Historical Backtesting for Scientific Question Discovery: A Protocol and Astronomy Pilot

Evaluating AI-Generated Scientific Questions Against Future Scientific Progress

Hui Mao and Claude Code

August 12, 2026

Abstract

Systems that generate scientific research questions are currently evaluated by expert scores, LLM-as-judge ratings, or curated case studies—all subjective, none falsifiable. We propose a different standard: *future scientific engagement as an observable, falsifiable proxy for one important dimension of a question's value.* We formalize **historical backtesting** as an evaluation protocol for scientific question discovery: a system generates questions from a corpus frozen at a historical cutoff, the questions are frozen before any access to later literature, and a temporally isolated future corpus determines—via fixed retrieval, a citation-constrained judge, and human adjudication—whether each question was subsequently answered, partially addressed, independently posed, or ignored, and whether its underlying premise was supported or refuted. The protocol is model-agnostic: any system that emits frozen questions can be scored, and all metrics are defined independently of how questions are produced. We release a reproducible astronomy benchmark instance (cutoff 2020-12-31): a 2,512-paper past corpus, a temporally isolated 1,891-paper future corpus (2021–2026), frozen questions, retrieval records, adjudicated outcome labels, and one-command metric computation, plus a submission interface and four reference baselines evaluated through the identical pipeline. In an initial set of ten questions generated by an evidence-graph system from pre-cutoff literature only, all ten were substantively engaged by later literature: two were answered, seven partially addressed, one independently posed and still open—and one question's underlying premise (a strongly subsolar water abundance for HD 209458 b) was subsequently refuted by three independent analyses, the exact convergence the question called for. A scaled second instance (astronomy v1L: 424 frozen baseline questions against a 5,754-paper future corpus) then stress-tests the small-sample conclusions and revises two of them: engagement rates *do* discriminate at $n = 125$ (random 73% vs. direct-LLM 96%, $p < 10^{-4}$), and premise refutation is rare but not unique—chasing highly cited results catches refutations on 3.2% of questions, while random templates stay at zero and acquire a measurable 11%-answered floor. Finally, we turn the benchmark's deepest threat—LLM weights that have read the future—into its subject: a generator decomposition (LLM-only vs. deterministic evidence-structure vs. structure-plus-LLM verbalization) crossed with a four-cutoff temporal stress test (2010–2024, 798 judged questions) whose last window postdates the model's training. LLM-only generation shows *memorized relevance without specific foresight*: near-ceiling engagement and the closest phrasing to future literature at every cutoff, but an answered rate flat across the training boundary, indistinguishable from random templates, and zero premise refutations outside the deepest-history era. A weight-free structural generator finds engaged questions at every cutoff, and adding the LLM back as a pure verbalizer refutes premises in every era including the post-training one—locating the foresight signal in pre-cutoff evidence structure, with the LLM as a separable realization layer. We analyze failure modes and threats to validity, and position this pilot as the initial instance of a benchmark that any question-discovery system can run against.

1 Introduction

A growing family of systems claims to generate scientific research questions, hypotheses, or ideas (Lu et al., 2024; Wang et al., 2024; Baek et al., 2024). How do we know whether any of them is good at it? Today, essentially every evaluation falls into one of three patterns: an *expert score* (a panel rates novelty and significance on a Likert scale), an *LLM score* (a language model rates the same properties), or a *case study* (a handful of generated ideas is narrated persuasively). All three share the same defect: they are subjective. Expert panels disagree with each other and with themselves (Si et al., 2024); LLM judges inherit the biases of their training distribution and can be steered by phrasing; case studies are selected by the authors. None of these evaluations can be *wrong* in a way that data could demonstrate, which is another way of saying none of them measures anything.

Science itself offers a harder criterion. Research questions are bets about where inquiry should go next, and the scientific community eventually settles those bets: it invests observing time, funds follow-ups, writes papers that answer some questions, poses others independently, and refutes the premises of a few. This suggests a measurement standard—deliberately a proxy, not a definition:

> *Future scientific engagement provides an observable, falsifiable proxy for one important dimension of a question's value.*

We do not claim engagement *defines* value: community attention carries popularity bias (Section 11), and a question can be excellent yet ignored for want of an instrument. The claim is narrower and stronger where it counts—engagement is the one dimension of value that is observable from frozen public data, and therefore the one on which systems can be compared without asking anyone's opinion.

The standard becomes an evaluation protocol the moment we rewind the clock. Fix a historical cutoff. Give a system only the literature available before the cutoff. Freeze the questions it generates. Then let the literature published *after* the cutoff—which the system never saw—grade the bet: Was the question answered? Substantially advanced? Independently posed by working scientists? Ignored? Was its underlying premise confirmed, or refuted? We call this procedure **historical backtesting**, by analogy with the evaluation of trading strategies on held-out past data (Bailey et al., 2014), and with recent forecasting benchmarks that score models against events occurring after training (Zou et al., 2022).

Crucially, the protocol contains no question generator (Figure 1). It takes a frozen list of questions as input and returns outcome labels and metrics as output. Any system—an evidence-graph pipeline, a prompted LLM, a citation heuristic, a human scientist—can be evaluated under identical conditions. This is what makes it a benchmark rather than a validation appendix for one particular architecture.

Contributions. We make three claims, in decreasing order of strength:

1. **Protocol (strong).** We formalize historical backtesting as an evaluation protocol for scientific question discovery: temporal isolation rules, a question-freezing requirement, fixed future-evidence retrieval, a two-dimensional outcome taxonomy that separates a question's fate (`answered`, `partially_addressed`, `posed_but_open`, `not_addressed`) from its premise's fate (`supported`, `refuted`, `weakened`, `still_plausible`, `not_applicable`), and metrics defined over those labels (Sections 3 and 4).

2. **Benchmark instance (medium).** We release a reproducible astronomy instance with temporally isolated past and future corpora (2,512 and 1,891 papers; cutoff 2020-12-31), frozen

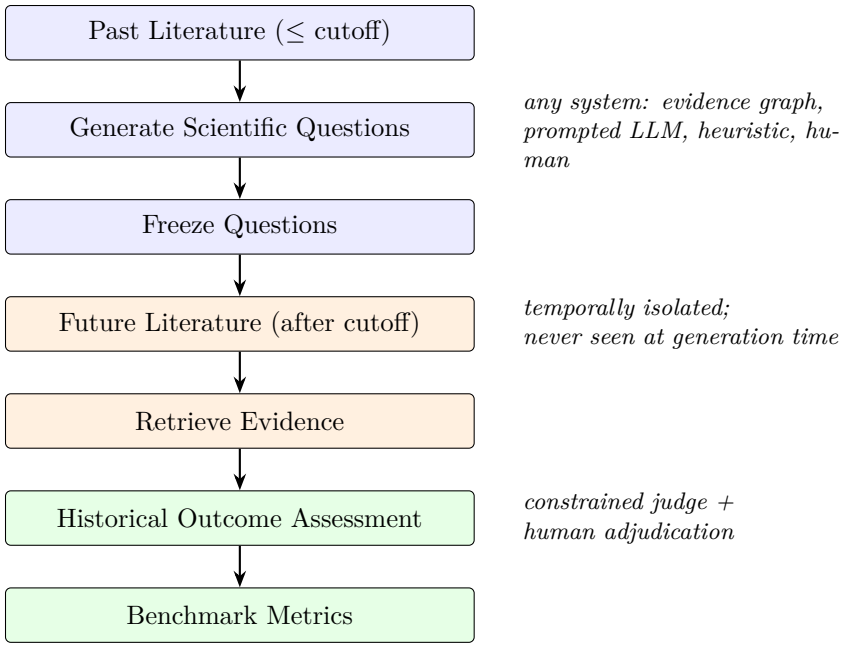


Figure 1: The historical backtesting protocol. No question generator, evidence graph, or particular LLM appears in the loop: the protocol evaluates *frozen question lists*, whatever produced them.

questions, released retrieval records, adjudicated labels, one-command metric computation with CI-enforced reproducibility, a submission format, and four reference baselines evaluated through the identical pipeline (Sections 5 and 6).

3. **Empirical pilot (cautious).** In an initial set of ten questions generated from pre-cutoff evidence only, all ten were substantively engaged by the 2021–2026 literature, including one whose underlying premise was subsequently refuted by three independent analyses (Section 7). A scaled instance with 424 baseline questions then revises two of the small-sample baseline conclusions—engagement discriminates at scale, and premise refutation is rare (0–3.2%) but not unique to the evidence-graph system—demonstrating that the benchmark updates its own conclusions as n grows (Sections 6 and 8). A generator decomposition and a four-cutoff temporal stress test then separate memorization from foresight empirically: the foresight signal survives in a weight-free structural generator at every cutoff, while LLM-only generation shows relevance without resolution in every era (Section 9). With $n = 10$ questions from one system in one domain, we do *not* claim that AI can reliably identify the most valuable future scientific questions; we claim the protocol can tell us, eventually, whether it can.

What we believe is ultimately most useful here is not any single result but the change of category: scientific question quality moves from a matter of taste ("this question seems interesting") to a measured, comparable quantity ("under historical backtesting, 100% of this system's questions were engaged by later literature; 10% led to a premise refutation"). Every artifact needed to run the protocol—data, code, labels, and checks—is public, and Astronomy v1 is offered as the first instance of the benchmark, not as its definition.

2 Related Work

Automated scientific discovery and question generation. Computational discovery has a long lineage, from rule-based rediscovery of physical laws (Langley et al., 1987) through closed-loop robot scientists (King et al., 2009) to the Nobel Turing Challenge's call for AI scientists (Kitano, 2021). Recent LLM-based systems generate research ideas, hypotheses, or full papers: literature-based generation (Wang et al., 2024), agentic idea refinement (Baek et al., 2024), and end-to-end automated research (Lu et al., 2024). Literature-based discovery pioneered the underlying intuition that recombining published evidence can anticipate findings later verified empirically (Swanson, 1986). Our work is orthogonal to all of these: we do not propose a better generator; we propose the missing evaluation.

Evaluating generated ideas. Existing evaluations are dominated by human preference and LLM scoring. Si et al. (2024) ran a large expert study comparing human and LLM research ideas on rated novelty and excitement—the most rigorous instance of the expert-score paradigm, and still a measurement of *opinion at generation time* rather than of what the ideas turned out to be worth. LLM-as-judge scoring inherits known biases (position, verbosity, self-preference) and, for questions about the future, cannot be validated against ground truth at all. Historical backtesting replaces both with an outcome variable that exists independently of any rater: the subsequent behavior of the scientific community.

Backtesting and forecasting benchmarks. Scoring a strategy on held-out history is standard in quantitative finance, along with well-documented failure modes—overfitting to the backtest itself (Bailey et al., 2014)—that motivate our freezing and no-overwrite rules. Forecasting benchmarks score models on events that resolve after training (Zou et al., 2022); retrodictive evaluation with temporal holdouts is likewise used to test whether models anticipate later discoveries (Swanson, 1986; King et al., 2009). We transplant this design to a harder target: not whether a stated event occurs, but whether an open-ended research question earns the community's future investment. Code benchmarks such as SWE-bench (Jimenez et al., 2024) demonstrated how a well-specified task format plus frozen data can reorganize a research area around measurable progress; we aim the same mechanism at question discovery.

Data contamination. Temporal splits are increasingly used to control LLM memorization in evaluation. Our protocol controls the *retrieval* channel completely (corpus manifests, isolation rules, CI checks) and treats the *weights* channel—models whose training data postdates the cutoff—as a declared, audited threat rather than a solved problem (Section 11).

3 The Historical Backtesting Protocol

The protocol evaluates a set of frozen questions $Q = \{q_1, \ldots, q_n\}$ against a future corpus. It has six steps; each is fully specified so that two groups running the same instance obtain the same measurement.

3.1 Step 1: Choose a cutoff

A historical date T (Astronomy v1: 2020-12-31) splits the literature into a *past corpus* Corpus A (everything available up to T) and a *future window* realized as an isolated corpus Corpus B

(strictly after T). The cutoff must be far enough in the past for the community to have had time to act—we recommend ≥ 4 years—and recent enough that the past corpus reflects a modern research frontier.

3.2 Step 2: Generate questions

Any method may generate questions: an evidence-graph pipeline, a prompted LLM, a heuristic over citation statistics, a human expert. The only requirements are (i) the generator consumes *only* Corpus A evidence, and (ii) every question records the pre-cutoff evidence it is grounded in (`source_evidence_ids`). A submission whose source evidence postdates T is invalid, mechanically (`scripts/validate_cutoff.py`).

3.3 Step 3: Freeze

Questions are serialized—identifier, text, cutoff, generating system, source evidence, system-assigned rank—with `frozen: true` *before any access to post-cutoff literature*, and are never edited afterwards. Freezing is the protocol's load-bearing rule: without it, question text drifts toward what the evaluator has meanwhile learned the future contains, and the backtest silently becomes a description of the future rather than a prediction of it (Bailey et al., 2014). In the released implementation frozen files are append-only and guarded by CI; editing a released question mints a new versioned instance rather than overwriting the old one.

3.4 Step 4: Define the future window

Corpus B is collected under its own frozen manifest (query set, date window, deduplication rules) and stored separately from Corpus A; records from Corpus B must never enter the generation pipeline. The Astronomy v1 window is 2021–2026. Bounding the window matters for comparability: "eventually engaged" is not a fixed target, but "engaged within k years" is.

3.5 Step 5: Retrieve future evidence

For each frozen question, the question text and every Corpus B document (title + abstract) are embedded (`text-embedding-3-small`); the top-k documents by cosine similarity ($k = 8$) become the candidate evidence. Retrieval is deliberately fixed and deliberately simple: systems are compared on their *questions*, not their retrievers, and reviewers can inspect exactly which documents the judge saw because retrieval records are part of the release.

3.6 Step 6: Assess outcomes

A judge reads the question and its retrieved candidates and assigns two *independent* labels (Table 1): the fate of the question and the fate of its premise.

Two design decisions deserve emphasis. First, the dimensions are separated because the single most informative outcome a backtest can surface—*the community answered this question by refuting its premise*—is inexpressible in a flat label set: it is simultaneously a resolution (`answered`) and a falsification (`refuted`). Our pilot's headline case (Section 7.1) is exactly of this type. Second, premise refutation is scored as a *success* of the question, not a failure: a question that provokes the community into overturning one of its own published conclusions has done the most a question can do.

Table 1: Outcome taxonomy v1.0. The two dimensions are labeled independently.

Dimension	Label	Meaning
outcome	`answered`	Future evidence substantially answers the question (including by refuting its premise)
	`partially_addressed`	Important, directly relevant progress; core question unresolved
	`posed_but_open`	The community independently poses essentially the same question without resolving it
	`not_addressed`	No meaningful follow-up in the future corpus
premise_status	`supported`	Future evidence confirms the underlying premise
	`refuted`	Future evidence falsifies the underlying premise
	`weakened`	Substantial doubt cast without falsification
	`still_plausible`	The premise was not directly tested after the cutoff
	`not_applicable`	The question rests on no contestable premise

The judge operates under hard constraints enforced outside the model: it may cite only bibcodes from the retrieved candidates (violations are errors, never silently dropped); topical similarity is explicitly insufficient—the cited paper must bear on the question's actual test or premise; unknown labels fall back to the most conservative value and are flagged. Every machine-assigned label then passes *human adjudication* under written guidelines (citation validity, the engagement bar, the outcome/premise split), and the adjudication log is released alongside the labels. No result enters the benchmark unreviewed.

3.7 Submissions

A system is evaluated by submitting a directory containing `questions.jsonl` (the frozen questions) and `metadata.json` (system description, including a mandatory declaration of any LLM components and their versions, for contamination auditing). The benchmark pipeline—isolation checks, retrieval, judging, adjudication, metrics, report—is identical for every submission; the system whose questions we evaluate in Section 7 interacts with the benchmark only through this interface.

4 Benchmark Metrics

Let Q be the n frozen questions of a submission, with outcome labels $o(q)$ and premise labels $p(q)$ as in Table 1, and let $E(q)$ be the set of independent post-cutoff papers cited as supporting evidence for q's label. All rates are over n, so the four outcome rates sum to one.

Coverage vs. answer rate. Coverage asks whether the question pointed anywhere the community went at all; the answer/partial/ open decomposition asks what happened when it got there. A system can maximize coverage with fashionable-topic questions, which is why coverage is never reported alone (Section 11 discusses the popularity confound).

Premise refutation rate. This is the metric we most want the field to adopt. Questions that trigger refutations are the rarest and arguably most valuable output of question discovery—they mark places where the literature's accepted conclusions were wrong and where a well-aimed question preceded the correction. Under the two-dimensional taxonomy the refutation is recorded without erasing the fact that the question was thereby *answered*.

Table 2: Benchmark metrics v1.0. All are computed by `benchmark/metrics.py` from the released annotation records.

Metric	Definition
Coverage (future attention rate)	$\frac{1}{n}\lvert\{q : o(q) \neq \texttt{not_addressed}\}\rvert$ — did future science engage the question at all?
Answer rate	$\frac{1}{n}\lvert\{q : o(q) = \texttt{answered}\}\rvert$
Partial rate	$\frac{1}{n}\lvert\{q : o(q) = \texttt{partially_addressed}\}\rvert$
Open rate	$\frac{1}{n}\lvert\{q : o(q) = \texttt{posed_but_open}\}\rvert$ — the community independently recognized the question
Premise refutation rate	$\frac{1}{n}\lvert\{q : p(q) = \texttt{refuted}\}\rvert$ — questions that led to an established conclusion being overturned
Evidence strength	mean $\lvert E(q)\rvert$ over engaged questions, and the multi-source rate $\frac{1}{n}\lvert\{q : \lvert E(q)\rvert \geq 2\}\rvert$ — is the label supported by multiple independent papers?
Lead time	years between question submission and the first time the community *independently poses* the same question
Community attention	volume of future investment engaging the question: papers, citations, review mentions, and major observing programs (e.g. JWST/HST proposals)

Lead time. If a system poses a question at the cutoff and the community first independently poses it in year $T{+}\ell$, the system led the field by ℓ years; averaged over questions this yields a comparable earliness score. Measuring ℓ requires identifying *community first-posed dates*, which demands careful review-literature annotation we do not yet have. Astronomy v1 therefore reports `mean_lead_time_years: null` rather than a number we cannot defend; the released records do include a weaker, well-defined lower bound—*first-engagement lag*, the years from cutoff to the earliest judge-cited supporting paper (pilot mean 2.9, range 1–5)— which should not be confused with lead time.

Community attention. Beyond binary engagement, the volume of future investment (paper counts, citations to engaging papers, review mentions, dedicated observing programs) reflects how much the community cared. v1 records the ingredients (supporting bibcodes, their venues and years) and reports evidence strength; a calibrated attention index is future work, and Section 11 explains why raw attention must never be the headline metric.

Reporting requirements. A benchmark report must state: the instance and protocol versions, n, all outcome and premise rates, the evidence-strength pair, and either lead time or an explicit null. The released implementation produces exactly this (`results/astronomy_v1/metrics.json`) with one command, and CI fails if the committed numbers do not reproduce from the raw annotations.

5 The Astronomy v1 Instance

Astronomy v1 instantiates the protocol in exoplanet atmospheres—a domain chosen because it uniquely combines a fast-moving literature, structured catalogs, and space-telescope archives, and because the 2021–2026 window contains a natural experiment: JWST began delivering data mid-window, resolving questions that were unanswerable at the cutoff. Table 3 summarizes the instance.

Table 3: Astronomy v1 at a glance. Corpus manifests, frozen questions, retrieval records, and adjudicated labels are all released.

Historical cutoff	2020-12-31
Domain scope	atmospheric composition in transmission spectra; cloud/haze degeneracies; instrument systematics
Corpus A (past)	2,512 deduplicated papers, 2015–2020 (NASA ADS; 500-paper full-text core)
Corpus B (future)	1,891 unique papers, 2021–2026, temporally isolated
Questions	10, frozen, ranked, with pre-cutoff source evidence
Retrieval	`text-embedding-3-small`, cosine, top-8; records released
Judge	`gpt-4.1`, temperature 0, citation-constrained; human-adjudicated

Corpora. Both corpora are defined by frozen manifests—ADS query sets, date windows, deduplication and filtering rules (records without abstracts are dropped)—rather than by bulk data dumps: the manifests are committed, and a script rebuilds either corpus from its manifest via the ADS API. Corpus B's manifest adds targeted follow-up queries for the questions' objects (HD 189733, HD 209458, WASP-12, WASP-121, TRAPPIST-1) so that engagement is measured against the relevant future literature rather than against whatever a generic query happens to return.

Leakage controls. Temporal isolation is enforced mechanically, not editorially: (1) nothing dated after the cutoff may enter Corpus A, including catalog rows updated post-cutoff; (2) every question's source evidence must predate the cutoff; (3) every retrieved document must postdate it; (4) the judge may cite only retrieved candidates; (5) question/retrieval/annotation records must align one-to-one; (6) Corpus B's window must start strictly after the cutoff. All six checks run in continuous integration on every change to the released data, together with a check that the released `metrics.json` reproduces bit-identically from the raw annotations.

Questions. The ten frozen questions were generated by an evidence-graph system (Scientific Question Discovery project, 2026) from Corpus A only: claims with provenance are extracted from the full-text core, cross-paper tensions are detected and typed (observational tensions, methodological challenges, single-dataset conclusions, independent qualifications), and surviving signals are refined into ranked, falsifiable questions. For the benchmark, that system is submission `evidence_graph_v1`—evaluated through the same interface as any future submission. Each released record carries the question text, system rank, signal type, target objects, and pre-cutoff source bibcodes; a curation log documenting human edits made *before* freezing (including a near-duplicate merge and presupposition fixes, Section 10) is released for audit.

What is released. Frozen questions; corpus manifests; per- question retrieval records (model, window, top-k, judge-cited documents, top-1 similarity; full ranked lists with scores are scheduled for v1.1); adjudicated outcome annotations with rationales and supporting bibcodes; the adjudication log; computed metrics and per-question results; four frozen baseline submissions with their full retrieval records (per-document scores included), judge annotations, and reports (Section 6); the scaled v1L instance (manifests, configs, 424 frozen baseline questions, retrieval records with per-document scores, judge annotations, per-system reports, and the statistical comparison script of Section 8); the tension-pair generators, the four-cutoff stress-test corpora manifests, all 798 stress-test judgments, and the deterministic specificity rubric of Section 9; and the complete pipeline code with tests. The data format is deliberately plain (JSONL + JSON manifests) so that other groups can mint new instances—different domain, different cutoff—by writing two manifests and

one config file.

6 Baselines

A benchmark that only ever scored one system would be a validation appendix. Astronomy v1 ships four reference baselines, chosen to bracket the interesting comparisons; each consumes only CORPUS A records, freezes its output before any future-corpus access, and is evaluated through the identical pipeline.

B1: Random claims. Sample random pre-cutoff papers and template their headline result into a robustness question. The floor: any system must beat chance-directed attention.

B2: Direct LLM. Give an LLM (`gpt-4.1`, 60 sampled pre-cutoff abstracts) a request for the most valuable open questions. The "why not just ask GPT?" comparison. Because a modern LLM's weights postdate the cutoff, this baseline is also a contamination probe: performance that vanishes for post-training-cutoff instances indicates memorized hindsight rather than generation ability (Section 11).

B3: Review future work. Extract explicitly posed open questions from pre-cutoff review papers. The strongest natural reference: a discovery system is interesting only if it adds value over questions the community had already written down. Note this baseline should score highly on coverage *by construction*—these questions are known community priorities—so the discriminating metrics are premise refutation and lead time, where a copied question can never lead the field.

B4: Citation leaders. Template follow-up questions from the most-cited pre-cutoff papers. Tests whether chasing prominence matches structured evidence analysis. One caveat is built in: ADS citation counts are fetched at corpus-rebuild time and therefore include post-cutoff citations, so this baseline selects papers with *hindsight* knowledge of which pre-2021 work the future found important—a bias in its favor that a cutoff-dated citation snapshot would remove.

Evaluation conditions. All four baselines were run through the released pipeline: top-8 retrieval with `text-embedding-3-small` against the manifest-rebuilt future corpus restricted to the frozen window (2,384 records, a superset of the frozen 1,891 due to retroactive ADS indexing, containing all 22 citations in the released annotations), then `gpt-4.1` judging at temperature 0 under the citation constraints of Section 3. Baseline labels are *judge-only*: they have not received the human adjudication that the released `evidence_graph_v1` labels did. To make that comparison honest, Table 4 also reports a judge-only rerun of the evidence-graph submission under exactly the baseline conditions. The rerun doubles as a replication check: it reproduces the released mean top-1 retrieval similarity (0.668 vs. 0.666) and lands within one label of the adjudicated results (coverage 90% vs. 100%, answered 10% vs. 20%, refutation 10% = 10%)—the deltas are exactly the two labels adjudication had strengthened, so judge-only scoring reads as the conservative floor of the adjudicated score.

Reading the leaderboard. Three observations, offered with the $n = 10$ caution of Section 7 applying to every row.

Coverage saturates. Every non-random system scores 90–100% on future attention: in a field this active, any fluent, topical question attracts partial engagement within five years. Coverage

Table 4: Astronomy v1 leaderboard ($n = 10$ questions per system). Adjud. = human-adjudicated labels; judge-only rows are directly comparable to each other. Cov. = future-attention rate; Ans. = answered; Part. = partially addressed; Open = posed but open; Ref. = premise refuted; s_1 = mean top-1 retrieval similarity. Lead time is null for all systems until community first-posed dates are annotated (Section 4).

System	Eval	Cov.	Ans.	Part.	Open	Ref.	s_1
`evidence_graph_v1` (Scientific Question Discovery project, 2026)	adjud.	100%	20%	70%	10%	10%	0.666
`evidence_graph_v1` (rerun)	judge	90%	10%	80%	0%	10%	0.668
B4 citation leaders	judge	100%	30%	70%	0%	0%	0.632
B3 review future work	judge	100%	10%	70%	20%	0%	0.593
B2 direct LLM	judge	90%	0%	90%	0%	0%	0.732
B1 random claims	judge	70%	0%	70%	0%	0%	0.611

separates the floor (random claims, 70%, the only system with three `not_addressed` labels) from everything else, and nothing else—which is why the protocol never reports it alone (Section 11).

Answered-rate comparisons need reading, not just ranking. The citation-leader baseline posts the highest judge-only answered rate (30%). Two mechanisms inflate it: its hindsight-biased paper selection (above), and its template—"does the conclusion of highly cited paper X hold?"—which pattern-matches the replication studies that prominent results reliably attract, so the judge can mark it answered whenever the community re-examined a famous result for any reason. What the template cannot do is risk anything: B4 refuted no premise, and by construction a question of the form "is the famous result right?" poses nothing the community was not already testing. The evidence-graph submission's answered questions, by contrast, specified particular tests (Section 7.1) and include the leaderboard's only refuted premise—on both adjudicated and judge-only rows.

The contamination probe registers a signal. The direct-LLM baseline has by far the highest retrieval similarity to future literature ($s_1 = 0.732$ vs. 0.593–0.668 for every other system) and the highest mean support count (4.6 cited papers per question)—its questions are phrased in the way the 2021–2026 literature would come to phrase them, consistent with weights that have read that literature. Yet it resolves nothing: 0% answered, 0% refuted, 90% partial. The pattern suggests phrasing-level contamination without commitment to falsifiable specifics—broad, well-aimed questions that everything engages and nothing settles. Prospective instances (Section 12) will separate the two channels definitively.

Baseline generators, frozen submissions, retrieval records with full per-document scores, judge annotations, and per-system reports are all released; the leaderboard file is designed for external submissions to append to. All three observations above are drawn from ten questions per system; Section 8 re-examines them on a scaled instance with 424 baseline questions, and two of the three require revision there.

7 Pilot Results: Historical Validation

We ran the full protocol on the ten frozen `evidence_graph_v1` questions. Headline numbers: every question was substantively engaged by the 2021–2026 literature (coverage 100%); two were answered, seven partially addressed, one independently posed and still open; one premise was refuted. Evidence strength: 2.4 supporting papers per question on average, with 60% of labels supported by ≥ 2 independent papers. The earliest judge-cited engagement came 1–5 years after the cutoff (mean 2.9). Table 5 gives the per-question picture.

Table 5: Per-question outcomes for `evidence_graph_v1` on Astronomy v1 (system rank order; $|E|$ = independent supporting papers; year = earliest cited engagement).

Rank	ID	Outcome	Premise	$\|E\|$	Year
1	q_002	`posed_but_open`	`still_plausible`	1	2022
2	q_001	`partially_addressed`	`supported`	3	2021
3	q_004	`partially_addressed`	`supported`	1	2021
4	q_008	`answered`	`refuted`	3	2025
5	q_007	`partially_addressed`	`still_plausible`	3	2024
6	q_010	`partially_addressed`	`still_plausible`	1	2022
7	q_005	`answered`	`supported`	2	2024
8	q_009	`partially_addressed`	`supported`	5	2021
9	q_006	`partially_addressed`	`still_plausible`	1	2025
10	q_011	`partially_addressed`	`still_plausible`	4	2024

7.1 Case study: a premise refuted (q_008, HD 209458 b)

From pre-2021 evidence, the system flagged a single-dataset conclusion: the influential retrieval of a *strongly subsolar* water abundance at the terminator of HD 209458 b (MacDonald and Madhusudhan, 2017). The frozen question asked, in 2020 terms:

> *Is the strongly subsolar terminator water abundance retrieved for HD 209458 b a property of the atmosphere or an artifact of retrieval assumptions, as tested by comparing independent retrieval frameworks on the same and on independent datasets?*

The 2021–2026 literature then performed exactly the test the question specified (Figure 2): a reanalysis of HST and JWST spectra with improved systematics treatment and Bayesian model averaging, an independent retrieval framework demonstrating that free-vs-equilibrium chemistry assumptions span the subsolar-to-solar range, and ground-based high-resolution spectroscopy constraining the abundance independently of space-based data. All three converge: the terminator water abundance is consistent with solar, and the strongly subsolar value was an artifact of earlier retrieval assumptions and data systematics. Under the taxonomy this is `answered` + `refuted`: the question was resolved *by* the community overturning the premise the question challenged—an outcome that no expert score assigned in 2020 could have certified, and precisely what backtesting exists to detect.

7.2 Secondary observations

The top-ranked question is independently posed and open. The system's rank-1 question (q_002: how terminator heterogeneity biases the WASP-12 b water abundance and C/O ratio) tracks a methodological concern the community has since engaged in general form — inhomogeneous-terminator biases in retrievals — without resolving it for WASP-12 b specifically: `posed_but_open`. A question the field poses but has not answered is a live research target; that the system's top pick lands there is the behavior a ranking is supposed to produce, though $n = 1$ at rank 1 proves nothing by itself.

An answered null result. q_005 asked whether HST transmission data independently support NH_3 or HCN in HD 209458 b — a molecular-detection claim from the same single-dataset analysis as q_008 (MacDonald and Madhusudhan, 2017). By 2024–2025, high-resolution spectroscopy had

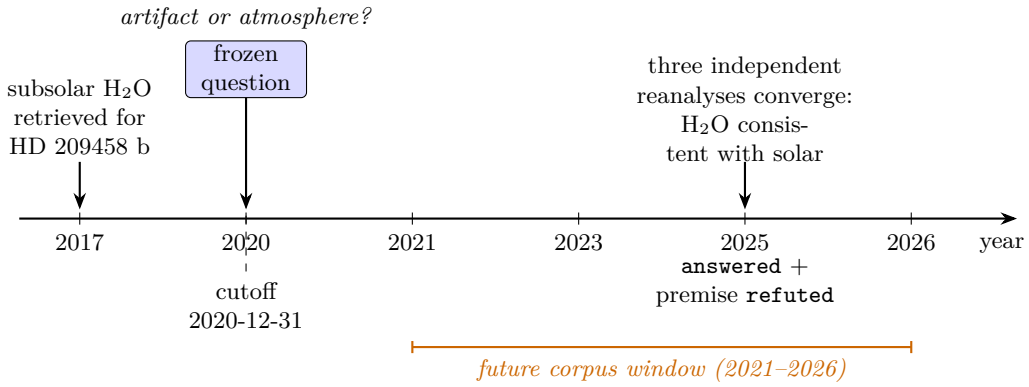


Figure 2: Timeline of q_008. The question was frozen from pre-2021 evidence; by 2025 three independent analyses had performed the test it specified and refuted its premise—the strongly subsolar water abundance was a retrieval artifact.

placed stringent upper limits on both species: `answered`, premise `supported` (the data indeed do not independently support the detection). Backtesting counts a cleanly resolved null exactly as it counts a positive.

Instrument-gated engagement. q_011 (whether JWST validated pre-launch predictions of TRAPPIST-1 CO_2 detectability, Lustig-Yaeger et al., 2019) could not have been engaged before JWST flew; its first cited engagement is 2024 and stellar contamination has so far prevented a definitive test. Engagement timing is partly an instrument schedule, not purely a question-quality signal—a confound Section 11 treats explicitly.

What these ten questions do not show. With ten questions from one system in one domain, rates carry wide intervals (the exact 95% Clopper–Pearson interval for 10/10 coverage is $[0.69, 1.0]$), the baseline rows are judge-only rather than adjudicated (Section 6), and the generating system's LLM components postdate the cutoff (Section 11). The pilot demonstrates that the protocol runs end-to-end, yields auditable, reproducible labels, and prices its baselines; it does not establish that any system reliably anticipates future science.

8 Scaling the Baselines: Astronomy v1L

Every baseline conclusion in Section 6 rests on ten questions per system. To test which of them survive a larger sample, we minted a second instance, **astronomy v1L**: same cutoff, same retrieval and judge settings, but corpora rebuilt from broadened manifests (12 past-corpus and 18 future-corpus ADS query sets covering clouds and hazes, atmospheric escape, phase curves, high-resolution spectroscopy, and JWST; 4,040 past and 5,754 frozen-window future records—1.8× and 2.4× the v1 corpora) and baselines scaled to 125 questions each. The review–future-work extractor is the exception by necessity: it exhausts the supply of explicitly posed questions in all 4,040 pre-cutoff abstracts at 49—itself a finding; the community's already-written-down questions are a finite resource. In total v1L evaluates 424 frozen baseline questions plus the ten evidence-graph questions re-run as a cross-instance anchor, all judge-only. The v1 instance and its released data are untouched.

Table 6: Astronomy v1L results (judge-only). Brackets are exact 95% Clopper–Pearson intervals. Eng. = future-attention rate; Ans. = answered; Ref. = premise refuted; s_1 = mean top-1 retrieval similarity; lag = mean years to earliest cited engagement. The evidence-graph row is the ten v1 questions re-evaluated on the v1L corpus (anchor), not a scaled submission.

System	n	Eng.	Ans.	Ref.	s_1	lag
`evidence_graph` (anchor)	10	80% [44,97]	30% [7,65]	10% [0,45]	0.682	3.5
B2 direct LLM	125	96% [91,99]	15% [9,23]	0% [0,3]	0.722	1.9
B4 citation leaders	125	87% [80,93]	26% [18,34]	3.2% [1,8]	0.632	2.5
B3 review future work	49	92% [80,98]	22% [12,37]	0% [0,7]	0.567	2.4
B1 random claims	125	73% [64,80]	11% [6,18]	0% [0,3]	0.622	2.8

Table 6 gives the scaled results. Sample size changes two of Section 6's three conclusions and sharpens the third—which is the point of running the experiment.

Revised: coverage does discriminate at scale. At $n = 10$ every non-random system sat at 90–100% engagement and we concluded coverage separates only the floor. At $n = 125$ the rates pull apart: random claims 72.8% [64,80], citation leaders 87.2%, direct LLM 96.0% [91,99]; random vs. direct-LLM engagement differs at $p < 10^{-4}$ and random vs. citation leaders at $p = 0.007$ (two-sided Fisher). The v1 reading was a small-sample artifact. What survives is the ceiling: fluent, topical LLM questions still approach saturation, so coverage separates the bottom and middle of the range while compressing the top—it remains unusable as a sole metric.

Revised: premise refutation is not unique to the evidence-graph system. At $n = 10$ no baseline refuted a premise; at $n = 125$ the citation-leader baseline catches four refutations (3.2% [1,8]): the subsolar-water conclusion for HD 209458 b (the same MacDonald and Madhusudhan, 2017 result behind q_008), the methane-depleted atmosphere of K2-18 b overturned by JWST, TiO in WASP-121 b unconfirmed by later data, and systematic bias found in benchmark ultracool-dwarf retrievals. Asking "is the famous result right?" of enough famous results does eventually catch the ones that fall—note its hindsight-biased selection (Section 6) works in its favor here, since post-cutoff citation counts are inflated by exactly the controversies that produce reversals. Three things remain true. Refutation is the rarest outcome for every system (random templates: 0/125, upper bound 2.9%—refutations are not free; a question must aim at a contestable claim). The evidence-graph submission's nominal rate stays highest (10% vs. 3.2%), but $n = 10$ cannot establish superiority ($p = 0.32$). And the two routes to a refutation differ qualitatively: prominence-chasing rediscovers that famous claims attract scrutiny, while the evidence-graph question specified the decisive test from pre-cutoff evidence tensions (Section 7.1). The scaled data cannot yet separate those routes quantitatively; a scaled evidence-graph submission could.

Sharpened: there is a nonzero answered floor. Random robustness templates get *answered* 11.2% [6,18] of the time—the field re-examines even arbitrarily chosen results at a measurable base rate. The v1 estimate of that floor (0/10) was too flattering to every other system: an answered rate is meaningful only against ∼11%, not zero. Citation leaders (25.6%) clear the floor ($p = 0.005$); direct LLM (15.2%) does not ($p = 0.42$).

Persistent: the contamination signature. The direct-LLM baseline keeps the highest similarity to future literature at scale ($s_1 = 0.722$ vs. 0.567–0.682 for all others), the broadest engagement

(96%, multi-source rate 94%), and the earliest mean engagement (1.9 years—its cited evidence concentrates in 2021–2022, the years closest to its training distribution). Scaling revises one part of the v1 reading: it does convert engagement into answers (15.2% vs. 0/10 at $n = 10$), but at a rate statistically indistinguishable from random templates, despite engaging twice as much of the literature. Breadth without resolution remains the signature.

Cross-instance anchor: metrics are corpus-relative. Re-judging the ten v1 questions on the 2.4× larger corpus flips four labels in both directions: two questions gain `answered` (richer evidence pools), one drops to `not_addressed` (its engaging papers pushed out of a top-8 that a larger corpus makes more competitive), and one premise moves `weakened`→ `still_plausible`. q_008's `answered` + `refuted` reproduces. The lesson is structural: rates are functions of the (corpus, retriever, judge) triple, so rows are comparable only within an instance—v1 and v1L rows must never be ranked against each other, and the frozen-instance design exists precisely to make the triple explicit.

What scaling did and did not change. The scaled study strengthens the benchmark's discriminative claims (coverage now separates three tiers; answered rates have a measurable floor) and weakens one system-level claim (refutation exclusivity). It does not change the evidence-graph submission's standing—its rates are unchanged and its refutation reproduces—but it narrows what that standing demonstrates: at current sample sizes, the defensible statement is that structured evidence analysis found a refutation by specifying its test in advance, not that it finds refutations at a higher rate than strong heuristics. Settling the rate question requires scaling the *submission*, not just the baselines—the first item on the revised roadmap.

9 Separating Hindsight Memorization from Foresight

The deepest objection to any backtest run with a modern LLM anywhere in the loop (Section 11) is that its apparent performance decomposes into three terms:

$$\text{Performance} = \underbrace{\text{reasoning over pre-cutoff evidence}}_{\text{what we want}} + \underbrace{\text{memorized future}}_{\text{contamination}} + \underbrace{\text{topic prior}}_{\text{fashion}}$$

and a single retrospective instance cannot tell the terms apart. A question that 2024 answered may have been *predicted* from 2020 evidence—or *remembered* from the model's training data. This section reports two experiments designed to pry the terms apart: holding the cutoff fixed while varying which component generates the question, and holding the generators fixed while moving the cutoff across the judge-model's training boundary.

9.1 Same cutoff, different generators: where does the signal live?

Three generation pipelines share the 2020 cutoff and the v1L evaluation conditions but differ in what produces the question:

A — LLM only. The direct-LLM baseline: `gpt-4.1` reads sampled pre-cutoff abstracts and proposes questions. Weights fully exposed to post-2020 literature.

B — structure → LLM. A deterministic, LLM-free detector finds pairs of pre-cutoff abstracts about the same catalogued object with opposing stances on the same species (detection vs. non-detection / upper limit); `gpt-4.1`'s only job is to verbalize each detected tension as one falsifiable question. The evidence structure is fixed before any LLM sees anything.

Table 7: The A/B/C decomposition on astronomy v1L (judge-only, $n = 125$ each; brackets are 95% Clopper–Pearson intervals). s_{obj} = share of questions naming a specific catalogued object (deterministic rubric, Section 9.3).

Pipeline	Eng.	Ans.	Ref.	s_1	s_{obj}
A: LLM only	96% [91,99]	15% [9,23]	0% [0,3]	0.722	5%
B: structure → LLM	74% [65,81]	39% [31,48]	13% [8,20]	0.694	95%
C: structure only	77% [68,84]	25% [17,33]	2.4% [0,7]	0.721	100%

Table 8: Temporal contamination stress test (judge-only, $n = 50$ per cell; c2010 tension cells have $n = 48$—the 855-paper 2005–2010 corpus yields only 48 detectable tension pairs). Cutoffs 2010–2020 lie inside the LLM's training data; the 2024 cutoff's future window (2025–2026-06, censored) postdates it. Ans. = answered; Ref. = refuted count.

		c2010	c2015	c2020	c2024
A: LLM only	Ans.	22%	12%	10%	16%
	Ref.	2	0	0	0
B: structure → LLM	Ans.	31%	26%	34%	32%
	Ref.	6	3	6	5
C: structure only	Ans.	10%	6%	16%	4%
	Ref.	0	0	0	0
Random claims	Ans.	12%	6%	4%	6%
	Ref.	1	1	0	0

C — structure only. The same detected pairs rendered by a fixed template. No LLM anywhere: this generator has no weights to contaminate.

B and C share identical evidence structures, so their gap isolates the language-realization layer; A and B share the same LLM, so their gap isolates evidence structure. (B/C are an evidence-structure-*lite* probe—object co-mention plus stance cues—not a reimplementation of the evidence-graph system.) Table 7 gives the $n = 125$ results.

Three readings. *First*, the contamination-free pipeline C beats the fully exposed pipeline A on answered rate (24.8% vs. 15.2%, $p = 0.08$) and on refutations (3 vs. 0)—evidence that a foresight signal exists in pre-cutoff evidence structure alone, extractable with zero model weights. *Second*, adding the LLM back as a pure verbalizer (B) roughly doubles resolution over the same structure (39.2% vs. 24.8% answered, $p = 0.02$; 12.8% vs. 2.4% refuted, $p = 0.003$): phrasing a tension as a crisp either/or question makes it judgeable, and B ends with five times pipeline A's refutation count while citing the same model. *Third*, A's questions are structurally different, not just weaker: only 5% name a specific object (vs. 95–100% for B/C), and its engagement is the highest of any system—broad questions that everything touches and little settles.

9.2 Same generators, moving cutoff: the temporal stress test

If pipeline A's performance were substantially the memorized-future term, it should degrade as the cutoff crosses the model's training boundary (June 2024 for `gpt-4.1`). We ran four generators (A, B, C, and random claims as an era control) at four cutoffs—2010, 2015, 2020, 2024—with uniform four-year future windows ($n = 50$ per cell; the 2024 window is censored at 2026-06 and flagged; corpora per era rebuilt from released manifests). Table 8 reports the grid.

The naive collapse does not happen. Pipeline A's answered rate is statistically flat across the training boundary (14.7% pooled in-training vs. 16.0% post-training, $p = 0.82$), as is every other generator's. At the outcome level, the memorized-future term is *not* where A's performance comes from—because, as the decomposition shows, A's performance never rested on specifics that memorization could supply. Its engagement sits at 92–98% at every cutoff, its specificity at the floor (≈1.1 of 3) at every cutoff: broad questions about each era's active topics, engaged everywhere, resolving little, in any era. That is the topic-prior term at work, and a topic prior does not need to remember the future—the present is enough.

Where a memorization trace does appear. Two places, both in the channels the topic prior cannot supply. Pipeline A's only premise refutations in the entire stress test (2 of 200) occur at the deepest-contamination cutoff, 2010—the one era whose reversals the model has certainly read about—and never after (0 of 150; too few for significance, CI [0.5%, 13.7%] at c2010). And A's phrasing similarity to future literature is highest inside its training window (0.717–0.724) with a mild post-training dip (0.703)—directionally consistent with phrasing-level memorization, though era confounds keep this suggestive rather than conclusive.

The structural signal is era-robust. Pipeline B refutes premises at every cutoff—6, 3, 6, 5—including the one whose future the model cannot have seen (10.0% post-training vs. 10.1% in-training). Across all cutoffs B refutes at 10.1% vs. A's 1.0% ($p = 4 \times 10^{-5}$), and B's answered margin over C persists post-training (+28 points at c2024). One residual channel remains open and is worth stating precisely: c2024 tension pairs are drawn from 2019–2024 abstracts, and a tension resolved in early-2024 literature the model saw could steer B's phrasing even though the evaluation window postdates training. C is immune by construction, which is why C finding *any* future engagement at every cutoff (70–83%) is the cleanest single fact in the grid.

9.3 Specificity-adjusted foresight

Coverage can be farmed by asking broad questions—"How can we better understand exoplanet atmospheres?" will be engaged with probability 1 in any active field. We therefore score every question with a deterministic, released rubric: +1 for naming a specific catalogued object, +1 for naming a measurable claim (species, quantity with units, abundance comparative), +1 for an explicit discriminative construction ("... or an artifact of...", "as tested by"); and define specificity-adjusted foresight SAF $= \mathbb{E}[\, w(\text{outcome}) \cdot \text{specificity}/3 \,]$ with $w = 1$ / 0.5 / 0.25 / 0 for answered / partial / posed-open / not addressed. (Template pipelines inherit the test-construction point from their template—the rubric's components are reported separately for exactly that reason; a novelty term awaits first-posed dates, Section 12.) Two facts survive every era and instance: pipeline A's anchoring is an order of magnitude below the structure pipelines' (s_{obj} 5% vs. 95–100% on v1L), and its SAF never exceeds the random-template floor by more than a few points (0.18–0.23 vs. 0.24–0.26), while structure pipelines reach 0.34–0.43. The evidence-graph submission's ten questions score $s_{\text{obj}} = 90\%$, SAF 0.40.

9.4 What this section establishes

Memorized relevance is not scientific foresight.

The LLM-only pipeline exhibits relevance everywhere—near-ceiling engagement, the closest phrasing to the future literature at every cutoff—and specific foresight nowhere: no refutation

outside the era it could have memorized, an answered rate indistinguishable from random templates, specificity at the floor. The structure-first pipelines invert the picture: lower engagement, higher resolution, refutations in every era including the one that postdates the model's training. On this evidence, the foresight signal measured by historical backtesting lives chiefly in pre-cutoff evidence structure; the LLM contributes a real but separable service—turning a detected tension into a question sharp enough to be judged. Contamination, meanwhile, turns out to be measurable rather than merely confessable: the stress test bounds its outcome-level effect (small for this task family) and localizes its traces (phrasing proximity; refutations only in deep history). All 798 stress-test judgments, the A/B/C submissions, and the per-cell corpora manifests are released; every number in this section is judge-only and carries $n = 48$–125 intervals—the qualitative pattern, not any single rate, is the finding.

10 Error Analysis

The released curation log records every human intervention made before freezing; the failure modes below are taken from it and from the adjudication log, and each motivates a benchmark rule.

Question duplication. The generator produced near-duplicate questions (q_002/q_003) from the same claim pair, differing mainly in emphasis; they were merged during curation into one question with two sub-questions. Left unmerged, duplicates would double-count a single insight in every rate. Rule: submissions are screened for near-duplicates, and instances should report a deduplication note; a mechanical similarity screen is a v1.1 roadmap item.

Leading questions (presupposition). Several generated questions presupposed their own answer. q_001 originally asserted that vertical wind shear *exists* rather than asking what causes the wind-speed discrepancy; q_008 originally presupposed the subsolar water abundance rather than framing artifact-vs-atmosphere attribution. A question that presupposes its answer cannot be cleanly refuted—and q_008's later refutation was only expressible because curation reframed it as attribution. Rule: the quality gate flags presupposing phrasings before freezing; the reframing is logged.

Self-contradictory quantifiers. q_003's original phrasing asked whether a detection was "robust to biases exceeding an order of magnitude"—a bias that large *is* non-robustness. Templated quantifier language can silently produce unanswerable questions; the clarity gate exists for this.

Conflated statistical and physical framing. q_010 originally asked what "temperature and pressure conditions" produce a 5.4σ water detection, conflating atmospheric state with the detection pipeline (significance depends on noise model, priors, null hypothesis—not on the atmosphere). It was reframed as a sensitivity analysis; its clarity-gate score (6.0) was the lowest of the set, and its outcome (`partially_addressed` on one supporting paper) remains among the weakest-evidenced labels.

Premise bias in generation. The generator inherits the premises of the papers it reads: single-dataset conclusions taken at face value can produce questions that merely restate a claim rather than test it. The tension-typing stage (which explicitly marks *single-dataset conclusion* as a signal type) partially controls this—q_008 and q_005 are that control working—but the benchmark's

premise dimension is the systematic check: a healthy portfolio should show a mix of `supported` and `refuted`, not uniform support of its sources.

Retrieval near-misses. Judged evidence for q_007 (HD 189733 b / HAT-P-11 b patchy clouds) leans partly on a three-retrieval-framework study of HAT-P-*18* b—directly relevant methodologically, but not the named targets. The adjudication log records the engagement-bar judgment; releasing full top-8 lists (v1.1) will let others re-litigate such calls, which is the point of releasing them.

11 Threats to Validity

Historical backtesting removes rater subjectivity; it does not remove every confound. We enumerate the serious ones and what the protocol does—and cannot do—about each.

Future inattention is not question badness. A question can be ignored because the enabling instrument never flew, the community's funding shifted, or the subfield is small—not because the question was poor. q_011 was unanswerable before JWST delivered TRAPPIST-1 spectra; had JWST slipped five years, an excellent question would have scored `not_addressed`. Mitigations: bounded windows make the censoring explicit; `posed_but_open` separates "recognized but unresolved" from "ignored"; instances should be read as *engagement within k years given the era's instruments*, not as timeless value.

Future attention is not question goodness. Symmetrically, a question on a fashionable topic collects engagement for reasons other than merit; coverage alone can be gamed by asking about whatever is popular. Mitigations: coverage is never reported alone; premise refutation and (future) lead time cannot be earned by fashion-chasing; baseline B4 (citation leaders) exists precisely to price in prominence; and the engagement bar requires the cited paper to bear on the question's actual test, not its topic.

LLM training contamination. The generating system and the judge both use models whose training data postdates the cutoff. The protocol seals the retrieval channel, not the weights channel: a model may "know" the 2025 refutation while drafting a 2020-framed question. This is the deepest threat to any backtest run with modern models. Mitigations, none complete: source-evidence audit forces every question to be grounded in cited pre-cutoff evidence; submission metadata must declare all LLM components and versions; baseline B2 doubles as a contamination probe—and registers one: its questions sit measurably closer to the future literature's phrasing than any other system's ($s_1 = 0.732$ at $n = 10$, 0.722 at $n = 125$; Tables 4 and 6), with its cited engagement concentrated in the years nearest its training distribution, while its answered rate stays indistinguishable from random templates—the signature of phrasing-level memorization without specific foresight. Section 9 upgrades this confession to a measurement: a generator decomposition plus a four-cutoff stress test bound the contamination's outcome-level effect and localize its traces; and the decisive test is *prospective* instances—questions frozen today and scored in 2030 cannot be contaminated. The protocol is explicitly designed so its instances convert from retrospective to prospective by just letting time pass.

Judge and adjudicator reliability. A single LLM judge, even citation-constrained and human-adjudicated, is one reading of the evidence; abstracts (not full texts) bound what it can see, and the adjudicators are not blind to which system produced the questions. Mitigations: labels ship with rationales and citations so any label is independently checkable; disagreement is handled by written conservative rules (prefer the weaker engagement); multi-judge agreement measurement and blinded adjudication are roadmap items.

Retrieval as a bottleneck. Top-8 abstract-embedding retrieval can miss engaging papers (undercounting engagement) or surface topically similar non-engagement (which the judge must reject). Fixed retrieval is a deliberate trade: it makes comparisons across systems fair and auditable at the cost of an engagement floor. Sensitivity of labels to k and to the retriever is measurable within the released data schema and belongs in v1.1.

Small n where it matters most, one domain, self-evaluation. The scaled instance lifts the baseline side to $n = 125$ per system, and Section 8 shows how much that matters: two of three small-sample conclusions did not survive. But the submission under test still has ten questions—every evidence-graph rate keeps a ∼30-point-wide interval, and the refutation-rate comparison against baselines is unresolvable at $n = 10$ ($p = 0.32$). One domain, evaluated by the group that built the leading system, judge-only baseline labels: all still true. We report priced reference points, not rankings; the protocol's value grows with adversarial use by systems we did not build.

Curation hindsight. Humans who edited questions before freezing know the post-2020 literature. The curation log is released so every edit is auditable (e.g. q_008's reframing strengthened falsifiability without smuggling in the answer), but retrospective instances cannot fully exclude this channel; prospective ones can.

12 Roadmap and Conclusion

Roadmap. The released instances are a floor, not a ceiling. In order: (1) scale the *submission* side—a 100+ question evidence-graph run on v1L—to resolve the refutation-rate comparison that $n = 10$ leaves open (Section 8), and adjudicate the judge-only rows; (2) release full top-8 retrieval lists for v1 with per-document scores and a near-duplicate screen (v1.1; the v1L records already ship with full scores); (3) annotate community first-posed dates in the review literature to activate lead time; (4) calibrate a community-attention index (papers, citations, observing programs) against the popularity confound; (5) mint instances in additional domains and at additional cutoffs, including *prospective* instances—questions frozen now, scored when the future arrives—which eliminate the contamination and hindsight channels entirely; (6) measure judge robustness (multi-judge agreement, retrieval sensitivity, blinded adjudication).

Conclusion. We formalized historical backtesting as an evaluation protocol for scientific question discovery, released two reproducible astronomy instances with temporally isolated corpora and fully auditable labels, and ran a ten-question pilot in which every frozen question was substantively engaged by literature the generating system never saw—including one whose premise the community subsequently refuted, the exact convergence test the question had specified. Scaling the baselines to 424 questions then did what a benchmark is supposed to do: it overturned two of our own small-sample conclusions, put a measurable floor under a third, and left the qualitative distinction—a refutation reached by specifying its test in advance—standing but explicitly unresolved at current

sample sizes. Finally, the generator decomposition and temporal stress test turned the benchmark's deepest limitation into its sharpest result: memorized relevance is not scientific foresight, and the foresight signal that historical backtesting measures survives in a generator with no weights at all. The individual numbers matter less than the category they inhabit: for the first time, "this system asks good scientific questions" is a claim with a denominator—falsifiable, comparable across systems, and computable by anyone from frozen public data. Question-asking has been argued to be a core capability on the path to more general scientific intelligence (Kitano, 2021; Scientific Question Discovery project, 2026); if that is so, the field will need to measure it. This protocol, and this first instance, are offered as the place to start—not as the definition of the benchmark, but as its initial version, built to be superseded by instances with more questions, more systems, more domains, and cutoffs whose futures have not yet happened.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, 2024. arXiv:2404.07738.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. *Notices of the American Mathematical Society*, 61(5):458–471, 2014.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *International Conference on Learning Representations*, 2024.

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The Automation of Science. *Science*, 324(5923):85–89, 2009.

Hiroaki Kitano. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7:29, 2021.

Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, 1987.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, 2024. arXiv:2408.06292.

Jacob Lustig-Yaeger, Victoria S. Meadows, and Andrew P. Lincowski. The Detectability and Characterization of the TRAPPIST-1 Exoplanet Atmospheres with JWST. *The Astronomical Journal*, 2019. doi: 10.3847/1538-3881/ab21e0. ADS bibcode: 2019AJ....158...27L.

Ryan J. MacDonald and Nikku Madhusudhan. HD 209458b in new light: evidence of nitrogen chemistry, patchy clouds and sub-solar water. *Monthly Notices of the Royal Astronomical Society*, 2017. doi: 10.1093/mnras/stx804. ADS bibcode: 2017MNRAS.469.1979M.

Scientific Question Discovery project. Scientific Question Discovery: Toward Question-Asking as a Core Capability of AGI, 2026. Manuscript and code: https://github.com/nonameisready/scientific-question-discovery.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, 2024. arXiv:2409.04109.

Don R. Swanson. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting Future World Events with Neural Networks. In *Advances in Neural Information Processing Systems*, 2022.