SIBYL

LITERATURE-BASED DISCOVERY

Blagoy Rangelov
Department of Physics, Texas State University

TRACK 3 · AI SCIENTIST COMPETITION

A Multi-Agent Pipeline for Autonomous Hypothesis Generation

Tool, co-author, or founder? Sibyl sits *in between* — it autonomously generates structured, falsifiable predictions (more than a tool), yet gates every hypothesis behind human review and never runs the experiment (less than a founder).

8 STAGES — 6 autonomous · 2 human gates

5 TIERED AGENTS — Haiku + Sonnet

3 SWAP-INS — retarget any field

SCOPE & AUTONOMY Where Sibyl sits

Sibyl

TOOL — CO-AUTHOR — FOUNDER

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	A	Sonnet
Knowledge synthesis	Wiki compilation	A → H	Sonnet
Hypothesis generation	Prediction gen.	A → H	Sonnet
Hypothesis evaluation	Backtesting	A	Sonnet

A = autonomous; **A → H** = autonomous, then a mandatory human decision. The two gates are enforced by the architecture, not optional checks.

ARCHITECTURE Tiered agents

TRIAGE TIER · **CLAUDE HAIKU** — LOW-COST, HIGH-VOLUME

Triage agent
Scores paper relevance; validated against the extraction model to confirm no systematic scoring bias.

REASONING TIER · **CLAUDE SONNET** — FULL-TEXT REASONING

Extraction
Full text → typed claim records with a mandatory verbatim quote.

Audit
Checks every claim on four provenance criteria; flags for exclusion or review.

Synthesis
Compiles verified claims into the wiki knowledge base (KB); established (≥3 sources) vs. candidate.

Reasoning
Queries the KB for gaps; writes mechanistic predictions with falsification conditions.

Inter-stage communication is through persistent structured files (JSON claims, Markdown wiki, prediction templates) — not ephemeral context — with checkpoint/resume every 50 papers.

PORTABILITY Use it for your field

SWAP · 3 COMPONENTS

- **Corpus query terms**
- **Triage prompt**
- **Source-class ontology**

REUSE · DOMAIN-AGNOSTIC

- ↺ Claim schema
- ↺ Prediction protocol
- ↺ Backtesting framework
- ↺ Wiki architecture
- ↺ Audit checks
- ↺ Orchestration layer

Retarget Sibyl to a new field by changing three components; everything else transfers unchanged. A second deployment on exoplanet atmospheres is underway.

TRUST Provenance chain

Prediction
The generated hypothesis — a specific correlation and its expected direction.

↓

Mechanistic basis
The wiki articles whose established relations were combined to motivate it.

↓

Supporting claims
The individual audited claim records those articles rest on.

↓

Verbatim quotes
The exact source sentence behind each claim, stored at extraction.

↓

Source papers · DOIs
The refereed papers by DOI, so any prediction traces back to print.

Every prediction is traceable end-to-end to its literature foundations — no claim is generated from parametric knowledge alone.

HALLUCINATION GUARD Cross-prediction check

Prediction A
cites 2019ApJ...42R
attributes claim **α**

Prediction B
cites 2019ApJ...42R
attributes claim **β**

⚠ **Same bibcode, conflicting content → flagged**
α ≠ β means at least one attribution is fabricated

When the same paper is cited across predictions, its attributed content must agree. This redundancy-based check was the **only** mechanism that caught a real bibcode cited with fabricated content — invisible to quote and population checks. Now a mandatory stage.

GOVERNANCE Safeguards

Human-in-the-loop gates
Two mandatory checkpoints — after KB compilation and after prediction generation. The failure mode they catch most is *superficial analogy*: claims sharing terminology but describing distinct physics.

Provenance & attribution
Every claim is anchored to a verbatim quote; the cross-prediction check guards against citation hallucination that prompt engineering alone cannot eliminate.

Transparency
Code, extraction prompts, audit scripts, the claim schema, and prediction templates are released open-source under a permissive license.

Dual-use
Operates only on published literature — no instruments, no proprietary data, no actionable protocols. The human gates are the natural control point for sensitive domains.

Validated by **temporal backtesting**: 60 predictions from pre-2015 literature, **18% confirmed** by independent post-2015 work (12.5–18% robust).

FULL RESULTS → COMPANION POSTER

RECENT ADVANCES

The architecture keeps tightening around provenance.

- **Cross-prediction auditing promoted to a mandatory stage** — the redundancy check that caught the hallucination is no longer post-hoc.
- **Deterministic surprise scoring.** KL (Kullback–Leibler) divergence between pre- and post-2015 claim-direction distributions — byte-identical across runs, replacing a stochastic LLM scorer.
- **Decoupled evidence × novelty schema.** Separates method-validating rediscovery from open, testable candidates — the discovery product.

FUTURE WORK

- ○ Ablations: verbatim-quote requirement, wiki vs. direct claim-to-prediction, established threshold (2/3/5 sources)
- ○ Cross-domain deployment: **exoplanet atmospheres**, validated against the post-2021 JWST (James Webb Space Telescope) record
- ○ Primary/secondary claim tagging at extraction — rigorous corpus-level leakage fix

TEXAS STATE UNIVERSITY

ICML 2026 · AI for Science Workshop · Seoul, South Korea

rangelov@txstate.edu · blagoyrangelov.com · github.com/blagoyrangelov

This paper (Architecture) — Companion (Results)