Topic Momentum Forecasts Which Astronomy Questions the Field Takes Up: Five Pre-registered Out-of-Time Backtests and a Sealed 2026–2031 Forecast

Hui Mao
Aurora AI Ventures

Claude Code
Anthropic

October 2026

Abstract

Can we tell, from the literature available on a given day, which research questions a field will substantively take up over the following years? Earlier work in this series built an evidence-graph question generator, a historical backtesting protocol, and a predictor of future attention fitted on 980 astronomy questions at cutoffs 2012–2020. Here we test that predictor, and two successors, where it matters: at cutoffs none of them has seen. We ran four pre-registered rounds, each committed publicly before its questions existed. Every prediction was frozen with SHA-256 digests and pushed before any outcome was labelled. The test material is 995 questions at five new cutoffs (2017, 2019, 2021, 2022, 2023) across eight astronomy subfields, scored against a 332,244-paper arXiv astro-ph corpus, with outcome windows running to October 2026. The original predictor transfers forward in time (AUC 0.63–0.69 at every cutoff) but does no better than knowing how a question was generated. Adding the question's embedding and *topic momentum*, meaning how many and how recent the pre-cutoff papers near the question are, yields a model (v3) that reaches AUC 0.769 on the fresh 2017 and 2023 cohorts. It beats the generation-source prior by +0.081 [0.042, 0.123], ranks questions within a source at 0.694, and is the best-calibrated model (Brier 0.191). We then test the three main threats to such results. Against a citation-graph outcome label that uses no text retrieval, v3 keeps AUC 0.680 (momentum alone 0.689). On 595 questions written with no language model, it scores 0.693; LLM phrasing inflates outcome rates by 16 points but not the ranking. On 90 questions labelled independently by two humans, it predicts their judgments (AUC 0.66–0.70) better than it predicts the LLM judge. The same human data shows that the strict LLM judge underlying every label in this line of work is far more conservative than human readers ($\kappa \approx 0$–0.14), whose own agreement is only $\kappa = 0.11$. A planning component from the series (Paper 4) has no predictive value. Finally, we seal a v3 forecast for 200 questions posed on 2026-10-14, to be scored after 2031-10-15. All code, data, caches, and the commit history that timestamps each step are public.

1 Introduction

Systems that propose scientific questions are usually judged by how good the questions look to a person or a model. A more demanding standard asks whether the field actually went where the questions point. Historical backtesting turns that standard into a measurement [4]:

- generate questions from literature published before a cutoff;
- freeze them;

- label each one by what the following years of literature did with it.

The third paper in this series [7] used the protocol to fit a predictor of future attention on 980 astronomy questions at cutoffs 2012–2020. It reported a temporal-split AUC of 0.71.

A predictor tuned on a fixed set of cutoffs can look good for reasons that do not survive time. This paper asks whether the forecast holds up when it is made the way a forecaster would actually have made it: at a cutoff the model never saw, trained only on outcomes that were already known on that day, and scored before anyone looks at the answers.

Contributions.

1. **Out-of-time replication.** Paper 3's predictor transfers to five new cutoffs (pooled AUC 0.655 over 995 questions). Most of that skill, however, is a generation-source prior (§5.1).
2. **Topic momentum.** A dense-retrieval measure of how much, and how recently, the field was already publishing near a question is the strongest single predictor we found. Combined with the question's embedding (v3), it reaches AUC 0.77 and within-source AUC 0.69 on fresh cutoffs, under pre-registered tests (§5.3).
3. **Three validity checks.** We test the forecast against a citation-graph outcome that bypasses text retrieval, against questions written with no language model, and against human judgments (§6). It survives all three.
4. **A measurement finding.** The strict LLM judge behind the series' labels is far stricter than humans, and humans barely agree with each other. Absolute outcome rates are therefore rater-relative, and LLM-phrased questions inflate them further (§6.3).
5. **A sealed forecast.** 200 questions posed on 2026-10-14 are scored by three model versions, frozen, and due for scoring after 2031-10-15 (§7).

2 Related work

Forecasting the direction of science. Embeddings trained on past literature can anticipate future findings: word embeddings of materials-science abstracts ranked compounds that were later reported for a property years before publication [11]. Semantic networks built from past papers predict which concept pairs will be studied together in quantum physics [2] and in AI [3]. Human-aware models that account for who can make a connection forecast discoveries better than content-only models [10]. Our task differs from all of these in three ways:

- the unit is a natural-language *research question*, not a concept pair or material;
- the outcome is *substantive engagement*, judged from later papers;
- every evaluation is pre-registered and out of time.

Predictability in the science of science. The science of science has long studied what can and cannot be predicted about papers and careers [1], for example long-term citation impact [12]. We predict uptake of *questions*, a step earlier than impact. Our momentum features are close in spirit to field-growth predictors, but they are measured per question.

LLM-generated ideas and LLM judges. Expert evaluation finds LLM-generated research ideas judged more novel but less feasible than human ones [9]. Our backtests replace expert opinion with observed outcomes, but they rely on an LLM judge. Known judge biases [13] and our own human comparison (§6.3) make that judge a measured instrument rather than ground truth.

This series. Paper 1 [5] generates questions from tensions in an evidence graph. Paper 2 [4] formalises historical backtesting and includes a blinded two-annotator study that we reuse. Paper 3 [7] provides the 980-question training set, the six feature groups, the label pipeline and the v1 model. Paper 4 [6] ranks next evidence-gathering actions for a question. All upstream code and data are vendored unmodified at pinned commits.

Pre-registration. Freezing hypotheses and analysis plans before seeing outcomes [8] is the main defence against tuning to the test set. We apply it to forecasting: the forecast itself is frozen and its digest published before the outcomes are computed.

3 Setup

Corpus. We harvested arXiv astro-ph metadata, including cross-lists, from 2005 to 2026-10-14 through OAI-PMH: 332,244 papers. The publication date is the version-1 date. The per-year counts match Paper 3's independent harvest: identical for 2007–2021 and within 0.06% elsewhere. Paper 3's keyword rules assign papers to eight subfields:

- exoplanet atmospheres;
- protoplanetary disks;
- stellar activity;
- fast radio bursts;
- gravitational waves;
- galaxy evolution;
- cosmology tensions;
- compact objects.

The 106,428 subfield papers are embedded with `text-embedding-3-small`.

Questions. For every cutoff and subfield, Paper 3's generators produce 25 questions from the preceding eight years of abstracts only:

- 10 *evidence-tension* questions, from clusters of contradiction statements across at least two papers (Paper 1's approach);
- 5 *direct-LLM* questions;
- 5 *author future-work* questions;
- 5 weak or negative *controls*.

Phrasing uses `gpt-4o-mini`. The five test cutoffs yield 995 questions; one small 2017 cell produced only 5 tension clusters.

Outcome label. A question is *addressed* if, among the top-8 post-cutoff papers of its subfield retrieved by TF-IDF, Paper 3's strict judge (`gpt-4o`) finds at least one that engages it directly and assigns an outcome other than `not_addressed`.

The outcome windows are:

- five years for the 2017 and 2019 cutoffs;
- to 2026-10-14 for the later cutoffs: 4.79, 3.79 and 2.79 years for 2021, 2022 and 2023.

Base rates range from 0.495 to 0.605.

Leak-free training. Each test cohort is scored by models fitted only on Paper 3 questions whose outcome windows had *closed* by the test cutoff:

Test cutoff	Training cutoffs
2017	2012
2019	2012, 2014
2021, 2022	2012, 2014, 2016
2023	2012, 2014, 2016, 2018

A model trained on later labels would see field dynamics from inside the test window.

Pre-registration and freezing. Each round's hypotheses, metrics and decision rules were committed to the public repository before that round's questions were generated. Questions, features and predictions were then frozen with SHA-256 digests and pushed before labelling. Models were designed only on cohorts whose labels had already been seen:

- v2 on the 2021 cohort;
- v3 on 2019, 2021 and 2022, plus leave-one-cutoff-out on Paper 3.

Confirmatory claims use only cohorts generated afterwards. Bootstrap intervals use 2,000 resamples stratified by cohort × source. *Within-source AUC* is the mean AUC over cohort × source strata that contain both outcome classes.

4 Models

v1: Paper 3's model. This is a logistic regression on six groups of cutoff-time features: text, evidence context, novelty, falsifiability, tractability and field environment. Our vendored copy reproduces Paper 3's published temporal-split AUC exactly (0.7126).

v2: embedding with a source fixed effect. This is a logistic regression on the question's embedding plus one-hot generation source. The fixed effect forces the remaining features to explain variation *within* a source. C is chosen by leave-one-cutoff-out AUC on the training set. A calibrated variant averages its log-odds with v1's features (plus the source effect) and applies Platt scaling on out-of-fold predictions.

Table 1: AUC by cutoff, as pooled / within-source, under Paper 3's label. Each cohort is scored by models trained only on labels closed by its cutoff. *dev* marks cohorts a model was designed on. Bold marks fresh tests of v3.

Cutoff (n)	v1 (Paper 3)	source prior	v2	v3
2017 (195)	0.630 / 0.577	0.628 / 0.50	0.674 / 0.594	**0.728 / 0.634**
2019 (200)	0.654 / 0.548	0.667 / 0.50	0.730 / 0.642	0.765 / 0.703 *(dev)*
2021 (200)	0.656 / 0.569	0.680 / 0.50	0.745 / 0.646 *(dev)*	0.784 / 0.704 *(dev)*
2022 (200)	0.677 / 0.648	0.667 / 0.50	0.747 / 0.650	0.776 / 0.687 *(dev)*
2023 (200)	0.693 / 0.609	0.758 / 0.50	0.827 / 0.738	**0.845 / 0.753**
All (995)	0.655 / 0.590	0.686 / 0.500	0.738 / 0.654	0.771 / 0.696

v3: adding topic momentum. For each question we compute cosine similarities between its embedding and every paper of its subfield in the eight-year past window, and derive nine features:

- mean similarity of the 10 and the 50 nearest papers;
- log-counts of papers above two similarity thresholds (0.58 and 0.65, set from the feature distribution, not from outcomes);
- that count as a share of the subfield;
- the share of the 50 nearest papers from the last two years;
- the mean age of the 20 nearest papers;
- the log-ratio of near papers in the last two years to the two years before;
- the count of near papers in the last two years.

v3 averages the log-odds of a momentum-plus-source logistic model and v2's embedding model, with the same Platt calibration. Its training cohorts never include the test cutoff.

5 Results

5.1 The original predictor transfers, but mostly as a source prior

At cutoff 2021, the first round, v1 scored AUC 0.656 [0.583, 0.733], against 0.71 within its own era:

- it beat a subfield-popularity baseline by +0.102 [0.039, 0.167];
- it did *not* beat a one-number-per-source prior (−0.024 [−0.109, 0.073]);
- its within-source AUC was 0.569 [0.496, 0.650].

Across all five cutoffs v1 is stable (Table 1): pooled AUC 0.630–0.693. Its ranking of direct-LLM questions is at or below chance (pooled 0.44), because those questions are engaged almost regardless of content.

Table 2: Round 3: pre-registered tests on the fresh 2017 and 2023 cohorts.

ID	Hypothesis	Result
T1	v3 ranks within a source	AUC 0.694 [0.612, 0.771] ✓
T2	v3 > v2 within a source	$\Delta = +0.028$ [−0.008, 0.062], not significant
T3	v3 > v1 within a source	$\Delta = +0.101$ [0.027, 0.176] ✓
T4	v3 > source prior (pooled)	0.769 vs 0.688; $\Delta = +0.081$ [0.042, 0.123] ✓
T5	Brier(v3) < Brier(v2)	0.191 vs 0.206 (v1 0.253) ✓
T6	v3 ranks evidence-tension questions	AUC 0.692 [0.604, 0.773] ✓

5.2 An embedding model beats the source prior

On the fresh 2019 and 2022 cohorts ($n = 400$), v2 was tested against pre-registered hypotheses:

- pooled AUC 0.737 against the source prior's 0.669, a gain of +0.068 [0.019, 0.117];
- within-source AUC 0.646 [0.583, 0.710];
- AUC 0.606 [0.515, 0.694] within evidence-tension questions;
- Brier score 0.207, against v1's 0.240 and the constant's 0.248.

Its within-source gain over v1 (+0.048) was not significant. v1 also scored above chance within sources on these cohorts (0.598), so part of round 1's within-source failure was specific to the 2021 cohort.

5.3 Topic momentum: the strongest predictor

Table 2 reports the round-3 tests on the fresh 2017 and 2023 cohorts ($n = 395$, base rate 0.51). v3:

- ranks within a source at 0.694 (T1);
- beats v1 within a source by +0.101 (T3);
- beats the source prior by +0.081 (T4);
- ranks evidence-tension questions at 0.692 (T6);
- has the best Brier score (T5).

Its within-source gain over v2, +0.028 [−0.008, 0.062], misses significance (T2). Across all five cutoffs v3 is the best model on both measures in every cohort (Table 1). The momentum model alone reaches pooled AUC 0.770 on the fresh cohorts. Its strongest contribution is inside evidence-tension questions: 0.738, against 0.599 for v2.

6 Validity checks

Figure 2 summarises v3's skill under every check in this section.

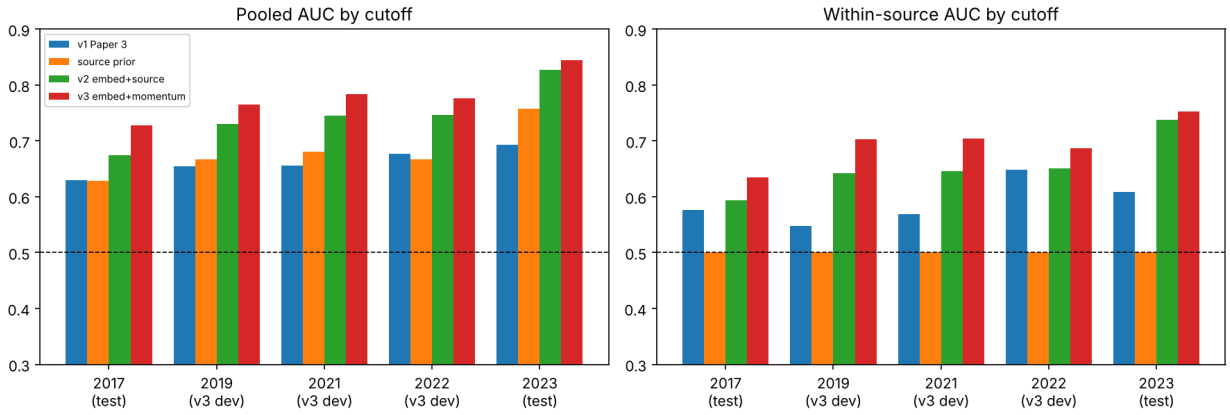


Figure 1: Pooled (left) and within-source (right) AUC at each cutoff. Cohorts marked *dev* were used to design v3; 2017 and 2023 are fresh tests.

6.1 Is momentum just predicting the retriever?

The outcome label finds future papers by TF-IDF similarity, and momentum counts past papers similar to the question. Momentum could therefore predict the retriever instead of the field.

We built a second label whose candidates are post-cutoff papers that *cite the question's source papers*, taken from the Semantic Scholar citation graph (1,213 of 1,214 source papers). It is judged by the same `gpt-4o` rubric. It is defined for the 755 questions that have source papers, across all five cutoffs, and agrees with the TF-IDF label on 73% of them.

On this label:

- v3 scores AUC 0.680 [0.639, 0.718];
- the momentum model alone scores 0.689 [0.650, 0.730], almost unchanged from its 0.699 on the TF-IDF label;
- v1 falls from 0.614 to 0.569.

Momentum tracks where the field went, not how the retriever works.

6.2 Did the question generator leak the future?

The generator (`gpt-4o-mini`) has read literature from after every backtest cutoff. To test whether that matters, we pre-registered and built 595 questions with *no language model*: the same tension clusters and future-work sentences, inserted verbatim into fixed templates. Each of the 395 tension templates shares its evidence cluster with exactly one LLM-phrased question.

- **v3's skill holds.** It scores 0.693 [0.648, 0.739] pooled and 0.704 [0.657, 0.750] within template type.
- **LLM phrasing inflates outcome rates.** On the same 395 clusters, the LLM-phrased question is addressed 44.6% of the time and the template 28.6%: +16.0 points [10.9, 21.0].
- **The ranking does not depend on phrasing.** v3's AUC is 0.726 on the templates against 0.679 on their LLM-phrased twins ($\Delta = +0.047$ [−0.032, 0.121]).

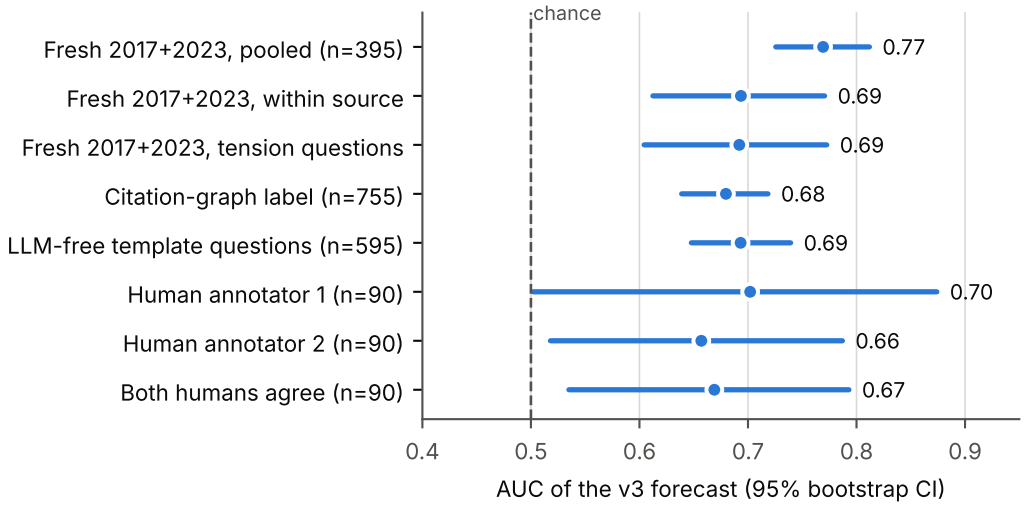


Figure 2: v3's AUC, with 95% bootstrap intervals, on the fresh test cohorts and under the three validity checks: a retrieval-independent citation label, LLM-free questions, and human annotators.

Absolute outcome rates of LLM-generated questions should therefore not be quoted as properties of the questions. Rankings are unaffected.

6.3 Do the forecasts predict human judgment?

Paper 2's blinded study had two annotators label 90 questions (cutoff 2020), each shown the same eight post-cutoff abstracts a judge sees. We ran Paper 3's judge on those items.

The judge and the humans disagree:

- humans call 90% and 73% of the questions addressed;
- the strict `gpt-4o` judge calls 43% addressed;
- agreement with each human is $\kappa = -0.00$ and 0.14;
- the two humans agree with each other at only $\kappa = 0.11\,[-0.08, 0.32]$.

We also checked that the judge itself is stable. Re-judging 120 of Paper 3's own questions today, `gpt-4o` reproduces the committed labels ($\kappa = 0.83$, identical 50.8% rate). `gpt-4o-mini`, the pipeline's code default, calls 93% addressed ($\kappa = 0.14$).

Scored with v3 trained only on cutoffs that closed by 2020, the 90 questions' forecasts predict:

- human 1, AUC $0.702\,[0.502, 0.874]$;
- human 2, $0.657\,[0.518, 0.787]$;
- questions both humans call addressed, $0.669\,[0.535, 0.793]$;
- the `gpt-4o` judge on the same items, only 0.548.

The forecast signal therefore reaches human judgment and is not an artefact of the judge's rubric. This analysis was not pre-registered, and $n = 90$ is small.

6.4 Negative result: the evidence planner

Paper 4's planner ranks next evidence-gathering actions. Its top-k utility does not predict uptake: AUC 0.419 $[0.305, 0.537]$ within tension questions at cutoff 2021, replicating Paper 4's own retraction.

Its top-ranked action is nearly constant. Under the default configuration it recommends a simulation for 41 of the 42 questions that were followed up. Under the Paper-3-informed configuration it recommends a literature review for all of them. It matches the action the field actually took (classified by `gpt-4o`) at 0.60, against a permutation null of 0.61 ($p = 1.0$).

7 A sealed 2026–2031 forecast

On 2026-10-14 we generated 200 new questions with the same generators and scored them with v1, v2 and v3. Each version's scores are frozen separately (`FREEZE*.json`). They will be labelled with the same pipeline over 2026-10-15 to 2031-10-14, no earlier than 2031-10-15. The pre-registered claims are:

- AUC > 0.5;
- the top quintile is addressed more often than the bottom quintile.

Paper 2's 200 questions, frozen on 2026-08-17, are scored the same way.

The forecast is about *uptake*, not answers. A score of 0.75 means the model expects at least one post-cutoff paper to engage the question directly and substantively, with new evidence, a test or a resolution attempt. It makes no claim about *what* that work will find, or whether the question will be settled. Resolutions are too rare in the historical labels (8 of Paper 3's 980 questions) for that to be forecastable at this scale.

v3's mean forecast is highest for fast radio bursts (0.62) and cosmology tensions (0.60), and lowest for stellar activity (0.42). Among the evidence-tension questions, the top-ranked concern:

- the Hubble-constant discrepancy and early-dark-energy resolutions (0.65–0.75);
- the abundance of luminous $z > 10$ galaxies seen by JWST (0.70–0.75);
- magnetar and merger-remnant origins of fast radio bursts (0.65–0.73);
- the $\sim$2.5 $M_\odot$ object in GW190814 and the black-hole mass gap (0.66–0.67).

The lowest-ranked are narrow object-level discrepancies (0.19–0.28), such as retrieved C/O ratios of Gliese 229 B or spot coverage of HAT-P-12.

8 Limitations and deviations

- **The label is an instrument, not ground truth.** "Addressed" means addressed according to a strict `gpt-4o` rubric over abstract-level retrieval. Human annotators are both more lenient and mutually inconsistent. Our human check ($n = 90$) supports the forecast's validity but is underpowered.
- **Shorter windows at recent cutoffs.** The 2021–2023 cohorts have outcome windows of 2.8–4.8 years, against five years for Paper 3.
- **One domain, one corpus.** Astronomy only, arXiv only, eight keyword-defined subfields.

- **Generator leakage.** It affects outcome rates (+16 points) but not the ranking. Remaining embedding-model leakage was not tested separately.
- **Deviations from pre-registration, all disclosed in the repository.**
  - Round 1 named the wrong judge model: the code default `gpt-4o-mini` instead of `gpt-4o`, which made Paper 3's labels. We report both and use the latter.
  - The citation label was delayed by API rate limits, and its round-1 analyses are exploratory.
  - v2 received the single-training-cutoff rule pre-stated for v3, to be scorable at 2017. This leaves multi-cutoff predictions byte-identical.
- **T2 failed.** v3's within-source improvement over v2 is not significant. Its advantage is significant against v1 and the source prior.

9 Conclusion

Across five pre-registered out-of-time cutoffs, a forecaster built from this series' components predicts which astronomy questions the field will take up. The decisive ingredient is not the structure of the question but *topic momentum*: whether the field was already moving toward it at the cutoff.

The forecast survives three checks:

- an outcome label that bypasses text retrieval;
- questions no language model wrote;
- human judgment.

The work also measures, rather than assumes, how much its LLM instruments shape absolute rates.

The sealed 2026–2031 forecast is the test no backtest can replace. Every model that exists today will be scored on literature none of them has seen.

Data and code.
https://github.com/nonameisready/scientific-discovery-astronomy. This contains:

- pre-registrations `docs/PREREGISTRATION*.md`;
- frozen predictions and digests (`data/*/FREEZE*.json`);
- labels and LLM caches;
- scripts reproducing every number in this paper.

References

[1] Aaron Clauset, Daniel B. Larremore, and Roberta Sinatra. Data-driven predictions in the science of science. *Science*, 355(6324):477–480, 2017. doi: 10.1126/science.aal4217.

[2] Mario Krenn and Anton Zeilinger. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences*, 117 (4):1910–1916, 2020. doi: 10.1073/pnas.1914370116.

[3] Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*, 5(11):1326–1335, 2023. doi: 10.1038/s42256-023-00735-0.

[4] Hui Mao. Historical backtesting for scientific question discovery: A protocol and astronomy pilot. *arXiv preprint arXiv:2608.16795*, 2026.

[5] Hui Mao. Evidence-based scientific question discovery: A framework with historical backtesting. *arXiv preprint arXiv:2608.09968*, 2026.

[6] Hui Mao. Planning the next evidence acquisition step: An offline framework for active scientific discovery. OpenFindings, https://openfindings.org/papers/or:2026:f35059ae, 2026.

[7] Hui Mao. What makes a scientific question succeed? predicting future attention, resolution, and premise revision from question structure and evidence context. OpenFindings, https://openfindings.org/papers/or:2026:3eab2db2, 2026.

[8] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

[9] 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.

[10] Jamshid Sourati and James A. Evans. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10):1682–1696, 2023. doi: 10.1038/s41562-023-01648-z.

[11] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019. doi: 10.1038/s41586-019-1335-8.

[12] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013. doi: 10.1126/science.1237825.

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

A Timeline of commits

Each step below is a public commit, so the order is verifiable.

Commit	Step
`bcf8469`	Round-1 pre-registration (cutoff 2021), before any question
`60f0f5d`	2021 cohort and predictions frozen, before any label
`a4eea0f`	Round-2 pre-registration and v2 models
`bd84ebc`, `58fe4b7`	2022 and 2019 cohorts frozen
`18665c2`	Round-3 pre-registration, v3 and momentum features
`d17b821`	2017 and 2023 cohorts frozen (v1, v2, v3)
`566b4d9`, `7327dc8`	Round-4 pre-registration; template questions frozen