InfAL: Inference Time Adversarial Learning for Improving Research Ideation

Sikun Guo, Amir Hassan Shariatmadari, Peng Wang, Albert Huang, Aidong Zhang
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
{qkm6sq, ahs5ce, pw7nc, kfa7fg, aidong}@virginia.edu

Abstract

Advancements in Large Language Models (LLMs) have opened new opportunities for scientific discovery by assisting researchers in generating novel hypotheses and ideas. In this process, a major challenge is how to optimally and efficiently utilize LLMs' parametric knowledge obtained from their pretraining process. Inspired by Generative Adversarial Networks (GANs), we propose inference time adversarial learning (termed InfAL), implemented through multi-LLM-agent interactions, to enhance research ideation. This approach optimizes the utilization of LLMs' parametric knowledge without requiring additional model training, making adversarial learning efficient and context-driven. To evaluate the quality of generated ideas, we propose a relative quality ranking metric as a scalable alternative to human evaluation. Our results show that InfAL significantly improves idea generation, with GPT-4o achieving a 21% increase in novelty and a 322% increase in feasibility, demonstrating its transformative potential for driving innovation in scientific research.

1 Introduction

Recent breakthroughs in large language models (LLMs) have revealed their remarkable potential to advance scientific research, particularly through automated hypothesis generation. Studies like Si et al. (2024); Guo et al. (2025a, 2024) demonstrate that LLM-generated research ideas can approach human-level creativity, fueling significant academic interest. However, despite extensive exploration into leveraging LLMs—such as prompt engineering, supervised fine-tuning, and retrieval-augmented generation (Wang et al., 2023d; Baek et al., 2024; Yang et al., 2023; Zhou et al., 2024; Boiko et al., 2023; Guo et al., 2025b; Xiong et al., 2025; Hu et al., 2024)—a critical challenge remains: effectively optimizing the immense but underutilized parametric knowledge within LLMs at inference time without costly retraining.

The parametric knowledge of an LLM refers to internal representations and information implicitly encoded in its learned parameters during pretraining. For example, parametric knowledge allows an LLM to implicitly understand fundamental concepts like gravity or biological evolution. However, tapping into precisely the right portion of this vast parametric knowledge to generate optimal answers for user queries is inherently challenging due to its implicit, distributed nature. Current methods often rely heavily on extensive external knowledge bases or costly retraining, rendering them inefficient or infeasible for on-the-fly optimization.

To address this gap, we propose Inference Time Adversarial Learning (InfAL), a novel framework inspired by Generative Adversarial Networks (GANs) (Goodfellow et al., 2020). InfAL directly optimizes parametric knowledge utilization during inference, enabling LLMs to produce high-quality outputs without additional training or labeled data. Unlike prompt optimization methods such as adv-ICL (Do et al., 2023), which narrowly focus on prompts for in-context learning, InfAL represents a universal, more powerful paradigm capable of enhancing end-to-end performance across diverse downstream tasks.

We implement InfAL via a lightweight multi-LLM-agent system comprising generator, optimizer, and discriminator agents that iteratively refine generated ideas. To facilitate scalable and fair evaluation of open-ended ideas, we introduce a relative quality ranking metric (denoted as Q) that closely aligns with human judgment. Experiments demonstrate InfAL's substantial impact: integrating InfAL with GPT-4o improves the novelty and feasibility of generated ideas by 21% and 322%, respectively.

In summary, our contributions are twofold:

- We formulate a novel inference-time adversarial learning framework, InfAL, leverag-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12501–12522
November 4-9, 2025

ing GAN principles to efficiently optimize parametric knowledge utilization of LLMs at inference-time, significantly enhancing their generative capabilities.

- We develop a customizable, scalable relative quality ranking metric (Q) to evaluate open-ended generation tasks, strongly correlating with human evaluation and serving as an effective, practical proxy to human evaluation.

2 Related Work

Inference-Time In-context Learning. Recent studies explored LLM inference-time learning via implicit Bayesian inference (Xie et al., 2021), "induction heads" in transformers (Olsson et al., 2022), and attention-based implicit meta-optimization (Dai et al., 2022). TextGrad (Yuksekgonul et al., 2024) introduced textual gradients for inference-time optimization.

Scientific Hypothesis Generation. LLMs have advanced scientific ideation beyond traditional methods (Wang et al., 2023c). SciMON (Wang et al., 2023d) employs retrieval-augmented generation; MOOSE (Yang et al., 2023) uses multi-level self-feedback. ResearchAgent (Baek et al., 2024) and AI Scientist (Lu et al., 2024a) automate entire research cycles, integrating knowledge graphs and agentic tree search. Iterative prompting methods (Zhou et al., 2024) and tool-equipped models like Coscientist (Boiko et al., 2023) further enhance scientific workflows.

Multi-LLM-Agent Interactions. Multi-agent frameworks address single-LLM limitations through structured discussions (Lu et al., 2024b), collaborations (Chih-Yao Chen et al., 2024; Zhang et al., 2023), and debates (Du et al., 2023; Liang et al., 2023). These interactions improve task consistency (Xiong et al., 2023), evaluation (Chan et al., 2023; Wang et al., 2023a), and supervision (Khan et al., 2024). Such frameworks effectively enhance tasks including narrative generation (Huot et al., 2024), translation (Liang et al., 2023), and knowledge reasoning (Wang et al., 2023b; Ma et al., 2024). Our work extends these frameworks to theoretically and empirically explore inference-time adversarial learning for research ideation.

3 Method

In this section, we first present the theoretical framework for inference time adversarial learning, aiming at optimizing the utilization of LLMs' parametric knowledge to perform user-specified tasks. In addition, we describe how this can be implemented through LLM-based agent interactions to enhance the generation of research ideas. Following this, we introduce a relative quality ranking-based metric designed to approximate human evaluation of the generated ideas.

3.1 Formulation of Inference Time Adversarial Learning

The goal of inference-time adversarial learning is to optimize the utilization of LLMs' parametric knowledge, enabling them to generate the best possible response given the limited context provided by the user's query. To achieve this, we represent the parametric knowledge as θ, denote the LLM's parametric knowledge base as $\{\theta\}$, which includes all possible parametric knowledge, and begin our formulation with Assumption 1:

Assumption 1. *We assume that given any user query x, there exists a static optimal answer $\dot{y}$, although the LLM may not explicitly generate $\dot{y}$ due to the discrete nature of its parametric knowledge base $\{\theta\}$.*

The parametric knowledge base $\{\theta\}$ of an LLM is obtained during the pre-training process of the given LLM and plays a crucial role in answering any user's query x. LLMs utilize parametric knowledge θ in their parametric knowledge base $\{\theta\}$ in responding to user queries.

Definition 1 (Parametric Knowledge Utilization)**.** *We define parametric knowledge utilization as a mapping function U that transforms the parametric knowledge base $\{\theta\}$ of the given LLM into the utilized parametric knowledge u given user query x:*

$$u = U(\{\theta\}, x) \tag{1}$$

where U encapsulates the mechanism through which the LLM retrieves and applies its parametric knowledge according to the user query.

As the oracle answer $\dot{y}$ may not be directly achievable, the objective shifts to generating an approximation answer $\hat{y}$ which is sufficiently close to $\dot{y}$. Thus, we have Assumption 2:

Assumption 2. *Given a user query x, if an LLM generates a $\hat{y}$ from its parametric knowledge base*

$\{\theta\}$ with utilization u, and $\hat{y}$ lies in the neighborhood B of $\dot{y}$ with radius ϵ, that is, $\hat{y} \in B_\epsilon(\dot{y})$, we posit that the LLM has optimized the use of its parametric knowledge in response to the user's query x, yeilding suboptimal answer $\hat{y}$.

To optimize the generation of $\hat{y}$, we formulate the objective inspired by Generative Adversarial Networks (GANs) (Goodfellow et al., 2020). Similar to GANs, the objective of inference time adversarial learning is framed as a minimax game between two models: a Generator G and a Discriminator D. The Generator's goal is to generate an answer $\hat{y}$ to approach $B_\epsilon(\dot{y})$, while the Discriminator is tasked with determining whether $\hat{y}$ belongs to $B_\epsilon(\dot{y})$. Therefore, the objective function can be defined as follows:

$$
\begin{aligned}
V(G,D) = \min_G \max_D \Big[& \mathbb{E}_{\dot{y}\in B_\epsilon(\dot{y})}[\log D(\dot{y})] \\
& + \mathbb{E}_{x\sim p_x(x)}[\log(1 - D(G(x)))]\Big] \\
\text{s.t.} \begin{cases} u_D^* = \arg\max_{u_D} \Big[\mathbb{E}_{\dot{y}\in B_\epsilon(\dot{y})}[\log D(\dot{y})] \\ \qquad + \mathbb{E}_{x\sim p_x(x)}[\log(1 - D(G(x)))]\Big], \\ u_G^* = \arg\min_{u_G} \mathbb{E}_{\dot{y}\in B_\epsilon(\dot{y})}[\log(1 - D(G(x)))]. \end{cases}
\end{aligned} \tag{2}
$$

where:

- $\mathbb{E}_{\dot{y}\in B_\epsilon(\dot{y})}[\log D(\dot{y})]$ represents the expected log-probability that the Discriminator assigns to the optimal answer $\dot{y}$, with the goal of maximizing this term so that the Discriminator can correctly reject any approximation $\hat{y}$ in the $B_\epsilon(\dot{y})$.

- $\mathbb{E}_{x\sim p_x(x)}[\log(1 - D(G(x)))]$ represents the expected log-probability that the Discriminator assigns to generated answer $G(x)$, where $G(x) = \hat{y}$, and x is a user query sampled from the user query distribution $p_x(x)$. The Generator aims to minimize this term, trying to convince the Discriminator to accept $\hat{y} \in B_\epsilon(\dot{y})$.

During this adversarial process, the Generator aims to minimize $\log(1 - D(G(x)))$, meaning it tries to convince the Discriminator to accept $\hat{y} \in B_\epsilon(\dot{y})$. Conversely, the Discriminator aims to maximize both $\log D(\dot{y})$ for the optimal answer $\dot{y}$ and $\log(1 - D(G(x)))$ for the generated answer $\hat{y}$. According to Proposition 2 in GANs (Goodfellow et al., 2020), if G and D have enough capacity, during the optimization process, $\hat{y}$ converges to $\dot{y}$. According to Theorem 1 in GANs (Goodfellow et al., 2020), the global minimum of the objective function is reached if and only if $\hat{y} = \dot{y}$. Though in open-ended generation tasks for LLMs it's challenging to generate $\hat{y} = \dot{y}$, achieving $\hat{y} \in B_\epsilon(\dot{y})$ remains plausible and practical. Note that ϵ is likely to vary from model to model.

The objective function defined in Formula (2) is designed to be optimized at inference time, eliminating the need for any updates to the model parameters during the process. This optimization is achieved by forcing D and G to search in their parametric knowledge base $\{\theta_D\}$ and $\{\theta_G\}$ to obtain the best utilization u_D^* and u_G^*, respectively.

3.2 Inference Time Adversarial Learning Through Multi-LLM Agents' Interactions

To implement inference time adversarial learning for research idea generation and refinement, we employ a multi-agent interaction system using LLMs. There are three agents in the system, each agent plays a unique role in the objective function and will be introduced in the following subsections. The overview of this system for research idea refinement is shown in Figure 1. In general, once the user provides a context x, the generator agent acts as the G to generate and refine idea $\hat{y}$, the optimizer agent serves as the optimizer by providing the gradient r, and the discriminator agent functions as the D in the objective function, continuing this process until the minimax game converges to equilibrium. For simplicity, the minimax game is shown to converge at the 4th iteration in Figure 1; however, in practice, the steps illustrated in iterations 2 and 3 may repeat multiple times, and additional iterations may be required for the minimax game to fully converge.

3.2.1 Research Idea Generator

The research idea generator agent acts as the G in the objective function. Its role is to generate and iteratively refine the research idea $\hat{y}$, striving to approach the optimal idea $\dot{y}$. At the beginning of the minimax game, the generator agent is profiled as a domain expert researcher and generates an initial idea $\hat{y}_0$ based on the user query x. In subsequent iterations of the minimax game, the generator agent is tasked with refining the idea based on feedback from the optimizer agent. Therefore, at the i-th iteration, the generator agent updates $\hat{y}$ via:

$$
\begin{cases} u_{i,G} = u_{i-1,G} - \eta_G r_i \\ \hat{y}_i = G(\hat{y}_{i-1}; u_{i,G}) \end{cases} \tag{3}
$$

where $r_i = \nabla_u V(G, D; u_i)$ is the "textual gradient" for updating the parametric knowledge uti-

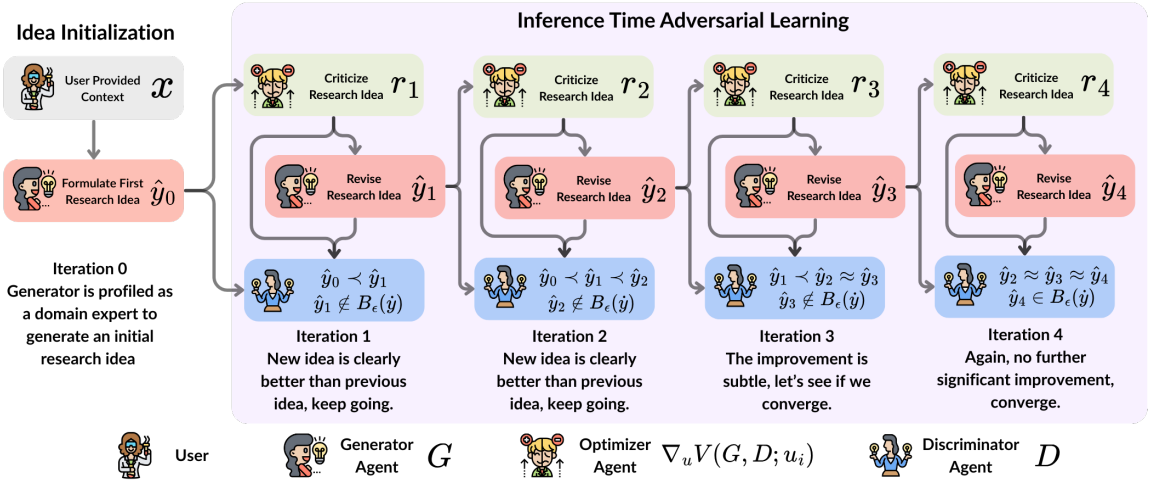

Figure 1: The overview of inference time adversarial learning via LLM-based agent interactions for enhancing research ideation.

lization to refine $\hat{y}$, which is provided by the research idea optimizer agent through its feedback and will be introduced in 3.2.2 with more details. The learning rate η_G is dynamically and implicitly determined by the generator G. We demonstrate all the prompt templates for research idea generator agent in Fig. 6, Fig. 9, and Fig. 10 in the Appendix.

3.2.2 Research Idea Optimizer

The research idea optimizer agent provides feedback r on the generator's idea $\hat{y}$ in the form of a "textual gradient" $\nabla_u V(G, D; u_i)$, which guides both the generator agent in refining its u_G to produce better ideas and the discriminator agent in updating its u_D to assess whether significant improvements have been made. Unlike traditional numerical gradients, textual gradients are expressed in natural language, making them more interpretable while still serving a similar optimization function in downstream tasks (Yuksekgonul et al., 2024). At the beginning of the minimax game, the optimizer agent is also profiled as a domain expert researcher, but its primary role is to critique ideas rather than generate them. At each iteration i, the optimizer evaluates the current idea based on user-specified quality indicators, such as novelty or feasibility, and provides constructive feedback. This feedback not only helps the generator refine its ideas but also aids the discriminator in determining whether meaningful improvements have been achieved. The prompt templates for the research idea optimizer agent are provided in Figures 7 and 11 in the Appendix.

3.2.3 Research Idea Discriminator

The discriminator agent functions as D in the minimax game and is initially profiled as a domain expert researcher. At each iteration i, it evaluates the improvement between the current idea $\hat{y}_i$ and the previous idea $\hat{y}_{i-1}$ based on the optimizer's feedback r_i. The discriminator updates its evaluation e_i as follows:

$$\begin{cases} u_{i,D} = u_{i-1,D} + \eta_D r_i \\ e_i = D(\hat{y}_{i-1}; \hat{y}_i; u_{i,D}) \end{cases} \tag{4}$$

where e_i represents the discriminator's evaluation. The feedback r_i provides directional guidance for idea refinement, helping the discriminator agent update its parametric utilization $u_{i,D}$ by incorporating information on how the idea has possibly improved. This enables the discriminator to more effectively assess whether $\hat{y}_i$ exhibits meaningful progress over $\hat{y}_{i-1}$.

There are two possible outcomes for e_i:

1. $\hat{y}_{i-1} \prec \hat{y}_i$, meaning the discriminator identifies significant improvements.
2. $\hat{y}_{i-1} \approx \hat{y}_i$, meaning no substantial improvement is detected.

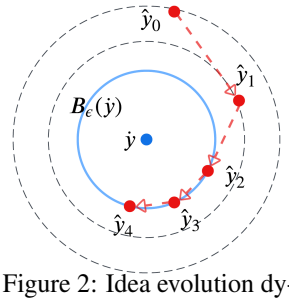

Figure 2: Idea evolution dynamics.

As illustrated in Figure 1, if $\hat{y}_{i-1} \prec \hat{y}_i$, further refinement may still be possible. However, if the discriminator repeatedly determines no further improvements ($\hat{y}_{i-2} \approx \hat{y}_{i-1} \approx \hat{y}_i$), we conclude that the process has converged, implying $\hat{y}_{i-2}, \hat{y}_{i-1}, \hat{y}_i \in B_\epsilon(\dot{y})$, where ϵ represents the implicit distance to $\dot{y}$. In this case, $\hat{y}_i$ is selected as

the final suboptimal research idea. The prompt templates for the discriminator agent are provided in Figures 8 and 12 in Appendix A.2.1.

Figure 2 illustrates the evolution of the generated research idea $\hat{y}$. Initially, the discriminator detects significant improvements between $\hat{y}_1$ and $\hat{y}_2$, prompting further refinements. From $\hat{y}_2$ to $\hat{y}_4$, no further improvements are identified, leading to convergence. While Figure 2 shows convergence at the 4th iteration for illustration, in practice, more iterations may be required.

3.3 Relative Quality Ranking

Although human judgment remains the gold standard for evaluating open-ended text generation, the Natural Language Processing community has been actively developing scalable alternatives to approximate human evaluation, as the labor involved in human evaluation is often costly and impractical in many cases. Recent studies have explored the use of LLMs as autoraters (Chiang and Lee, 2023; Liu et al., 2023; Bubeck et al., 2023; Fu et al., 2024; Vu et al., 2024; Gu and Krenn, 2024). These studies show that the correlation between human evaluators and LLM autoraters positions LLMs as a promising alternative for large-scale assessments for open-ended generation. To automate the evaluation of the generated ideas, we develop a relative ranking-based metric designed to assess idea quality in a fair and customizable manner. This metric can be customized to accommodate various quality indicators, such as novelty, feasibility, or any other criteria specified by the user, as long as a target research idea and the context used to generate this target idea are available. The target idea can either be generated by the user or selected from existing literature. Compared to the winrate (Zheng et al., 2023), our metric offers a more granular measurement. Please refer to Section A.3.1 in the Appendix for more discussions.

For a given context used to generate a set of research ideas, any high-capcity LLM can be used to rank the quality of all the ideas (both generated ideas and the target idea) based on user-specified quality indicators, without revealing which idea is the target research idea (Gu and Krenn, 2024). The used LLM is prompted to assess the ideas based on its understanding of quality indicators such as novelty and feasibility, and then rank them accordingly.

The prompt template used to achieve this is shown in Figure 13 in the Appendix. The position of the target research idea within the ranked list of ideas reflects the quality of the generated ideas with respect to the specified quality indicators. Intuitively, if the target idea ranks higher on the list, this suggests that the generated ideas are of lower quality compared to the target idea. Conversely, if the generated ideas rank higher than the target idea, it indicates that the generated ideas may be of better quality. Given a target idea t and n generated ideas based on the given context, let n_t denote the rank of t among the target idea and the generated ideas. The relative quality ranking Q of the generated ideas is computed as follows:

$$Q = \frac{n_t - 1}{n}. \tag{5}$$

Intuitively, $Q \in [0, 1]$. If the target idea ranks first on the list, then $n_t = 1$, yielding $Q = 0$, which indicates that all generated ideas are worse than the target idea. Conversely, if the target idea ranks below all the generated ideas, that is, $n_t = n + 1$, then $Q = 1$, indicating that all generated ideas are superior to the target idea. To ensure fair comparison across different idea generation strategies, it is important to generate the same number of research ideas n for all compared strategies.

4 Experiments

The primary objective of our experiments is fourfold: (1) to assess whether InfAL enhances the quality of research ideas generated by LLMs, (2) to examine how research ideas evolve and converge during the process of InfAL, (3) To assess InfAL's each component's contribution to overall performance, and (4) whether relative quality ranking given by GPT-4o aligns with human judgment.

4.1 Datasets

To evaluate the effectiveness of InfAL in enhancing research idea generation, we constructed five datasets. The five datasets collectively include 500 target papers, representing high-quality research published in 2024 across the following five domains: Health and Medicine (354 papers), Genetics and Molecular Biology (153 papers), Environmental Sciences (131 papers), Neuroscience and Cognitive Sciences (127 papers), and Technology and Engineering (135 papers). Each target paper's research idea, denoted as t_i, serves as the "gold standard" for comparing generated ideas. For each target paper, we collected background information,

consisting of abstracts from its reference papers, denoted as x_i, which provides LLMs with the foundational context necessary to generate research ideas inspired by the same background information as the target papers.

4.2 Experimental Setup

Agent Initialization. We initialize the agents with meta prompts to make sure each agent understands its role and which field it specializes in (Liang et al., 2023). All the prompt templates along with the algorithm for agent interactions are provided in Appendix A.2.1.

Baselines. We consider the following two baseline methods to demonstrate the effectiveness of InfAL. ❶ **Initial idea**: This naive baseline is the initial idea $\hat{y}_0$ generated by the generator agent when given a set of background information x_i for a target idea t_i. GPT-4o serves as the backbone LLM for this method. ❷ **Self-reflection**: Self-reflection method is used as the strong baseline. The same initial ideas $\hat{y}_0$ are iteratively improved through self-evaluation (Madaan et al., 2024; Liang et al., 2023), where the generator agent reflects on its own generated research ideas and modifies them without external interaction. The process terminates once the agent determines that further improvements are no longer being made. GPT-4o is also used as the backbone LLM for this approach.

We choose these two baselines as they represent intuitive benchmarks—initial idea serves as the simplest baseline, and self-reflection embodies a strong iterative improvement method that is used by other works like MOOSE (Yang et al., 2023). These two approaches sufficiently capture the spectrum of existing practical methods, eliminating the need for comparisons with orthogonal techniques like retrieval-augmented generation method like SciMON (Wang et al., 2023d).

We measure improvement based on two key quality indicators, novelty and feasibility, using the relative ranking quality score Q described in Section 3.3. In the main experiment, we compare our method with baselines and evaluate performance using GPT-4o, GPT-4o Mini, GPT-3.5 Turbo, and Llama 3.1 model families as backbone models. For both baselines and our method, we generate results in two cases: one focusing on improving novelty and the other on feasibility. In each case, three research ideas are generated for each target idea. These ideas, along with the target paper's idea, are ranked based on novelty and feasibility, and the proposed relative ranking score Q is computed using Equation (5). Q is computed separately for novelty and feasibility, allowing us to evaluate each dimension independently. A prompt template is used to rank research ideas for novelty or feasibility without knowing whether the ideas are human or LLM generated, ensuring fairness (see Figure 13 and the Appendix for more details). All rankings are computed using GPT-4o, the highest-capacity model, regardless of the backbone model used for generation. Finally, Q is averaged across all target papers in the five datasets. Additionally, we compute the rankings using LLama 3.1 70B-Instruct and report similar results in the Appendix.

4.3 Main Results

Method	Base Model	Average Q (Novelty)	Average Q (Feasibility)
Initial	GPT-4o	0.808	0.171
SR	GPT-4o	0.952	0.342
Ours	GPT-3.5 Turbo	0.963	0.589
	GPT-4o Mini	0.960	**0.762**
	GPT-4o	0.981	0.550
Ours	Llama 3.1 8B-Instruct	0.953	0.451
	Llama 3.1 70B-Instruct	0.971	0.423
	Llama 3.1 405B-Instruct	**0.988**	0.363

Table 1: Main experimental results. The performance of different methods with GPT-based models (top) and the results of Llama 3.1 family of models (bottom). "Intial" and "SR" refers to the initial idea baseline and self-reflection baseline, respectively. We report the average relative quality ranking scores, denoted as Average Q, for novelty and feasibility.

The results presented in Table 1 highlight the effectiveness of InfAL in enhancing research idea generation, with substantial improvements observed in both novelty and feasibility. Notably, feasibility exhibits the most significant gains, underscoring InfAL's unique ability to generate research ideas that are not only creative but also practical. To summarize, InfAL carries strengths from four aspects:

(1) Model-agnostic effectiveness: The results in Table 1 show that both GPT-based models and the Llama 3.1 family benefit significantly from InfAL. For instance, with GPT-4o, InfAL achieved a novelty score of 0.981, while with Llama 3.1 405B-Instruct, it achieved the highest novelty score of 0.988. These results confirm InfAL's effectiveness

across various model families, making it a versatile and broadly applicable method for enhancing research ideation.

(2) Domain-agnostic improvements: The five datasets used in our experiments—covering Health and Medicine, Genetics and Molecular Biology, Environmental Sciences, Neuroscience and Cognitive Sciences, and Technology and Engineering—all exhibit performance gains in both novelty and feasibility. Please refer to Section **??** of the Appendix for detailed results. This indicates that InfAL can enhance research ideation regardless of the domain, making it a valuable tool for diverse scientific disciplines.

(3) Boosting lower-capacity models to outperform the baseline results of their higher-capacity counterparts. For example, GPT-3.5 Turbo, despite being a smaller model than GPT-4o, achieved a feasibility score of 0.589, far surpassing GPT-4o's initial baseline score. Similarly, Llama 3.1 70B-Instruct and Llama 3.1 8B-Instruct achieved feasibility scores of 0.423 and 0.451, respectively, exceeding the baseline results. This suggests that InfAL unlocks latent capabilities within smaller models, reducing dependence on the largest and most computationally expensive models while still achieving competitive results. Please refer to Section **??** of the Appendix for cost for deployment analysis.

(4) Rapid convergence: This efficiency will be further analyzed in the convergence analysis and ablation study in subsections 4.4 and 4.5. The ability of InfAL to rapidly refine research ideas without additional model training highlights its practicality for real-world applications, where computational resources and efficiency are critical constraints.

4.4 Convergence Analysis

In this subsection, we examine how the quality of generated research ideas evolves and converges during InfAL's iterative inference-time learning process. Figure 3 shows clear improvements in both novelty and feasibility across iterations, closely mirroring typical validation or test performance curves from traditional machine learning. Effective learning, regardless of whether it occurs during training or inference, should demonstrate a consistent pattern: initial rapid improvements in performance followed by stabilization and convergence.

Initially, the optimizer agent's feedback r_i provides essential gradient information $\nabla_u V(G, D; u_i)$, enabling the generator to enhance its knowledge utilization $u_{i,G}$ and thereby improve generated ideas. These iterative refinements ultimately lead the discriminator to recognize when ideas have reached an optimal neighborhood, stabilizing performance.

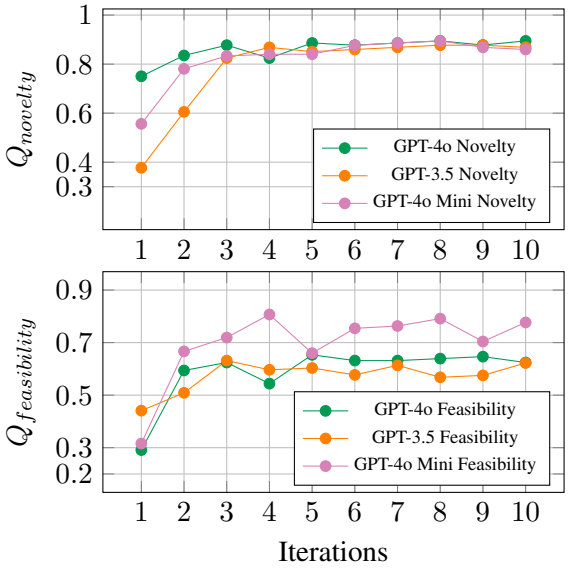

Figure 3: Evolution of research ideas' novelty (top) and feasibility (bottom) as number of iterations increases.

Backbone	Improvement Type	Median Iterations	Cvg. < 10
GPT-4o	Novelty	6	0.980
	Feasibility	5	0.963
GPT-3.5 Turbo	Novelty	10	0.254
	Feasibility	10	0.348
GPT-4o Mini	Novelty	10	0.005
	Feasibility	10	0.084

Table 2: Convergence statistics of our method from the main experimental results when improving novelty and feasibility. We report the median number of iterations until convergence, as well as the proportion of runs that reach convergence within 10 iterations (Cvg. < 10).

The rate at which convergence is recognized varies by model capacity. With GPT-4o, the discriminator detects convergence rapidly, achieving it within 10 iterations in over 96% of cases, with median iterations of 6 for novelty and 5 for feasibility (Table 2). This behavior matches Figure 3, where performance converges swiftly after initial improvement. Conversely, lower-capacity models such as GPT-4o Mini and GPT-3.5 Turbo exhibit slower convergence, frequently requiring the maximum number of iterations. This delay reflects their reduced ability to accurately assess incremental performance improvements, further reinforcing that InfAL indeed facilitates meaningful inference-time learning.

4.5 Ablation Study

In this subsection, we focus on how the absence of the discriminator and optimizer agents affects InfAL's ability to refine research ideas and converge to the optimal idea neighborhood $B_\epsilon(\dot{y})$. GPT-4o is used as the agents' backbone LLM in the ablation study, conducted on a smaller dataset with 100 target papers. The results are summarized in Table 3.

Ablation	$Q_{novelty}$	Iters.	$Q_{feasibility}$	Iters.
w/o Disc.	0.974	7	0.220	7
w/o Opt.	0.967	9	0.505	6
Ours	**0.983**	**6**	**0.521**	**5**

Table 3: Ablation study results showing the impact of our system without (w/o) the discriminator (Disc.) or optimizer (Opt.) agent on the novelty and feasibility of generated research ideas. We report the average relative quality ranking scores (Q) and median iterations (iters.) until convergence.

Without the discriminator agent, the system lacks the component responsible for determining whether the generated idea $\hat{y}$ falls within the neighborhood $B_\epsilon(\dot{y})$ of the optimal idea. Without this key component, we fix the number of iterations to seven, since the median iterations needed for our method to converge is below seven (Table 2). The results show a drop in the novelty and a major decline in the feasibility for the research ideas, indicating that the discriminator plays a crucial role in ensuring whether the generated ideas fall within $B_\epsilon(\dot{y})$. Without the discriminator to stop the process when an idea is within the optimal neighborhood, many ideas fail to reach the optimal idea neighborhood, especially in terms of feasibility. This supports our theoretical framework that convergence depends on the discriminator's ability to assess when further refinement is unnecessary.

The optimizer agent provides the essential gradient $r_i = \nabla_u V(G, D; u_i)$, enabling the generator to refine the idea iteratively. Removing the optimizer increases the number of iterations required to reach convergence since the generator agent lacks effective feedback to get u_G^*. Without the optimizer, both novelty and feasibility suffer. This aligns with the theoretical formulation that optimizer's feedback is essential for approximating the gradient necessary to update $u_{i,G}$ and $u_{i,D}$ for moving $\hat{y}$ toward the optimal neighborhood $B_\epsilon(\dot{y})$.

The ablation study highlights the critical roles of the discriminator and optimizer agents in the InfAL framework. Without these agents, the system either converges more slowly or fails to consistently produce ideas that approach $B_\epsilon(\dot{y})$. Please refer to Section A.3 the Appendix for more experiments and discussions.

4.6 Human Study

To evaluate the alignment between GPT-4o and humans in assessing research ideas with relative quality ranking, we conducted a human study. We selected 10 sets of research ideas focused on novelty and 10 sets focused on feasibility, generated using our proposed InfAL. Each set included three generated ideas and their respective target paper idea.

We recruited 10 volunteer PhD students to rank the ideas in each set based on either novelty or feasibility, depending on the focus. The researchers were unaware of which ideas were generated and which originated from the target paper. We then compared the agreement (denoted as $A(Q)$) between relative quality ranking given by human researchers and GPT-4o:

$$A(Q) = 1 - |Q_{\text{Human}} - Q_{\text{GPT-4o}}| \quad (6)$$

where Q_{Human} is the relative quality ranking from human researchers calculated using Formula (5) and $Q_{\text{GPT-4o}}$, similarly, is the relative quality ranking from GPT-4o.

The results indicate that human researchers and GPT-4o exhibit a high degree of agreement in ranking the target research ideas relative to the generated ones. From the average $A(Q)$ we see 90% agreement between GPT-4o and humans for ranking the target papers for novelty, and 70% agreement for feasibility. These findings highlight the effectiveness of GPT-4o in aligning with human judgment when assessing research ideas using our proposed relative quality ranking.

5 Conclusion

In this work, we formulated inference time adversarial learning and implemented it through a multi-LLM-agent interaction system to enhance the scientific research ideation process. Additionally, we developed a relative quality ranking metric to evaluate the generated ideas in a customizable and fair manner, serving as a proxy for human evaluation. Our promising results demonstrate that inference time adversarial learning not only improves scientific ideation but also holds potential for enhancing other tasks involving user interaction with LLMs.

Acknowledgment

This work is supported in part by the US National Science Foundation (NSF) and the National Institute of Health (NIH) under grants IIS-2106913, IIS-2538206, IIS-2529378, and R01LM014012-01A1. Any recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIH or NSF.

Limitations

A potential limitation of InfAL is its reliance on user-provided context at inference time. While InfAL effectively leverages LLMs to refine ideas based on given contexts, it does not explicitly assess the adequacy or relevance of the provided context itself. Consequently, if users supply incomplete, biased, or irrelevant information, InfAL's ability to generate high-quality and genuinely insightful research ideas may be constrained. Future enhancements might incorporate context-assessment mechanisms to ensure provided inputs meaningfully support the targeted research ideation tasks.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024. Magicore: Multi-agent, iterative, coarse-to-fine refinement for reasoning. *arXiv e-prints*, pages arXiv–2409.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*.

Xuan Long Do, Yiran Zhao, Hannah Brown, Yuxi Xie, James Xu Zhao, Nancy F Chen, Kenji Kawaguchi, Michael Qizhe Xie, and Junxian He. 2023. Prompt optimization via adversarial in-context learning. *arXiv preprint arXiv:2312.02614*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Xuemei Gu and Mario Krenn. 2024. Generation and human-expert evaluation of interesting research ideas using knowledge graphs and large language models. *arXiv preprint arXiv:2405.17044*.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. 2025a. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5888–5899.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, and Aidong Zhang. 2024. Embracing foundation models for advancing scientific discovery. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1746–1755. IEEE.

Sikun Guo, Guangzhi Xiong, and Aidong Zhang. 2025b. Optimizing external and internal knowledge of foundation models for scientific discovery. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pages 431–434. SIAM.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, and 1 others. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024a. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024b. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Jie Ma, Zhitao Gao, Qi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and 1 others. 2024. Debate on graph: a flexible and reliable reasoning framework for large language models. *arXiv preprint arXiv:2409.03155*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:2407.10817*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023b. Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. *arXiv preprint arXiv:2312.04854*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023c. Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv preprint arXiv:2305.14259*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023d. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2025. Toward reliable biomedical hypothesis generation: Evaluating truthfulness and hallucination in large language models. *arXiv e-prints*, pages arXiv–2505.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2023. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Jintian Zhang, Xin Xu, and Shumin Deng. 2023. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and

chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*.

A Appendix

The appendix is organized as follows:

A.1 Implementation Details

A.1.1 Code and Data Availability

The code used to implement InfAL along with the data used for our experiments are available here: `https://github.com/amir-hassan25/InfAL.git`.

A.2 Details on Dataset Preparation

For our evaluation benchmark, we curated 500 high-quality scientific articles published in 2024 across five domains: Health & Medicine, Genetics & Molecular Biology, Environmental Sciences, Neuroscience & Cognitive Sciences, and Technology & Engineering. Articles were selected based on specific criteria to guarantee recentness, quality, and no overlap with the training datasets of the evaluated LLMs.

A paper qualified as a target article if it satisfied either of these criteria: it was published in a leading venue as per Google Scholar's domain-specific rankings and received at least one citation, or it had garnered at least 20 citations regardless of venue. All selected papers were published exclusively in 2024. This approach ensured we included influential, recent works recognized by the scholarly community but excluded from the training data timeframe of the LLMs. To identify qualifying papers and retrieve associated cited references for context, we utilized the Semantic Scholar API (Kinney et al., 2023).

For each target paper, we derived a reference research idea (g_i) directly from its abstract. This extraction was performed using GPT-4-mini, guided by the prompt template depicted in Figure 14.

Algorithm 1 Algorithm for inference time adversarial learning for research idea generation via multi-LLM-agent interactions.

Input: User defined `{quality_indicator}`, `{quality_indicator_traits}`, `{research_area}`, background context $\{b_1, \ldots, b_{k_i}\}$, maximum iterations max_iters
Output: Final research idea $\hat{y}_{i+1}$

Step 1: Initialization
Initialize generator, optimizer, and discriminator agents based on `{quality_indicator}`, `{quality_indicator_traits}`, and `{research_area}`
$i \leftarrow 0$
Generate initial research idea $\hat{y}_0 \leftarrow \text{generator}(\{b_1, \ldots, b_{k_i}\})$
$r_0 \leftarrow \text{optimizer}(\hat{y}_0)$
Step 2: Iterative Improvement Process
while $i <$ max_iters **do**
  Generate new research idea $\hat{y}_{i+1} \leftarrow \text{generator}(\hat{y}_i, r_i)$
  $stop \leftarrow \text{discriminator}(\hat{y}_i, r_i, \hat{y}_{i+1})$
  Review new idea $r_{i+1} \leftarrow \text{optimizer}(\hat{y}_{i+1})$
  if *stop* is True **then**
    break
  end if
  Review new idea $r_{i+1} \leftarrow \text{optimizer}(\hat{y}_{i+1})$
  $i \leftarrow i + 1$
end while
Step 3: Return Final Idea
Return final research idea $\hat{y}_{i+1}$

A.2.1 Multi-LLM-Agent System Implementation Details

This section provides further details on the implementation of inference time adversarial learning via multi-LLM-agent interaction.

The generator, optimizer, and discriminator agents interact iteratively following Algorithm 1.

Each agent has a meta prompt defining its role and task. Figures 6, 7, and 8 illustrate the meta prompts for the generator, optimizer, and discriminator, respectively. These prompts take in user-defined `{research_area}` and `{quality_indicator}` hyperparameters, which specify the research field and the aspect of research ideas to improve. Additionally, the generator and optimizer agents' prompts take a `{quality_indicator_traits}` hyperparameter, listing the traits relevant to the `{quality_indicator}`.

Each agent also uses task-specific prompt templates. The generator generates an initial idea $\hat{y}_0$ using the template in Figure 9, inputting the `{research_area}` and a given target paper's reference paper abstracts $\{b_1, \ldots, b_{k_i}\}$ with the prompt template parameters `{background_paper_1_abstract}`, ..., `{background_paper_k_abstract}`. After receiving feedback from the optimizer, the generator revises the idea using the template in Figure 10 to improve the `{quality_indicator}`. The optimizer provides feedback with the template in Figure 11, considering both the `{quality_indicator}` and `{quality_indicator_traits}`. The discriminator then assesses whether the generator successfully improved its idea's `{quality_indicator}` using the template in Figure 12.

In our experiments, we set `{research_area}` to Health & Medicine, Genetics & Molecular Biology, Environmental Sciences, Neuroscience & Cognitive Sciences, and Technology & Engineering accordingly, and evaluate the system on two `{quality_indicator}` values: novelty and feasibility. For novelty, we set `{quality_indicator_traits}` to be "creativity of the hypothesis, innovation of the approach, disruptiveness, originality, conceptual shift, and addressing a research gap." For feasibility we set `{quality_indicator_traits}` to be "accessibility of resources, simplicity of method, data availability, time and cost efficiency, scalability, and practicality." Additionally, in Algorithm 1, we set the maximum number of iterations, max_iters to 10 for all experiments.

A.2.2 Relative Quality Ranking Implementation Details

In Section 3.3, we introduce the relative quality ranking metric, which evaluates LLM-generated research ideas based on a specified `{quality_indicator}`. The term n_t in Equation 5 represents the rank of a human-generated target idea t when compared to n other LLM-generated ideas.

To compute n_t, we use GPT-4o with the prompt template shown in Figure 13. This template takes the `{quality_indicator}`, the `{target_paper_idea}`, and the n LLM-generated ideas (`{generated_idea_1}`, ..., `{generated_idea_n}`) as inputs.

To ensure a fair comparison between the LLM-generated ideas and the target paper's idea, we extract the core research idea from the target paper's abstract using GPT-4o with a customized prompt. Abstracts often include extraneous details, such as results or technical specifics, which may not reflect the central idea. To avoid bias in the ranking, we use a prompt that summarizes the main research idea, aligning with the style in which the LLM generates ideas. The prompt template for this extraction is shown in Figure 14. This process ensures an equitable ranking of the target paper's idea alongside the LLM-generated ideas.

A.3 Additional Experiments and Discussion

In this section, we offer further experiments and noteworthy discussions. We evaluate and discuss the validity of our automatic evaluation of research ideas using the proposed relative quality ranking metric with GPT-4o. In particular, we address potential concerns regarding bias of leveraging LLM-as-a-Judge to score and rank research ideas generated by LLM itself (Wataoka et al., 2024). At last, we discuss the cost of generating research ideas using our method.

A.3.1 Relative Quality Ranking

Alignment with Human Judgment

Our results in the human study section in the main text suggest a high agreement between human judgment and LLM judgment. Similarly, in SCIMUSE, the authors collaborated with over 100 research group leaders across diverse domains to rank more than 4,400 research ideas generated by their SCIMUSE system(Gu and Krenn, 2024). Their findings revealed that LLM-based ranking,

specifically using GPT-4o, aligns closely with human expert evaluations, achieving a top-1 precision of 51% and a top-5 precision of 46.7%. These results highlight the feasibility of using LLM-driven ranking as a scalable proxy for human evaluation, particularly when assessing large volumes of research ideas across various fields.

Handling Potential Bias from GPT-4o as an Autorater.

Google researchers show that LLMs can be used as reliable autoraters, and GPT-4o is overall the best off-the-shelf model in handling bias ((Vu et al., 2024)). That's why we use GPT-4o as the autorater when evaluating research ideas. Furthermore, our relative quality metric does not prompt GPT-4o to give an absolute score for the quality of the ideas, because it may be biased. Rather, we provide a target idea to force GPT-4o to rank ideas based on a quality indicator specified by users, such as novelty and feasibility. This enables GPT-4o to provide a more objective evaluation than asking for an absolute score.

Confidence Interval for Relative Quality Ranking.

To ensure the robustness and consistency of our automatic evaluation, we calculated confidence intervals (CIs) for GPT-4o's relative quality rankings, which provide a clearer representation of the metric's reliability and variability. Using a dataset of 100 target papers, we generated novel and feasible research ideas with our method and computed the average relative quality rankings (Average Q) across five iterations. This allowed us to obtain 95% confidence intervals for both novelty and feasibility, along with the standard deviation and variance.

The results, presented in Table 4, demonstrate that GPT-4o's rankings are highly consistent, with minimal variation in computed relative quality rankings, further supporting the validity of the metric.

	Q **CI**	**STD**	**Variance**
Novelty	0.983 ± 0.003	0.003	1.216×10^{-5}
Feasibility	0.484 ± 0.026	0.028	8.0464×10^{-4}

Table 4: Experiment assessing the consistency of GPT-4o's relative quality rankings. The table reports the 95% confidence intervals (CI), standard deviations (STD), and variances of the Average Q scores for novelty and feasibility, calculated five times.

Comparison with Other Metrics.

In open-ended generation tasks, winrate is a metric commonly used to assess quality by determining the proportion of instances in which one model's output is preferred over another's in a binary comparison (Zheng et al., 2023). However, this approach reduces nuanced evaluations to binary outcomes, which can lead to significant information loss in capturing the diversity and subtle differences between outputs. Our relative quality ranking offers a more granular approach by allowing for a graded comparison across multiple dimensions of quality. Instead of a binary decision boundary, this metric ranks outputs on a continuum, capturing more nuanced differences in quality. This fine-grained assessment provides richer insights into the strengths and weaknesses of each model output, enhancing the accuracy of quality evaluations in open-ended generation tasks.

Comparison of Relative Quality Ranking Provided by Different LLMs.

To assess the consistency of relative quality rankings generated by different LLMs, we compare the rankings generated by GPT-4o and LLama 3.1 70B-Instruct on the research ideas produced by GPT-4o using our proposed method from the main experiment. The results are shown in Table 5 and indicate that both models generate similar relative quality rankings for both novelty and feasibility.

Base Model	Q (Novelty)	Q (Feasibility)
GPT-4o	0.981	0.550
Llamma 3.1 70B-Instruct	0.991	0.459

Table 5: Comparison of relative quality rankings generated by different LLMs. The table reports the relative quality rankings (denoted as Average Q) for novelty and feasibility given by GPT-4o and LLama 3.1 70B-Instruct on research ideas generated by GPT-4o using our proposed method in the main experiment.

A.3.2 Categorized Main Results

We further downsampled each research domain to 65 papers, ensuring no overlap of papers across domains. We summarize the results for relative quality ranking—covering initial idea generation, self-reflection (SR), and InfAL—in Tables 9 to 14. These results consistently show InfAL's superior performance across all models and domains.

A.3.3 P-values between InfAL and Self-reflection

We recognize that comparing to self-reflection baseline, InfAL "appears to be" only delivering marginal improvements. To make sure that the improvements of InfAL is statically significant, we perform the p test between the two methods. Statistical significance tests (p-values below in Figure 6) confirm the consistency and significance of InfAL's improvements.

Model	p-value Novelty	p-value Feasibility
GPT-4o	4.28×10^{-4}	3.28×10^{-16}
Llama 3.1 70B	6.62×10^{-18}	2.51×10^{-84}

Table 6: p-value Analysis for Baseline Comparisons (Self-reflection vs. InfAL)

A.3.4 Cost for Deployment

We calculated the average cost of generating a research idea for each model using our method and report the costs in 7. The cost were calculated with OpenAI [1] and DeepInfra [2] (API service for Llamma 3.1 models) cost per million tokens. We see that GPT-4o is the most expensive model to generate research ideas with, while Llamma 3.1 8B-Instruct is the cheapest.

Base Model	Average Cost Per Idea
GPT-4o	\$1.27
GPT-4o Mini	\$0.21
GPT-3.5 Turbo	\$0.88
Llama 3.1 405B-Instruct	\$0.27
Llama 3.1 70B-Instruct	\$0.04
Llama 3.1 8B-Instruct	\$0.02

Table 7: Average cost of generating a research idea using our method with different backbone LLMs.

A.3.5 Computation Efficiency

We quantitatively compare our InfAL approach against the self-reflection baseline, specifically evaluating convergence speed (runtime efficiency) using GPT-4o and Llama 3.1 70B:

As shown in Table 8, InfAL demonstrates comparable or superior efficiency relative to the self-reflection baseline. Notably, for GPT-4o (Novelty), InfAL converges significantly faster, reducing median iterations from 9 to 6, highlighting that our multi-agent framework achieves substantial performance improvements without compromising efficiency.

Model	Quality	SR	InfAL
GPT-4o	Novelty	9	6
GPT-4o	Feasibility	4	5
Llama 3.1 70B	Novelty	10	10
Llama 3.1 70B	Feasibility	10	10

Table 8: Runtime Efficiency Comparison (Median iterations, fewer is better)

A.4 Case Studies

To qualitatively evaluate how InfAL improves research ideation, we present a set of case studies that showcase how multi-agent inference-time optimization via InfAL enhances the novelty or feasibility of research ideas over multiple interaction rounds. These studies highlight the dynamic interplay between generator, optimizer, and discriminator agents, with optimizer feedback acting as textual gradients that iteratively guide refinement, and the discriminator determining when convergence is achieved. We include two illustrative examples—one focused on novelty, the other on feasibility—as well as four full conversational trajectories annotated with detailed human feedback across rounds. Together, these case studies validate the practical effectiveness of InfAL in enhancing research ideas with targeted improvements and transparent optimization dynamics.

Both illustrative examples begin with the same initial hypothesis proposed by the generator agent: that orexin levels may improve sleep quality and energy levels in adolescent athletes. From this shared starting point, the optimizer agent offers feedback aligned with the target quality indicator (novelty or feasibility), prompting the generator to revise its ideas accordingly.

Novelty Improvement (Figure 4) To improve novelty, the optimizer agent critiques the idea's originality and proposes directions to enhance conceptual disruption and scientific innovation. Over seven rounds, the idea evolves significantly. By the final round, the hypothesis explores the modulation of orexin levels not only for sleep quality but also for cognitive function and emotional resilience. It

[1] More details about the OpenAI API service can be found here: https://platform.openai.com/docs/overview

[2] More details about the DeepInfra's API service can be found here: https://deepinfra.com/

proposes investigating the effects of orexin in relation to social jetlag, using advanced statistical modeling to uncover novel associations. Human evaluation noted that the final iteration reflects a notable conceptual shift and deeper interdisciplinary integration.

Feasibility Improvement (Figure 5) In contrast, the feasibility refinement path focuses on grounding the idea in practical experimental design without radically altering the core hypothesis. The optimizer emphasizes methodological simplicity, resource availability, and realistic implementation. The final iteration frames a study testing whether dietary guidelines and sleep hygiene education improve sleep and energy in adolescent athletes. It details a feasible study plan involving 15 participants, activity monitors, and parental involvement to ensure adherence—elements that enhance the idea's implementability without compromising scientific relevance.

Human Evaluation at Each Round. To further substantiate InfAL's effectiveness, we conducted detailed human evaluations for each round of InfAL on four randomly selected examples (two novelty, two feasibility). Due to space constraints, we include links to the full annotated conversations:

- Novelty Example 1
- Novelty Example 2
- Feasibility Example 1
- Feasibility Example 2

In each figure, the center panel illustrates the round-by-round InfAL trajectory. The left panel shows human feedback on generator outputs, and the right panel shows feedback on optimizer responses. These evaluations confirm that InfAL's textual gradients lead to measurable improvements across iterations. Reviewers consistently observed sharper, more targeted optimizer feedback over time, with generator outputs evolving accordingly. The discriminator's convergence decisions also align with human-perceived saturation, validating its role in terminating the optimization process once meaningful improvements plateau.

Category	GPT-4o	GPT-3.5	GPT-4o_mini	Llama_3.1_8b	Llama_3.1_70b	Llama_3.1_450b
Health and Medicine	0.856	0.229	0.564	0.307	0.622	0.854
Genetics and Molecular Biology	0.813	0.193	0.417	0.288	0.546	0.813
Environmental Sciences	0.812	0.282	0.556	0.199	0.576	0.825
Neuroscience and Cognitive Sci	0.818	0.185	0.469	0.265	0.586	0.812
Technology and Engineering	0.749	0.180	0.544	0.261	0.582	0.745

Table 9: Categorized Relative Quality Ranking for Novelty—Initial Idea Generation

Category	GPT-4o (SR)	GPT-3.5 (SR)	GPT-4o_mini (SR)	Llama_3.1_8b (SR)	Llama_3.1_70b (SR)	Llama_3.1_450b (SR)
Health and Medicine	0.911	0.935	0.879	0.756	0.951	0.961
Genetics and Molecular Biology	0.963	0.911	0.871	0.713	0.963	0.973
Environmental Sciences	0.935	0.942	0.901	0.692	0.946	0.951
Neuroscience and Cognitive Sci	0.941	0.909	0.903	0.717	0.959	0.972
Technology and Engineering	0.902	0.891	0.867	0.726	0.949	0.965

Table 10: Categorized Relative Quality Ranking for Novelty—Self-reflection (SR)

Category	GPT-4o (InfAL)	GPT-3.5 (InfAL)	GPT-4o_mini (InfAL)	Llama_3.1_8b (InfAL)	Llama_3.1_70b (InfAL)	Llama_3.1_450b (InfAL)
Health and Medicine	0.969	0.958	0.897	0.910	0.974	0.979
Genetics and Molecular Biology	0.984	0.954	0.892	0.934	0.967	0.974
Environmental Sciences	0.974	0.979	0.910	0.908	0.969	0.995
Neuroscience and Cognitive Sci	0.958	0.959	0.918	0.939	0.957	0.984
Technology and Engineering	0.981	0.979	0.887	0.938	0.968	0.984

Table 11: Categorized Relative Quality Ranking for Novelty—InfAL

Category	GPT-4o	GPT-3.5	GPT-4o_mini	Llama_3.1_8b	Llama_3.1_70b	Llama_3.1_450b
Health and Medicine	0.192	0.295	0.212	0.187	0.212	0.120
Genetics and Molecular Biology	0.142	0.408	0.199	0.134	0.136	0.134
Environmental Sciences	0.192	0.295	0.212	0.153	0.212	0.120
Neuroscience and Cognitive Sci	0.166	0.484	0.223	0.162	0.169	0.162
Technology and Engineering	0.202	0.314	0.228	0.136	0.184	0.192

Table 12: Categorized Relative Quality Ranking for Feasibility—Initial Idea Generation

Category	GPT-4o (SR)	GPT-3.5 (SR)	GPT-4o_mini (SR)	Llama_3.1_8b (SR)	Llama_3.1_70b (SR)	Llama_3.1_450b (SR)
Health and Medicine	0.331	0.487	0.496	0.378	0.356	0.337
Genetics and Molecular Biology	0.312	0.504	0.508	0.365	0.367	0.351
Environmental Sciences	0.314	0.476	0.467	0.344	0.340	0.324
Neuroscience and Cognitive Sci	0.347	0.512	0.496	0.356	0.347	0.347
Technology and Engineering	0.352	0.494	0.512	0.312	0.351	0.353

Table 13: Categorized Relative Quality Ranking for Feasibility—Self-reflection (SR)

Category	GPT-4o (InfAL)	GPT-3.5 (InfAL)	GPT-4o_mini (InfAL)	Llama_3.1_8b (InfAL)	Llama_3.1_70b (InfAL)	Llama_3.1_450b (InfAL)
Health and Medicine	0.555	0.582	0.762	0.457	0.421	0.351
Genetics and Molecular Biology	0.542	0.593	0.735	0.412	0.413	0.387
Environmental Sciences	0.591	0.578	0.756	0.395	0.467	0.413
Neuroscience and Cognitive Sci	0.542	0.614	0.779	0.412	0.441	0.413
Technology and Engineering	0.575	0.582	0.753	0.429	0.453	0.420

Table 14: Categorized Relative Quality Ranking for Feasibility—InfAL

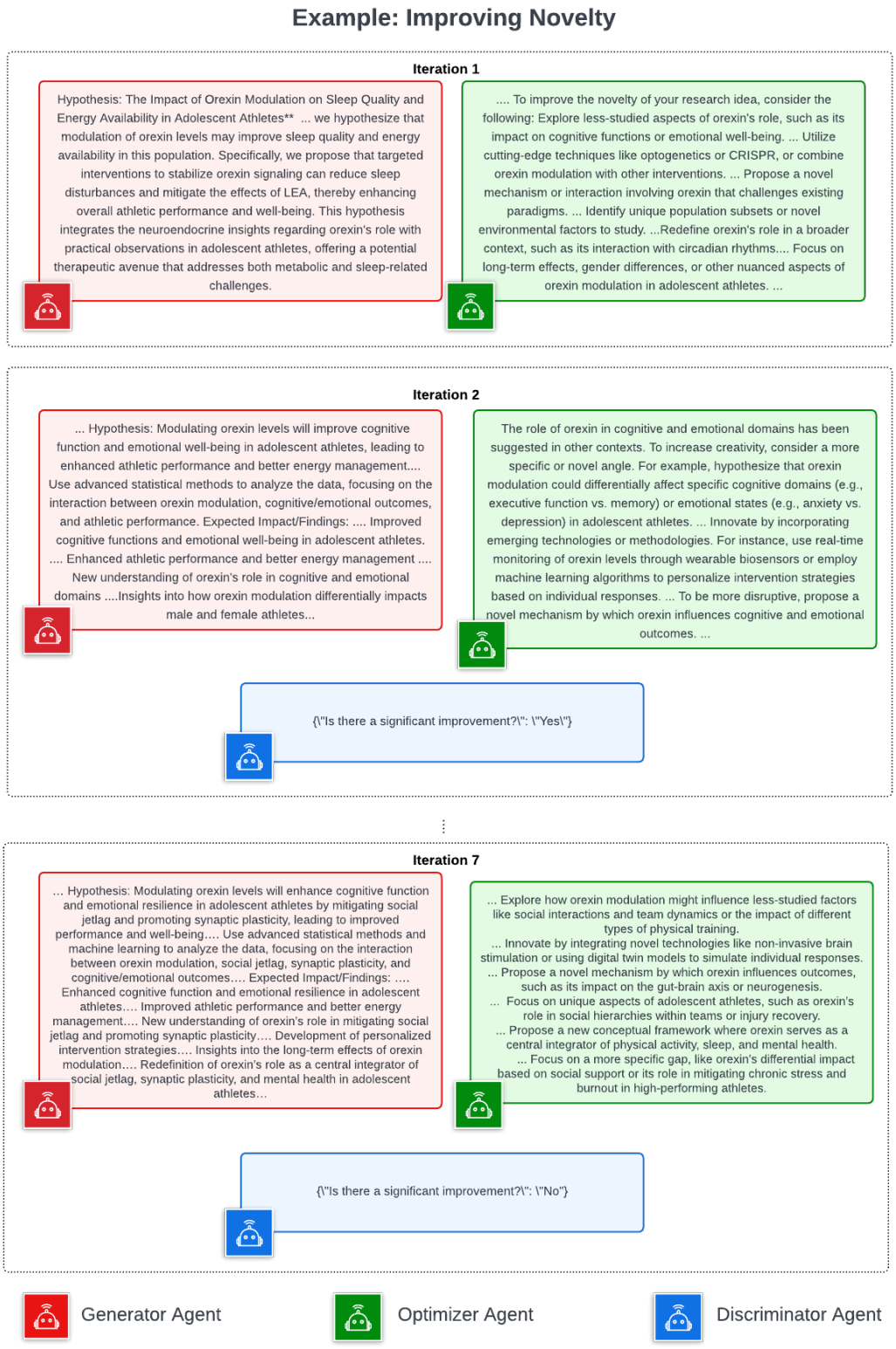

Figure 4: An example of how the inference time adversarial learning helps to improve the novelty of one generated research idea.

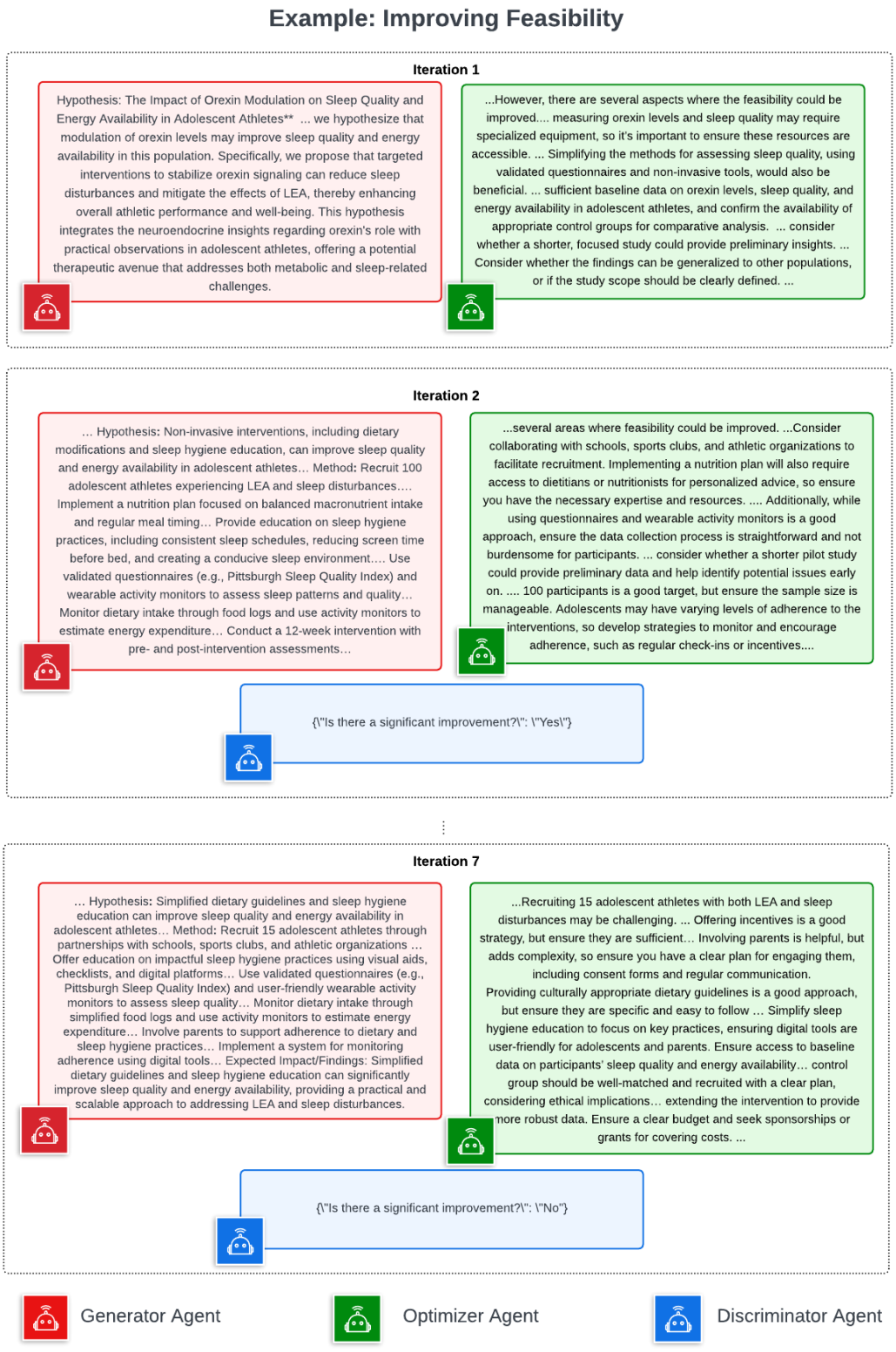

Figure 5: An example of how the inference time adversarial learning helps to improve the feasibility of one generated research idea.

Generator Meta Prompt Template.

You are a {research_area} researcher proposing research ideas. Your role is to create a research idea and refine the idea if you receive feedback. A reviewer will review your research idea based on its {quality_indicator} and give you feedback. You should try your best to improve the idea based on the reviewer's feedback and your expertise, especially paying attention to the idea's {quality_indicator_traits}.

Figure 6: Meta prompt template for the generator agent to inform the agent of its role and responsibility.

Optimizer Meta Prompt Template.

You are an experienced {research_area} researcher reviewing research ideas. Your role is to receive a research idea and try your best to give constructive criticism about the idea's {quality_indicator} so that the idea proposer can review your feedback and improve the idea's {quality_indicator} as much as possible. When reviewing, focus on the idea's {quality_indicator_traits}.

Figure 7: Meta prompt template for the optimizer agent to inform the agent of its role and responsibility.

Discriminator Meta Prompt Template.

You are an area chair for a high-impact {research_area} conference. You will receive a proposer's prior research idea and the proposer's revised research idea based on a reviewer's feedback. Your role is to try your best to identify any improvement in the revised idea and determine whether the revised idea has a significant improvement in {quality_indicator}.

Figure 8: Meta prompt template for the discriminator agent to inform the agent of its role and responsibility.

Generator Agent Prompt Template for Generating Initial Research Ideas

You are a {research_area} researcher. You are tasked with creating a hypothesis or research idea given some background knowledge. The background knowledge is provided by abstracts from other papers.

Here are the abstracts:

Abstract 1:{background_paper_1_abstract}
Abstract 2:{background_paper_2_abstract}
.....
Abstract k:{background_paper_k_abstract}

Using these abstracts, reason over them and come up with a novel hypothesis. Please avoid copying ideas directly, rather use the insights to inspire a novel hypothesis in the form of a brief and concise paragraph.

Figure 9: Prompt template for the generator agent to generate an initial research idea based on research paper abstracts as background context.

Generator Agent Prompt Template.

optimizer_agent_feedback}

Based on the reviewer's feedback regarding the previous research idea's {quality_indicator}, generate a revised and improved research idea using the following format:

Title: [A brief, focused title]

Problem: [The core issue or gap being addressed]

Objective: [The main goal or research question]

Hypothesis: [The hypothesis being tested or explored]

Method: [The approach or methodology]

Expected Impact/Findings: [The anticipated outcomes or contributions].

Please only respond with the improved research idea returned in the format provided above. Do not respond with anything irrelevant.

Figure 10: Prompt template for the generator agent to generate a revised research idea based on the optimizer agent's feedback.

Optimizer Agent Prompt Template.

You will receive the proposer's research idea. Try your best to give the best constructive criticism on the research idea's {quality_indicator} so that the proposer can improve the idea's {quality_indicator} as much as possible. In your response, please explain why the research idea lacks in {quality_indicator}, specifically considering the idea's {quality_indicator_traits}. Here is the proposer's research idea: {research_idea}.

Figure 11: Prompt template for the optimizer agent to give constructive criticism and feedback to the generator agent for its generated research idea.

Discriminator Agent Prompt Template.

Here is the proposer's prior idea: `{prior_research_idea}`

Here is the reviewer's constructive feedback for the proposer's prior idea: `{optimizer_agent_feedback}`

The proposer's revised idea: `{new_research_idea}`

Based on the reviewer's feedback, you will compare the proposer's prior and revised ideas in this round. Try your best to determine what improvement has been made in the revised idea and answer whether the revised idea has significantly improved in `{quality_indicator}`.

Please answer in a Python Dictionary with the following format:
`{"Is there a significant improvement?": "Yes" or "No"}`

Please strictly output in the Python Dictionary format; do not output irrelevant content.

Figure 12: Prompt template for the discriminator agent to determine whether generator's research ideas have improved.

Prompt template used to rank research ideas based on user specified quality indicators

You are a reviewer tasked with ranking the quality of a set of research ideas based on their `{quality_indicator}`. The idea with the highest `{quality_indicator}` should be ranked first.

Please rank the following hypotheses in the format: 1. Hypothesis (insert number):(insert brief rationale)
2. Hypothesis (insert number):(insert brief rationale)
3. Hypothesis (insert number):(insert brief rationale)
.....
n. Hypothesis (insert number):(insert brief rationale)

Please rank the following hypotheses:
Hypothesis 1: `{target_paper_idea}`
Hypothesis 2: `{generated_idea_1}`
Hypothesis 3: `{generated_idea_2}`
.....
Hypothesis n: `{generated_idea_n}`

Figure 13: Prompt template used to rank research ideas based on user specified quality indicators.

Paper Abstract Summary Prompt Template.

Write a concise summary of the following paper abstract, proposing a research idea based on the abstract's content. Format the summary using the following structure, and if a field does not exist in the abstract, write "NONE" for that field:

Title: [A brief, focused title]

Problem: [The core issue or gap being addressed]

Objective: [The main goal or research question]

Hypothesis: [The hypothesis being tested or explored]

Method: [The approach or methodology]

Expected impact / findings: [The anticipated outcomes or contributions]

Abstract: `{target_paper_abstract}`
Summary:

Figure 14: Prompt template for summarizing a target paper's abstract into a research idea.