SIBYL
LITERATURE-BASED DISCOVERY

Blagoy Rangelov
Department of Physics, Texas State University

TRACK 1 · ORIGINAL RESEARCH

Temporal Backtesting for Literature-Based Scientific Discovery with LLM Agents

A multi-agent LLM pipeline mines the literature for *implicit connections*, generates falsifiable predictions from pre-cutoff papers, and scores them against held-out future work — *backtesting, applied to scientific discovery.*

14.4K PAPERS MINED — 11K structured claims
60 PREDICTIONS — pre-2015 literature
18% CONFIRMED — 12.5–18% robust

MOTIVATION The gap

Refereed astrophysics publications exceed twenty thousand per year, and no researcher can track the *implicit connections* buried across them — correlations that follow from combining published results but that no one has had time to test. Existing literature-based discovery counts term co-occurrence or ranks semantic neighbours; none combines LLM reasoning over full text with structured, provenanced claims and a test against future literature. We evaluate Sibyl the way a quantitative strategy is evaluated: by backtesting.

METHOD How Sibyl works

▸ TRAINING + VALIDATION CORPORA

1. **Corpus assembly** — *ADS API · hybrid query* — 14,400 papers
2. **Full-text fetch** — *ar5iv HTML mirror* — 8,800 texts
3. **Abstract triage** — *Claude Haiku* — scores
4. **Claim extraction** — *Claude Sonnet · typed JSON + verbatim quote* — 11K claims
4.5. **Provenance audit** — *Claude Sonnet · added post-deployment* — flags

▸ TRAINING CORPUS ONLY

5. **Wiki knowledge base** Human gate — *Claude Sonnet* — 38 articles
6. **Prediction generation** Human gate — *Claude Sonnet · Level-2 + falsification* — 60 preds

▸ VALIDATION CORPUS

7. **Backtesting** — *Claude Sonnet* — status

EVALUATION Temporal backtesting

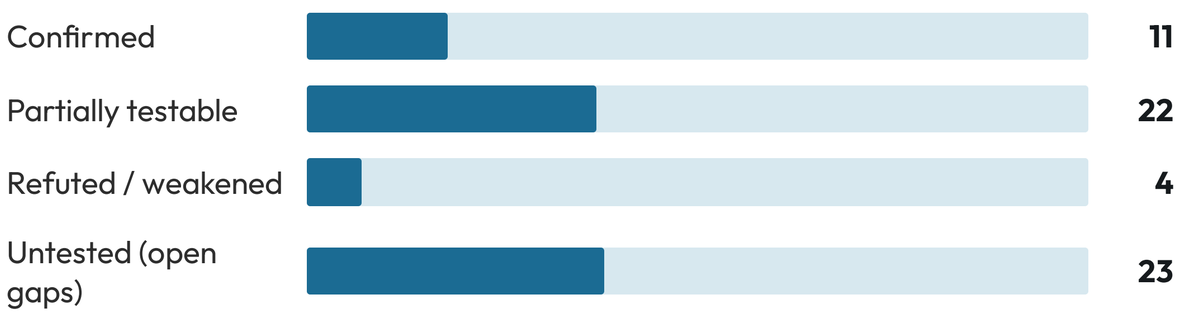
We split the corpus at a fixed 2014 cutoff, generate predictions from pre-cutoff literature alone, and score them against post-2015 publications. A prediction counts as *confirmed* only when the post-cutoff result matches the predicted direction, the data source was unavailable before the cutoff, and the confirming publication is independent of the prediction. We note that the cutoff places three transformative programs — LIGO O1, NICER, and eROSITA — entirely within the validation window, so the pipeline cannot have seen their results.

EVIDENCE Confirmed predictions

ID	PREDICTION	CONFIRMING EVIDENCE
007	HMXB L_X/SFR ∝ $Z^{-0.64…-0.8}$ (low-metallicity enhancement)	Lehmer+19; Fornasini+19
010	Cyclotron line energy–luminosity relation reverses sign above critical L	Tsygankov+16; Doroshenko+17
011	HMXB L_X lags SFR by 30–60 Myr	Lehmer+17, +24
025	NS radio faintness correlates with magnetic field strength	van den Eijnden+21, +24
035	BH hard-state coronal kT_e anti-correlates with disk truncation R_{in}	Kajava+17 (direction)

Five of eleven confirmations, spanning four X-ray binary (XRB) subtype families (Be, BH transients, LMXB, accreting NS) — signal across the domain, not one well-studied system. All generated from pre-2015 literature only.

OUTCOMES Prediction portfolio



The 23 untested predictions are research directions the field logically implied but never pursued — Sibyl as a systematic gap-finder.

ROBUSTNESS Sensitivity

PROVENANCE FILTER	CONF.	TOTAL	RATE
Full set	11	60	**18%**
Excl. hallucinated citation	10	60	17%
Excl. all validation-era leakage	7	56	**12.5%**

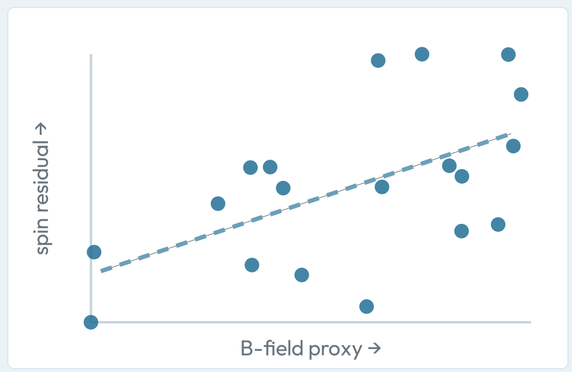
The drop from 18% to 12.5% under the most conservative filter is modest — the confirmation rate is not fragile to the identified methodology issues.

DIRECT TEST PRED-001, measured

Spin-period residual vs. magnetic-field proxy across 19 accreting NS systems (Staubert+19) — the predicted **positive** direction, significant at 2σ.

ρ = +0.556
p = 0.013 · N = 19

Without the empirical scaling law: ρ = +0.397, p = 0.092 — same direction, underpowered at N = 19 (min detectable |ρ| = 0.605).



PROVENANCE AUDIT Honest about failure

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%

A post-hoc audit checks every extracted claim against its cited paper — quote, source class, formula, and sample size. We note that the 24% population-flag rate is substantially *corpus contamination*: non-XRB papers admitted by a broad ADS query drive 74% of the flags, and none of the eleven confirmed predictions drew on the contaminated pool. Among claims with valid source-class labels the hard-flag rate is **1.1%** and the verified rate is **66%** (N = 4,921 at triage ≥ 0.7).

ERROR TAXONOMY Four failure modes, ordered by fixability

01 · SCHEMA
Underspecification
Predictions stored wiki paths, not source bibcodes — provenance was indirect. Fixed with a mandatory *primary_bibcodes* field.
SCHEMA FIX

02 · CORPUS
Contamination
Broad keywords admitted 196 non-XRB papers → 459 flagged claims. Clean pool is 91.6%; tighten the query.
QUERY REFINEMENT

03 · LEAKAGE
Validation-era leakage
4 of 11 confirmed predictions had post-cutoff papers in pre-cutoff sections. An instruction gap, not a reasoning failure — fixed with a cutoff-date table.
INSTRUCTION CONSTRAINT

04 · HALLUCINATION
Citation hallucination
A real bibcode cited with fabricated content. Invisible to quote/population checks — caught only by the **cross-prediction consistency check**.
ARCHITECTURAL SAFEGUARD

RECENT ADVANCES

Confirmation validates the method; candidate yield is the product.

- **Decoupled evidence × novelty schema.** Separates rediscovery (the validation control) from open, testable candidates. Latest run: **66 candidates + 5 validating rediscoveries** from 88 predictions, exploring 30 of 38 knowledge-base articles.
- **Live citation-leakage detection.** Per-source restatement detection distinguishes a paper's own findings from restatements of pre-cutoff work — closing the leakage path the audit surfaced.
- **Deterministic surprise scoring.** Surprise = KL (Kullback–Leibler) divergence between pre- and post-2015 claim-direction distributions — byte-identical across runs, replacing a stochastic LLM scorer (run-to-run ρ = 0.77, top-10 overlap 0.13).

FUTURE WORK

- ○ Second domain: **exoplanet atmospheres** — validated against the post-2021 JWST (James Webb Space Telescope) record
- ○ Primary/secondary claim tagging at extraction — a rigorous, corpus-level leakage fix
- ○ Ablations: empirical vs. LLM scorer; exhaustive vs. genetic selection; random-pair baseline
- ○ Single-expert blinded baseline; deeper run to measure true exploration depth

TEXAS STATE UNIVERSITY

ICML 2026 · AI for Science Workshop · Seoul, South Korea
rangelov@txstate.edu · blagoyrangelov.com · github.com/blagoyrangelov

This paper (Results)
Companion (Architecture)