SIBYL: A Multi-Agent Pipeline for Autonomous Hypothesis Generation with Temporal Backtesting

Blagoy Rangelov [1]

Abstract

We present SIBYL, a multi-agent LLM pipeline that autonomously produces falsifiable predictions from published scientific literature, covering literature synthesis, knowledge representation, hypothesis generation, and hypothesis evaluation. The system uses a tiered agent architecture with two mandatory human-in-the-loop checkpoints and an automated provenance audit, evaluated through a temporal backtesting framework. In a proof-of-concept deployment on X-ray binary astrophysics, the system generated 60 predictions from pre-2015 literature, of which 11 (18%) were confirmed by independent post-2015 publications (12.5% under the most conservative provenance filters). These are preliminary results from an ongoing project.

1. Scope and Autonomy

SIBYL operates on the *literature-to-hypothesis* segment of the scientific workflow. Table 1 maps the pipeline stages to the standard discovery cycle and characterizes the decision-making mode at each stage.

The system is **autonomous** in corpus assembly, text processing, triage, claim extraction, provenance auditing, and backtesting. It is **semi-autonomous** at two junctures: after knowledge base compilation and after prediction generation, where a domain expert reviews the outputs before the pipeline proceeds. These are not optional checks—they are mandatory gates enforced by the pipeline architecture.

The system does *not* cover experimental design execution, instrument operation, data collection, or manuscript drafting. It occupies the "ideation through evaluation" arc, producing structured predictions that a human scientist can then pursue through observation or experiment.

[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).

Table 1. Pipeline stages mapped to scientific workflow phases. **A** = autonomous, **H** = human decision required. Stage 4.5 was added after initial deployment revealed systematic extraction errors.

Workflow phase	Pipeline stage	Mode	Agent
Literature search	Corpus assembly	A	Python + API
Reading	Full-text fetch	A	Python
Relevance filtering	Abstract triage	A	Haiku
Information extraction	Claim extraction	A	Sonnet
Quality assurance	Provenance audit (4.5)	A	Sonnet
Knowledge synthesis	Wiki compilation	A→**H**	Sonnet
Hypothesis gen.	Prediction gen.	A→**H**	Sonnet
Hypothesis eval.	Backtesting	A	Sonnet

2. Architecture and Tooling

Tiered agent design. The pipeline uses different LLM models matched to task complexity and cost:

- **Triage agent** (Claude Haiku): classifies paper relevance. We validated against the extraction agent to confirm no systematic scoring bias.
- **Extraction agent** (Claude Sonnet): processes full-text articles into structured claim records using a typed JSON schema with mandatory verbatim quote provenance.
- **Audit agent** (Claude Sonnet): checks every claim against four provenance criteria (quote verification, population consistency, formula provenance, sample size accuracy) and flags failures for exclusion or human review.
- **Synthesis agent** (Claude Sonnet): compiles verified claims into a wiki-style knowledge base, distinguishing established ($\geq$3 sources) from candidate correlations.
- **Reasoning agent** (Claude Sonnet): queries the knowledge base to identify gaps between known results, generates mechanistic hypotheses, and writes structured predictions with falsification conditions.

Orchestration. An orchestrating layer manages stage sequencing, checkpoint/resume logic (saving state every 50 papers), and the human review gates. All inter-stage communication occurs through persistent structured files (JSON claims, Markdown wiki articles, structured prediction templates), not ephemeral context windows.

Provenance chain. Every prediction traces back through the reasoning chain: prediction → mechanistic basis (wiki articles) → supporting claims → verbatim quotes → source papers with DOIs. This end-to-end provenance enables auditing any prediction to its literature foundations. In addition, a **cross-prediction consistency check** verifies that when the same source paper is cited in multiple predictions, the attributed content is consistent across all citations. We note that this check was not designed a priori—it emerged from the pipeline's own redundancy and proved to be the only mechanism that detected a hallucinated citation in which a real bibcode was cited with fabricated content.

Domain adaptation. Three components are domain-specific: the corpus query terms, the triage prompt, and the source-class ontology. All other infrastructure transfers to new domains without retraining or architectural changes.

3. Evaluation Roadmap

Temporal backtesting (completed). The primary evaluation uses a temporal train/validation split. Predictions are generated from pre-cutoff literature only and scored against post-cutoff publications. In the proof-of-concept domain (X-ray binary astrophysics, 2014 cutoff), the system generated 60 predictions, of which 11 (18%) were independently confirmed by post-2015 publications. A provenance audit identified methodology issues in a subset of confirmed predictions (validation-era leakage into hypothesis framing, one hallucinated citation), and sensitivity analysis shows the confirmation rate is robust at 12.5–18% under progressively conservative filters. Confirmation requires consistency with the predicted direction, a data source not available before the cutoff, and independence of the confirming publication.

Ablation studies (planned). We plan to evaluate the contribution of individual pipeline components: (i) claim extraction with vs. without the verbatim quote requirement, (ii) prediction generation with vs. without the wiki knowledge base (direct claim-to-prediction vs. synthesized knowledge), and (iii) varying the established/candidate threshold (2, 3, or 5 sources required).

Cross-domain generalization (planned). To validate domain-agnostic claims, we will deploy the pipeline on exoplanet atmospheric characterization, a field with an entirely different physical ontology and a natural post-2021 (JWST) validation window.

We note that subsequent work decouples evidence from novelty, classifying each prediction as a method-validating rediscovery or an open, testable candidate, so that the confirmation rate is read as a validation control rather than the primary discovery metric.

4. Governance and Safeguards

Human-in-the-loop gates. The two mandatory checkpoints ensure that no prediction enters the scientific record without expert validation. In our experience, the failure mode caught at these gates is *superficial analogy*—the LLM connecting claims that share terminology but describe physically distinct phenomena. We note that the incident that triggered the provenance audit stage was caught at this checkpoint.

Provenance and attribution. Every prediction is traceable to source literature through the provenance chain. The system does not generate claims from parametric knowledge alone, and the cross-prediction consistency check provides an architectural safeguard against citation hallucination that cannot be eliminated through prompt engineering alone.

Transparency. The full pipeline code, extraction prompts, audit scripts, claim schema, and prediction templates will be released as open-source software under a permissive license.

Dual-use considerations. The system operates on published scientific literature and does not access proprietary data, control instruments, or produce actionable protocols for hazardous materials. The demonstration domain (astrophysics) poses no dual-use risk. For future deployment in sensitive domains, the human checkpoint gates provide a natural control point for institutional review.

Acknowledgements

We thank the anonymous reviewer for constructive feedback that improved this paper. This work made use of the NASA ADS and the ar5iv accessible HTML mirror of arXiv.

Impact Statement

SIBYL demonstrates that semi-autonomous AI systems can extract predictive signal from scientific literature, validated through temporal backtesting. The provenance audit protocol—particularly the cross-prediction consistency check for detecting citation hallucination—addresses what we consider a concerning failure mode for AI systems in scientific contexts. The mandatory human oversight gates, provenance chains, and planned open-source release ensure that the system augments rather than replaces scientific judgment. We advocate for temporal backtesting and automated provenance auditing as standard evaluation protocols for any AI system claiming to generate scientific hypotheses.