Evidence-Based Scientific Question Discovery:
A Framework with Historical Backtesting

Hui Mao
Aurora AI Ventures

July 29, 2026

Abstract

Current AI systems are optimized for answering questions; the scientific enterprise is bottlenecked earlier, at *discovering the questions worth investigating*. We present a framework that turns a traceable, reproducible, scope-controlled research corpus into ranked, falsifiable research questions: evidence is represented as provenance-carrying claims; cross-paper tensions are detected, typed, and human-adjudicated; surviving signals are refined into questions and ranked by a two-stage protocol separating scientific priority from execution priority. We instantiate the framework on exoplanet atmospheres, a domain that uniquely combines literature, structured catalogs, and space-telescope archives. In a historical backtest, all questions generated from evidence available before 2021 were substantively engaged by the 2021–2026 literature the system never saw: two were answered—including one whose premise the community later explicitly refuted—and the top-ranked question is independently posed and still open. These results suggest that systematic question discovery from evidence tensions surfaces the questions working scientists subsequently invest in.

1 Introduction

> *"What are the important problems of your field? . . . If you do not work on an important problem, it's unlikely you'll do important work."*
>
> — Richard Hamming, *You and Your Research* (1986)

Hamming's question is about asking, not answering. Current AI is optimized for the opposite: benchmarks, training objectives, and deployment all reward answering questions someone else has already posed. Yet the scientific enterprise is bottlenecked earlier in the loop—someone must notice that a question is worth asking. Recognizing a productive question—an unexplained discrepancy, a conclusion resting on a single dataset, a methodological assumption quietly doing load-bearing work—is a distinct capability from answering it, and arguably the rarer one. We take the position that this capability, *scientific question discovery*, is a core requirement for general scientific intelligence, and that it can be studied rigorously today.

Studying question discovery requires discipline that question answering does not. An answer can be checked against ground truth; a question's value only reveals itself later, through the community's engagement with it. This paper contributes a methodology that makes question discovery measurable despite that asymmetry:

1. **A framework** (Section 3) that represents published evidence as provenance-carrying claims, detects and types cross-paper tensions in an evidence graph, refines surviving signals into falsifiable questions, and ranks them with a two-stage protocol that separates scientific priority from execution priority.

2. **A corpus discipline** (Section 4) in which every experiment is bound to a manifest that pins queries, versions, and—critically—a time boundary. Data need not live in the repository, but the queries, versions, time boundaries, and processing code that produce it must.

3. **A historical validation protocol** (Section 4) that exploits the time boundary: questions are generated from evidence available before a cutoff and judged against the literature published after it, which the system never saw. This prevents the system from peeking at the future and converts question value into a measurable outcome.

4. **An end-to-end instantiation** (Section 5) on exoplanet atmospheres, chosen because it uniquely combines papers, structured catalogs (NASA Exoplanet Archive), and space-telescope archives (JWST/HST via MAST), allowing the full loop—evidence, questions, falsification, ranking—to run end to end.

In the historical backtest, every question generated from pre-2021 evidence was substantively engaged by the 2021–2026 literature; two were answered, including one whose premise the community later explicitly refuted—the strongest form of question-discovery success, since the system had isolated a published conclusion that was in fact an artifact awaiting correction.

2 Related Work

Automated scientific discovery. Computational discovery has a long lineage: BACON and its successors rediscovered empirical laws from data (Langley et al., 1987), and the Robot Scientist line (Adam, Eve) closed the loop from hypothesis generation through physical experimentation in functional genomics (King et al., 2004, 2009). Contemporary "agentic-scientist" systems extend this ambition with large language models: the AI Scientist generates ideas and drafts papers end to end (Lu et al., 2024), Google's AI co-scientist generates and tournament-ranks research hypotheses with multi-agent debate (Gottweis et al., 2025), and deep-research agents synthesize literature into reports (OpenAI, 2025). Across this line, the object produced is a hypothesis, an experiment, or a paper; the selection of *which question deserves the loop* is either inherited from the human operator or made implicitly inside a prompt. Our system makes that selection step itself the object of study, with its own representations, human checkpoints, and—critically—an outcome-based evaluation.

Literature understanding and scientific knowledge bases. Our evidence layer builds on infrastructure for machine-readable science: literature graphs such as Semantic Scholar (Ammar et al., 2018) and OpenAlex (Priem et al., 2022), structured claim resources like the Open Research Knowledge Graph (Jaradeh et al., 2019), scientific claim verification (Wadden et al., 2020), science-specialized language models (Taylor et al., 2022), and retrieval-augmented literature agents (Lála et al., 2023; Skarlinski et al., 2024). We differ in what the representation is *for*: claims carry verification tiers and source locations, and relations between claims are typed tension edges (*contradicts, qualifies, challenges-method, explains-discrepancy, complements, supports*) adjudicated by a human—an empirical necessity, since our review found only 1 of 16 top tension candidates to be a genuine contradiction (Section 5.3).

Open-ended learning and intrinsic motivation. Curiosity-driven agents formalize "interestingness" via intrinsic motivation (Oudeyer and Kaplan, 2007; Schmidhuber, 2010), novelty search (Lehman and Stanley, 2011), and the broader open-endedness program (Stanley et al., 2017). Those signals are computed from an agent's own experience stream; ours are computed from the evidential

Table 1: Positioning relative to representative systems. "Evidence tensions" = questions/hypotheses grounded in explicit cross-paper evidential structure; "historical validation" = outcome-based evaluation against literature published after a data cutoff.

System	Hypotheses	Questions	Tensions	Backtest
BACON (Langley et al., 1987)	✓	–	–	–
Robot Scientist (King et al., 2009)	✓	–	–	–
AI Scientist (Lu et al., 2024)	✓	–	–	–
AI co-scientist (Gottweis et al., 2025)	✓	–	–	–
PaperQA2 (Skarlinski et al., 2024)	–	–	–	–
LLM ideation (Si et al., 2024)	✓	✓	–	–
This work	–	✓	✓	✓

structure of the scientific record—contradiction, single-source support, unexamined assumptions—which is where working scientists actually look for questions.

LLM research ideation and its evaluation. Recent studies evaluate LLM-generated research ideas with expert panels, finding above-human novelty ratings alongside feasibility concerns (Si et al., 2024); LLM-as-judge protocols and their biases are themselves under scrutiny (Zheng et al., 2023). Panel evaluation asks contemporaries to *predict* a question's value. Our historical backtest removes that prediction: questions are generated from a time-frozen corpus and scored by what the field actually did in the following five years.

3 Framework

The framework transforms existing scientific knowledge into ranked, falsifiable questions through six stages (Figure 1). Throughout, the design principle is *traceability*: every artifact carries the provenance needed to audit or reproduce it.

3.1 Evidence representation

Knowledge enters as three layers, collected in order of increasing cost: a *literature* layer (what has been claimed), a *catalog* layer (what objects exist, with parameters and uncertainties), and an *observation* layer (where the underlying data lives—metadata only; data products are fetched on demand after a question is selected). Every raw record is stored immutably inside a provenance envelope (*source*, *retrieved-at*, *query-version*, *source-id*, *payload*), then normalized into relational tables.

The atomic unit is the **claim**: a tuple

$$c = (\textit{text},\ \textit{evidence},\ \textit{assumptions},\ \textit{uncertainty},\ \textit{objects},\ \textit{datasets},\ \textit{location},\ \textit{tier})$$

where *location* points back into the source document and *tier* records the verification provenance: `human_gold` (annotated by a domain reviewer), `model_matched_gold` (model extraction semantically matching a gold claim; the canonical text remains the human's, both provenances are kept), or `model_only` (unreviewed model extraction, default confidence 0.75). Model claims that duplicate gold claims are absorbed into a single canonical node rather than added as siblings—duplicate nodes would both inflate the graph and manufacture spurious tension.

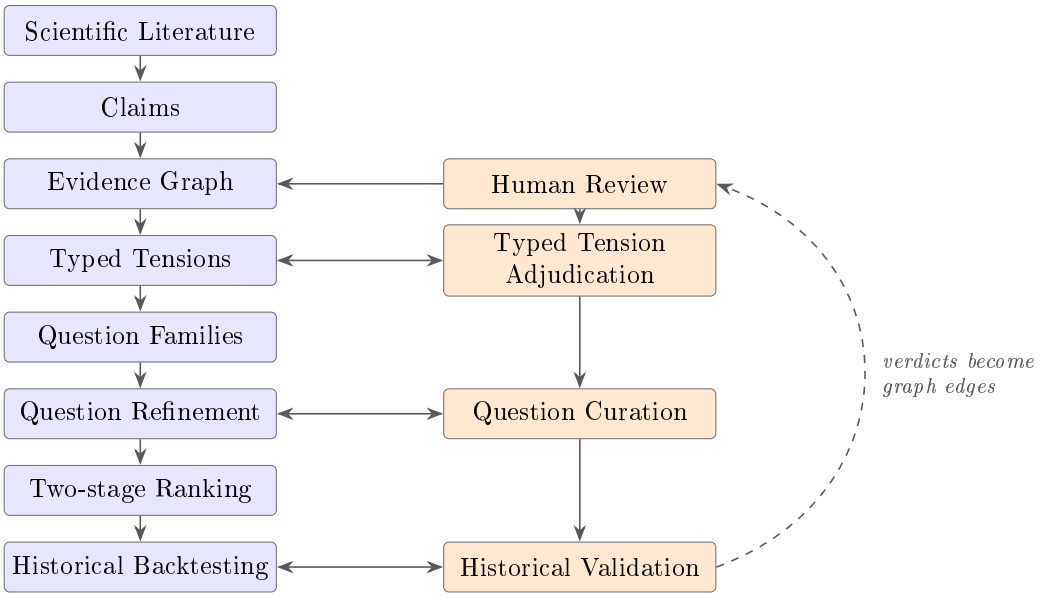


Figure 1: Architecture: a dual loop. The automated pipeline (left, blue) turns a time-frozen corpus into ranked questions; the human loop (right, amber) adjudicates tension types, curates question phrasing, and validates outcomes. Human verdicts are committed as data and feed back into the evidence graph, keeping expert judgment cheap, auditable, and reusable.

3.2 Evidence graph and typed tensions

Claims, objects, datasets, and assumptions become nodes; deterministic rules add edges (*uses-dataset*, *measures*, *assumes*), each recording the rule that created it. Cross-paper claim pairs that share an object *and* a topic term become *tension candidates*—deliberately a weak, high-recall rule. The key representational commitment is that candidate tensions are **typed by human adjudication** before they may drive question generation, using the vocabulary

$$\{\textit{contradicts},\ \textit{qualifies},\ \textit{challenges-method},\ \textit{explains-discrepancy},\ \textit{complements},\ \textit{supports}\}.$$

This vocabulary is an empirical result of the calibration review (Section 5.3): most high-scoring candidates are not contradictions, and a graph that collapses "adds limiting conditions" into "negates" will amplify false conflicts as it grows. Figure 2 shows the instantiation of this machinery on the one confirmed contradiction in our corpus.

3.3 Question generation

Questions are generated only from prioritized signal classes: (P1) human-confirmed observational tensions, (P2) methodological challenges, (P3) qualifications across independent datasets, and (P4) single-dataset conclusions from trusted claims. Unreviewed tension candidates are excluded by design. Signal assembly is deterministic; a language model only phrases the question, and every candidate carries its evidence trail (claim identifiers, papers, adjudication basis) plus an archival-data availability count—the first falsification screen. Human editorial decisions (merges into question families, rewrites, rejections) are recorded as committed curation records and re-applied deterministically.

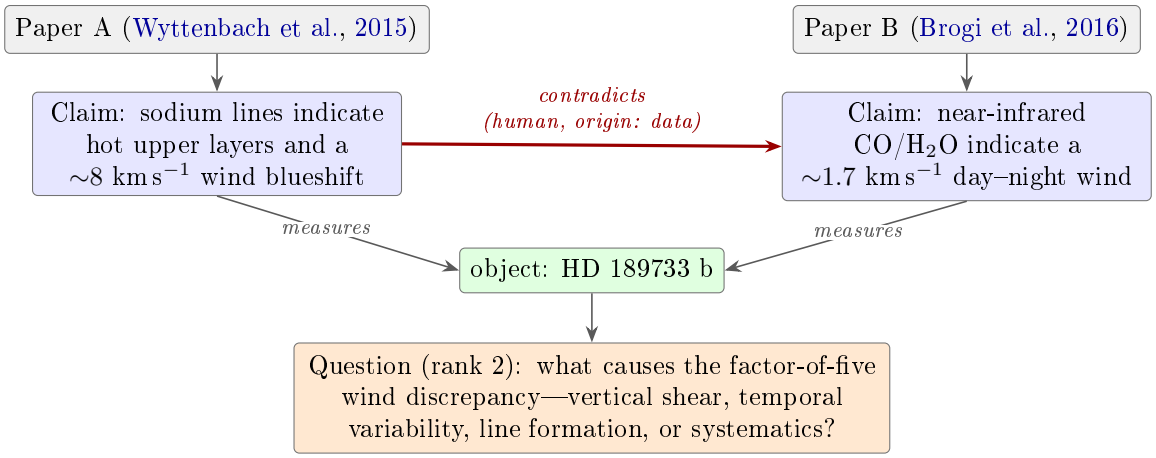


Figure 2: The evidence graph around the one confirmed contradiction: two claims from different instruments measure the same object, disagree on a shared quantity, survive human adjudication as a genuine data-origin tension, and become the pipeline's top tension-driven question. The 15 other reviewed candidates received non-contradiction types and correspondingly weaker question priority.

3.4 Two-stage ranking

Stage A is a quality gate: evidence grounding plus a six-criterion clarity rubric (presupposed conclusions, statistics/physics conflation, explicit comparison, falsifiable outcome, bounded scope, stateable failure condition). Clarity gates rather than scores—in calibration it scored uniformly and only added a constant. Human-curated questions bypass the model gate, whose verdict is retained as advisory.

Stage B ranks survivors by

$$\textit{scientific priority} = 0.35\,S + 0.25\,T + 0.20\,F + 0.10\,N + 0.10\,G,$$

where S (significance) is judged by a model but hard-capped by signal tier (P1: 10, P2: 9, P3: 8.5, P4: 7.5) so that single-dataset robustness checks cannot rank beside confirmed tensions; T (tension strength) is structural; F (feasibility) averages judged subcomponents (availability of the *right* data rather than raw record counts, data independence, analysis readiness, reanalysis sufficiency); N (novelty) is deliberately down-weighted—the system generates questions from existing evidence tensions, so surface novelty is not the point—and blends a judge with a corpus-similarity check; G is expected information gain. *Execution priority* (feasibility) is reported separately: abundant archives must not outrank a confirmed cross-instrument tension, and a question can be scientifically first while operationally harder.

4 Corpus Discipline and Evaluation Protocol

4.1 Corpus manifests

Every experiment must know exactly what the system has seen. A *corpus manifest* pins the scope, source queries, per-source field lists, record caps, selection rules, and—critically—a `cutoff_date`. Manifests are frozen at collection time and referenced by experiments; a change of scope mints a new corpus identifier rather than editing a manifest in place. The repository commits manifests, collectors, schemas, curation and adjudication records, and small samples; bulk data is regenerable from those.

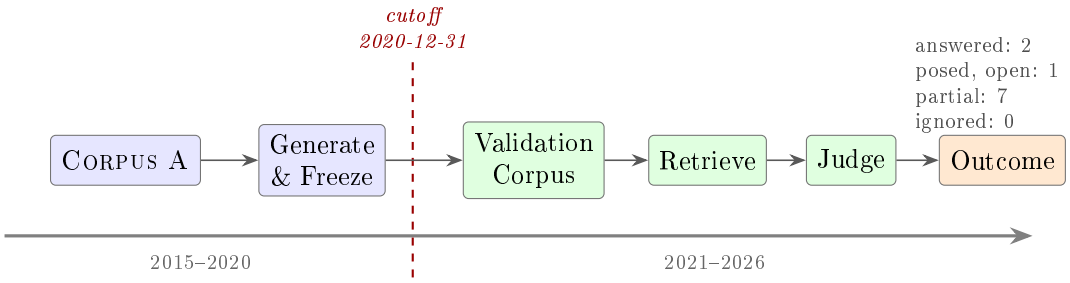


Figure 3: Historical backtesting. Questions are generated from a corpus frozen at the cutoff, then judged against post-cutoff literature the pipeline never saw. The right panel shows the observed outcome distribution for the ten ranked questions (Section 5.5).

4.2 Historical validation

The cutoff enables a backtest that removes guesswork about the future from evaluation:

1. Corpus A contains only evidence available up to the cutoff (here 2020-12-31), enforced end to end: literature windows, catalog snapshots, and a contamination guard that filters every normalized table by corpus identifier.
2. The full pipeline runs on Corpus A to produce ranked questions.
3. A *validation corpus* of post-cutoff literature (2021–2026), never seen by the pipeline, is collected under its own manifest. For each ranked question, the most relevant post-cutoff abstracts are retrieved by embedding similarity and a judge classifies the question's fate—`answered`, `partially_addressed`, `posed_but_open`, or `not_addressed`—citing only retrieved records, and states whether the question's premise was validated, refuted, or left untested.

A question the community independently posed, answered, or refuted after the cutoff is evidence the pipeline surfaces questions worth asking; a premise the community later refuted is the strongest success mode, since the system isolated a published conclusion that was an artifact awaiting correction. Figure 3 summarizes the protocol.

4.3 Question quality rubric

Human evaluation shares one rubric with the ranking stage: novelty, feasibility, significance, and clarity, with clarity applied as a gate (Section 3.4). All rubric prompts, gate criteria, weights, and per-component scores are committed, making every ranking auditable.

5 Experiments

We instantiate the framework on exoplanet atmospheres—specifically, atmospheric composition interpretation in transmission spectra, cloud/haze degeneracies, and instrument systematics. All numbers below are produced by the committed pipeline under the frozen manifest `exoplanet_atmospheres_v1` (cutoff 2020-12-31).

Table 2: CORPUS A at a glance (manifest `exoplanet_atmospheres_v1`, cutoff 2020-12-31).

Literature metadata (deduplicated)	2,512 papers
from NASA ADS / arXiv	1,639 / 1,277
Core full-text set	500 papers
Citation edges	51,616
Catalog objects (NASA Exoplanet Archive)	6,324 planets
Observation metadata (MAST, JWST+HST spectra)	306,729 records
Paper–object links	2,364
Catalog planets with archival spectra	553

5.1 Corpus construction

Table 2 summarizes CORPUS A. Literature metadata came from NASA ADS (primary) and arXiv (supplement) under four topic queries; cross-source deduplication by arXiv identifier removed 404 duplicate records. A core set of 500 papers was selected by greedy, year-stratified optimization over internal citation in-degree, keyword coverage, and target diversity (83–84 papers per year, 2015–2020). Object mentions were linked to the NASA Exoplanet Archive by boundary-anchored name matching tolerant of space/hyphen variants; naive substring matching produced systematic false positives ("WASP-1" firing inside "WASP-12b") that inflated links by 2.3× before the fix.

5.2 Claim extraction: calibration before scale

A domain reviewer first annotated 20 claim-dense core papers by hand, yielding 37 gold claims with source locations, assumptions, and uncertainties. Model extraction (from abstracts in this version) was then calibrated against the gold set over three prompt revisions: the initial prompt over-split coupled measurements, occasionally copied first-person phrasing, and preferred easily-copied quantitative results over interpretive conclusions (e.g., it captured "mostly subsolar H_2O" across ten hot Jupiters (Pinhas et al., 2019) but dropped the paper's load-bearing interpretation of subsolar oxygen or supersolar C/O). The final prompt enforces 2–5 load-bearing claims with a priority order (conclusion > interpretation > methodological challenge > decisive quantity > secondary measurement), merging of coupled measurements, normalized third person, and an omission self-check. Under it, all reviewer-cited misses were recovered, first-person copying fell to zero, and no paper exceeded five claims (mean 3.68 across 353 claims from 100 papers; 4 methods papers correctly yielded none).

Semantic deduplication against the gold set (same paper, compatible objects, same conclusion polarity, embedding cosine ≥ 0.62) matched all 37 gold claims to a model counterpart (similarity 0.63–0.97, median 0.76), producing a canonical table of 353 claims: 37 `model_matched_gold` and 316 `model_only`.

5.3 Tension detection and adjudication

The high-recall candidate rule produced 161 cross-paper tension candidates over the canonical claims (807-node, 1,237-edge graph). The reviewer adjudicated the top-ranked candidates; the resulting 16 typed verdicts are themselves a finding (Table 3): only *one* candidate is a genuine observational contradiction—the HD 189733 b wind-velocity discrepancy, $\sim$8 km s^{-1} from optical sodium (Wyttenbach et al., 2015) versus $\sim$1.7 km s^{-1} from near-infrared CO/H_2O (Brogi et al., 2016). The modal relations are *supports* and *complements* (e.g., terminator transmission versus dayside

Table 3: Human adjudication of the 16 reviewed tension candidates.

supports	7
complements	4
challenges-method	2
qualifies	1
explains-discrepancy	1
contradicts (confirmed observational tension)	1

emission water results for HD 209458 b probe different geometries and do not conflict), followed by *challenges-method* (a 1D-retrieval bias result (MacDonald et al., 2020) limits the interpretation of a water abundance (Kreidberg et al., 2015) without negating the detection). An early binary confirm/reject protocol would have discarded most of these relationships or, worse, upgraded them to contradictions.

5.4 Question generation and two-stage ranking

From the prioritized signals (1 confirmed tension, 2 methodological challenges, 1 independent qualification, 7 trusted single-dataset conclusions), the generator produced 11 candidates; editorial curation merged one near-duplicate pair into a question family with two sub-questions and rewrote three questions (one had conflated statistical significance with physical conditions; one presupposed the tension's favored explanation; one was reframed for historical validation). The Stage-A gate additionally caught the presupposition defect independently. After curation, all 10 questions pass the gate.

Stage-B ranking places the terminator-heterogeneity family and the wind-discrepancy question statistically tied at the top (scientific priority 8.13 vs. 8.13, separated by 0.005) with opposite execution profiles: the heterogeneity family is reanalysis-ready against 2,643 archival HST spectra of WASP-12 (execution priority 9.1), whereas the confirmed tension is scientifically first-equal but operationally harder (execution 6.9—the needed simultaneous multi-band data may not exist). Significance caps hold the seven P4 robustness questions to a 6.6–7.1 band, below both tension-driven questions.

5.5 Historical validation

The validation corpus contains 1,891 unique 2021–2026 papers (collected under its own manifest; a contamination guard keeps Corpus A tables bit-identical before and after collection). Table 4 lists all ten questions with their verdicts: 2 answered, 1 posed but open, 7 partially addressed, 0 not addressed.

Three outcomes deserve emphasis. First, **a premise refuted**: the question of whether HD 209458 b's strongly subsolar terminator water abundance (MacDonald and Madhusudhan, 2017) is atmospheric reality or retrieval artifact was answered by 2025 reanalyses using independent retrieval frameworks with improved systematics treatment, which find solar-consistent water—the pipeline had isolated a published conclusion that was in fact an artifact awaiting correction. Second, **the top question is independently posed and open**: the terminator-heterogeneity family is actively pursued by the community without resolution. Third, **the reframed historical-validation case behaves as designed**: JWST/NIRSpec PRISM programs are testing the pre-launch TRAPPIST-1 CO_2 detectability prediction (Lustig-Yaeger et al., 2019), with stellar contamination (Rackham et al.,

Table 4: Historical validation of all ten ranked questions (generated from pre-2021 evidence, judged against 2021–2026 literature). Premise: whether the post-cutoff literature validated, refuted, or left untested the evidential premise behind the question.

Rank	ID	Question (abridged)	Verdict	Premise
1	q_002	WASP-12 b: terminator heterogeneity vs. water evidence, abundance, C/O	posed, open	untested
2	q_001	HD 189733 b: cause of factor-of-five wind discrepancy (Na vs. CO/H_2O)	partial	validated
3	q_004	HD 189733 b: conditions for subsolar H_2O with robust CO	partial	validated
4	q_008	HD 209458 b: subsolar water—atmospheric reality or retrieval artifact?	**answered**	**refuted**
5	q_007	Patchy-cloud solar-composition models under independent retrievals & data	partial	untested
6	q_010	WASP-121 b: water detection sensitivity to clouds, contamination, priors, null model	partial	untested
7	q_005	HD 209458 b: independent support for NH_3/HCN across cloud scenarios	**answered**	validated
8	q_009	Conditions for patchy-cloud fits without high mean molecular weight	partial	validated
9	q_006	WASP-12 b: does C/O > 1 exclusion survive without equilibrium chemistry?	partial	untested
10	q_011	TRAPPIST-1: have JWST observations validated pre-launch CO_2 detectability predictions?	partial	untested

2018) currently blocking a decisive verdict—the judge correctly reports the premise as untested rather than refuted.

That zero questions were ignored by the subsequent literature is the central quantitative result: questions surfaced systematically from evidence tensions in pre-2021 knowledge are the questions the field subsequently invested in.

5.6 Failure analysis

The seven *partial* verdicts are not one failure mode but three, and each points at a specific pipeline improvement.

Blocked by a shared upstream obstacle. For the TRAPPIST-1 prediction test (rank 10) and the M-dwarf-adjacent cloud questions (rank 5), the post-cutoff literature engaged exactly the right test but could not reach a verdict because *stellar contamination* limits the achievable precision. Notably, the pipeline's own corpus contained the warning (Rackham et al., 2018): had the falsification screen consulted the evidence graph for known systematics on the target's host star, these questions could have been generated with the obstacle named and a contamination-robust test demanded. Failure source: the falsification stage checks data *availability* but not data *sufficiency*.

Questions demanding joint analyses no single study performs. The wind-discrepancy question (rank 2) needs contemporaneous optical–near-infrared spectroscopy; the subsolar-H_2O-with-CO question (rank 3) needs both species constrained in one dataset. The community advanced each piece separately. These questions are well posed but implicitly assume a coordinated analysis;

an execution-aware refinement step could split them into staged sub-questions the way the curated rank-1 family already is.

Questions broader than any single study answers. The class-level conditions question (rank 8) and the five-factor sensitivity question (rank 6) span parameter spaces the literature answers instance by instance. The clarity gate checks that scope is *bounded*, not that it matches the granularity at which the field publishes; a granularity criterion would have flagged both.

Why zero *not-addressed*. Beyond genuine question quality, two selection effects contribute: tensions among highly-cited claims concern targets the community already observes intensively, and the validation corpus was collected with target-specific queries that guarantee topical coverage. The baseline experiments proposed in Section 6 (random future-work questions; graph-free LLM questions over the same corpus) are designed to measure exactly how much of the engagement rate these effects explain.

6 Discussion

What the backtest does and does not show. Historical validation measures whether generated questions align with where the field actually went—strong evidence of question quality, but conservative in one direction: a question the community never engaged might still be valuable (the field can be collectively wrong or resource-constrained). Conversely, engagement is partially confounded by obviousness; a question everyone was about to ask scores well. The premise-refuted outcome is the cleanest signal, since it requires the system to have flagged a specific published conclusion as fragile before the community overturned it. Scaling from 10 questions to hundreds, and comparing against baselines (questions sampled from review-paper future-work sections; questions generated without the evidence graph), is the necessary next step before strong claims.

Human adjudication is load-bearing—by design. Tension typing and question curation are human steps, and the reviewed sample shows why: only 1 of 16 top candidates was a true contradiction. The framework's contribution is to make expert judgment cheap and auditable (typed verdicts over ranked candidates, committed as data) rather than to eliminate it. Automating adjudication against the growing corpus of human verdicts—with the 33% early-precision figure as the baseline to beat—is a measurable follow-on task.

6.1 Threats to validity

Single domain (astronomy only). Exoplanet atmospheres was chosen because literature, catalogs, and observation archives align unusually well. The framework's interfaces are domain-agnostic, but the tension topic vocabulary and falsification screens are domain-supplied; the engagement rate may not transfer to fields with looser evidential structure. Porting to a second domain is the direct test.

Abstract-only extraction. Claims are extracted from abstracts. Calibration showed abstracts carry the load-bearing conclusions, but assumptions and uncertainty statements often live in the body; full-text ingestion (with the chunk-level source locations the schema already mandates) should raise claim coverage and adjudication quality.

Single human adjudicator. All tension verdicts and curation decisions come from one domain reviewer. Verdicts are committed verbatim and re-appliable, so inter-annotator agreement can be measured by replaying the same candidates past additional reviewers—but it has not yet been.

Judge bias. The validation judge is an LLM; engagement judgments ("substantive" vs. topical) retain subjectivity even with citations restricted to retrieved records. The extraction, ranking, and validation stages also share a model family, risking correlated blind spots. Mitigations in place: forced bibcode-grounded evidence, human spot-checks of verdicts; mitigations still needed: independent judges, human re-scoring of a verdict sample.

Small sample, no baseline. Ten questions support existence claims, not rates. The engagement-rate confounds identified in the failure analysis (highly-cited targets, targeted validation queries) require the baseline comparisons described above before the headline numbers can be read as effect sizes.

Implications for general scientific intelligence. The results support a specific, testable position: question-asking competence can be decomposed into auditable stages—evidence representation, tension detection, refinement, prioritization—each of which can be measured and improved independently, with historical backtesting as the end-to-end score. If artificial general intelligence is to include the scientist's core skill of knowing *what to ask*, benchmarks of this form—rather than open-ended ideation judged by contemporaneous panels—are the right instrument for tracking progress toward it.

7 Conclusion

We presented a framework that treats scientific question discovery as a first-class, measurable capability: evidence becomes provenance-carrying claims; cross-paper tensions are detected and typed under human adjudication; prioritized signals become falsifiable questions; and a two-stage protocol ranks scientific priority separately from execution priority. A corpus discipline built on frozen, time-bounded manifests makes the whole loop reproducible and enables historical backtesting.

Instantiated on exoplanet atmospheres, the pipeline generated 10 questions from evidence available before 2021. The 2021–2026 literature—never seen by the system—engaged every one of them: two were answered, one with its premise explicitly refuted; the top-ranked question is independently posed and remains open. Question discovery, these results suggest, is not an ineffable spark but a capability that can be engineered, audited, and scored—and therefore one on which progress toward general scientific intelligence can be tracked.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in Semantic Scholar. In *Proceedings of NAACL-HLT (Industry Track)*, 2018.

M. Brogi, R. J. de Kok, S. Albrecht, I. A. G. Snellen, J. L. Birkby, and H. Schwarz. Rotation and Winds of Exoplanet HD 189733 b Measured with High-dispersion Transmission Spectroscopy. *The Astrophysical Journal*, 2016. doi: 10.3847/0004-637X/817/2/106. ADS bibcode: 2016ApJ...817..106B.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of K-CAP*, 2019.

Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004.

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, et al. The automation of science. *Science*, 324(5923):85–89, 2009.

Laura Kreidberg, Michael R. Line, Jacob L. Bean, Kevin B. Stevenson, Jean-Michel Désert, Nikku Madhusudhan, Jonathan J. Fortney, Joanna K. Barstow, et al. A Detection of Water in the Transmission Spectrum of the Hot Jupiter WASP-12b and Implications for Its Atmospheric Composition. *The Astrophysical Journal*, 2015. doi: 10.1088/0004-637X/814/1/66. ADS bibcode: 2015ApJ...814...66K.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

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

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011.

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.

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.

Ryan J. MacDonald, Jayesh M. Goyal, and Nikole K. Lewis. Why Is it So Cold in Here? Explaining the Cold Temperatures Retrieved from Transmission Spectra of Exoplanet Atmospheres. *The Astrophysical Journal*, 2020. doi: 10.3847/2041-8213/ab8238. ADS bibcode: 2020ApJ...893L..43M.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025.

Pierre-Yves Oudeyer and Frédéric Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1:6, 2007.

Arazi Pinhas, Nikku Madhusudhan, Siddharth Gandhi, and Ryan MacDonald. H₂O abundances and cloud properties in ten hot giant exoplanets. *Monthly Notices of the Royal Astronomical Society*, 2019. doi: 10.1093/mnras/sty2544. ADS bibcode: 2019MNRAS.482.1485P.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Benjamin V. Rackham, Dániel Apai, and Mark S. Giampapa. The Transit Light Source Effect: False Spectral Features and Incorrect Densities for M-dwarf Transiting Planets. *The Astrophysical Journal*, 2018. doi: 10.3847/1538-4357/aaa08c. ADS bibcode: 2018ApJ...853..122R.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*, 2024.

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

Kenneth O. Stanley, Joel Lehman, and Lisa Soros. Open-endedness: The last grand challenge you've never heard of. O'Reilly Radar, 2017.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

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 EMNLP*, 2020.

A. Wyttenbach, D. Ehrenreich, C. Lovis, S. Udry, and F. Pepe. Spectrally resolved detection of sodium in the atmosphere of HD 189733b with the HARPS spectrograph. *Astronomy & Astrophysics*, 2015. doi: 10.1051/0004-6361/201525729. ADS bibcode: 2015A&A...577A..62W.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.