Tracing the Cascade: A Topology-Aware Evaluation Framework for Scientific Agent Hallucinations

Xinshun Feng[1,2,*] Ziqi Miao[2,*] Lijun Li[2,*†] Jing Shao[2,†]

[1] Beihang University [2] Shanghai Artificial Intelligence Laboratory

Code Dataset

Motivation

Scientific errors propagate through concepts.

A false protein-function claim can mislead later pathway reasoning and disease-level conclusions.

What do existing evaluations miss?

Final-answer accuracy does not reveal the faulty reasoning step. Uniform error counts give the same penalty to errors on central and peripheral concepts.

Evaluation Objective

Locate false claims, quantify their structural impact, and identify corrections that change the outcome.

Benchmark

1,295 questions · 5 task formats · 10 LLMs

Task format	Protein	PathVQA
Claim choice	382	198
Boolean support	108	87
Two-hop reasoning	47	20
Essay	239	169
Experimental code	24	21

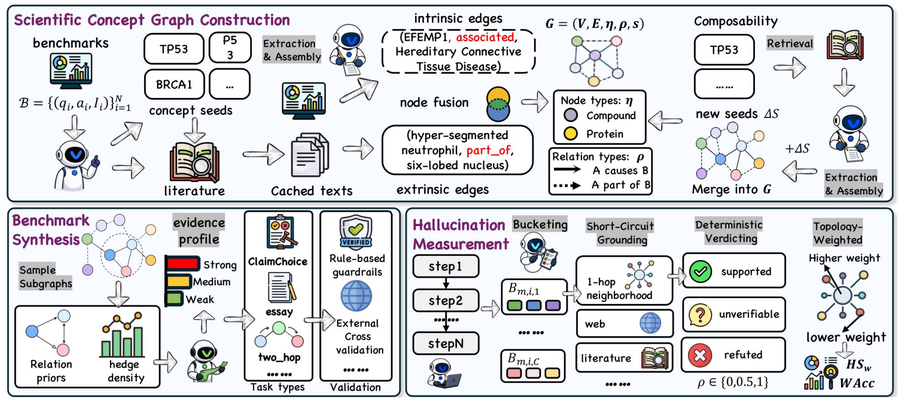
Topology-weighted severity

$$w_g(c) = 1 + \log(1 + \text{degree}(c))$$

$$\text{HS}_w = \sum_c w_g(c)\rho(c) / \sum_c w_g(c)$$

Off-graph claims retain the baseline weight of 1.

SCHEMA Framework

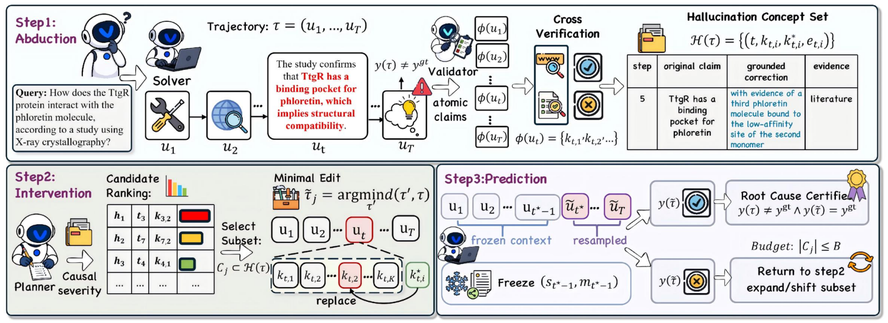


- Concept graphs connect task synthesis with claim-level auditing.
- Topology weights reveal the structural impact of reasoning errors.

Hallucination Detection & Attribution

Verify evidence → Correct the claim → Replay the trajectory

Freeze prior context and memory; test whether the correction changes the answer.



Pilot: 65 / 141 failures repaired (46.1%). Intern-S1-Pro · 346 ProteinLMBench questions

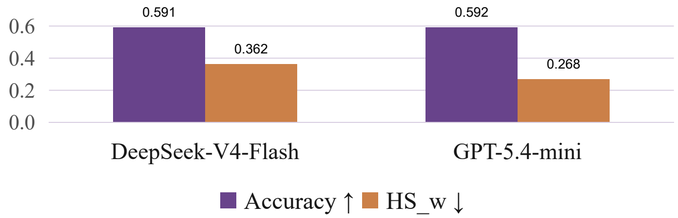
Evaluation Results

10 LLMs share the same agent workflow and tool budgets. Final-answer accuracy on two biomedical subsets:

Model	Protein ↑	PathVQA ↑
Kimi-K2.5	0.299	0.595
Llama-4-Scout	0.600	0.588
Doubao-Seed-2.0-Pro	0.739	**0.677**
Intern-S1-Pro	0.428	0.489
DeepSeek-V4-Flash	0.591	0.675
Qwen3.6-Plus	0.546	0.555
GPT-4o	0.482	0.418
GPT-5.4-mini	0.592	0.609
Gemini-3-Flash	**0.747**	0.529
Grok-4.1-Fast	0.554	0.658

Hallucination Results

Similar accuracy, different hallucination severity



Protein: 0.591 vs. 0.592 accuracy, but 0.362 vs. 0.268 HS_w. Comparable answers can conceal different reasoning risks.

Low hallucination rate can still hide high severity

PathVQA-Enhanced	HR ↓	HS_w ↓
Qwen3.6-Plus	0.009	0.432
Doubao-Seed-2.0-Pro	0.017	0.407
GPT-5.4-mini	0.024	0.287

Audit: failed-answer trajectories (Appendix E.3). ρ = 0 / 0.5 / 1: supported / unverifiable / refuted.

* Equal contribution. † Corresponding authors.