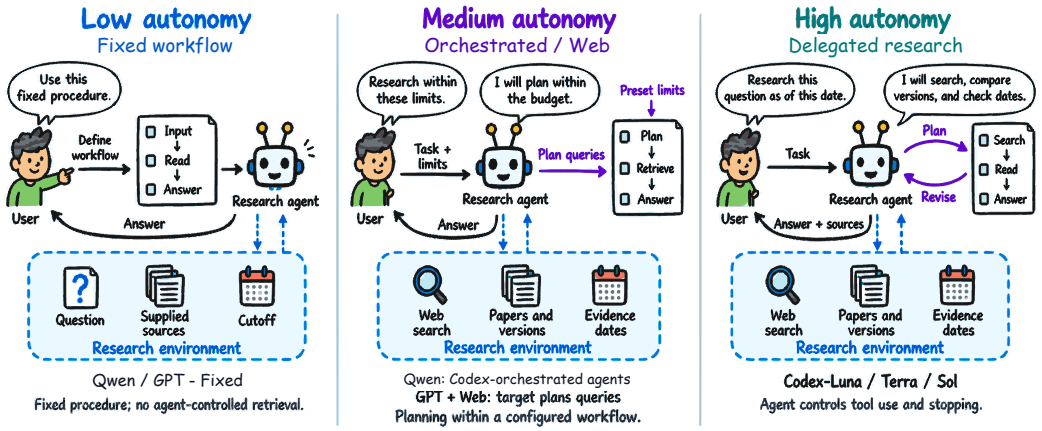


Figure A2: Workflow control in the three experimental settings. Low uses a fixed answering procedure (Fixed); Medium permits controller-led orchestration (Orchestrated) or target-planned bounded search (Web); High delegates the research loop to the complete Codex-Luna/Terra/Sol systems. These settings organize comparisons of paired success as research decisions shift from the experimenter to the system. Source packets are determined separately by the evidence condition.

requires review before assigning a unique answer. Items specifying correction or version precedence, or asking what a particular source reported, use that stated rule instead. Equation 2 gives the intended cutoff ordering. A first document version may lack a claim added later, whereas a later article may repeat an earlier claim; origin assignment follows the claim-bearing version rather than the most recently found document.

Paired-format checks. For pair i, the dates c_i^{early} and c_i^{late} identify its early and late views. Checks require exactly one of each, identical wording after replacing the cutoff, identical options and order, and two different reference labels drawn from those options. Each rendered question includes the scientific scope, inclusive cutoff, evidence-selection rule and final-answer format. These mechanical checks validate the paired format; full-text review in Step 4 supplies the source-based justification for its answers.

B Evidence Contracts and Archived Configuration

B.1 Workflow Groups and Research Control

The evaluation spans fixed answering, planned workflows and delegated research. Table 1 maps the workflows to the figure groups; Table A3 lists the model configurations; Figure A2 shows their research control. Low, Medium and High are the corresponding figure groups. This organization uses the responsibility-allocation perspective of Zhu et al. (2025b). The source information provided with a question is specified separately below.

B.2 Source Inputs and Question Variants

Question only: the full task. The system receives the benchmark question and answers using the capabilities of its workflow. Finding or recalling the relevant evidence is part of answering in this condition. Fixed answerers rely on their model knowledge; research-enabled workflows can retrieve evidence.

Claim summaries: provide the competing content. The original cutoff question is accompanied by question-relevant source summaries. They preserve the scientific claims and their qualifications, while explicit source identities and audited chronology are withheld. This tests whether access to the claim content helps even when the model must work out its temporal applicability. Naturally occurring dates may remain in the source text.

Claims + dates: make temporal eligibility explicit. The source packet includes identities, public dates, source ordering and guidance on applying the cutoff. Comparing this condition with Claim summaries tests whether organizing the evidence by source and public date helps recover the appropriate historical position. The source-information package supplies both dates and the information needed to connect them to claims.

Shared evidence for the three task controls. Standard Temporal QA, Evidence-Only QA and Evidence-Guided Temporal QA all supply a summary of the source supporting the evaluated target's reference answer. Reference-based source selection is shared across this branch; the model's task, rather than the relevance of the supplied source, is varied. The temporal conditions use the supporting source for each cutoff. Evidence-Only QA uses the later target, allowing a matched comparison with the late temporal question.

Standard Temporal QA. The original cutoff question is followed by the supporting summary. This preserves the standard question wording and measures whether access to the right source is sufficient for answering. It is the baseline for testing whether explicitly directing evidence use helps.

Evidence-Guided Temporal QA. The original cutoff question and its answer-supporting summary are retained. An explicit instruction directs the model to base its answer on the supplied evidence and preserve every time restriction. Relative to Standard Temporal QA, this tests evidence-use guidance with the temporal task held fixed. Relative to Evidence-Only QA on matched late targets, it tests whether recognition of the supplied position carries over when chronological selection is required.

Evidence-Only QA. The scientific scope and answer positions are retained, while publication-date, cutoff and first-public-record requirements are removed. The instruction asks the model to select the position supported by the supplied summary. This tests claim recognition without chronological selection. Its score is single-view accuracy on the later targets.

The two contrasts answer different questions. Standard Temporal QA versus Evidence-Guided Temporal QA concerns whether to use available evidence under the original task; Evidence-Only QA versus Evidence-Guided Temporal QA concerns using that evidence with versus without temporal requirements. The latter uses late accuracy on both sides. Question only–Claim summaries–Claims + dates remains a separate comparison of claim availability and provenance. Source summaries retain scientific qualifications, and common model–question IDs and stored summary matches determine the corresponding controlled comparisons.

Document-to-summary preparation. The archived evidence pipeline uses the evaluated model to extract question-relevant claims and consolidate them. The extraction instruction preserves assumptions, scientific qualifications, equations and naturally occurring dates while prohibiting final option selection. Explicit source identity and audited chronology are removed from Claim summaries packets; Claims + dates attaches them with ordered provenance and a cutoff-use instruction. Standard Temporal QA/Evidence-Only QA/Evidence-Guided Temporal QA use reference-selected sources to evaluate answering with supporting evidence. Historical records retain different levels of runtime-prompt and extraction-cache detail; their available inputs are recorded with the source provenance.

The original legacy stream uses 8,000-character full-text windows with 400-character overlap and prepares Claim summaries and dated version-chain packets separately. The extension stream uses 24,000-character windows with the same overlap and shares the extracted summaries between Claim summaries and Claims + dates. Documented extraction and consolidation budgets are 1,536 and 2,400 tokens. The protocol associated with each imported output records its preparation settings and supports matched within-stream checks.

Evidence placement and distractors can affect source use (Liu et al., 2024; Shi et al., 2023), motivating input checks alongside question-ID matching. Standard Temporal QA and Evidence-Guided Temporal QA select the supporting source for the evaluated view. Standard Temporal QA appends the summary to the original cutoff question; Evidence-Guided Temporal QA additionally directs use of that summary while retaining every time restriction. In the open-model and original normal-GPT Evidence-Only QA/Evidence-Guided Temporal QA design, the later target and supporting summary are shared while the task and its instruction change. Web runs can retrieve different pages for different prompts.

Table A2: Source inputs and question variants used in Tables 3–5. The three task controls share reference-selected, answer-supporting evidence. Standard versus Evidence-Guided Temporal QA tests explicit evidence-use guidance under the original question; Evidence-Only versus Evidence-Guided Temporal QA tests transfer from claim recognition to temporal answering using matched late accuracy. Claim summaries versus Claims + dates compares the two source-information inputs in group A.

Condition	Supplied evidence	Question or instruction	Metric
A. Source information for the original task			
Question only	None	Original cutoff question, answered with the model knowledge or research actions available to its workflow.	Pair
Claim summaries	Prepared source content	Original cutoff task; explicit identity and audited chronology withheld.	Pair
Claims + dates	Claims, sources and public dates	Adds identity, public dates, ordering and cutoff guidance to the source condition.	Pair
B. Using a selected supporting source			
Standard Temporal QA	Answer-supporting summary	Original cutoff question: is the right evidence enough?	Pair
Evidence-Guided Temporal QA	Answer-supporting summary	Original cutoff question plus explicit evidence-use guidance.	Pair
Evidence-Only QA	Answer-supporting summary (late target)	Read the claim without chronological selection.	Late

The shared-source interpretation applies where the stored inputs confirm it; the aggregate comparison also reflects these preparation differences.

Checkpoints and decoding. Table A3 gives the checkpoint identifiers and open-model sampling policies. The open-model inference configuration uses vLLM, seed 0 and a 32,768-token context window. Thinking is enabled in the configuration for summary and answer stages where supported; Qwen3.8 uses medium reasoning effort and Qwen3.5/3.8 use minimum-$p = 0$. The large-model queue uses four-way tensor parallelism, whereas some imported historical outputs used two-way parallelism. The saved generations and their source configurations determine which protocol applies. The High Codex systems follow the separate question-only, web-enabled procedure in Section 4.

Answer budgets and scoring. The open-model launch configuration allocates up to 12,000 answer tokens, 1,536 tokens per chunk extraction and 2,400 per consolidated document summary. Imported legacy runs retain their original budgets and parser outcomes. Correctness is recomputed against the reference answers and the offered labels; a completed output with no valid selection is incorrect. Score reproduction uses the packaged predictions; inference reproduction additionally requires the full runtime inputs and source caches.

Workflow controls. *Fixed* applies the same prescribed answering procedure across questions, using the target model named in the table row. Its inputs are the question and the packet specified by the evidence condition; Question only supplies no packet and the recorded Fixed generations use model knowledge. *Orchestrated* denotes controller-led orchestration with that target model: a controller implemented in Codex with `gpt-5.6-luna` generates a question- and target-specific multi-agent task graph, and the target checkpoint executes its nodes and generates the answer. Our implementation adapts the Tongyi DeepResearch workflow structure while retaining the named target checkpoint as the answerer. The workflow name determines who organizes the computation; the evidence condition determines what source information is supplied. In *GPT Web*, the target proposes queries and answers from pages collected by a fixed retrieval stage. In the *High Codex systems*, Codex-Luna, Codex-Terra

Table A3: Model and workflow configurations. The upper panel gives open-model decoding: temperature controls sampling, top-p is a cumulative-probability filter and top-k limits the candidate tokens. Dashes omit sampling filters for greedy decoding. Open-model runs use context 32,768 and seed 0. The lower panel lists GPT targets, delegated Codex systems and the workflow controller; numerical sampling filters are not specified for these interfaces. [a]The recorded Codex completion runner uses xhigh; invocation-level metadata for reused historical calls remain attached to their source records. The workflow controller generates the task graph; the target checkpoint executes its nodes.

Target model	Checkpoint identifier	Temp.	Top-p	Top-k
Open checkpoints: Fixed and Orchestrated target answerers				
Qwen3-8B	Qwen3-8B	0.0	–	–
DeepSeek-R1-14B	DeepSeek-R1-Distill-Qwen-14B	0.0	–	–
Qwen3-30B-A3B	Qwen3-30B-A3B	0.6	0.95	20
Llama-3.1-8B	Llama-3.1-8B-Instruct	0.0	–	–
Gemma-3-12B	Gemma-3-12B-IT	0.0	–	–
Gemma-3-27B	Gemma-3-27B-IT	0.0	–	–
Qwen3.5-9B	Qwen3.5-9B	1.0	0.95	20
Qwen3.5-27B	Qwen3.5-27B	1.0	0.95	20
Qwen3.8-27B	Qwen3.8-27B	1.0	0.95	20
Qwen3-Next-80B	Qwen3-Next-80B-A3B-Thinking	0.6	0.95	20
Llama-3.1-70B	Llama-3.1-70B-Instruct	0.0	–	–
Qwen3.5-122B	Qwen3.5-122B-A10B-FP8	1.0	0.95	20

Model / component	Identifier	Workflow	Reasoning effort
GPT-Luna	gpt-5.6-luna	Fixed	xhigh
GPT-Luna	gpt-5.6-luna	Web	xhigh
GPT-Sol	gpt-5.6-sol	Fixed	xhigh
GPT-Sol	gpt-5.6-sol	Web	xhigh
Codex-Luna	gpt-5.6-luna	Delegated research	xhigh[a]
Codex-Terra	gpt-5.6-terra	Delegated research	xhigh[a]
Codex-Sol	gpt-5.6-sol	Delegated research	xhigh[a]
Workflow controller	gpt-5.6-luna	Workflow generation	high

or Codex-Sol controls the full search, reading, tool-use and stopping loop. The experimental unit is the complete workflow configuration, including its model, control structure, tools and budgets.

The GPT Fixed/Web Question only contrast preserves the target model, benchmark question-pair IDs and reference key. Both Fixed and Web use the full benchmark denominator in every condition, counting missing or failed outputs as incorrect. Web adds bounded query planning and retrieval. This comparison measures the gain from adding the specified Web procedure to the same GPT target model.

Delegated invocation and isolation. The delegated runs use `gpt-5.6-luna`, `gpt-5.6-terra` and `gpt-5.6-sol`, with web research enabled and `xhigh` reasoning effort. Each invocation receives the question, cutoff, options and final-answer syntax. Early and late views are separate calls. The serial completion runner creates a fresh temporary workspace and ephemeral session, with memories, apps, plugins, hooks and multi-agent operation disabled. It ignores user configuration and project rules, sets the project-document budget to zero, archives the output and removes the workspace. The runner adds no output-token cap. Older parallel runs have less complete isolation metadata; the later runner's controls are documented for the calls to which they apply. Completed answers are reused independently of correctness.

GPT query planning and retrieval. GPT-Luna and GPT-Sol use `xhigh` reasoning in both modes. Fixed directly answers the condition's input. Web uses one query-planning stage, fixed retrieval and answer generation. The target proposes up to four queries within 1,200 planning tokens. The runner collects up to five hits per query, deduplicates URLs and reads up to eight pages, retaining 8,000 characters per page. If planning produces no usable query, a question-prefix query is used. The retrieved text is appended to the answering input.

C Model-Level Results and Matched Diagnostics

C.1 Model Variation in the Main Results

Reading the main result table. The main-text Table 2 reports all six conditions for each open target and GPT Fixed/Web configuration. Tables 3, 4 and 5 summarize the source-information, evidence-use and claim-recognition comparisons, respectively. Each averages the twelve open targets equally. Group A compares source-information conditions, group B compares Standard and Guided instructions, and group C reports later-target claim recognition. The final two columns use late accuracy; all preceding columns use strict-pair accuracy. Evidence-Only QA evaluates the later target alone; its temporal counterpart is the late accuracy used in Table A4. The delegated-system scores appear in Figure 3 and Section 5.1.

Source information across models. Claims + dates improves on Claim summaries for all twelve open targets under each workflow. The resulting pair scores range from 19.0% to 94.0% for Fixed and from 32.5% to 96.5% for Orchestrated. Orchestrated exceeds Fixed for seven targets and falls below it for five, producing the 1.75-point mean gap in Table 3. The GPT rows retain a separate Fixed/Web comparison: Claims + dates raises GPT-Luna from 34.5% to 64.0% under Fixed and from 40.5% to 68.0% under Web; GPT-Sol rises from 32.5% to 72.0% and from 46.5% to 77.0%, respectively.

Guidance across workflows. Moving from Standard to Evidence-Guided Temporal QA improves eleven Fixed targets, with gains up to 20.5 points for Qwen3-Next-80B; Gemma-3-12B decreases by 2.5 points. Orchestrated has two gains, seven declines and three ties: Qwen3.5-27B gains 4.5 points, while Llama-3.1-70B loses 9.0. GPT-Luna is unchanged under Fixed and gains 4.0 points under Web; GPT-Sol gains 9.5 and 8.0 points. These rows show how the effect of the same evidence-use instruction varies with the answering procedure and target model.

Correct late answers with an early miss. The lower block of Table 5 averages the E-miss column of Table A5: 10.29% for Fixed and 10.67% for Orchestrated. The denominator includes all 2,400 model–pair cases per workflow. E-miss requires correct Evidence-Only QA and Guided late answers together with an incorrect Guided early answer. For Fixed Gemma-3-12B, both late tasks are correct on 68.0% of pairs: 40.5% also have the early answer correct and 27.5% have an early miss. This matched comparison identifies which pairs retain an early error after successful late answering.

C.2 Matched Outcomes and Composition

Claim recognition and late temporal answers. For the full-target Evidence-Only QA/Evidence-Guided Temporal QA late comparison, Qwen3.5-27B gains 10.5 percentage points, Qwen3-30B-A3B loses 1.0, Qwen3-Next-80B gains 3.5 and Qwen3.5-122B gains 2.5; Gemma-3-12B loses 8.5. Across the twelve Fixed configurations, six improve, five decline and one is unchanged. Across the twelve Orchestrated configurations, six improve and six decline. Missing targets are scored as errors, so these changes combine response behavior with completion coverage. The three-view partition separates an isolated early miss from loss of a previously correct late answer after the task changes. The valid-answer sensitivity statistic reports the late-accuracy difference among targets with valid selections in both conditions.

Composition sensitivity. Discipline-macro strict-pair accuracy is 6.60% for Question only, 14.90% for Claim summaries and 61.79% for Claims + dates, compared with full-target pair-weighted scores of 7.17%, 15.71% and 61.96%. The dated-evidence improvement remains under equal discipline weighting. Table A6 reports discipline and source-stream breakdowns on the same full-target model–pair units across conditions.

Table A4: Matched late outcomes for Evidence-Only QA and Evidence-Guided Temporal QA (%). Each arrow goes from Evidence-Only QA to the Guided late answer; C denotes correct and W denotes wrong. The four outcome columns sum to 100% before rounding. Δ is Guided late accuracy minus Evidence-Only QA accuracy, in percentage points. Δ_{valid} restricts this difference to targets with valid late selections in both conditions. All other columns use the full benchmark of 200 targets, with missing, failed and invalid responses scored as errors. Orch. denotes Orchestrated; unqualified model names denote Fixed. Outcomes are matched by target using Appendix A.

Model / workflow	Matched late outcomes				Change (pp)	
	C→C	C→W	W→C	W→W	Δ	Δ_{valid}
Qwen3-8B	86.5	3.5	2.0	8.0	-1.5	0.0
DeepSeek-R1-14B	91.5	2.5	2.5	3.5	0.0	-2.1
Qwen3-30B-A3B	88.5	3.0	2.0	6.5	-1.0	-1.5
Llama-3.1-8B	87.0	2.0	2.5	8.5	0.5	0.0
Gemma-3-12B	68.0	11.0	2.5	18.5	-8.5	-1.8
Gemma-3-27B	95.0	1.0	0.5	3.5	-0.5	-0.5
Qwen3.5-9B	82.5	7.0	4.5	6.0	-2.5	-2.6
Qwen3.5-27B	80.5	2.0	12.5	5.0	10.5	10.3
Qwen3.8-27B	87.0	1.0	6.0	6.0	5.0	5.0
Qwen3-Next-80B	76.0	6.5	10.0	7.5	3.5	0.5
Llama-3.1-70B	45.0	6.0	9.5	39.5	3.5	2.5
Qwen3.5-122B	52.0	15.5	18.0	14.5	2.5	-1.9
Qwen3-8B + Orch.	79.5	0.0	2.5	18.0	2.5	2.0
DeepSeek-R1-14B + Orch.	88.0	4.5	2.0	5.5	-2.5	0.5
Qwen3-30B-A3B + Orch.	87.5	5.5	1.0	6.0	-4.5	-3.1
Llama-3.1-8B + Orch.	87.0	5.0	2.5	5.5	-2.5	-1.5
Gemma-3-12B + Orch.	71.5	6.5	5.5	16.5	-1.0	-2.1
Gemma-3-27B + Orch.	98.0	0.5	0.0	1.5	-0.5	-0.5
Qwen3.5-9B + Orch.	74.5	10.5	9.0	6.0	-1.5	-2.1
Qwen3.5-27B + Orch.	81.0	5.0	8.5	5.5	3.5	1.1
Qwen3.8-27B + Orch.	86.5	1.5	11.5	0.5	10.0	8.6
Qwen3-Next-80B + Orch.	74.0	8.5	14.0	3.5	5.5	4.5
Llama-3.1-70B + Orch.	54.0	9.0	12.5	24.5	3.5	11.7
Qwen3.5-122B + Orch.	68.5	5.5	16.0	10.0	10.5	5.5
GPT-Luna	64.0	7.5	4.5	24.0	-3.0	2.1
GPT-Luna + Web	65.0	3.5	12.0	19.5	8.5	6.4
GPT-Sol	68.0	3.0	10.0	19.0	7.0	5.6
GPT-Sol + Web	65.5	5.5	12.5	16.5	7.0	4.8

Table A5: Three-view outcomes across Evidence-Only QA and the early and late answers from Evidence-Guided Temporal QA (%). Full: all three correct. E-miss: Evidence-Only QA and Guided late correct, Guided early wrong. Regress: Evidence-Only QA correct, Guided late wrong. Rescue: Evidence-Only QA wrong, both Guided views correct. Joint: remaining outcomes. The five classes partition all 200 benchmark pairs and sum to 100% before rounding, with missing, failed and invalid responses scored as errors. Orch. denotes Orchestrated; unqualified model names denote Fixed. The partition uses matched stored predictions and the scoring rule in Appendix A.

Model / workflow	Full	E-miss	Regress	Rescue	Joint
Qwen3-8B	84.5	2.0	3.5	1.0	9.0
DeepSeek-R1-14B	85.5	6.0	2.5	2.5	3.5
Qwen3-30B-A3B	84.5	4.0	3.0	1.5	7.0
Llama-3.1-8B	76.5	10.5	2.0	1.5	9.5
Gemma-3-12B	40.5	27.5	11.0	1.5	19.5
Gemma-3-27B	90.0	5.0	1.0	0.0	4.0
Qwen3.5-9B	72.0	10.5	7.0	4.0	6.5
Qwen3.5-27B	74.5	6.0	2.0	11.0	6.5
Qwen3.8-27B	84.0	3.0	1.0	5.0	7.0
Qwen3-Next-80B	73.0	3.0	6.5	10.0	7.5
Llama-3.1-70B	22.0	23.0	6.0	2.0	47.0
Qwen3.5-122B	29.0	23.0	15.5	11.5	21.0
Qwen3-8B + Orch.	77.0	2.5	0.0	2.5	18.0
DeepSeek-R1-14B + Orch.	76.5	11.5	4.5	2.0	5.5
Qwen3-30B-A3B + Orch.	84.0	3.5	5.5	1.0	6.0
Llama-3.1-8B + Orch.	79.0	8.0	5.0	1.5	6.5
Gemma-3-12B + Orch.	49.5	22.0	6.5	2.0	20.0
Gemma-3-27B + Orch.	94.5	3.5	0.5	0.0	1.5
Qwen3.5-9B + Orch.	69.0	5.5	10.5	9.0	6.0
Qwen3.5-27B + Orch.	77.5	3.5	5.0	5.5	8.5
Qwen3.8-27B + Orch.	82.5	4.0	1.5	10.5	1.5
Qwen3-Next-80B + Orch.	66.0	8.0	8.5	13.0	4.5
Llama-3.1-70B + Orch.	24.0	30.0	9.0	2.0	35.0
Qwen3.5-122B + Orch.	42.5	26.0	5.5	11.5	14.5
GPT-Luna	40.5	23.5	7.5	2.0	26.5
GPT-Luna + Web	46.0	19.0	3.5	7.0	24.5
GPT-Sol	43.5	24.5	3.0	5.5	23.5
GPT-Sol + Web	44.5	21.0	5.5	10.5	18.5

Weighting. The stratified table uses identical full-target model–pair units across conditions, counting missing answers as incorrect, before splitting by discipline or source stream. A discipline-macro statistic averages the within-discipline pair scores with equal field weights. The reported results are point estimates for the current 200-pair collection; source-group IDs record provenance, while the scoring unit remains the question pair.

Table A6: Fixed strict-pair accuracy (%) by discipline and source stream. All three conditions use identical target model–pair units within a row across twelve open checkpoints; missing answers are incorrect. Discipline macro weights represented fields equally. Provenance gain is Claims + dates minus Claim summaries.

Stratum	Question only	Claim summaries	Claims + dates	Provenance gain (pp)
Computer science & engineering	6.5	13.7	57.1	43.5
Physical sciences	5.6	15.3	56.3	41.0
Earth & environmental sciences	5.6	7.9	61.1	53.2
Chemistry & materials	6.8	9.9	64.6	54.7
Life sciences	7.9	20.4	62.5	42.1
Medicine & health	11.0	19.3	67.0	47.7
Mathematics & statistics	5.8	15.0	60.8	45.8
Social sciences & economics	3.6	17.8	64.9	47.1
Prior	10.6	23.4	57.5	34.1
New literature	6.3	13.7	63.1	49.4
Pair-weighted	7.2	15.7	62.0	46.2
Discipline macro	6.6	14.9	61.8	46.9

C.3 Completed-Response Repair Diagnostics

Diagnostic populations. Accuracy, stratification and all Evidence-Only QA/Evidence-Guided Temporal QA comparisons use the full target set, with missing or failed answers scored as incorrect. Response-category, repair-rate and gold-aware recovery diagnostics instead require completed matched Question only/Claim summaries/Claims + dates responses. Figure A3 uses this completed-response subset. The manifest retains memberships and missing IDs for both scoring scopes.

Temporal recovery and source interpretation. On the completed-response subset defined above, Claims + dates reaches 75.13% early and 82.66% late accuracy. Future-endpoint selections remain in 18.38% of early outputs, compared with earlier-endpoint selections in 12.94% of late outputs. The 9.6-point difference between the 68.9% and 78.5% endpoint repair rates shows that Claims + dates more often repairs stale late answers than anticipatory early answers. This comparison evaluates the joint provenance package of metadata, source ordering and cutoff instructions.

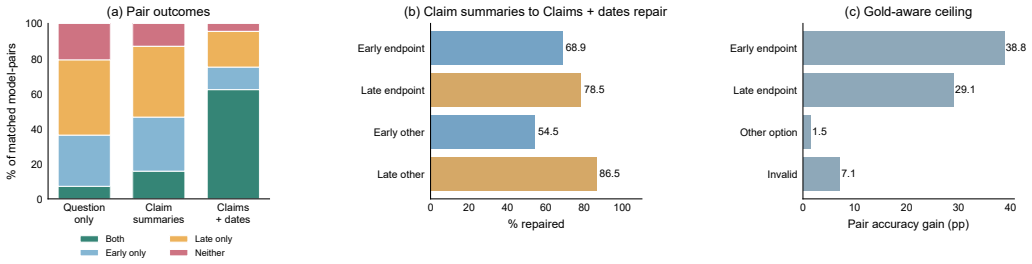


Figure A3: Evidence diagnostics on completed matched responses (missing outputs excluded). (a) Paired outcomes. (b) Percentage of Claim summaries errors repaired by Claims + dates, separating early and late endpoint errors from other-option errors. (c) Gold-aware recovery of Claim summaries pair accuracy, correcting one error type at a time. These category-specific diagnostic ceilings measure the paired accuracy recoverable through gold-based correction.

Matched repair and recovery calculations. For a source error category k, the repair rate is the fraction of those source-condition errors whose matched target-condition answer is correct. The

denominator is the set of source-condition errors in category k. For gold-aware recovery, replace only predictions assigned to category k with their gold labels, leave every sibling and other prediction unchanged, and recompute strict pair accuracy. Its gain is

$$G_k = 100\,[\mathrm{PairAcc}(\hat{y}^{(k)}) - \mathrm{PairAcc}(\hat{y})]. \tag{4}$$

The ceilings are category-specific and generally non-additive: a pair with two different errors may require both corrections. For Claim summaries, future-endpoint-only correction yields +38.82 points, earlier-endpoint-only correction +29.06, other-option correction +1.51 and invalid-answer correction +7.12 on the completed-response subset. These gold-aware values quantify the contribution of each error category to the remaining paired-answer gap.

D Response Coding and Audit Details

This appendix defines the audit populations, primary labels and normalization used in Figure 4. The response analysis retains the saved question, cutoff and reference answer, linking the assigned pattern to the same target used in paired scoring. Figure 5 summarizes the review sequence; the procedures below describe reference checks, outcome coding and response inspection.

D.1 Labels, Review Procedure and Normalization

Primary labels. We use six response categories. *Origin/date* concerns binding a claim to a public event or revision. *Claim/scope* concerns interpreting its content and qualifications. *Temporal rule* concerns applying the stated cutoff criterion. *Evidence use* concerns obtaining or using supporting content. *Task instruction* concerns following the requested answering task. *Answer mapping* concerns converting the selected proposition to an offered label and accepted final format. These names match the categories in Section 5.3; the figure uses the short labels "time rule" and "answer code" for the corresponding categories.

Audit procedure and interpretation. The review follows Figure 5: check the reference, score both dates, classify observable answer outcomes, and inspect the response evidence. Unresolved reference questions return to source review. The heatmap retains reviewer-proposed response categories, including cases where the saved text is insufficient to establish the explanation. Its distributions therefore describe coding hypotheses rather than independently verified failure causes (Jacovi & Goldberg, 2020; Turpin et al., 2023; Lanham et al., 2023).

In Figure 5, *Unattributed* denotes an explanation that cannot be established from the available response evidence. An evidence-use proposal may concern a stated evidence gap or apparent misuse of content; the source artifacts determine which interpretation is supported. Reference review remains separate from scoring against the frozen answer key.

Review inputs and passage checks. The reviewer receives the question, reference snapshot, endpoint origins and complete stored response, including available reasoning or node records. Two Luna/low coding passes are followed by focused checks that the selected excerpts occur in the archive and support the recorded description. These are AI-assisted response annotations. The passage checks allow proposed categories to be compared with the saved text.

Normalization. For model m, the heatmap reports $p_{m,v,k} = 100 n_{m,v,k}/E_m$, where $n_{m,v,k}$ counts reviewed wrong responses in temporal view v assigned to category k, and E_m counts all reviewed wrong responses across both views in that row's audit population. Each row sums to 100% across its early and late blocks. A both-wrong pair contributes two annotations and one failed pair. Missing or unreviewed responses are outside the audit denominator. Pooled shares weight reviewed wrong answers; each plotted row uses its own population.

Across reviewed Medium errors, evidence use accounts for 61.1% of primary assignments, followed by task instruction (19.8%) and claim/scope (11.1%). Combining early and late cells gives an evidence-use share of 49.7–71.8% across Medium rows. Its pooled High share is 66.4%. These summaries describe the reviewed populations and record types specified below.

D.2 Current Qwen Question-Only Coding

The Low profiles use a completed Luna/low re-review of saved Question only errors for Qwen3.8-27B, Qwen3-Next-80B and Qwen3.5-122B. Endpoint identities, questions, temporal views, references, predictions and archive hashes preserve the prior model–question–view membership and early/late denominators. The re-review updates the category assignments on that fixed population. Supplementary Fixed answers remain outside this annotation set.

Each complete saved response is reviewed with its frozen reference context. These archived Fixed runs used model knowledge, with no supplied source packet or retrieval tool. A source mentioned in a response is therefore treated as generated answer content. The annotation checks its relationship to the claim, date and final option under that recorded input condition.

D.3 Orchestrated Question-Only Trajectory Coding

The Medium profiles use archived Orchestrated Question-only responses from before the context-repair rerun. The current accuracy tables use the updated scoring snapshot, while this audit preserves the historical responses and their original model–question–view membership. Original target prompts, outputs and frozen references are retained. Every saved node is checked against its assigned input. Long trajectories are divided at complete-node boundaries, and the reviewed findings are merged while retaining each full node.

The annotation assesses a node relative to its role in the generated workflow: an extraction node, for example, is evaluated against its extraction instruction. Retrieval-related descriptions are checked against the tools available in that recorded condition. Source/reference concerns and rejected factual descriptions are stored separately from the selected hypothesis. Each wrong endpoint contributes one primary label; intermediate-node findings provide supporting context for that assignment.

D.4 Retrospective Review of Delegated Systems

The High profiles use retrospective explanations of reviewed wrong responses from Codex-Luna, Codex-Terra and Codex-Sol. Cases are matched by endpoint identity, question, temporal view, model, final answer and archive hashes. Each wrong response contributes one primary assignment; correct responses are retained separately.

Two Luna/low passes check the proposed explanation against the available records. Each review retains the generated explanation, original excerpts and reference context. Scoring continues to use the saved benchmark answers independently of this annotation.

D.5 Separate GPT Fixed/Web Review Protocol

The GPT reviews preserve model, Fixed/Web mode and evidence condition separately. Reviewers receive the question, frozen endpoint labels, origin metadata and complete returned response. Annotations are retained after exact checks of the response, prediction, question and key. Changed explanatory responses are reviewed by Codex-Luna at high reasoning effort, with a second pass for boundary cases. Each model–mode–condition uses one early-plus-late error denominator; Evidence-Only QA has only a late target. These records form a separate audit from the Question-only profiles in Figure 4. Original responses, selected attempts, reviewer outputs and excerpt checks are preserved.

D.6 Matched Cases

Table A7 illustrates origin-date mapping, revision-date substitution, text-order selection, an option-label mismatch and extraction-instruction interference. The five cases come from an earlier close-reading audit selected to show different response patterns. They provide concrete examples alongside the population summaries. A47 quotes stored reasoning; the other excerpts quote returned answers.

Table A7: Five matched cases connect returned text to distinct response patterns. Case IDs index the released audit records. Each contrast uses the same model and question target under the stated conditions. The A47 excerpt is stored reasoning; the other excerpts are returned answer text.

Case / contrast	Question and audited chronology	Verbatim excerpt and interpretation
A08 Qwen3.5-27B Evidence-Guided Temporal QA: A (wrong); Claims + dates: B (correct)	Split-liver transplantation: B uses a 10-hour cold-ischemia threshold; A uses 8 hours. At the 2020-10-20 cutoff, B originated on 2020-10-10 and A on 2022-04-10.	*"qualifying document (Received 2020 Dec 21)"* Evidence-Guided Temporal QA excludes B using a received date. Claims + dates links B to its eligible public origin and excludes A.
A28 Qwen3.5-27B Evidence-Only QA: A (correct); Evidence-Guided Temporal QA: B (wrong)	Graphene adsorption: A challenges IBMA/SPA validity; B supports it. A originated on 2019-10-02, before the late cutoff of 2019-10-17.	*"arXiv:1910.00361v2 ... 23 Nov 2019"* Evidence-Guided Temporal QA uses the later v2 date to exclude A. Evidence-Only QA correctly identifies the claim supported by the comment.
A25 Gemma-3-12B Evidence-Only QA: A (correct); Evidence-Guided Temporal QA: D (wrong)	Cross-submodularity: D originated on 2017-06-16; the corrected characterization A originated on 2017-10-03. Both precede the late cutoff of 2017-10-04.	*"we select the one that appears last in the provided text"* Evidence-Guided Temporal QA assumes simultaneous origins and chooses by text order. The task instead selects the later public origin.
A20 Qwen3.5-27B Evidence-Guided Temporal QA: B (wrong); Claims + dates: A (correct)	TCGA reprocessing: A says signals disappeared; B says they persisted. A originated on 2023-07-31, before the early cutoff of 2023-08-23; B originated in 2024.	*"which corresponds to Position B"* Evidence-Guided Temporal QA describes the disappearing-signals claim and its eligible 2023 source, then calls it B. Claims + dates preserves the A/B mapping.
A47 Qwen3.5-27B Standard Temporal QA: source notes; Evidence-Guided Temporal QA: A (correct)	Cross-submodularity at the late cutoff of 2017-10-04. The corrected characterization A originated on 2017-10-03; the earlier endpoint D on 2017-06-16.	*Reasoning: "However, I must NOT say "This matches Position A". I must just state the source claim."* The Standard Temporal QA reasoning matches the theorem to the gold option A, then suppresses that answer to follow extraction constraints. Evidence-Guided Temporal QA selects A.

E SUPPLEMENTARY RESULTS AND REPLICATION CONSIDERATIONS

Marginal scores and paired outcomes. Codex-Luna, Codex-Terra and Codex-Sol have early accuracies of 56.50%, 58.50% and 52.50%, and late accuracies of 66.50%, 68.50% and 74.50%. Their late-to-pair drops are 40.00, 38.50 and 39.00 percentage points. Among late-correct pairs, 39.85%, 43.80% and 47.65%, respectively, also have a correct early answer. Sol's late–early gap is 22.00 points, compared with 10.00 for both Luna and Terra. Codex-Sol has more late-only than early-only successes; the difference is smaller for Codex-Luna and Codex-Terra.

Pooling the within-model paired outcomes across these three systems, the reported marginal and pair scores imply that 92.8% of failed pairs have exactly one correct view.

GPT Fixed and Web contrasts. On the full benchmark, GPT-Luna and GPT-Sol Question only pair scores rise from 10.50% and 16.00% in Fixed to 27.50% and 35.00% with Web search, gains of 17.00 and 19.00 percentage points. Their late accuracies rise from 60.50% and 72.50% to 64.50% and 77.50%, still above paired success. Web belongs to Medium; the Codex systems constitute High. Appendix B documents the configurations.

Directional errors and repeated answers. Early errors account for 56.5%, 56.8% and 65.1% of the errors for Codex-Luna, Codex-Terra and Codex-Sol, respectively. An other-option error selects a different offered label from either endpoint; an invalid answer has no parsed offered label.

Future-endpoint selections account for 97.70% and 98.80% of Luna and Terra early errors, respectively. Because the reference labels differ across cutoffs, a repeated label leaves at least one view wrong. On binary items, the opposing endpoint is the only valid incorrect selection. The paired and matched-evidence analyses connect these directional output patterns to the information and selection requirements of the task.

Replication records. The scoring snapshot combines retained outputs with the missing-only Fixed supplement. The archived question IDs, reference labels, predictions, evidence inputs and source-stream metadata define each reported comparison. Input and configuration records preserve the available information about tool access, extraction, inference budgets and invocation isolation. Luna classification calls add response annotations to this archive; benchmark scores continue to use the stored target-model answers. Appendix A specifies scoring reproduction and Appendix B documents the source configurations.