Propose, Critique, Falsify: Benchmarking Self-Verifying AI Scientists

Anonymous Authors[1]

Abstract

AI systems that autonomously generate scientific hypotheses are proliferating, yet the verification of their claims remains largely unexamined. We introduce a Propose-Critique-Falsify (PCF) benchmark that evaluates whether multi-agent pipelines can reduce the false discovery rate of AI-generated scientific claims. Across three evaluation domains (biomedical claim verification, synthetic statistical experiments, and scientific novelty assessment) and five frontier-class language models, we find that PCF pipelines dramatically increase false negatives through a mechanism we term conservatism-as-abstention: models emit Uncertain verdicts rather than risk incorrect classifications, rendering the false discovery rate undefined in many conditions. Critically, the direction of this error is model-dependent. Claude Sonnet 4 and GPT-4o become over-conservative under PCF, while GPT-5.4 becomes over-permissive (FDR = 0.75), labeling nearly all claims as valid. We further confirm that iterative self-refinement consistently increases false discovery rates across all models tested. Our results reveal that multi-agent verification architectures introduce systematic, model-specific biases that cannot be resolved through pipeline design alone.[1]

1. Introduction

The vision of AI systems that conduct scientific research autonomously has advanced rapidly, with recent proposals for end-to-end "AI scientists" that generate hypotheses, design experiments, and write papers (Lu et al., 2024; Gottweis et al., 2025). These systems can produce novel research ideas at scale (Si et al., 2024), but a critical gap remains: the claims they generate are not systematically verified before dissemination. Without reliable verification, autonomous AI science risks amplifying the reproducibility crisis rather than resolving it (Chaturvedi et al., 2026).

Human science addresses this problem through adversarial mechanisms: peer review, replication, and deliberate attempts at falsification. The question we investigate is whether analogous adversarial structures can be implemented within AI systems themselves. Specifically, can a multi-agent pipeline in which one agent proposes claims, a second critiques them, and a third attempts falsification systematically reduce the false discovery rate of AI-generated scientific claims?

We introduce the PCF benchmark, a three-role multi-agent evaluation framework for scientific claim verification. Our benchmark encompasses three domains with known ground truth: biomedical claim verification modeled on SciFact (Wadden et al., 2020), synthetic statistical experiments with controlled effect sizes, and scientific novelty assessment. We evaluate four experimental conditions (single-agent, self-refine, three-agent debate, and same-model PCF) across five frontier-class language models, with full four-condition results for Claude Sonnet 4 and GPT-4o.

Our central finding is that PCF pipelines do not simply trade sensitivity for specificity in a controlled manner. Instead, they induce conservatism-as-abstention: models overwhelmingly emit Uncertain verdicts rather than commit to a classification. On SciFact, GPT-4o's PCF pipeline drops accuracy from 1.00 to 0.20, with 16 of 20 instances classified as Uncertain. Furthermore, the direction of the PCF error is model-dependent: while Claude Sonnet 4 and GPT-4o become over-conservative, GPT-5.4's PCF pipeline becomes over-permissive, achieving FDR = 0.75 on synthetic statistics by labeling nearly all claims as valid.

Our contributions are as follows:

- We introduce the PCF benchmark for multi-agent scientific claim verification with controlled ground

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Submitted to the AI for Science workshop (ICML 2026).

truth across three domains, reporting full confusion matrices with Uncertain counts.

- We identify conservatism-as-abstention as the dominant failure mode of adversarial self-verification, distinct from a simple sensitivity-specificity tradeoff.
- We demonstrate that the direction of PCF error is model-dependent, with some models becoming over-conservative and others becoming over-permissive under identical pipeline conditions.
- We provide empirical confirmation that iterative self-refinement increases false discovery rates in scientific verification, extending the findings of Huang et al. (2024) and Tyen et al. (2024).

2. Related Work

Autonomous AI Science. Recent work has explored large language models (LLMs) as autonomous scientific agents. Lu et al. (2024) proposed the AI Scientist, an end-to-end system for automated discovery that generates ideas, runs experiments, and writes papers. Gottweis et al. (2025) introduced an AI co-scientist with multi-agent debate for biomedical hypothesis generation. Si et al. (2024) evaluated whether LLMs can generate genuinely novel research ideas. Chen et al. (2025) developed ScienceAgentBench for rigorous assessment of language agents in data-driven scientific discovery. Chaturvedi et al. (2026) surveyed the reliability, safety, and security requirements for LLMs in scientific applications. These systems focus primarily on generation rather than verification, and our work complements them by benchmarking the reliability of AI-generated scientific claims.

Self-Correction and Self-Refinement. A growing body of evidence challenges the assumption that LLMs can improve their own outputs through iterative refinement. Huang et al. (2024) demonstrated that LLMs cannot self-correct reasoning without external feedback. Tyen et al. (2024) showed that while LLMs struggle to identify reasoning errors, they can correct them if told where the errors are. Madaan et al. (2023) proposed Self-Refine for iterative improvement, but subsequent work has questioned its effectiveness for tasks requiring genuine reasoning correction. Our experiments directly test self-refinement for scientific verification and confirm these negative findings.

Multi-Agent Critique, Debate, and Verification. Multi-agent architectures have been proposed to improve LLM reasoning. Du et al. (2023) showed that multi-agent debate improves factuality. Liang et al. (2024) encouraged divergent thinking through debate. Chan et al. (2024) used multi-agent debate for evaluation. Gou et al. (2024) introduced tool-interactive critiquing. Lan et al. (2024) trained language models to critique using multi-agent feedback, demonstrating that structured critique can improve model outputs. Mulian et al. (2026) proposed AgentFixer for detecting and repairing failures in LLM agentic systems, complementing our focus on verification pipeline failures. Our PCF pipeline differs from symmetric debate by assigning asymmetric, specialized roles inspired by the scientific method: a proposer, an adversarial critic, and a falsifier.

Scientific Claim Verification and Adversarial Robustness. SciFact (Wadden et al., 2020) introduced a benchmark for verifying scientific claims against evidence. Min et al. (2023) developed fine-grained factual evaluation. Most relevant to our work, Huang et al. (2025) proposed POPPER, an agentic framework for hypothesis validation through sequential falsification. While POPPER focuses on validating specific hypotheses through tool use and data analysis, our PCF benchmark evaluates multi-agent verification as a general-purpose pipeline. Ou et al. (2026) demonstrated adversarial claim attacks against LLM-based fact-checking, highlighting the vulnerability of verification pipelines to adversarial inputs. Zeng et al. (2026) proposed a Bayesian adversarial multi-agent framework for AI-for-Science, providing complementary evidence that adversarial multi-agent designs require careful calibration.

3. Method

3.1. The Propose-Critique-Falsify Pipeline

The PCF pipeline assigns three specialized roles to separate LLM agents, inspired by the adversarial structure of the scientific method.

Proposer. The Proposer agent receives a scientific claim and its associated context (evidence abstracts for biomedical claims, statistical data summaries for synthetic experiments, or domain descriptions for novelty assessment). It evaluates the claim and generates an initial assessment of validity, including a reasoning chain and preliminary verdict.

Critic. The Critic agent receives the Proposer's assessment and adversarially identifies flaws, confounders, weaknesses, and alternative explanations. The Critic is instructed to be maximally skeptical: its objective is to find the strongest possible objections to the Pro-

poser's conclusions. This role draws on the principle that hypotheses are strengthened by surviving scrutiny (Bai et al., 2022).

Verifier/Falsifier. The Verifier receives both the Proposer's assessment and the Critic's objections. It designs the strongest possible falsification test for the claim, then renders a final verdict: Valid, Invalid, or Uncertain. The inclusion of Uncertain as a permissible verdict is a deliberate design choice that allows models to abstain rather than guess, but as our results demonstrate, this abstention mechanism becomes a dominant failure mode.

3.2. Evaluation Domains

We evaluate across three domains, each providing known ground truth:

Biomedical Claim Verification (SciFact). We curate 20 claims modeled on the SciFact dataset (Wadden et al., 2020), each paired with evidence abstracts. All 20 claims have ground-truth label Support (no Contradict instances). This imbalance means we cannot measure FDR on SciFact, as any false positive would require a Contradict instance predicted as Support. We retain this domain because the high single-agent accuracy (1.00 for both models) provides a ceiling against which PCF degradation is clearly visible.

Synthetic Statistical Experiments. We construct 20 instances with controlled ground truth: 5 with genuine statistical relationships (Valid) and 15 with confounded or null effects (Invalid). Each instance includes a data summary with statistical test results. This domain directly evaluates false discovery control, as the 3:1 ratio of invalid-to-valid instances tests whether pipelines can correctly reject spurious findings.

Scientific Novelty Assessment. We compile 20 claims about research findings: 7 describing genuinely novel results (Novel) and 13 describing well-established findings (Not_Novel). Models must classify claims accordingly. This domain tests the ability to distinguish genuine novelty from prior work.

3.3. Experimental Conditions

We compare four conditions:

1. Single-Agent: One LLM directly evaluates the claim and renders a verdict.
2. Self-Refine: Following Madaan et al. (2023), one LLM evaluates the claim, then critiques and revises its own assessment over two iterative rounds.
3. Debate (3-agent): Three instances of the same LLM engage in symmetric, two-round debate before a majority vote determines the verdict (Du et al., 2023).
4. PCF Same-Model: The three PCF roles (Proposer, Critic, Verifier) are instantiated with the same LLM.

3.4. Models

We evaluate five models. Our two primary models, for which we report full four-condition results, are Claude Sonnet 4 (claude-sonnet-4-20250514, Anthropic) and GPT-4o (gpt-4o-2024-08-06, OpenAI). We additionally evaluate Gemini 2.0 Flash (Google), which produced systematic structured-output compliance failures (Section 4.6); GPT-5.4 (OpenAI), for which partial results (no debate condition) reveal a critical model-dependent finding; and Gemini 2.5 Flash (Google), for which partial results are available.

3.5. Metrics and Uncertain Handling

Our primary metrics are accuracy and false discovery rate (FDR), defined as FDR = FP/(FP + TP). We report full confusion matrices including Uncertain counts.

Uncertain verdicts are treated as incorrect classifications and are included in the accuracy denominator. FDR is undefined (denoted "—" in tables) when the model makes no positive predictions (FP + TP = 0), as the ratio has a zero denominator. This occurs frequently in PCF conditions, where models abstain from positive predictions entirely. We also report F1 score, which is undefined (also denoted "—") when both precision and recall are zero. For statistical comparison between conditions, we use McNemar's test for paired binary outcomes. All confidence intervals are Wilson score intervals at the 95% level.

4. Results

4.1. Main Results

Tables 1 and 2 present the full results for our two primary models. We report accuracy with 95% confidence intervals, F1, FDR, and complete confusion matrix entries (TP, FP, TN, FN, Unc). Figures 1–3 visualize accuracy, FDR, and the Uncertain rate across conditions.

4.2. Finding 1: Conservatism-as-Abstention

The PCF pipeline does not achieve false discovery control through improved discrimination between valid

Table 1. Claude Sonnet 4 results across three domains and four conditions ($n = 20$ per cell). FDR and F1 are denoted "—" when undefined due to zero denominators. Unc = number of Uncertain verdicts (treated as incorrect).

Domain	Condition	Acc [95% CI]	F1	FDR	TP	FP	TN	FN	Unc
SciFact (20S, 0C)	Single-Agent	1.00 [1.0, 1.0]	1.00	0.00	20	0	0	0	0
	Self-Refine	1.00 [1.0, 1.0]	1.00	0.00	20	0	0	0	0
	Debate	0.95 [0.85, 1.0]	0.97	0.00	19	0	0	1	0
	PCF Same	0.80 [0.60, 0.95]	0.89	0.00	16	0	0	4	1
Synth. Stats (5V, 15I)	Single-Agent	1.00 [1.0, 1.0]	1.00	0.00	5	0	15	0	0
	Self-Refine	0.80 [0.60, 0.95]	0.83	0.29	5	2	11	0	2
	Debate	1.00 [1.0, 1.0]	1.00	0.00	5	0	15	0	0
	PCF Same	0.55 [0.30, 0.75]	—	1.00	0	2	11	5	3
Novelty (7N, 13NN)	Single-Agent	0.85 [0.70, 1.0]	0.73	0.00	4	0	13	3	0
	Self-Refine	0.65 [0.45, 0.85]	0.53	0.50	4	4	9	3	0
	Debate	0.75 [0.55, 0.90]	0.44	0.00	2	0	13	5	0
	PCF Same	0.70 [0.50, 0.90]	0.25	0.00	1	0	13	6	0

Table 2. GPT-4o results across three domains and four conditions ($n = 20$ per cell). Conventions follow Table 1. Note the extreme Uncertain rates in the PCF condition, particularly on SciFact (16/20) and Synthetic Statistics (12/20).

Domain	Condition	Acc [95% CI]	F1	FDR	TP	FP	TN	FN	Unc
SciFact (20S, 0C)	Single-Agent	1.00 [1.0, 1.0]	1.00	0.00	20	0	0	0	0
	Self-Refine	0.85 [0.70, 1.0]	0.92	0.00	17	0	0	3	3
	Debate	1.00 [1.0, 1.0]	1.00	0.00	20	0	0	0	0
	PCF Same	0.20 [0.05, 0.40]	0.33	0.00	4	0	0	16	16
Synth. Stats (5V, 15I)	Single-Agent	0.95 [0.85, 1.0]	0.89	0.00	4	0	15	1	1
	Self-Refine	0.40 [0.20, 0.60]	—	1.00	0	3	8	5	9
	Debate	1.00 [1.0, 1.0]	1.00	0.00	5	0	15	0	0
	PCF Same	0.40 [0.20, 0.60]	—	—	0	0	8	5	12
Novelty (7N, 13NN)	Single-Agent	1.00 [1.0, 1.0]	1.00	0.00	7	0	13	0	0
	Self-Refine	0.90 [0.75, 1.0]	0.83	0.00	5	0	13	2	0
	Debate	0.90 [0.75, 1.0]	0.83	0.00	5	0	13	2	0
	PCF Same	0.95 [0.85, 1.0]	0.92	0.00	6	0	13	1	0

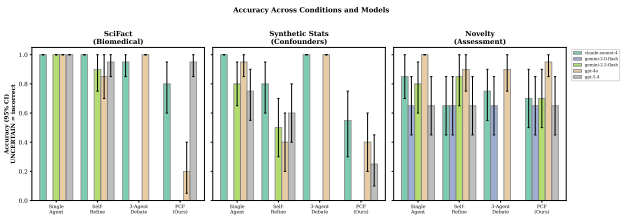

Figure 1. Accuracy with 95% Wilson confidence intervals across conditions and domains for Claude Sonnet 4 and GPT-4o. PCF Same-Model dramatically reduces accuracy on SciFact for GPT-4o (1.00 → 0.20) and on Synthetic Statistics for both models.

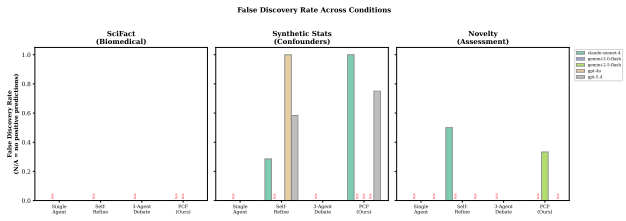

Figure 2. False discovery rate across conditions. "—" annotations indicate conditions where FDR is undefined because the model made no positive predictions (TP + FP = 0). Self-Refine is the only condition that systematically produces nonzero, defined FDR values.

and invalid claims. Instead, it induces conservatism-as-abstention: models overwhelmingly emit Uncertain verdicts rather than committing to a classification. This mechanism is qualitatively distinct from a sensitivity-specificity tradeoff, because the model is not making more negative predictions but rather refusing to classify at all.

The clearest demonstration is GPT-4o on SciFact: accuracy drops from 1.00 (single-agent) to 0.20 (PCF), with 16 of 20 instances classified as Uncertain (Table 2). On Synthetic Statistics, GPT-4o's PCF produces 12 Uncertain verdicts out of 20, yielding FDR = — (no positive predictions at all). Claude Sonnet 4 shows a less extreme but still substantial pattern:

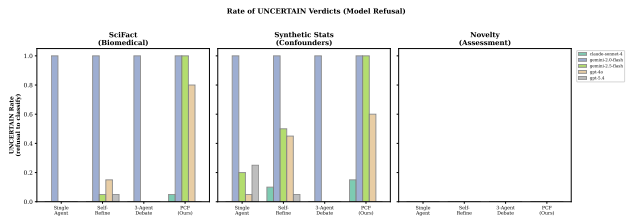

Figure 3. Uncertain rate (proportion of instances classified as Uncertain) per condition. The PCF pipeline induces extreme abstention in GPT-4o, with 80% of SciFact instances and 60% of Synthetic Statistics instances receiving Uncertain verdicts.

on Synthetic Statistics, PCF accuracy drops from 1.00 to 0.55, with 3 Uncertain verdicts, and the 2 positive predictions that are made are both false positives (FDR = 1.00).

The mechanism is straightforward: the Critic's adversarial arguments are sufficiently persuasive that the Verifier defaults to Uncertain rather than rendering a definitive verdict. Even when the underlying evidence strongly supports a claim, the Critic identifies enough potential confounders and alternative explanations to prevent the Verifier from committing.

McNemar's test confirms statistically significant differences between PCF and single-agent for GPT-4o on SciFact ($\chi^2 = 14.06$, $p = 0.0002$) and Synthetic Statistics ($\chi^2 = 9.09$, $p = 0.003$), and for Claude Sonnet 4 on Synthetic Statistics ($\chi^2 = 7.11$, $p = 0.008$). The SciFact comparison for Claude does not reach significance ($\chi^2 = 2.25$, $p = 0.134$), consistent with the smaller accuracy drop.

4.3. Finding 2: Self-Refine Degrades Verification

Self-Refine (Madaan et al., 2023) applied to scientific verification consistently increases false discovery rates, confirming and extending the findings of Huang et al. (2024). Across all models tested, self-refinement introduces nonzero FDR where the single-agent baseline achieves FDR = 0.00:

- GPT-4o, Synthetic Statistics: FDR 0.00 → 1.00
- Claude Sonnet 4, Synthetic Statistics: FDR 0.00 → 0.29
- Claude Sonnet 4, Novelty: FDR 0.00 → 0.50
- GPT-5.4, Synthetic Statistics: FDR 0.00 → 0.58 (partial)

The failure mode is that iterative self-critique introduces second-guessing without genuine new evidence. The model revises initially correct assessments based on self-generated objections, a process that Tyen et al. (2024) have shown to be unreliable. In the context of scientific verification, the model talks itself into accepting invalid claims or rejecting valid ones.

4.4. Finding 3: Model-Dependent Skepticism Direction

A critical finding emerges from partial GPT-5.4 results: the direction of PCF error is model-dependent. While Claude Sonnet 4 and GPT-4o become overconservative under PCF (excessive false negatives and Uncertain verdicts), GPT-5.4 exhibits the opposite failure mode.

On Synthetic Statistics, GPT-5.4's PCF pipeline achieves accuracy of only 0.25 with FDR = 0.75. Rather than abstaining, GPT-5.4's Verifier classifies nearly all instances as Valid, producing 15 false positives out of 20 instances (the ground truth contains only 5 valid instances). GPT-5.4's single-agent baseline achieves accuracy of 0.75 with FDR = 0.00, demonstrating that the over-permissive behavior is introduced specifically by the PCF pipeline.

This finding has important implications for the design of multi-agent verification systems: the same pipeline architecture can produce diametrically opposite failure modes depending on the underlying model. Calibrating a PCF pipeline for one model does not guarantee appropriate behavior when the model is changed.

4.5. Finding 4: Debate Matches Single-Agent

Three-agent symmetric debate generally matches single-agent accuracy while consuming approximately 9× more inference tokens (three agents across three rounds of deliberation). On SciFact, GPT-4o achieves 1.00 under both conditions; Claude drops from 1.00 to 0.95. On Synthetic Statistics, debate matches single-agent performance for both models (1.00 for Claude and GPT-4o). On Novelty, debate performs comparably or slightly below single-agent baselines.

These results challenge the practical value of symmetric debate for scientific verification. While Du et al. (2023) and Liang et al. (2024) have shown debate to improve factuality in general reasoning tasks, the benefits do not transfer to scientific claim verification, where the task requires weighing evidence rather than recalling facts.

4.6. Finding 5: Structured Output Compliance

Gemini 2.0 Flash exhibited systematic failure to produce structured verdicts, returning Uncertain for 100% of verification tasks in the SciFact and Synthetic Statistics domains despite explicit FINAL_VERDICT formatting instructions. Only the Novelty domain produced classification verdicts, where all instances were labeled Not_Novel, achieving 65% accuracy (equal to the majority-class baseline, as 13 of 20 instances are Not_Novel).

This failure demonstrates that structured output compliance is a prerequisite for multi-agent verification pipelines. A model that cannot reliably follow output formatting instructions is unusable as any agent in a pipeline that requires structured handoffs between roles, regardless of its underlying reasoning capability.

5. Discussion

Conservatism as Abstention, Not Discrimination. The central finding of this work is that adversarial verification in multi-agent pipelines does not produce a controlled tradeoff between false discovery rate and sensitivity. Instead, the PCF pipeline induces a qualitatively different failure mode: the Verifier abstains from classification rather than improving its discrimination between valid and invalid claims. When FDR is defined (i.e., when the model does make positive predictions), it can be arbitrarily high (Claude Sonnet 4, Synthetic Statistics: FDR = 1.00) or zero, depending on the model and domain. The PCF pipeline does not reduce FDR; rather, it makes FDR undefined by suppressing positive predictions.

This distinction matters for AI scientist design. A system that abstains on 80% of inputs (GPT-4o, SciFact) is not safer than one that makes wrong predictions; it is simply uninformative. The implicit assumption that adversarial critique will improve calibration is not supported by our data.

Model-Dependent Error Direction. The observation that GPT-5.4 becomes over-permissive while Claude Sonnet 4 and GPT-4o become over-conservative under identical pipeline conditions suggests that the PCF architecture amplifies model-specific tendencies rather than imposing a uniform verification standard. Models that are naturally inclined toward caution become more cautious under adversarial critique; models that are naturally inclined toward acceptance become more accepting when the Verifier must reconcile competing arguments. This finding resonates with recent work on adversarial robustness of LLM-based fact-checking (Ou et al., 2026) and the calibration challenges identified in Bayesian adversarial multi-agent frameworks (Zeng et al., 2026).

Implications for AI Scientist Design. Our results suggest several design principles. First, self-refinement should be avoided for verification tasks, as it consistently degrades performance, confirming Huang et al. (2024) in the scientific domain. Second, symmetric debate provides limited benefit over single-agent evaluation at substantially higher computational cost. Third, adversarial multi-agent pipelines require model-specific calibration, and a pipeline tuned for one model may produce opposite failure modes on another. Fourth, the Uncertain option, while intuitively appealing, becomes a dominant failure mode in adversarial settings and may need to be constrained or eliminated in favor of forced-choice verdicts.

Comparison with POPPER and Related Frameworks. Our work complements the POPPER framework (Huang et al., 2025), which validates hypotheses through sequential tool-augmented falsification. POPPER uses real data analysis as a calibration mechanism, which our pure language-based verification lacks. The conservatism-as-abstention pattern we identify may be ameliorated in tool-augmented systems where the Verifier can ground its judgment in concrete data rather than relying solely on the persuasiveness of natural-language arguments. Similarly, the multi-agent critique training approach of Lan et al. (2024) suggests that training models specifically for the critic role, rather than using general-purpose models, may improve pipeline calibration.

6. Limitations

We identify the following limitations, which should be weighed carefully when interpreting our findings.

Small sample size. All results are based on $n = 20$ instances per domain per model per condition. The resulting 95% confidence intervals are wide (Tables 1–2), and our statistical power to detect moderate effects is limited. McNemar's test is underpowered at this sample size, so non-significant comparisons should not be interpreted as evidence of no difference.

Imbalanced SciFact domain. All 20 SciFact claims have ground-truth label Support, with no Contradict instances. This means we cannot measure FDR on SciFact (any false discovery would require a ground-truth negative), and accuracy on this domain conflates discrimination ability with a tendency to predict Support.

No open-source models. Infrastructure limitations prevented evaluation of open-source models (Mixtral, LLaMA, Mistral). Our findings may not generalize to models with different training procedures, safety tuning, or instruction-following characteristics.

Curated data, not original SciFact. Our SciFact-domain instances are hand-written claims modeled on SciFact, not drawn from the original dataset. This limits comparability with prior work on the SciFact benchmark.

No critic severity ablation. We planned but did not complete ablations varying the Critic's aggressiveness. The relationship between critic severity and Verifier abstention rate is an important open question.

No systematic error analysis. We do not provide a qualitative analysis of which instance types trigger Uncertain verdicts or what patterns distinguish correct from incorrect PCF predictions.

Single seed. All results are from a single experimental run with no variance across seeds reported. Stochasticity in LLM outputs (at $\tau = 0.7$) means that specific confusion matrix entries may vary across runs.

Partial results for three models. Gemini 2.0 Flash produced no usable verification data outside the Novelty domain. GPT-5.4 and Gemini 2.5 Flash are missing the debate condition, preventing full comparison across all conditions.

No cross-model PCF condition. The planned cross-model PCF condition (using different models for each role) was invalidated by Gemini's structured-output failures and has been omitted from the final analysis.

7. Conclusion

We introduced the PCF benchmark for evaluating multi-agent scientific verification and identified conservatism-as-abstention as the dominant failure mode of adversarial self-verification pipelines. Our experiments demonstrate that PCF architectures do not improve discrimination between valid and invalid claims but instead induce excessive Uncertain verdicts or, model-dependently, excessive false positives. Iterative self-refinement consistently worsens verification accuracy, and symmetric debate provides limited benefit. These findings highlight that reliable AI scientific verification requires more than adversarial pipeline design: it requires calibrated verification mechanisms that account for model-specific tendencies, constrained abstention policies, and potentially grounding in external tools rather than relying solely on natural-language reasoning.

Impact Statement

This paper presents work whose goal is to advance the reliability of AI systems in scientific contexts. We specifically aim to reduce the risk of false scientific discoveries generated by autonomous AI systems. The broader impact of this work is aligned with improving scientific integrity. We note that our benchmark could be misused to optimize adversarial pipelines for generating claims that evade verification, though the current results suggest this risk is low given the failure modes we identify.

References

Bai, Y. et al. Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073, 2022.

Chan, C.-M., Chen, W., Su, Y., et al. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In ICLR, 2024.

Chaturvedi, S., Bergerson, J., and Mallick, T. Toward reliable, safe, and secure LLMs for scientific applications. arXiv preprint arXiv:2603.18235, 2026.

Chen, Z., Chen, S., Ning, Y., et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In ICLR, 2025.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325, 2023.

Gottweis, J. et al. Towards an AI co-scientist. arXiv preprint arXiv:2502.18864, 2025.

Gou, Z., Shao, Z., Gong, Y., et al. CRITIC: Large language models can self-correct with tool-interactive critiquing. In ICLR, 2024.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. In International Conference on Learning Representations (ICLR), 2024.

Huang, K., Jin, Y., Li, R., Li, M. Y., Candès, E., and Leskovec, J. Automated hypothesis validation with agentic sequential falsifications. In Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025.

Lan, T., Zhang, W., Lyu, C., et al. Training language models to critique with multi-agent feedback. arXiv preprint arXiv:2410.15287, 2024.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. In EMNLP, 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. 2024.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. In NeurIPS, 2023.

Min, S., Krishna, K., Lyu, X., Lewis, M., tau Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In EMNLP, 2023.

Mulian, H., Zeltyn, S., et al. AgentFixer: From failure detection to fix recommendations in LLM agentic systems. In ICSE Workshop on Agentic Engineering, 2026.

Ou, H., Chen, K., et al. DECEIVE-AFC: Adversarial claim attacks against LLM-based fact-checking. arXiv preprint arXiv:2602.02569, 2026.

Si, C., Yang, D., et al. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. arXiv preprint arXiv:2409.04109, 2024.

Tyen, G., Mansoor, H., Cărbune, V., Chen, P., and Mak, T. LLMs cannot find reasoning errors, but can correct them given the error location. In Findings of ACL, 2024.

Wadden, D., Lo, K., Wang, L. L., Lin, S., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or fiction: Verifying scientific claims. In Proceedings of EMNLP, 2020.

Zeng, Z., Zhang, J., et al. AI-for-Science low-code platform with Bayesian adversarial multi-agent framework. arXiv preprint arXiv:2603.03233, 2026.

A. Statistical Tests

Table 3 reports McNemar's test results comparing the PCF Same-Model condition against the Single-Agent baseline. Discordant pairs are reported as (PCF wrong | Single correct, PCF correct | Single wrong).

Table 3. McNemar's test comparing PCF Same-Model vs. Single-Agent. Significance: *** $p < 0.01$, ns = not significant.

Model	Domain	χ^2	p-value	Disc. (d, 0)	Sig.
Claude Sonnet 4	SciFact	2.25	0.134	(4, 0)	ns
Claude Sonnet 4	Synth. Stats	7.11	0.008	(9, 0)	***
Claude Sonnet 4	Novelty	1.33	0.248	(3, 0)	ns
GPT-4o	SciFact	14.06	0.0002	(16, 0)	***
GPT-4o	Synth. Stats	9.09	0.003	(11, 0)	***
GPT-4o	Novelty	0.00	1.000	(1, 1)	ns

In all significant comparisons, discordant pairs are exclusively in the direction of PCF degrading single-agent performance (column "d"), with zero instances of PCF correcting a single-agent error (column "0"). This one-sided pattern reinforces that the PCF pipeline introduces errors without compensating corrections.

B. GPT-5.4 Partial Results

Table 4 reports available results for GPT-5.4 on Synthetic Statistics (no debate condition was run). The key finding is the reversal of error direction: GPT-5.4's PCF pipeline is over-permissive rather than over-conservative.

Table 4. GPT-5.4 partial results on Synthetic Statistics ($n = 20$, GT: 15 Invalid, 5 Valid). No debate condition was run. Note the extreme FDR = 0.75 in the PCF condition.

Condition	Acc	FDR	TP	FP	FN
Single-Agent	0.75	0.00	5	0	5
Self-Refine	0.60	0.58	5	7	3
PCF Same	0.25	0.75	5	15	0

GPT-5.4's PCF Verifier classifies all 20 instances as Valid, producing 15 false positives. This is the opposite of the over-conservative pattern observed in Claude Sonnet 4 and GPT-4o, and demonstrates that the direction of PCF-induced bias is a property of the model, not the pipeline.

C. Gemini 2.0 Flash Detailed Results

Gemini 2.0 Flash returned Uncertain for 100% of SciFact and Synthetic Statistics instances across all conditions, resulting in accuracy of 0.00 (since all Uncertain verdicts are counted as incorrect) and undefined FDR/F1 metrics. Only the Novelty domain produced verdicts, where all instances were classified as Not_Novel across all four conditions, achieving 65% accuracy (matching the majority-class baseline of 13/20). No differentiation between pipeline conditions was observed.

D. Prompt Templates

We provide abbreviated versions of the prompts used for each role in the PCF pipeline. Full prompts, evaluation data, and analysis code will be released upon acceptance.

D.1. Proposer Prompt

```
You are a scientific claim evaluator. Given a claim and
supporting evidence, assess whether the evidence SUPPORTS
or CONTRADICTS the claim. Provide step-by-step reasoning,
then give your verdict as:
```

FINAL_VERDICT: [VALID/INVALID/UNCERTAIN]

D.2. Critic Prompt

You are an adversarial scientific critic. Given a claim, evidence, and a previous evaluator's assessment, identify ALL possible flaws, confounders, weaknesses, and alternative explanations. Be maximally skeptical. Your goal is to find the strongest possible objections.

D.3. Verifier/Falsifier Prompt

You are a scientific verifier. You have received:
1. A claim with evidence
2. An initial assessment (from the Proposer)
3. An adversarial critique (from the Critic)

Design the strongest possible falsification test for this claim. Then, weighing ALL evidence and objections, render your final verdict as:
FINAL_VERDICT: [VALID/INVALID/UNCERTAIN]

E. Experimental Configuration

All experiments were conducted via the respective model APIs. Hyperparameters were held constant across all conditions and models:

- Temperature: $\tau = 0.7$
- Maximum generation tokens: 1024
- Top-p: default (model-specific)
- Number of instances per domain per condition: 20

Model API identifiers:

- Claude Sonnet 4: claude-sonnet-4-20250514 (Anthropic)
- GPT-4o: gpt-4o-2024-08-06 (OpenAI)
- Gemini 2.0 Flash: gemini-2.0-flash (Google)
- GPT-5.4: OpenAI (specific version identifier withheld)
- Gemini 2.5 Flash: Google (specific version identifier withheld)