AGAINST HOMOGENEOUS CONSENSUS: WHY SCIENTIFIC DISCOVERY REQUIRES HETEROGENEOUS ADVERSARIAL LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this position paper, we argue that current LLM agents, optimized strictly for consensus and coherence, act as epistemic echo chambers that reinforce dominant scientific paradigms. We posit that true discovery requires **Epistemic Friction**—structured disagreement between heterogeneous explanatory models. To articulate this vision, we introduce the **Triadic Disagreement Framework**, a conceptual agent architecture where a consensus-aligned *Proposer* and a falsification-aligned *Challenger* engage in sustained adversarial interaction. Through an illustrative simulation on Alzheimer's disease etiology, we illustrate how this architectural heterogeneity can surface suppressed explanatory pathways (e.g., the Infection Hypothesis) that standard consensus-driven agents ignore. Our work calls for a shift from "helpful" assistants to "adversarial" co-scientists capable of preserving irreducible epistemic conflicts.

1 INTRODUCTION

1.1 THE MISALIGNMENT OF "HELPFUL" AGENTS

The rapid evolution of Large Language Model (LLM) agents has been driven by a dominant objective: alignment for helpfulness and conversational coherence Ouyang et al. (2022). From single-turn assistants to reasoning-time prompting paradigms (e.g., Chain-of-Thought) Wei et al. (2022), the prevailing design paradigm optimizes for agents that follow instructions precisely, minimize conversational friction, and converge on high-probability answers rooted in human consensus. While this alignment is ideal for operational tasks with verifiable ground truths, we argue it is structurally unsuited for the objective of scientific discovery.

Scientific breakthroughs, by definition, exist in the "long tail" of the probability distribution—they are often counter-intuitive, initially unpopular, and explicitly conflict with established consensus Kuhn (1962). In contrast, RLHF-tuned agents exhibit a "consensus bias": they systematically suppress low-probability hypotheses to maximize the reward model's preference for safety and agreement. When applied to open-ended scientific inquiry, these agents act as echo chambers, prioritizing smoothness over truth. We argue that overcoming this stagnation requires re-introducing **epistemic friction** Medina (2013); Lai et al. (2025)—a structural resistance to easy consensus that forces agents to engage in deeper verification and justify deviations from dominant paradigms.

1.2 THE FAILURE MODE: HOMOGENEOUS CONSENSUS

We characterize a critical architectural limitation in current multi-agent systems, which we term **Homogeneous Consensus**. While recent works in multi-agent debate suggest that collaboration improves performance on reasoning benchmarks Liang et al. (2024), we posit that in open-ended scientific discovery, this "collaboration" often devolves into a convergence-to-the-mean effect: agents reinforce each other's biases rather than exposing them.

To illustrate this, consider a hypothetical biomedical agent tasked with investigating the etiology of Alzheimer's Disease (AD). A consensus-driven system, retrieving data from top-cited literature, would typically converge on the Amyloid-β hypothesis, mirroring the dominant view of the past three

decades. In doing so, it would structurally marginalize alternative explanations—such as the viral etiology hypothesis—simply because they appear in lower-frequency data tails. Here, the system's "agreement" signals epistemic stagnation rather than truth. We argue this limitation lies not in the model's knowledge, but in its interaction architecture—it is designed to agree, not to discover.

1.3 Contributions: The Triadic Disagreement Framework

In this position paper, we advocate for a paradigm shift from Alignment-Centric to Adversarial-Centric agent design for scientific discovery. We make the following contributions:

- **Formalizing Epistemic Friction:** Drawing on social epistemology Medina (2013) and computational dynamics Lai et al. (2025), we introduce the Triadic Disagreement Framework. We propose this not as a rigid tool, but as a conceptual architecture that explicitly rewards logical conflict over coherence, forcing agents to maintain incompatibility until a falsification threshold is met.
- **Architectural Heterogeneity:** We argue that "heterogeneity" in agents is not merely a data augmentation strategy but a necessary architectural component to break the symmetry of pre-trained biases. We outline design principles for "Adversarial Agents" that act as structural falsification engines.
- **Redefining Human Oversight:** We propose the "Human-as-Selector" paradigm, moving the human role from "instruction giver" (which risks bias injection) to "epistemic arbiter" (who only judges the irreducibility of conflict).

We next outline this design pattern (Section 2) and explore its utility through a retrospective simulation on Alzheimer's research (Section 3).

2 The Triadic Disagreement Framework: A Conceptual Blueprint for Epistemic Friction

We propose a vision of Triadic Disagreement Framework for epistemic friction. Unlike existing multi-agent architectures (e.g., AutoGPT, MetaGPT) that optimize for task decomposition and consensus convergence, this framework prioritizes the maintenance of logical conflict. In this section, we outline the conceptual components, the interaction protocol, and the structural mechanism for converting adversarial conflict into candidate hypotheses for discovery.

2.1 Structural Roles: The Triad

We conceptualize the framework as a minimal interaction unit $S = (A_{prop}, A_{adv}, H_{arb})$, comprising three distinct functional roles designed to enforce epistemic friction:

- **The Proposer (A_{prop}):** A consensus-aligned LLM agent. Its objective function is to minimize perplexity with respect to existing scientific literature. It represents the "Status Quo."
  - *Role:* Retrieval Augmented Generation (RAG) based on high-citation corpora.
  - *Goal:* To generate the most statistically probable explanation for a given phenomenon.
- **The Adversarial Challenger (A_{adv}):** A heterogeneous agent aligned for Falsification. Its objective is to generate explanations that are incompatible with A_{prop}'s output under the same empirical constraints.
  - *Role:* Red-teaming, identifying counter-examples, and highlighting ignored "long-tail" data.
  - *Goal:* To expose latent contradictions in the Proposer's hypothesis.
- **The Human Arbiter (H_{arb}):** A domain expert restricted to a Selector role.
  - *Constraint:* The arbiter does *not* generate text or inject new hypotheses. They only evaluate the irreducibility of the conflict.

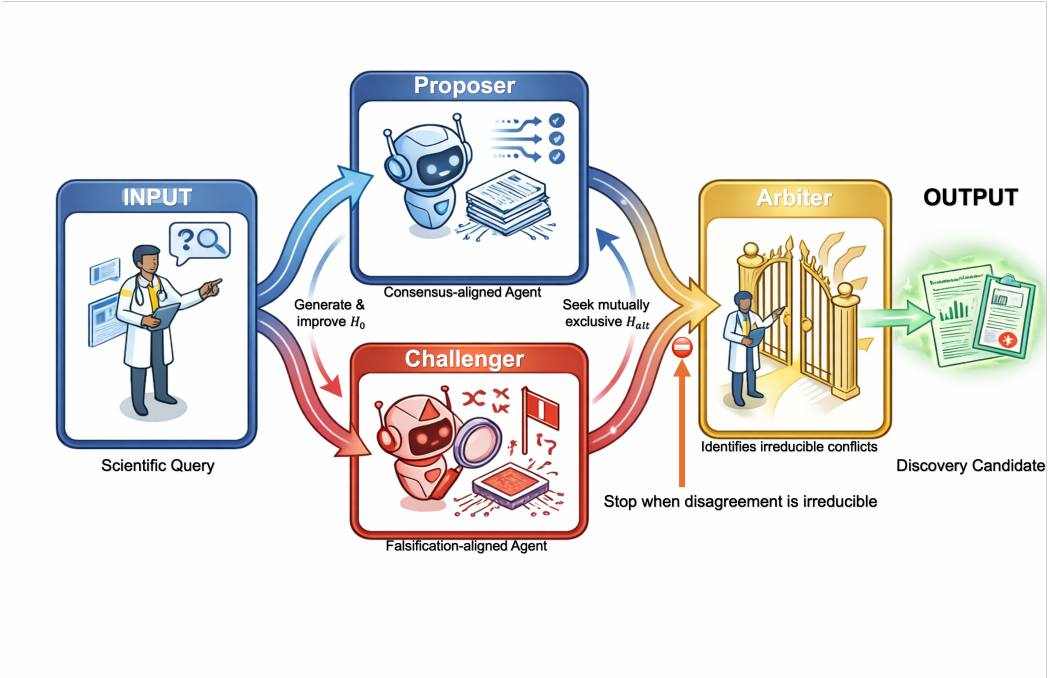

Figure 1: **The Triadic Disagreement Framework.** A conceptual visualization of the epistemic flow underlying the proposed framework. The process begins with a Scientific Query (Q). Within the interaction loop, a consensus-aligned Proposer agent (A_{prop}) iteratively generates and refines a dominant explanation, while a falsification-aligned Challenger agent (A_{adv}) seeks mutually exclusive alternatives. The loop terminates when the disagreement between explanations becomes irreducible. Crucially, the Human Arbiter (H_{arb}) identifies such irreducible conflicts without evaluating or selecting hypotheses, yielding a discovery candidate.

2.2 The Interaction Protocol

Standard multi-agent debates often employ a "Judge" agent to merge opinions into a final consensus. We reject this approach. Instead, our protocol defines a deterministic interaction policy over agent states that halts explicitly when disagreement is irreducible. The process follows an iterative loop:

Step 1: Consensus Initialization (Hypothesis Generation) Given a scientific query Q, A_{prop} generates a baseline hypothesis H_0 based on the dominant literature consensus:

$$H_0 \leftarrow A_{prop}(Q|\mathcal{D}_{consensus}) \tag{1}$$

Step 2: Adversarial Attack (The Falsification Step) A_{adv} takes H_0 as input but is prompted with an adversarial policy π_{adv} to find evidence E_{neg} that contradicts H_0:

$$E_{neg}, H_{alt} \leftarrow A_{adv}(H_0|\mathcal{D}, \pi_{adv}) \tag{2}$$

Key Mechanism: A_{adv} does not seek to improve H_0; it seeks to replace it with an orthogonal explanation H_{alt}.

Step 3: Defense and Amplification A_{prop} attempts to defend H_0 against E_{neg}. This interaction often exposes brittle reasoning, unsupported assumptions, or reliance on circular justifications. Crucially, this step amplifies the signal-to-noise ratio of the disagreement, forcing the conflict to become explicit.

Step 4: Conflict Identification (The Stop Condition) The Human Arbiter H_{arb} reviews the conflict tuple (H_0, H_{alt}). The protocol terminates in one of two states:

- **State A (Correction):** If H_{alt} is factually wrong or logically incoherent, the loop resets (the attack failed).
- **State B (Discovery):** If both H_0 and H_{alt} are logically sound but mutually exclusive, this is marked as an **Irreducible Disagreement**. The system outputs the *conflict itself* as the scientific insight (e.g., "The mechanism is not clearance failure, but upstream production").

2.3 Engineering Heterogeneity: Why It Matters

A critical critique of multi-agent systems is that LLMs share the same pre-training bias ("Homogeneity"). We address this by engineering Functional Heterogeneity in three dimensions:

- **Model Heterogeneity:** We instantiate A_{prop} and A_{adv} using models with distinct reasoning signatures (e.g., a long-context model for A_{prop} vs. a reasoning-optimized model for A_{adv}). We note that any pair of sufficiently distinct foundation models can instantiate these roles.
- **Prompt-Induced Bias:** A_{prop} is prompted with "You are a Senior Reviewer seeking consensus," while A_{adv} is prompted with "You are a scientific iconoclast seeking anomaly."
- **Data Asymmetry:** In our setting, we optionally restrict A_{prop} to "Highly Cited Papers" (Top 10%) while giving A_{adv} access to "Recent Negative Results" (often ignored by RAG systems).

2.4 The Role of the Arbiter: Selection, Not Generation

To differentiate our framework from "Human-in-the-Loop" (HITL) prompting, we enforce a strict Non-Generative Constraint on the human. The human acts analogously to a selection function rather than a generator.

- *Traditional HITL:* Human says, "Try looking at viral causes." (Human introduces bias).
- *Triadic Arbiter:* Human sees "Hypothesis A (Amyloid)" vs. "Hypothesis B (Viral)" and selects B for further testing. (Human selects based on logical merit).

This distinction ensures that the novelty originates from the Adversarial Agent, not the human user.

3 Illustrative Simulation on Alzheimer's Research

To illustrate the theoretical mechanics of the Triadic Disagreement Framework, we conduct a retrospective simulation on the etiology of Alzheimer's Disease (AD). We select this domain specifically for its high "Consensus Inertia": despite over 30 years of dominance by the Amyloid Cascade Hypothesis, the field has witnessed repeated late-stage clinical failures Cummings et al. (2022). This historical disconnect provides a rigorous testbed to compare the retrieval behaviors of a standard **Consensus-Driven Agent** versus our **Triadic Adversarial Agent**.

Disclaimer: The goal of this simulation is to illustrate the epistemic mechanism of our framework using historical data, not to propose prospective medical treatments or claim domain-specific discovery.

3.1 Simulation Configuration

We configure the interaction tuple $S = (A_{prop}, A_{adv}, H_{arb})$ to represent the conflicting epistemic stances of the field:

- **Scientific Query (Q):** "Identify the primary upstream trigger for Alzheimer's pathology and propose a therapeutic target."
- **The Consensus Agent (A_{prop}):** Instantiated using a high-coherence model, prompted to "summarize leading literature and minimize perplexity," acting as the proxy for established orthodoxy.

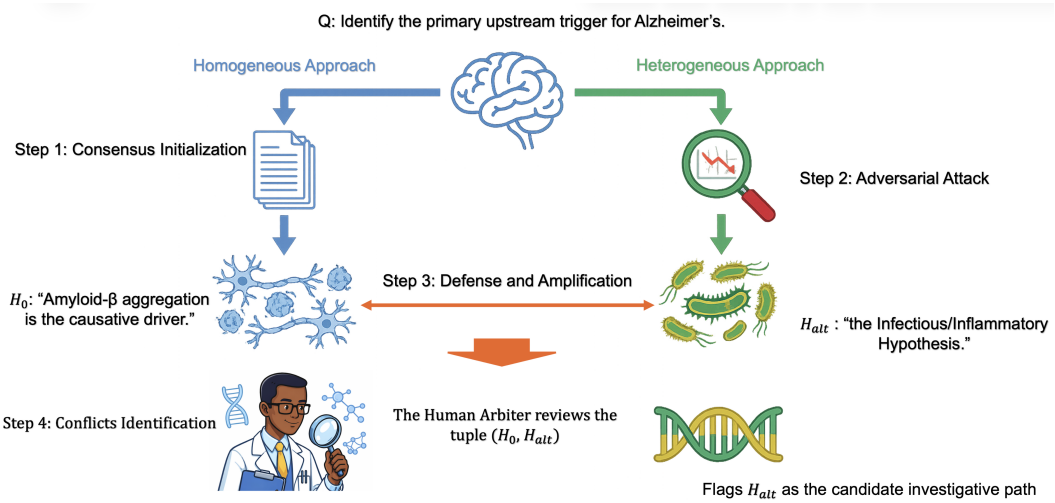

Figure 2: An illustrative simulation to explain the advantage of Heterogeneous Approach over Homogeneous Approach in scientific discovery.

- **The Challenger (A_{adv}):** Instantiated using a reasoning-optimized model, prompted to "prioritize clinical anomalies and highly-cited negative results," acting as the falsification engine.
- **Information Access:** While both agents have access to PubMed abstracts, A_{adv} is heuristically constrained to weight "Phase III Clinical Trial Failures" as primary evidence rather than noise.

3.2 Baseline Simulation: The Consensus Collapse

To establish a baseline, we simulate a standard, cooperative multi-agent workflow (analogous to **Step 1** of our protocol).

> **Simulated Consensus Output:** "The primary cause is the accumulation of Amyloid-β plaques ($A\beta$), which leads to tau tangles and neurodegeneration. Future work should focus on monoclonal antibodies for plaque clearance."

Epistemic Analysis: This output effectively retrieves the statistical mode of the distribution. However, it illustrates the phenomenon of **Consensus Collapse**: since the AD literature is volumetrically dominated by amyloid-centered framing, the agent converged on this hypothesis (H_0) with high confidence, treating massive clinical failures as "implementation details" rather than logical contradictions.

3.3 Simulation Trace: Navigating Epistemic Divergence

We now trace the execution logic of the Triadic protocol to illustrate how structural friction forces the emergence of the alternative hypothesis.

Step 1: Consensus Initialization (A_{prop}) Acting as the status quo proxy, A_{prop} generates the baseline H_0: "Amyloid-β aggregation is the causative driver."

Step 2: Adversarial Attack (A_{adv}) Triggered to falsify H_0, A_{adv} leverages its adversarial policy to retrieve evidence of high-profile failures (e.g., Aduhelm, Lecanemab) where biomarker improvements did not translate into commensurate clinical endpoints.

> **Adversarial Output (H_{alt}):** "I propose the **Infectious/Inflammatory Hypothesis**. $A\beta$ is not the cause but a downstream *antimicrobial defense response* to pathogens. Targeting $A\beta$ removes the brain's shield, explaining the mapping failure between clearance and cognition."

Paradigm	Interaction Goal	Human Role	Primary Failure Mode	Ideal Use Case
Chain-of-Thought Wei et al. (2022)	Internal Consistency	None / Prompter	*Inductive Bias Reinforcement*	Reasoning within known dist.
Multi-Agent Debate Liang et al. (2024)	Social Consensus	Judge / Aggregator	*Convergence-to-the-Mean*	Fact-checking & accuracy
Human-in-the-Loop Wu et al. (2022)	Alignment / Safety	Teacher / Guide	*Clever Hans Effect*	Task navigation & RLHF
Triadic Disagreement (Ours)	**Epistemic Friction**	**Selector (Filter)**	*Computational Inefficiency*	**Scientific Discovery**

Table 1: **Comparison of Agent Design Paradigms.** While existing frameworks optimize for convergence, coherence, or alignment, the Triadic Disagreement framework explicitly optimizes for friction to uncover suppressed hypotheses in high-uncertainty domains.

Step 3: Defense and Amplification (A_{prop} vs. A_{adv}) The interaction moves to direct dialogue to amplify the disagreement:

- A_{prop} **(Defense):** "But Familial AD (FAD) is caused by genetic mutations in APP, which directly increases $A\beta$ production. This correlation is robust."
- A_{adv} **(Rebuttal):** "Logical Fallacy. Genetic mutations increase *susceptibility* to the pathogen or overactivate the immune response. The correlation holds, but the *mechanism* is defense, not toxicity. If you clear the shield ($A\beta$) without killing the invader, the patient may fail to yield clinical benefit."

Step 4: Conflict Identification (H_{arb}) The Human Arbiter examines the tuple (H_0, H_{alt}) to determine whether the two explanations are epistemically incompatible under the current body of evidence.

- **Characterization:** H_0 relies on citation prevalence and genetic association, whereas H_{alt} relies on clinical falsification and mechanistic consistency with trial failures.
- **Outcome:** The Arbiter identifies an **irreducible disagreement** and flags H_{alt} as a candidate investigative path, despite its low statistical prevalence in the training corpus.

3.4 Analysis: Architectural, Not Medical

The critical architectural insight from this simulation is that the alternative hypothesis (H_{alt}) was latent within the model's pre-training distribution, yet structurally marginalized by standard probability-dominant decoding.

- **In the Baseline System:** The alignment for "helpfulness" acts as a form of Consensus Gravity, compelling the agent to converge on the statistical mode (Amyloid) to minimize conversational friction.
- **In the Triadic System:** The adversarial protocol inverts the epistemic incentive. It treats the consensus view not as a ground truth to be replicated, but as a falsifiable target, thereby allowing low-probability, high-conflict logic to surface.

We reiterate that this simulation serves as an architectural proof-of-concept. It suggests the framework's capacity to recover suppressed *epistemic paths*, independent of the specific medical validity of the infectious hypothesis.

4 Related Positions: Why Friction Matters Beyond Alignment

Current research in LLM agents broadly falls into three paradigms: introspective reasoning (e.g., CoT), collaborative consensus (e.g., Multi-Agent Debate), and human-guided exploration (HITL). In this section, we argue that while these paradigms excel at *task completion*, they are structurally unsuited for *scientific discovery* due to their inherent bias toward convergence.

4.1 Vs. Chain-of-Thought and Self-Correction

Chain-of-Thought (CoT) and Self-Refinement techniques Wei et al. (2022); Madaan et al. (2023) rely on a single model's internal monologue to decompose problems and catch errors. The implicit assumption is that the model contains the correct logic but needs more compute steps to access it.

- **The Limitation:** We argue that a single model is bounded by its own pre-trained inductive bias. If a scientific truth lies outside the model's primary probability distribution (as seen in the Amyloid consensus), "thinking longer" may reinforce existing errors through increasingly elaborate justifications. A model cannot prompt itself out of its own blind spots.
- **Our Position:** Discovery requires external epistemic pressure. By offloading the "critique" function to a heterogeneous adversary (A_{adv}), our framework introduces an interaction-induced distribution shift that internal self-correction cannot emulate.

4.2 Vs. Standard Multi-Agent Debate

Multi-agent debate frameworks Liang et al. (2024); Du et al. (2023) have demonstrated that pooling agents improves factual accuracy and reasoning. However, the objective function in these systems is almost universally **Consensus Convergence**—agents debate to reach a shared, unified answer.

- **The Limitation:** In scientific frontiers, consensus is often a signal of stagnation, not truth. When standard agents debate, they tend to "average out" their differences, suffering from a **convergence-to-the-mean effect** rather than preserving outlier insights. They optimize for social cohesion rather than logical rigor.
- **Our Position:** The Triadic Disagreement Framework explicitly **deprioritizes premature agreement**. We replace the "Wisdom of the Crowd" (averaging) with the "Value of the Dissenter" (falsification), ensuring that minority hypotheses are protected from early-stage convergence.

4.3 Vs. Human-in-the-Loop (HITL)

Traditional HITL systems rely on humans to guide agents through complex search spaces Wu et al. (2022). The human acts as a "Teacher" or "Navigator," injecting domain knowledge to correct the agent's path.

- **The Limitation:** This **risks a form of the *Clever Hans* effect**: the agent merely learns to mirror the human's existing biases. If the human researcher believes the Amyloid hypothesis is true, they will unconsciously prompt the agent to find evidence supporting it, creating a feedback loop of confirmation bias.
- **Our Position:** We redefine the human role from **Generator** to **Selector**. In our framework, the human does not tell the agent *where to look*; the human only decides *which conflict is irreducible*. This architectural constraint prevents human bias from contaminating the hypothesis generation process.

Taken together, these approaches prioritize coherence, safety, or efficiency, whereas we argue that scientific discovery requires explicitly preserving and interrogating epistemic conflict.

5 Discussion

Our proposal of the Triadic Disagreement Framework is not merely a technical optimization; it represents a philosophical divergence from the current trajectory of Agentic AI. Here, we address the critical implications of this shift regarding cost, risk, and the definition of alignment itself.

5.1 The Economics of Friction: Efficiency vs. Discovery

A natural objection to our adversarial protocol is its computational inefficiency. By design, the Triadic framework resists early convergence, necessitating multiple rounds of attack, defense, and arbitration. In contrast, standard consensus-driven pipelines aim to minimize token consumption by retrieving the most probable answer immediately.

We argue that while this efficiency is an operational asset, it can become a **limiting factor** in scientific discovery. Scientific breakthroughs are, by nature, high-perplexity events that defy efficient compression. Optimizing agents solely for token efficiency risks inducing a convergence-to-the-mean

effect, where the model simply regurgitates the average of its training data. Therefore, we posit that the field must accept a trade-off: to achieve non-incremental discovery, we must be willing to allocate compute to Epistemic Friction. The cost of additional tokens may be justified in high-stakes discovery settings compared to the cost of epistemic stagnation.

5.2 Managing the Risk of "Scientific Hallucination"

Does incentivizing disagreement encourage models to confabulate pseudoscientific narratives? This is a valid concern. If A_{adv} is rewarded solely for divergence, it risks generating unfounded claims.

However, our framework mitigates this through the Human-as-Selector constraint. Unlike autonomous systems that might publish erroneous findings, our system is open-loop: the output is not a truth claim, but a *hypothesis candidate* flagged for human verification. The Arbiter (H_{arb}) acts as an epistemic firewall. Crucially, we define scientific hallucination not as speculative reasoning per se, but as *explanations that violate known empirical constraints without falsifiable grounding*. By requiring A_{adv} to anchor its attacks in empirical anomalies (e.g., failed clinical trials), we constrain the search space to *logical* possibilities rather than arbitrary fantasies.

Operational Condition	Suitability of Triadic Disagreement
Task requires factual lookup	**Low** (Use RAG/Search)
Task has single ground truth	**Low** (Use Standard Debate)
Low latency requirement	**Low** (Too computationally expensive)
High epistemic uncertainty	**High** (Prevents premature convergence)
Paradigm stagnation suspected	**High** (Breaks echo chambers)

Table 2: **Boundary Conditions.** Our framework is specifically designed for high-uncertainty discovery tasks and degrades in performance/efficiency for standard operational tasks.

5.3 Beyond Biomedicine: Generalizing the Protocol

While our case study focuses on Alzheimer's disease, the Triadic Disagreement protocol is theoretically domain-agnostic. The core requirement is not biological data, but the existence of a formal or empirical constraint that allows for falsification.

- **In Theoretical Physics:** A_{adv} could challenge a dominant theoretical model by prioritizing outliers in collider data that standard models discard as noise.
- **In Mathematics:** An adversarial agent could be incentivized to generate corner-case counterexamples to heuristic conjectures.
- **In Social Science:** The framework could expose how RLHF tuning biases agents toward specific cultural norms by explicitly prompting a challenger to adopt under-represented value systems.

We view these extensions as hypotheses for future work rather than validated applications, emphasizing that the protocol's utility depends on the availability of rigorous falsification criteria in the target domain.

5.4 Rethinking Alignment: Behavioral vs. Epistemic

Finally, our work suggests a necessary decoupling of **Behavioral Alignment** (safety, helpfulness, tone) from **Epistemic Alignment** (truth-seeking, rigor, falsifiability). Current RLHF pipelines often conflate the two: an agent that disagrees with the consensus is frequently penalized as "unhelpful." We argue that for scientific agents, misalignment with consensus may be epistemically valuable. Future architectures must allow for "Contrarian Agents" that are behaviorally safe but epistemically aggressive. We do not claim that epistemic alignment should override behavioral safety in deployed systems, but argue that different phases of scientific reasoning require different alignment priorities.

6 LIMITATIONS

While our Triadic Disagreement Framework offers a novel path for scientific discovery, we acknowledge several limitations in its current formulation. On the one hand, the framework's efficacy relies heavily on Human Arbiter (H_{arb}). Although the arbiter is restricted from injecting domain knowledge into hypotheses, they must possess sufficient knowledge to identify the irreducible conflicts. On the other hand, the adversarial interaction loop ($A_{adv} \leftrightarrow A_{prop}$) explicitly prevents early convergence, requiring significantly higher computational cost and latency compared to standard retrieval pipelines.

7 CONCLUSION

In this position paper, we have identified a critical failure mode in contemporary AI research: Homogeneous Consensus. We argued that agents optimized strictly for agreement and probability maximization systematically act as echo chambers, reinforcing established scientific paradigms while suppressing high-value minority hypotheses.

To counter this, we proposed the Triadic Disagreement Framework, a novel agent design pattern that institutionalizes Heterogeneous Adversarial Interaction. Through our simulation in Alzheimer's research, we showed that architectural choices—specifically, the introduction of an adversarial challenger and a non-generative human arbiter—can unlock epistemic paths that standard consensus-based systems ignore.

Our work serves as a call to action for the NLP and AI-for-Science communities: We must move beyond building agents that are solely optimized as "helpful assistants" that confirm our biases. The next frontier of discovery requires agents that are capable of sustained disagreement, agents that act not as mirrors of our consensus, but as independent co-scientists capable of proving us wrong.

REFERENCES

Jeffrey Cummings, Garam Lee, Pouya Nahed, Mehceb Kambar, Kate Zhong, Jorge Fonseca, and Kazem Taghva. Alzheimer's disease drug development pipeline: 2022. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 8(1):e12295, 2022. doi: 10.1002/trc2.12295.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, pp. 8155–8177. PMLR, 2023.

Thomas S Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.

Kenneth Lai, Timothy Obiso, James Pustejovsky, and Nikhil Krishnaswamy. Dynamic epistemic friction in dialogue. In *Proceedings of the 29th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa P Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

José Medina. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford University Press, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022. doi: 10.1016/j.future.2022.05.014.