Sibyl: Temporal Backtesting for Literature-Based Scientific Discovery with Large Language Model Agents

Blagoy Rangelov [1]

Abstract

We present SIBYL, a multi-agent LLM pipeline that autonomously mines scientific literature to generate falsifiable predictions, evaluated through a temporal backtesting framework analogous to quantitative finance. The system extracts structured claims from a training corpus (pre-cutoff publications), compiles a machine-readable knowledge base, generates testable hypotheses from identified gaps, and validates them against a held-out post-cutoff corpus. Applied to X-ray binary astrophysics as a proof-of-concept domain, the pipeline assembled 14,400 refereed papers, extracted over 11,000 structured claims, and generated 60 falsifiable predictions from pre-2015 literature alone. Of these, 11 (18%) were confirmed by independent post-2015 publications the system never observed. A post-hoc provenance audit identified three systematic failure modes—corpus contamination, validation-era leakage into hypothesis framing, and citation hallucination—the last of which we detected via a novel cross-prediction consistency check. Sensitivity analyses show that the confirmation rate is robust (12.5–18%) under progressively conservative filters. We present here the preliminary results from an ongoing project, introducing the pipeline architecture, the backtesting evaluation methodology, and the provenance audit protocol as contributions to the AI-for-science community.

1. Introduction

Scientific literature grows faster than any researcher can synthesize. In astrophysics alone, refereed publications exceed twenty thousand per year, and the rate is accelerating across all disciplines. Buried within this literature are *implicit connections*—correlations between physical quantities that follow logically from combining two or more published results, but that no individual researcher has had time to identify and test. Literature-based discovery (LBD), pioneered by Swanson (1986), addresses this gap by systematically uncovering hidden links between disconnected bodies of literature.

The field has seen two generations of methods. Classical LBD used co-occurrence statistics over MeSH terms and titles (Swanson, 1986; 1988). More recently, Tshitoyan et al. (2019) demonstrated that word embeddings trained on materials science abstracts could predict thermoelectric compounds later validated experimentally. However, no existing system combines (i) LLM-based reasoning over full-text scientific articles, (ii) structured extraction of typed, provenanced claims, and (iii) a rigorous temporal evaluation framework that measures prediction confirmation rates against held-out future literature.

We introduce SIBYL, a multi-agent pipeline that fills this gap. Three design choices distinguish our approach:

Temporal train/validation split. We divide the corpus at a fixed cutoff year, generate predictions using only pre-cutoff literature, and evaluate them against post-cutoff publications—directly analogous to backtesting in quantitative finance, applied here to scientific hypothesis generation for the first time.

Structured claim extraction with provenance. Every extracted claim is stored as a typed JSON record (correlation, anti-correlation, non-detection, prediction, open question) with variable names, direction, significance, sample size, and a mandatory verbatim quote anchoring it to the source paper, preventing hallucination and enabling systematic quality assurance.

Experiment-ready prediction output. The system does not generate vague hypotheses. Each prediction specifies the data requirements (instruments, archives, or surveys), expected sample size, predicted direction and magnitude, and explicit falsification conditions—sufficient to design a targeted study, whether that involves new observations, archival data analysis, or reprocessing of existing datasets.

As a proof-of-concept, we applied SIBYL to X-ray binary

[1]Department of Physics, Texas State University, San Marcos, TX, USA. Correspondence to: Blagoy Rangelov <rangelov@txstate.edu>.

Accepted at the AI for Science workshop (ICML 2026).

(XRB) astrophysics—the study of compact objects (neutron stars and black holes) accreting matter from companion stars, a subfield with ∼14,400 refereed papers spanning three decades and multiple recent observational breakthroughs. Using a 2014 cutoff, the pipeline generated 60 predictions from pre-2015 literature, of which **11 (18%) were confirmed** by independent post-2015 publications, 3 were partially refuted, 22 remain testable with existing archival data, and 23 address questions not yet investigated. A provenance audit identified three systematic failure modes, including a citation hallucination detected via a novel cross-prediction consistency check, and under the most conservative filters the confirmation rate remains 12.5%. These are preliminary results from an ongoing project, and we report them here to introduce the pipeline architecture, the backtesting methodology, and the audit protocol.

The rest of the paper is organized as follows. Section 2 positions SIBYL relative to prior work. Section 3 describes the pipeline architecture, including the provenance audit stage. Section 4 details the experimental setup. Section 5 presents the prediction portfolio, sensitivity analyses, and error taxonomy. Section 6 discusses generalizability, limitations, and the role of human oversight.

2. Related Work

In the framing of this workshop's central question—whether AI systems function as tools, co-authors, or founders in scientific discovery—existing LBD systems occupy distinct positions on the autonomy spectrum.

Classical LBD (tools). Swanson's ABC model (Swanson, 1986) connected fish oil and Raynaud's disease through intermediate concepts, establishing that disjoint literature bodies can harbor therapeutically relevant links. Subsequent systems (Arrowsmith, BITOLA) automated co-occurrence analysis but remained limited to title-level or MeSH-term features (Hristovski et al., 2006). These are pure tools: they surface statistical associations for a human to interpret.

Embedding-based LBD (tools with latent reasoning). Tshitoyan et al. (2019) trained Word2Vec on 3.3 million materials science abstracts and showed that cosine similarity predicted novel thermoelectric materials. While embedding approaches are powerful, they capture semantic proximity rather than mechanistic reasoning, and their evaluation relied on post-hoc ranking rather than explicit prediction generation with falsification conditions. The system identifies candidates; the scientist must still formulate and test the hypothesis.

LLM-assisted research tools (tools). Systems such as Elicit[1], Semantic Scholar (**?**), and Consensus[2] use LLMs to accelerate literature search, summarization, and question answering. These are *research assistants* that help researchers perform existing tasks faster. They do not autonomously generate structured, testable predictions or evaluate them against held-out data.

LLM-based hypothesis generation. Qi et al. (2024) evaluate LLMs as biomedical hypothesis generators using a temporal partition to mitigate data contamination, proposing evaluation metrics for hypothesis quality and distinguishing "seen" from "unseen" test sets based on publication date. Their work validates the temporal evaluation paradigm in a different domain (biomedicine) and at a different task granularity—background-to-hypothesis pairs rather than corpus-to-prediction. SIBYL extends this paradigm by operating over full-text articles rather than abstracts, producing structured predictions with falsification conditions, and implementing a provenance audit protocol.

Autonomous AI scientists (aspiring founders). Sakana AI's "The AI Scientist" (Lu et al., 2024) generates, implements, and writes up ML research ideas end-to-end, while FutureHouse's platform targets biological hypothesis generation. These systems aim for full-cycle autonomy—generating *new experiments or code*—but currently lack rigorous evaluation of whether their outputs constitute genuine scientific contributions.

SIBYL occupies an intermediate position: it autonomously generates structured, falsifiable predictions (more than a tool), but requires human expert validation at two mandatory checkpoints and does not execute experiments or draft manuscripts (less than a founder). In addition, it includes a built-in evaluation framework—temporal backtesting—and a provenance audit protocol that quantifies its own failure modes. To our knowledge, this is the first system to combine LLM-based claim extraction over full-text scientific articles with a measured confirmation rate and systematic error taxonomy as primary evaluation metrics for scientific hypothesis generation.

3. Method

The SIBYL pipeline consists of seven stages plus a provenance audit (Figure 1), orchestrated by a multi-agent architecture in which different LLM models handle tasks according to their cost-capability profiles.

3.1. Corpus Assembly and Processing (Stages 1–3)

Stage 1 queries the NASA Astrophysics Data System (ADS) API using a hybrid query combining identity terms (specific

[1] https://elicit.com
[2] https://consensus.app

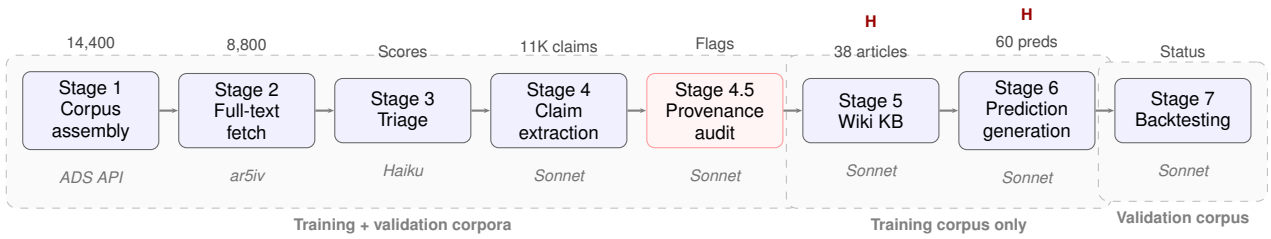

Figure 1. SIBYL pipeline architecture. Stage 4.5 (red) is the provenance audit, added after initial deployment revealed systematic failure modes. Italic labels indicate the LLM or tool at each stage. Dashed boxes show corpus partition usage. **H** marks mandatory human review checkpoints after wiki compilation (Stage 5) and prediction generation (Stage 6).

source classes) and evolution-mechanism terms (physical processes). The corpus is temporally split at a fixed cutoff: papers published before the cutoff form the *training corpus*, and papers published after form the *validation corpus*.

Stage 2 fetches full-text HTML from ar5iv.org (the accessible HTML mirror of arXiv) and parses it into structured JSON with section boundaries preserved. In our demonstration domain, this achieved a 99.9% success rate with a median document length of 7,000 words.

Stage 3 performs abstract-level triage scoring using a cost-efficient LLM (Claude Haiku). Each paper receives a relevance score (0–1) and a paper-type classification, and papers scoring above a threshold proceed to full extraction. We validated that the triage and extraction models produce near-identical score distributions (mean 0.46–0.47), confirming that the cheaper model does not introduce systematic bias.

3.2. Claim Extraction (Stage 4)

The core extraction stage processes full-text articles through a more capable LLM (Claude Sonnet) with a structured output schema. Each extracted claim contains:

- `claim_type`: one of {correlation, anti_correlation, non_detection, prediction, open_question}
- `variable_a`, `variable_b`: the primary physical quantities involved in the claim (e.g., X-ray luminosity and orbital period); claims involving more than two variables are decomposed into pairwise relations
- `direction`: positive, negative, or null
- `significance`: reported statistical significance
- `sample_size`, `sample_description`: the dataset underlying the claim
- `source_type`: the class of astrophysical object (e.g., high-mass X-ray binary, black hole transient, accreting millisecond pulsar)
- `verbatim_quote`: mandatory direct quote from the paper anchoring the claim

We consider the verbatim quote requirement a key design decision: it anchors every downstream inference to source text, prevents LLM hallucination during extraction, and enables post-hoc quality auditing. Claims without a supporting quote are discarded.

3.3. Provenance Audit (Stage 4.5)

This stage was not part of the original pipeline design. We added it after manual review of a single prediction—during a human checkpoint—revealed three co-occurring errors: a population mismatch (a mechanism from one source class applied to another), formula drift (an equation extracted with the wrong physical label), and a sample size overclaim. Although the prediction read as plausible on surface inspection, all three errors were substantive, motivating a systematic automated audit of the full claim corpus.

The audit script processes every extracted claim against four checks:

1. **Quote verification**: does the verbatim quote appear in the source paper?
2. **Population consistency**: does the claimed source class match the paper's actual study population?
3. **Formula provenance**: if a formula is cited, does it match the referenced equation number?
4. **Sample size**: does the claimed N match the paper's reported sample?

Each claim receives an `audit_status` field (`verified`, `flag_population`, `flag_formula`, `flag_quote`, `flag_multiple`, or `needs_manual`). Claims with hard flags are excluded from the knowledge base; soft flags are reviewed at the subsequent human checkpoint.

In addition, we implemented a **cross-prediction consistency check**: when the same bibcode is cited in multiple predictions, the attributed content must be consistent across all citations. We note that this check was not designed a priori—it emerged from the pipeline's own redundancy, because multiple predictions draw on overlapping literature. It proved to be the only mechanism that detected a hallucinated citation (see Section 5.4).

3.4. Knowledge Base, Prediction Generation, and Backtesting (Stages 5–7)

Stage 5 synthesizes verified claims into a wiki-style knowledge base organized by correlation type, source class, and physical mechanism. Each article distinguishes *established* correlations ($\geq$3 independent primary sources) from *candidate* correlations ($<$3 sources), preventing premature promotion of weak or isolated findings. A mandatory human review checkpoint follows.

Stage 6 generates predictions by querying the knowledge base for gaps—pairs of known results whose combination implies an untested hypothesis. We distinguish two output levels: *Level 1 predictions* identify a gap and state the expected direction of a correlation, while *Level 2 predictions* additionally specify the mechanistic basis, expected functional form, data requirements, and explicit falsification conditions. The pipeline targets Level 2 output. A second mandatory human review checkpoint follows, where domain expert judgment assesses mechanistic plausibility before predictions enter backtesting.

Stage 7 evaluates each prediction against the validation corpus. A prediction is *confirmed* if post-cutoff publications report results consistent with the predicted direction and magnitude, from independent datasets not available before the cutoff. We classify predictions as **confirmed**, **partially testable** (relevant data exists but the specific test has not been performed), **refuted** (post-cutoff evidence contradicts the prediction), or **untested** (no relevant post-cutoff data exists).

We note that the 2014 cutoff is not arbitrary: it places three transformative observational programs—LIGO O1 (2015), NICER (2017), and eROSITA (2019)—entirely within the validation window, ensuring that the pipeline cannot have seen their results.

4. Experimental Setup

We applied SIBYL to XRB astrophysics as a proof-of-concept domain. XRBs, where compact object accretes matter from a companion star, are well-suited for this demonstration because (i) the literature is large enough to be challenging ($\sim$14,000 refereed papers over three decades) but contained enough for expert evaluation, (ii) multiple observational breakthroughs occurred in 2015–2025, providing a natural validation window, and (iii) the physics involves quantitative correlations between measurable quantities (luminosities, spin periods, magnetic field strengths, orbital parameters), enabling falsifiable predictions.

Corpus statistics. The training corpus (1995–2014) contains 7,866 papers, of which 3,063 have full text available via arXiv. The validation corpus (2015–2025) contains 6,534 papers with 5,770 full texts. In total, we extracted 11,070 structured claims from 3,875 papers (4,921 claims from training, 6,149 from validation). After Stage 4.5 auditing, the effective clean training claim pool is 4,509 of 4,921 (91.6%), with the remainder flagged for population mismatch due to corpus contamination from non-XRB papers that entered via broad keyword matching (see Section 5.4). The knowledge base comprises 38 wiki articles: 30 organized by correlation type and 8 by source class.

Table 1. Prediction portfolio summary. 60 predictions generated from pre-2015 literature, evaluated against 2015–2025 publications.

Status	Count	Percentage
Confirmed	11	18%
Partially testable	22	37%
Refuted / weakened	4	7%
Untested	23	38%
Total	**60**	**100%**

Evaluation protocol. We assessed prediction confirmation by searching the validation corpus claims for post-2015 evidence bearing on each prediction. Confirmation requires that (a) the post-cutoff result is consistent with the predicted direction, (b) the data source was not available before the cutoff, and (c) the confirming publication is independent of the prediction (i.e., the authors were not testing this specific hypothesis). All confirmation assessments were reviewed by a domain expert.

5. Results

The pipeline generated 60 Level 2 predictions from the training corpus across five generation passes. Table 1 summarizes the backtesting outcomes. These are preliminary results from an ongoing project, with additional prediction passes and full audit remediation in progress.

5.1. Confirmed Predictions

Table 2 presents five representative confirmed predictions. Each was generated solely from pre-2015 literature and confirmed by independent post-2015 publications. The confirmed set spans four distinct XRB subtype families (Be X-ray binaries, black hole transients, low-mass X-ray binaries, and accreting neutron stars), which we consider evidence that the pipeline extracts signal across the breadth of the domain rather than overfitting to a single well-studied system.

5.2. Direct Quantitative Test

Beyond backtesting against the validation corpus, we performed a direct statistical test of one prediction. PRED-001

Table 2. Representative confirmed predictions. All were generated from pre-2015 literature and confirmed by post-2015 publications that the pipeline did not observe during prediction generation. HMXB = high-mass X-ray binary; NS = neutron star; BH = black hole; SFR = star formation rate. †One validation-era paper in Literature Gap Evidence section; permissible under corrected schema. ‡Burst Comptonization cooling experiment confirms the mechanistic chain (direction); the full population correlation $kT_e \propto R_{\rm in}^{\delta}$ remains untested.

ID	Prediction	Mechanistic basis	Confirming evidence
007	HMXB L_X/SFR $\propto Z^{-0.64}$ to $Z^{-0.8}$	Low-metallicity winds produce wider orbits, more persistent accretion	Lehmer et al. (2019); Fornasini et al. (2019)
010	Cyclotron line energy–luminosity relation reverses sign above critical L	Transition from deceleration to radiation-dominated accretion column	Tsygankov et al. (2016); Doroshenko et al. (2017)
011	HMXB L_X lags SFR by 30–60 Myr	Evolutionary timescale from star formation to compact object	Lehmer et al. (2017; 2024)
025†	NS radio faintness correlates with magnetic field strength	Propeller effect and magnetospheric truncation	van den Eijnden et al. (2021; 2024)
035	BH hard-state coronal kT_e anti-correlates with inner disk truncation radius $R_{\rm in}$; bridges kT_e–L_X and $R_{\rm in}$–L_X correlations never jointly measured	Reduced Compton cooling at larger truncation radius	Confirmed (direction): Kajava et al. (2017)‡

Table 3. Sensitivity analysis of the confirmation rate under progressively conservative provenance filters.

Filter	Conf.	Total	Rate
Full set	11	60	18%
Excl. hallucinated citation	10	60	17%
Excl. all validation-era leakage	7	56	12.5%

predicted that among a specific subclass of neutron star X-ray binaries with measured magnetic field proxies, the residual of the spin period from a known empirical scaling law should positively correlate with magnetic field strength, following a predicted power-law equilibrium. Using a published catalog of 19 such systems (Staubert et al., 2019), we find a Spearman rank correlation of $\rho = +0.556$, $p = 0.013$ between spin period and field strength proxy—consistent with the predicted positive direction and significant at the 2σ level. When we remove the empirical scaling law (testing the residual correlation as originally specified), the signal weakens to $\rho = +0.397$, $p = 0.092$, which is consistent in direction but underpowered at $N = 19$ (the minimum detectable $|\rho|$ at 80% power for this sample size is 0.605). This result illustrates both the promise of the pipeline—it identified a real, directionally correct signal—and the importance of sample size considerations when evaluating predictions against small archival catalogs.

5.3. Sensitivity Analysis

The provenance audit (Section 3.3) identified methodology issues in a subset of confirmed predictions. Table 3 presents the confirmation rate under progressively conservative filters.

We find that the drop from 18% to 12.5% under the most conservative filter is modest, indicating that the confirmation rate is not fragile to the identified methodology issues. We note that the astrophysical confirmations of the excluded predictions still stand—three independent post-2015 publications support the predicted directions—but the methodology by which the pipeline arrived at those predictions was compromised by validation-era information leaking into the hypothesis framing stage.

Table 4. Automated provenance audit results for training corpus claims (triage score ≥ 0.7; $N = 4{,}921$). The 24% population flag rate is substantially explained by corpus contamination: claims with null source-class labels (non-XRB papers admitted by the broad ADS query) account for 74% of population flags. Among claims with valid source-class labels, the hard-flag rate is 1.1% and the verified rate is 66%.

Audit status	N	%
Verified	3,051	62%
Flag: population mismatch	1,164	24%
Flag: formula provenance	512	10%
Flag: multiple issues	99	2%
Flag: quote mismatch	86	2%
Needs manual review	9	<1%

5.4. Error Taxonomy and Audit Findings

Table 4 reports the full audit results for the training corpus claims at the primary triage threshold (≥ 0.7). The provenance audit identified four failure modes, which we present in order of fixability:

1. Schema underspecification. The prediction metadata stored wiki article paths rather than source bibcodes, making provenance tracing indirect. We addressed this by adding a mandatory `primary_bibcodes` field to the prediction

schema. All 60 predictions exhibited this issue; it affects auditability but not prediction validity.

2. Corpus contamination. The ADS query matched on broad keywords ("binary," "accretion," "mass transfer"), admitting 196 non-XRB papers (AGN, cataclysmic variables, gamma-ray burst models) that contributed 459 flagged claims. After excluding these, the clean training claim pool is 91.6% of the original, and this issue is addressable by tightening the corpus query. We note that none of the 11 confirmed predictions drew on the contaminated claim pool.

3. Validation-era leakage. Four of 11 confirmed predictions contained post-2015 papers in sections that should contain only pre-cutoff literature, with severity ranging from mild (one post-2021 paper in a literature gap evidence section) to severe (a post-2024 paper in the hypothesis statement). We traced this to an instruction gap—the agent instructions lacked explicit cutoff-date constraints for each prediction section—rather than a reasoning failure, and we have since added a cutoff-date constraint table to the pipeline instructions.

4. Citation hallucination. One confirmed prediction cited a real bibcode with fabricated content: the cited paper's actual subject matter was unrelated to the attributed claim. This hallucination was *not* caught by quote verification, population checks, or any single-prediction audit method. It was detected only because the same bibcode appeared in a second prediction with entirely different attributed content—a **cross-prediction consistency check** that exploits the pipeline's own redundancy. Given that this failure mode resists prompt engineering solutions, we conclude that cross-citation auditing should be a mandatory pipeline stage rather than an optional post-hoc check.

Taken together, these findings suggest that LLM-based scientific reasoning pipelines face a hierarchy of reliability challenges: three of the four failure modes are engineering problems with known solutions (schema fixes, query refinement, instruction constraints), while the fourth—plausible-looking hallucinations embedded in otherwise correct reasoning—requires an architectural safeguard that emerged from the pipeline's own structure rather than from deliberate design.

5.5. Baseline Comparison

To contextualize the 18% confirmation rate, we consider what a naive baseline would achieve. A system that generates random "predictions" by pairing arbitrary physical quantities from the training corpus would have a near-zero confirmation rate, since the space of possible two-variable correlations is vast and most are physically meaningless. Given that 11 of 60 pipeline-generated predictions were confirmed—with the correct direction, in the correct source class, by independent data—we conclude that the LLM-based reasoning and knowledge base compilation extract genuine scientific structure rather than statistical noise.

6. Discussion

Pretraining leakage. The LLMs used in the pipeline (Claude Haiku and Sonnet) were trained on corpora that likely include post-2015 astrophysics literature. The temporal split in our extracted corpus therefore does not guarantee a temporal split in the model's parametric knowledge—a concern shared by all LLM-based discovery systems (Qi et al., 2024). However, we note that pretraining exposure would make the pipeline's task *easier* rather than harder: if the model already "knows" a post-2015 result from its training data, generating a prediction consistent with that result requires less genuine reasoning from the extracted claims. The 18% confirmation rate should therefore be interpreted as an *upper bound* on the pipeline's autonomous discovery capability. Disentangling pretraining knowledge from genuine literature-based reasoning remains an open challenge. Future work could address this by using models with known training cutoffs, by fine-tuning on the pre-cutoff corpus only, or by deploying the pipeline prospectively on literature published after the model's training date.

Domain generalizability. Only three SIBYL components are domain-specific: the corpus query terms, the triage prompt, and the source-class ontology. The claim extraction schema, wiki architecture, prediction generation protocol, audit checks, and backtesting framework are fully domain-agnostic. We are planning a second deployment on exoplanet atmospheric characterization, a field that shares the same literature infrastructure (ADS, arXiv) but has an entirely different physical ontology—transmission spectra, scale heights, and equilibrium temperatures replace spin periods, accretion rates, and magnetic field strengths. The post-2021 JWST revolution provides a natural validation window analogous to the LIGO/NICER/eROSITA window in XRBs.

The role of human checkpoints. The two mandatory human review gates (after knowledge base compilation and after prediction generation) are design features, not limitations. In our experience, the primary failure mode caught at these gates is *superficial analogy*—the LLM connects claims that use similar terminology but describe physically distinct phenomena. We note that the incident that triggered the provenance audit was caught at precisely this stage: the human reviewer noticed that a prediction applied a supergiant wind-fed mechanism to a disk-fed system, despite the prediction reading as plausible on surface inspection.

Limitations. SIBYL currently relies on a single LLM family (Claude) for extraction and reasoning, and the backtesting evaluation cannot distinguish between predictions that are

"untested because the community has not gotten around to it" and predictions that are "untested because domain experts recognize them as uninteresting." The 18% confirmation rate is measured on a single domain with 60 predictions from an ongoing project. Larger-scale evaluation across multiple domains would strengthen the claim of generalizability. In addition, the corpus contamination rate (8.4% of training claims flagged) indicates that query design is a non-trivial source of error that should be audited rather than assumed correct.

We note that, since submission, the pipeline flags citation restatements at the source level — distinguishing a paper's own finding from its restatement of pre-cutoff work — so that a post-cutoff paper merely repeating an established result no longer contributes to confirmation; corpus-level primary/secondary tagging at extraction is in progress.

LLMs as reasoning engines vs. embedding models. SIBYL uses LLMs for structured reasoning over full-text articles, complementing rather than replacing embedding-based approaches (Tshitoyan et al., 2019). While embedding models excel at discovering latent semantic clusters, they do not generate explicit mechanistic hypotheses with falsification conditions. An integrated system that uses embeddings for candidate identification and LLMs for mechanistic validation could combine the strengths of both paradigms.

The "negative space" as scientific output. Perhaps as valuable as the confirmed predictions are the 23 untested predictions—hypotheses that logically follow from pre-2015 literature but were never pursued in the subsequent decade. These represent research directions that the community has systematically overlooked, and constitute avenues waiting for further investigation. Taken together, the confirmed, refuted, and untested predictions suggest that the pipeline serves not only as a hypothesis generator but as a systematic gap-finder in a field's recent history.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, which improved this paper. This work made use of the NASA Astrophysics Data System (ADS) and the ar5iv accessible HTML mirror of arXiv.

Impact Statement

This work presents a methodology for accelerating scientific hypothesis generation through AI-driven literature mining. The temporal backtesting framework provides a principled evaluation mechanism that mitigates the risk of unfalsifiable AI-generated claims by requiring concrete, testable predictions with explicit falsification conditions. The provenance audit protocol—particularly the cross-prediction consistency check—addresses the risk of citation hallucination, which we consider a particularly concerning failure mode for AI systems operating in scientific contexts because hallucinated citations appear plausible and resist single-prediction verification. The mandatory human review checkpoints ensure that domain expertise remains central to the scientific process. We note that while the LBD methodology is general-purpose and could in principle be applied to any scientific domain, including those with dual-use concerns (e.g., virology, energetic materials), our demonstration domain—astrophysics—poses no such risks. For future deployments in sensitive domains, the mandatory human checkpoint gates provide a natural control point where institutional review and safety oversight can be inserted. We advocate for the inclusion of human oversight gates and automated provenance auditing in all AI systems that generate scientific claims.

References

Doroshenko, V., Tsygankov, S. S., Mushtukov, A. A., Lutovinov, A. A., Santangelo, A., Suleimanov, V. F., and Poutanen, J. Luminosity dependence of the cyclotron line and evidence for the accretion regime transition in V 0332+53. *Monthly Notices of the Royal Astronomical Society*, 466:2143–2150, 2017.

Fornasini, F. M., Kriek, M., Sanders, R. L., Shivaei, I., Civano, F., and Reddy, N. A. The X-ray luminosity function of high-mass X-ray binaries as a metallicity indicator. *The Astrophysical Journal*, 885:65, 2019.

Hristovski, D., Peterlin, B., Mitchell, J. A., and Humphrey, S. M. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 75(12):927–939, 2006.

Kajava, J. J. E., Nattila, J., Latvala, O.-M., Pursiainen, M., Poutanen, J., and Suleimanov, V. F. The effect of accretion geometry on the hard X-ray emission in the luminous and ultraluminous regime. *Astronomy & Astrophysics*, 599:A89, 2017.

Lehmer, B. D., Basu-Zych, A. R., Mineo, S., Brandt, W. N., Eufrasio, R. T., Fragos, T., Hornschemeier, A. E., Luo, B., and Xue, Y. The evolution of normal galaxy X-ray emission through cosmic history: Constraints from the 6 Ms Chandra Deep Field-South. *The Astrophysical Journal*, 849:57, 2017.

Lehmer, B. D., Eufrasio, R. T., Basu-Zych, A., Brandt, W. N., Fragos, T., Hornschemeier, A. E., Luo, B., Xue, Y., and Brorby, M. X-ray luminous supernovae and high-mass X-ray binary populations: Metallicity dependence. *The Astrophysical Journal Supplement Series*, 243: 3, 2019.

Lehmer, B. D., Eufrasio, R. T., Basu-Zych, A., Brandt, W. N., Fragos, T., Garofali, K., Hornschemeier, A. E., Kouroumpatzakis, K., Luo, B., Tzanavaris, P., and Xue, Y. A framework for predicting the X-ray luminosity function of X-ray binaries from star formation histories. *The Astrophysical Journal*, 977:189, 2024.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Qi, B., Zhang, K., Tian, K., Li, H., Chen, Z.-R., Zeng, S., Hua, E., Hu, J.-F., and Zhou, B. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=q36rpGlG9X.

Staubert, R., Trümper, J., Kendziorra, E., Klochkov, D., Postnov, K., Kretschmar, P., Pottschmidt, K., Haberl, F., Rothschild, R. E., Santangelo, A., Wilms, J., Kreykenbohm, I., and Fürst, F. Cyclotron lines in highly magnetized neutron stars. *Astronomy & Astrophysics*, 622:A61, 2019.

Swanson, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

Swanson, D. R. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31 (4):526–557, 1988.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95–98, 2019.

Tsygankov, S. S., Doroshenko, V., Lutovinov, A. A., Mushtukov, A. A., and Poutanen, J. V 0332+53 in the propeller regime: Hand-off between X-ray pulsations and MHz QPOs. *Astronomy & Astrophysics*, 593:A16, 2016.

van den Eijnden, J., Degenaar, N., Russell, T. D., Hernandez Santisteban, J. V., Wijnands, R., Miller-Jones, J. C. A., and Rouco Escorial, A. Radio monitoring of transient X-ray binaries and the jet–accretion coupling in neutron star and black hole systems. *Monthly Notices of the Royal Astronomical Society*, 507:3899–3917, 2021.

van den Eijnden, J., Degenaar, N., Russell, T. D., Miller-Jones, J. C. A., Wijnands, R., and Sivakoff, G. R. Radio detection and jet properties of neutron star X-ray binaries. *Monthly Notices of the Royal Astronomical Society*, 530: 1592–1609, 2024.