

Enhancing Research Idea Generation through Combinatorial Innovation and Multi-Agent Iterative Search Strategies

Shuai Chen¹, Chengzhi Zhang²

¹*shuaichen@njust.edu.cn*, ²*zhangcz@njust.edu.cn*

Department of Information Management, Nanjing University of Science and Technology, Nanjing, 210094 (China)

Abstract

Scientific progress relies on the continuous emergence of innovative discoveries. However, the exponential growth in scientific literature has increased the cost of information filtering, making it significantly more challenging for scientists to identify innovative research directions. Although artificial intelligence (AI) methods have shown potential in tasks such as research idea generation and hypothesis formulation, the ideas they produce are often repetitive and simplistic. Combinatorial innovation theory posits that new entities arise from the recombination of existing elements, offering a novel approach to addressing these challenges.

This study draws on combinatorial innovation theory and the Delphi method to introduce a multi-agent iterative planning and search strategy into the research idea generation process, aiming to enhance the diversity and novelty of generated ideas. The strategy integrates iterative knowledge search with a large language model (LLM)-based multi-agent system to iteratively generate, evaluate, and refine research ideas. Experiments conducted using data from the field of natural language processing demonstrate that the multi-agent iterative planning and search strategy outperforms state-of-the-art methods in terms of diversity and novelty, showcasing its potential to generate high-quality research ideas. This study not only validates the effectiveness of the multi-agent iterative search strategy but also provides a theoretical explanation, grounded in combinatorial innovation theory and methodologies, for its ability to improve research idea generation performance. It offers new perspectives for future work in this domain.

Introduction

Over the past few decades, the volume of scientific literature has experienced exponential growth, reflecting the vigorous expansion of research activities and the continuous advancement of science and technology. However, the sheer magnitude of scientific publications has imposed significant temporal and cognitive burdens on scientists as they endeavor to filter and assimilate relevant information. Concurrently, this phenomenon has exacerbated the issue of redundancy in scientific research (Larivière et al., 2008), leading to substantial inefficiencies in the allocation of research resources. These compounding factors have collectively contributed to the escalating challenges scientists face in pursuing innovative research endeavors. Recently, large language models (LLMs) have demonstrated remarkable performance across a variety of challenging tasks, including mathematical proof (Yang et al., 2024), information retrieval (Ajith et al., 2024), and solving specific research problems through code generation (Lu et al., 2024; Schmidgall et al., 2025; Yuan et al., 2025). These models have even shown the potential to generate innovative research ideas (Baek et al., 2024; X. Gu & Krenn, 2024; Kumar et al., 2024; Lu et al., 2024; Si et al., 2024). However, despite evidence suggesting that LLMs can produce novel research concepts, their outputs often exhibit a high degree

of redundancy(Si et al., 2024). This issue necessitates additional effort from researchers to filter and deduplicate generated content, thereby hindering their broader application in academic research. Several studies have attempted to address this challenge through various approaches. For instance, (Baek et al., 2024) employed knowledge graph construction, (Si et al., 2024) utilized keyword-based searches for specific knowledge, and (Hu et al., 2024) adopted iterative knowledge search strategies. Nevertheless, these methods remain limited in critical ways. On the one hand, they often focus narrowly on knowledge within a single domain, failing to adequately integrate insights from multiple related fields. This significantly constrains the breadth of knowledge sources and the diversity of problem-solving perspectives. On the other hand, these studies have not sufficiently addressed the potential biases introduced by relying on a single large language model.

In light of these considerations, this study introduces combinatorial innovation theory and the multi-agent iterative planning and search strategy to the task of research idea generation. This strategy leverages knowledge planning and search mechanisms to integrate multi-domain knowledge, supported by a large language model-based multi-agent system. It simulates the expert survey method (Delphi method)(Linstone & Turoff, 1975) commonly employed in innovation practices, iteratively generating, evaluating, and refining research ideas. Specifically, the large language model is assigned the role of an expert with a specific disciplinary background to simulate the Delphi method discussion process in real-world research scenarios. Experiments conducted on a dataset of academic papers in the field of natural language processing demonstrate that the proposed method outperforms baseline approaches across key metrics, including diversity, novelty, and quality scores. Furthermore, the study provides an explanation for the enhanced performance of the strategy in generating research ideas, drawing on combinatorial innovation theory and methodological applications. This offers novel insights and perspectives for future research on idea generation. The study addresses the following two research questions:

RQ1: Can the multi-agent iterative planning and search strategy enhance the diversity and novelty of research ideas generated by large language models?

RQ2: Can combinatorial innovation theory and methodological approaches guide the task of generating research ideas using large language models?

The contributions of this study are threefold:

First, this paper proposes a multi-agent iterative planning and search strategy, which is applied to the task of generating research ideas using large language models. The strategy is evaluated through role-playing simulations with real-world data, and the final outputs are assessed objectively.

Second, the study conducts comprehensive experiments to evaluate the multi-agent iterative planning and search strategy. These experiments include comparisons with baseline methods, assessments of different team configurations, variations in the number of iterations, and ablation studies of individual modules. The results demonstrate that the proposed strategy significantly enhances the quality of research idea generation, outperforming existing baseline methods.

Third, this paper provides a theoretical explanation for the improved performance of the multi-agent iterative planning strategy in generating research ideas, drawing on combinatorial innovation theory and methodological applications. This not only offers new insights into the mechanisms underlying the strategy's success but also provides a novel tool and perspective for future research on idea generation using large language models.

Related work

This section reviews related work from three perspectives: (1) Generating Research Ideas Using Large Language Models; (2) Prompt Engineering for Logical Reasoning in Large Language Models; (3) Combinatorial Innovation Theory and Methodological Approaches.

Generating Research Ideas Using Large Language Models

In recent years, a growing body of research has demonstrated that large language models (LLMs) possess the capability to generate novel and innovative scientific research ideas, a phenomenon that has garnered significant attention from scholars. Among these studies, some researchers have adopted approaches such as retrieving relevant papers based on research topics(Lu et al., 2024) or directly utilizing the references of target papers(Guo et al., 2024), embedding these materials into the contextual prompts of LLMs to stimulate the generation of related research ideas. Others have first retrieved relevant papers as a knowledge base and enhanced idea generation by retrieving related knowledge during the process(Si et al., 2024) , a method known as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). Additionally, some scholars have constructed scientific knowledge graphs, employing co-occurrence entity search techniques to integrate retrieved entities into LLM prompts, thereby generating unique and novel research ideas(Baek et al., 2024; X. Gu & Krenn, 2024). IdeaSynth(Pu et al., 2024) introduced human expertise into the research idea generation process, demonstrating that human-AI collaboration outperforms single LLM baselines. VIRSCI(Su et al., 2024) further incorporated multi-agent collaboration into the idea generation process, utilizing LLMs to simulate real-world scientific collaboration scenarios, thereby opening new avenues for generating research ideas. (Li et al., 2024) employed a two-stage approach involving supervised fine-tuning and reinforcement learning to enhance the feasibility, novelty, and effectiveness of research ideas generated by LLMs.(T. Gu et al., 2024) deconstructed paper knowledge into distinct innovative components, leveraging LLMs to combinatorially generate innovative research ideas.

Although existing research has shown that LLMs can produce ideas that are more novel than those written by human experts, it has also highlighted the issue of excessive redundancy in generated ideas (Si et al., 2024) . While Nova (Hu et al., 2024) proposed an iterative planning and search method to reduce the repetition rate of LLM-generated ideas, this study adopts a multi-agent iterative planning and search perspective to further enhance the diversity and novelty of research ideas generated by LLMs.

Prompt Engineering for Logical Reasoning in Large Language Models

Prompt engineering has become an indispensable technique for extending the capabilities of large language models (LLMs) (Sahoo et al., 2024), and the logical reasoning abilities of LLMs are a focal point in the field of artificial intelligence. Consequently, how to leverage prompt engineering to enhance the logical reasoning capabilities of LLMs has become a central focus of scholarly research. Chain-of-Thought (CoT) prompting (Wei et al., 2022) addresses complex problems such as mathematical word problems and commonsense reasoning by presenting reasoning pathways as examples to LLMs, thereby improving their interpretability. Subsequently, (Kojima et al., 2022) proposed Zero-Shot Chain-of-Thought prompting, discovering that simply appending the phrase 'Let's think step by step' to a question enables LLMs to generate a reasoning chain, from which more accurate answers can be extracted. However, creating high-quality Chain-of-Thought examples is time-consuming and labor-intensive. To address this, (Zhang et al., 2022) introduced Auto Chain-of-Thought prompting, which automatically guides LLMs to generate reasoning chains and employs diverse sampling to enhance robustness. (X. Wang et al., 2022) proposed the Self-consistency prompting method, which samples multiple reasoning chains from the LLM's decoder and aggregates them to identify the most consistent answer, significantly improving the performance of Chain-of-Thought methods. Following this, (Zhou et al., 2022) introduced Least-to-Most (LtM) prompting, incorporating planning into prompt engineering by decomposing problems into subproblems and solving them sequentially, thereby enhancing LLMs' ability to tackle complex reasoning tasks. (Yao et al., 2024) proposed the Tree of Thoughts framework, enabling LLMs to explore multiple reasoning paths and self-evaluate before determining the next steps.

While these studies have improved the logical reasoning capabilities of LLMs to some extent, they are limited by the internal knowledge of LLMs and lack interaction with external environments, often leading to hallucinations. To overcome this limitation, (Trivedi et al., 2022) proposed a method combining Chain-of-Thought with external knowledge retrieval, enhancing LLMs' ability to solve knowledge-intensive tasks. Unlike previous approaches that separate reasoning and action in LLMs, ReAct (Yao et al., 2022) allows LLMs to simultaneously generate reasoning trajectories and task-specific actions, fostering synergy between reasoning and action. Specifically, ReAct interacts with external knowledge retrieval tools to address hallucinations and error propagation, thereby improving the factual accuracy of LLM-generated content. In contrast to the linear reasoning chains of LLMs, human thinking is non-linear. To address this, (Besta et al., 2024) introduced Graph of Thoughts prompting, which enables dynamic interaction, backtracking, and evaluation of ideas generated by LLMs, allowing for the aggregation and combination of thoughts from different branches and moving beyond the linear structure of Tree of Thoughts.

Given that a single LLM may be influenced by various biases, leading to inaccuracies in its generated or evaluated outputs (Liusie et al., 2023; P. Wang et al., 2023), many scholars have proposed techniques and architectures for multi-agent LLM systems, such as role-playing (N. Wu et al., 2023), debate (Chan et al., 2023), and voting (Zhu

et al., 2024). Therefore, this study attempts to integrate multi-agent systems with iterative planning to address the complex task of research idea generation.

Combinatorial Innovation Theory and Methodological Approaches

The question of how innovation arises has long been a topic of interest among scholars. Although many theories on innovation are based on human creative activities, they also provide valuable guidance for large language models (LLMs) in engaging in creative endeavors.

Schumpeter proposed that innovation is combinatorial in nature, suggesting that new entities emerge through the recombination of existing elements (Schumpeter & Swedberg, 2021). Boden shares a similar perspective, arguing that novel ideas arise from associating familiar concepts in new ways (Boden, 2004). This mechanism is particularly well-suited for LLMs, which can explore vast knowledge spaces to recombine information and generate novel outputs (T. Gu et al., 2024). In the field of scientometrics, researchers have already begun to explore the application of combinatorial innovation in scientific contexts (Lee et al., 2015; Shi & Evans, 2023; Uzzi et al., 2013). However, the innovation process is not linear but rather cyclical and iterative, often involving continuous "generation-evaluation" loops (Sharpies, 2013). (Sadler-Smith, 2015) further divides the creative process into four stages—preparation, incubation, insight, and verification—providing a new perspective for understanding creativity.

Combinatorial innovation theory and methodologies offer critical guidance for the design of the approach proposed in this study: **A Step-by-Step Research Idea Generation Process:** Initial ideas are conceptualized, iteratively refined, and finally summarized to deepen and perfect research directions. **Systematic Cross-Domain Knowledge Exploration:** A planning-based approach is employed to extensively search knowledge across different domains, using combinatorial knowledge prompts to leverage LLMs' ability to integrate diverse information. **Multi-Agent Simulated Brainstorming and Evaluation Mechanism:** A multi-agent system simulates the Delphi method to conduct brainstorming sessions, where each agent proposes ideas and an evaluator iteratively assesses them, generating research ideas of greater value.

Data and Methodology

This section provides a detailed exposition of the entire workflow of the multi-agent iterative planning and search strategy. The process comprises four key steps: (1) Dataset Construction; (2) Initial Research Idea Generation; (3) Iterative Refinement of Ideas; (4) Abstract Generation. The framework of the proposed methodology is illustrated in Figure 1.

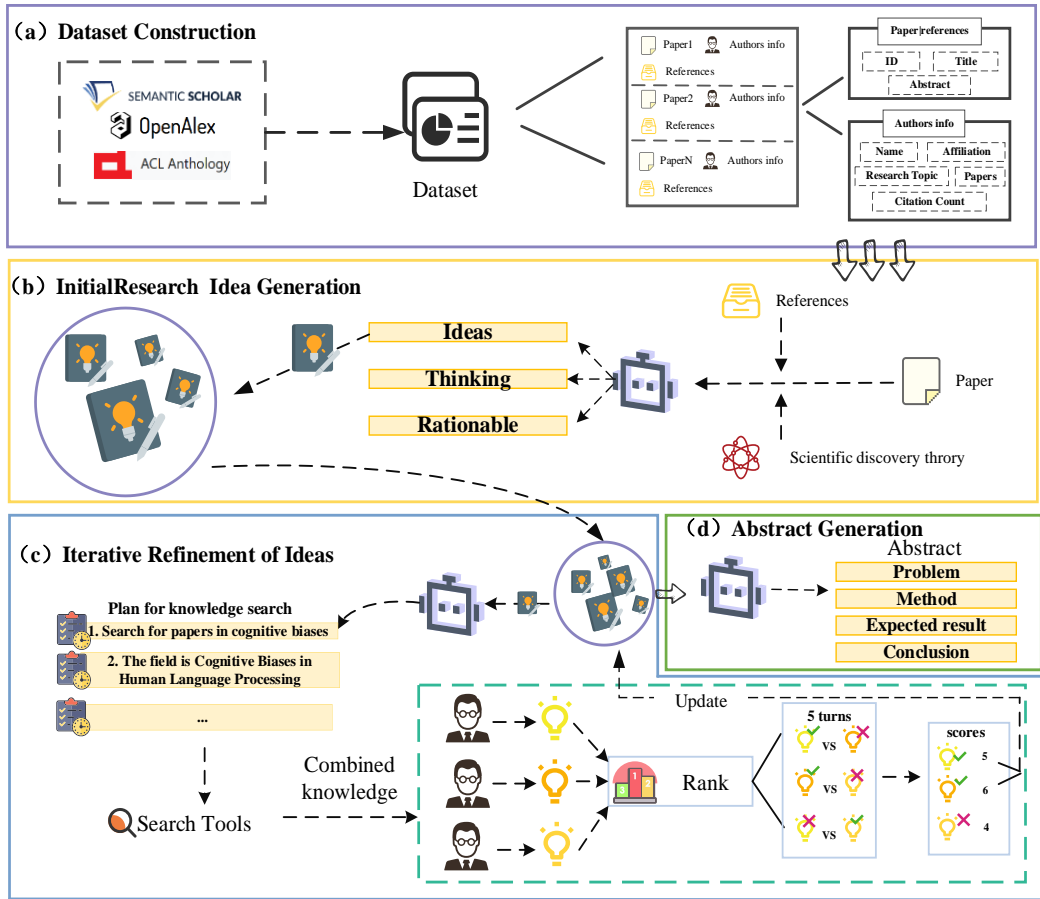


Figure 1. Framework of this study.

Dataset Construction

The data in this study is primarily utilized for two purposes: (1) the generation of initial research ideas and (2) the construction of multi-agent background information. To achieve this, the study requires access to target papers, their references, and information about the authors of the target papers. Research by (Guo et al., 2024) has demonstrated that high-quality papers significantly enhance the quality of research ideas generated by large language models (LLMs). Therefore, this study selects long papers from the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)¹ as the initial corpus. However, a single paper database is insufficient to meet the data requirements of this study. Consequently, during the data collection process, this study leverages multiple data sources, including the ACL Anthology Corpus¹, OpenAlex²(Priem et al., 2022) and Semantic Scholar³(Kinney et al., 2023), to gather the necessary data.

Ultimately, we successfully collected 675 target papers along with their corresponding 22,647 references. To ensure data quality, further filtering was

¹ <https://aclanthology.org/>

² <https://openalex.org/>

³ <https://www.semanticscholar.org/>

applied to exclude target papers with fewer than 10 citations, fewer than 20 references, or missing author information. After rigorous screening, the final dataset consists of 144 target papers, 6,153 references, 953 author profiles, and 25,906 papers published by the corresponding authors.

In this dataset, the data fields for the target papers and their references include the paper titles and abstracts. The author information fields encompass research interests, affiliated institutions, publication counts, citation counts, and the papers they have published. Additionally, to protect privacy, sensitive information such as names in the dataset has been appropriately anonymized.

Initial Research Idea Generation

This study begins by randomly selecting a target paper to define the direction for research idea generation and to determine the scale of the multi-agent team. Drawing on (Sadler-Smith, 2015) framework, which divides the creative process into four stages—preparation, incubation, insight, and verification—the initial research idea generation phase aims to prepare and incubate ideas, laying the groundwork for subsequent iterations by the agents to produce truly novel ideas. To this end, an initial idea generation module is designed, emphasizing diversity and novelty as foundational principles. Upon receiving the input paper, the large language model (LLM) utilizes its references and scientific discovery theories to generate ideas.

To enhance the scientific rigor and diversity of the initial research ideas, this study adopts an approach inspired by Nova (Hu et al., 2024), employing ten scientific discovery methods to constrain and stimulate the LLM. These methods guide the LLM to generate innovative ideas based on the input paper and its references. For example, leveraging Pierce's hypothetico-deductive method, the model starts with facts and propositions, formulates a hypothesis or premise, and then conducts logical reasoning to derive conclusions. By analyzing the relationships between premises, the validity and truth value of the conclusions can be assessed.

In alignment with the creative process, the study utilizes the internal knowledge of the LLM to stimulate idea generation, ensuring that the model comprehends the input paper and its references, evaluates them, and provides reasoning and thought processes to maintain interpretability (Wei et al., 2022). Finally, 15 initial research ideas are generated, forming an idea pool to facilitate subsequent iterations.

To formalize the prompting process, this study defines P as the target paper, L as its references, T as the scientific method theory, and R as the generated research idea. Thus, the initial research idea generation can be expressed as:

$$R = f(P, L, T) \quad (1)$$

Where f represents the large language model, leveraging its language comprehension capabilities to generate research ideas. The prompt templates and examples for initial research idea generation are provided in Appendix Tables 1 and 2.

Iterative Refinement of Ideas

Previous methods have predominantly relied on keyword-based searches or co-occurrence of entity concepts to incorporate external knowledge. However, these approaches exhibit significant limitations, such as inaccurate or overly broad search results, which hinder the ability of large language models (LLMs) to engage in deep reasoning (Hu et al., 2024).

To effectively address these shortcomings, this study integrates planning principles into the knowledge search phase of research idea generation. Specifically, the LLM is utilized to meticulously plan and design knowledge search tasks, which are then executed sequentially using external academic search APIs. Ultimately, knowledge from diverse domains that is closely related to the research idea is combinatorially integrated into the LLM's prompts, providing more targeted and novel composite knowledge for idea generation. The prompt templates and examples for knowledge planning and search are provided in Appendix Tables 3 and 4.4.

The multi-agent system constructed in this study comprises multiple agents, each endowed with background knowledge of real-world scientists, denoted as $S = [s_1, s_2, \dots, s_n]$, where represents the entire scientific agent team and n 为 denotes the team size. The background knowledge of these scientific agents is derived from the author team information of the target papers. In the iterative process of research idea generation, these agents simulate the Delphi method, a widely recognized practice in innovation. Specifically, upon acquiring new knowledge, each agent proposes its own research ideas and conducts self-evaluation and scoring based on best practices from AI conference reviews (e.g., ICLR and ACL) (Si et al., 2024). The scoring criteria are provided to each agent as contextual prompts. Detailed scoring guidelines can be found in Appendix Table 5.

The research ideas generated by each scientific agent are evaluated for their creative quality using a Swiss System Tournament and a zero-shot large language model (LLM) ranker. The ranker employs a pairwise comparison approach to determine which idea is superior. Each idea undergoes five rounds of comparison, with a score of 1 point awarded for each win. Empirical evidence suggests that this quality assessment method outperforms direct comparison approaches (Lu et al., 2024). Ultimately, ideas scoring 5 points or higher are selected as the final output of the current iteration. Additionally, the negative feedback recorded during the comparison process is carried forward, along with the selected final ideas, into the next iteration. The prompting process for each scientific agent can be expressed as:

$$R_i = f(R_t, K, B) \quad (2)$$

Where, R_i represents the research idea generated by the i -th scientific agent, R_t denotes the research idea generated in the t -th iteration, K signifies the new knowledge acquired through planning and search, and B represents the feedback from the research ideas generated in the t -th iteration. The prompt templates and examples for research idea generation can be found in Appendix Tables 6 and 7, while the prompts for research idea comparison are provided in Appendix Table 8.

In each iteration, newly generated research ideas replace the older ones. Through this mechanism, the agents in this study are able to conduct more in-depth research exploration, significantly expanding the boundaries of the search space.

Abstract Generation

After T iterations, the final research ideas are established. In this process, the study draws on the summary generation method proposed by VIRSCI (Su et al., 2024). Specifically, the finalized research ideas are input into the large language model (LLM) with a rigorously defined summary format (including aspects such as objectives and problems, methods, expected results, and conclusions), ensuring that the research ideas are presented in a detailed and structured manner. Additionally, since the summaries will subsequently be compared with reference paper abstracts for evaluation, outputting the research ideas in summary form is both practical and aligned with the assessment requirements. The prompt templates and examples for research idea summary generation can be found in Appendix Tables 9 and 10.

Experiments and Results Analysis

This section conducts comprehensive experiments to evaluate the effectiveness of the multi-agent iterative planning and search strategy, followed by an in-depth analysis and interpretation of the results.

Experimental Setup

Large Language Model Configuration

We implement the proposed method within the multi-agent application framework Agentscope⁴(Gao et al., 2024). The large language model (LLM) employed in this study is DeepSeek-V3⁵, which has demonstrated superior performance across multiple benchmarks compared to other open-source models such as Qwen2.5-72B⁶ and Llama-3.1-405B⁷. Additionally, its performance is on par with world-leading proprietary models, including GPT-4o⁸ and Claude-3.5-Sonnet⁹ (Liu et al., 2024).

Baselines

To demonstrate the effectiveness of the proposed method, this study selects state-of-the-art approaches as baselines, including AI-Researcher¹⁰(Si et al., 2024). This method introduces an end-to-end framework for generating research ideas using large language models (LLMs) and demonstrates that LLM-generated ideas are more novel than those produced by human experts.

⁴ <https://github.com/modelscope/agentscope>

⁵ <https://platform.deepseek.com/>

⁶ <https://huggingface.co/Qwen/Qwen2.5-72B>

⁷ <https://huggingface.co/meta-llama/Llama-3.1-405B>

⁸ <https://openai.com/index/hello-gpt-4o/>

⁹ <https://www.anthropic.com/claude/sonnet>

¹⁰ <https://github.com/NoviSci/AI-Researcher>

Evaluation Metrics

Drawing on the evaluation methodologies of AI-Researcher(Si et al., 2024) and Nova(Hu et al., 2024) , We assesses the research ideas generated by LLMs from three perspectives: quality score, diversity, and novelty.

(1) Quality Score: The quality of research ideas is evaluated using a Swiss System Tournament and a zero-shot LLM ranker. Specifically, the ranker employs a pairwise comparison approach to determine which idea is superior. Each idea undergoes five rounds of comparison, with 1 point awarded for each win. This quality assessment method has been empirically shown to outperform direct comparison or scoring approaches (Lu et al., 2024) . Ideas scoring above 5 points are considered high-quality. The quality score is ultimately measured by the proportion of high-quality ideas, calculated as follows:

$$HightScoreRatio = \frac{\sum_i^n I(s_i \geq 5)}{n} \quad (3)$$

Where, n represents the total number of generated research ideas, s_i denotes the score of the i -th idea, and $I(s_i \geq 5)$ is an indicator function that equals 1 when $s_i \geq 5$ and 0 otherwise.

(2) Novelty: We employ semantic similarity to assess the novelty of research ideas generated by large language models. Specifically, we first use a text embedding model to convert the generated research ideas and relevant literature into vector representations, then calculate the similarity between them. If the similarity falls below a predefined threshold, the idea is considered novel. This approach has been widely adopted in the evaluation of research idea generation(Hu et al., 2024; Kumar et al., 2024; Si et al., 2024). Additionally, the all-MiniLM-L6-v2¹¹ model is used for embedding, with a cosine similarity threshold of 0.5 to determine similarity. The novelty score is calculated as follows:

$$Novelty = \frac{\sum_{i=1}^n I(\max_j sim(a_i, r_{ij}) < \theta)}{n} \quad (4)$$

Where, n represents the total number of generated research ideas, $sim(a_i, r_{ij})$ denotes the cosine similarity between the i -th idea a_i and its related literature r_{ij} , $I(\square)$ is an indicator function that returns 1 if the condition is true and 0 otherwise.

(3) Diversity: Similar to(Hu et al., 2024; Si et al., 2024) , the diversity of generated research ideas is measured by the proportion of unique ideas. Specifically, the same similarity metric used for novelty assessment is applied, with a duplication threshold set at 0.8. The diversity score is calculated as follows:

¹¹ <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

$$Diversity = \frac{1}{n} \sum_{i=1}^n I(\max_{j \neq i} sim(i, j) < threshold) \quad (5)$$

Where, n represents the total number of generated research ideas, $sim(i, j)$ denotes the cosine similarity between the i -th idea and j -th idea, $threshold$ represents the similarity threshold, and $I(\square)$ is an indicator function that returns 1 if the condition is true and 0 otherwise.

During the evaluation process, this method randomly selects 5 papers for each team size ranging from 2 to 8 members, totaling 35 papers, and generates 525 initial research ideas. Each baseline method also produces 5 sets of data, resulting in 75 research ideas for evaluation. The final assessment is based on the average proportion of high-quality ideas, average novelty, and average diversity scores.

Throughout the experiments, the multi-agent iterative planning and search strategy, after three iterations, cumulatively generated 2,027 research ideas. Specifically, the first iteration produced 568 ideas, the second iteration generated 656 ideas, and the third iteration yielded 803 ideas. In detail, teams of 8 members contributed 126 ideas, 7-member teams generated 113 ideas, 6-member teams produced 107 ideas, 5-member teams contributed 97 ideas, 4-member teams formed 79 ideas, 3-member teams created 77 ideas, and 2-member teams generated 75 ideas. The trend in the average number of ideas per team size is illustrated in Figure 2. Clearly, larger team sizes result in a greater number of research ideas after filtering through the LLM's self-evaluation process.

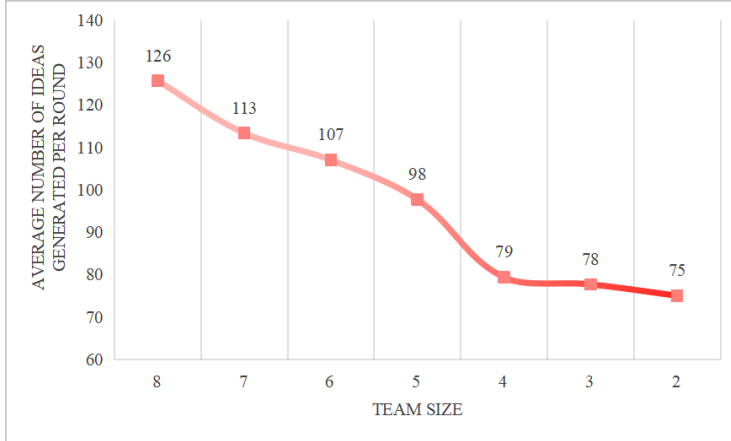


Figure 2. Average Number of Ideas Generated per Team per Iteration.

Comparison with Baseline Methods

We answer RQ1 in this section. Following the methodology of AI-Researcher (Si et al., 2024), we fully replicate their approach and, to ensure consistency with our method, generate 5 sets of data to obtain 75 research ideas for analysis. For our proposed method, we use the average performance metrics across iterations to ensure

a fair comparison. The results are presented in Figure 3. Our method outperforms AI-Researcher in both the average diversity ratio and the proportion of high-quality ideas, while also demonstrating a slight advantage in the average novelty ratio. These findings indicate that the multi-agent iterative planning and search strategy can effectively enhance the diversity and novelty of research ideas generated by large language models.

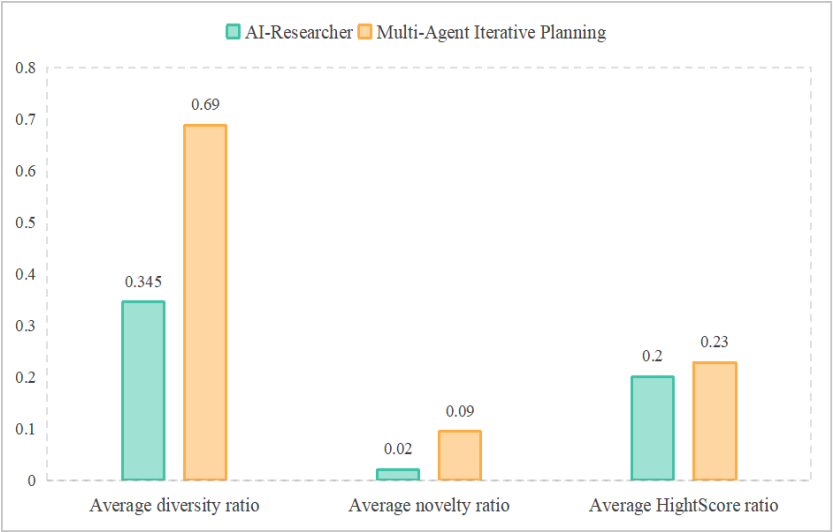


Figure 3. Comparison with Baseline Methods.

Impact of Agent Team Size on Performance Metrics

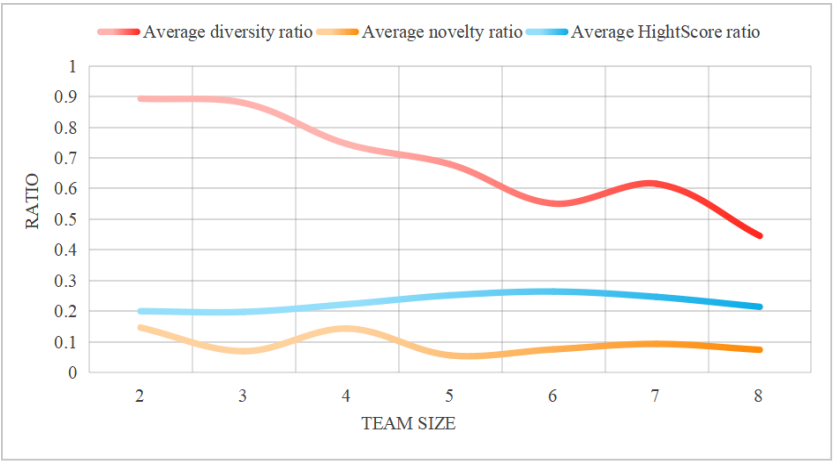


Figure 4. Trend of Metrics Across Different Team Sizes.

We examine the impact of varying agent team sizes on performance metrics by analyzing the best-performing third iteration results. As shown in Figure 4, for diversity, as the team size increases from 2 to 8, the uniqueness ratio exhibits an overall declining trend, starting from a relatively high level and gradually decreasing.

This suggests that larger team sizes may lead to a reduction in uniqueness, which is likely related to the inherent knowledge limitations of large language models (LLMs). Generating more content increases the likelihood of similarity, indicating that expanding the scale of multi-agent systems does not necessarily enhance the uniqueness of LLM-generated content. This reflects a trade-off between quality and uniqueness.

For novelty, no clear trend is observed in relation to team size. However, the overall values remain relatively low and stable, indicating that team size has an insignificant impact on novelty. This further suggests that the proposed method cannot improve novelty by scaling up the number of agents.

The proportion of high-quality ideas fluctuates between 0.2 and 0.3 as the team size varies from 2 to 8, without showing a clear linear increase or decrease. However, in local variations:

Small teams (team size of 2-3): The proportion of high-quality ideas is relatively low, around 0.2. This may be due to limited resources and manpower in smaller teams, making it difficult to achieve high performance across all aspects.

Medium-sized teams (team size of 4-7): The proportion of high-quality ideas increases and stabilizes around 0.25. At this scale, teams may achieve a better balance in personnel allocation and collaboration, leading to improved overall performance.

Large teams (team size of 8): The proportion of high-quality ideas drops back to around 0.2. This may be attributed to increased management complexity and communication costs in larger teams, which can negatively impact overall efficiency and quality.

These findings align with the conclusion that an optimal team size can facilitate the generation of impactful research (L. Wu et al., 2019).

Impact of Iteration Count on Performance Metrics

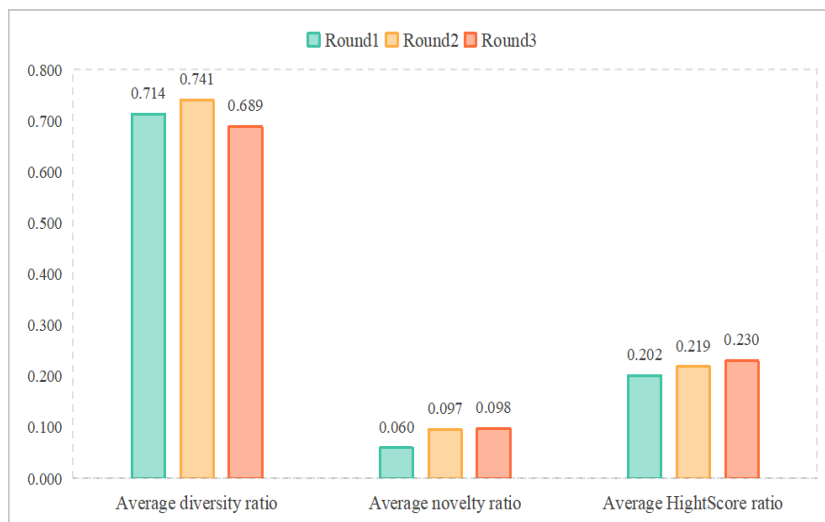


Figure 5. Impact of Iteration Count on Average Metrics.

As shown in Figure 5, the number of iterations has a significant impact on all metrics. The average diversity ratio peaks during the second iteration and then slightly declines. The average novelty ratio shows a notable improvement in the second iteration, with a marginal increase in the third iteration. Meanwhile, the proportion of high-quality ideas gradually rises with each iteration. These results suggest that the proposed method retains potential for generating high-quality ideas, though it exhibits some limitations in terms of novelty and diversity.

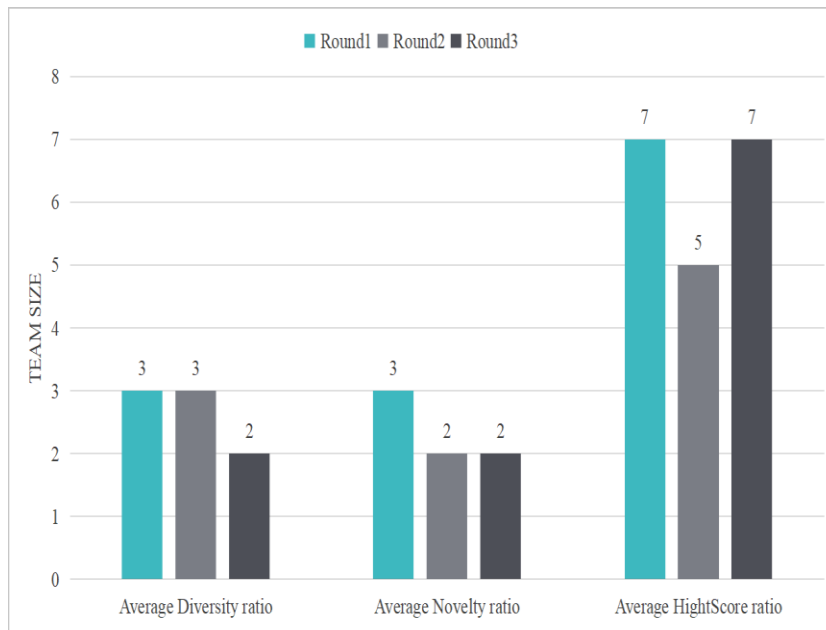


Figure 6. Variation in Team Size Corresponding to the Best Metrics per Iteration.

Across different iteration counts, as illustrated in Figure 6, the best performance in diversity and novelty metrics consistently occurs in smaller teams, while the highest proportion of high-quality ideas is consistently achieved by teams of 5-7 members. This indicates that the multi-agent strategy holds promise for enhancing the quality of research ideas generated by large language models.

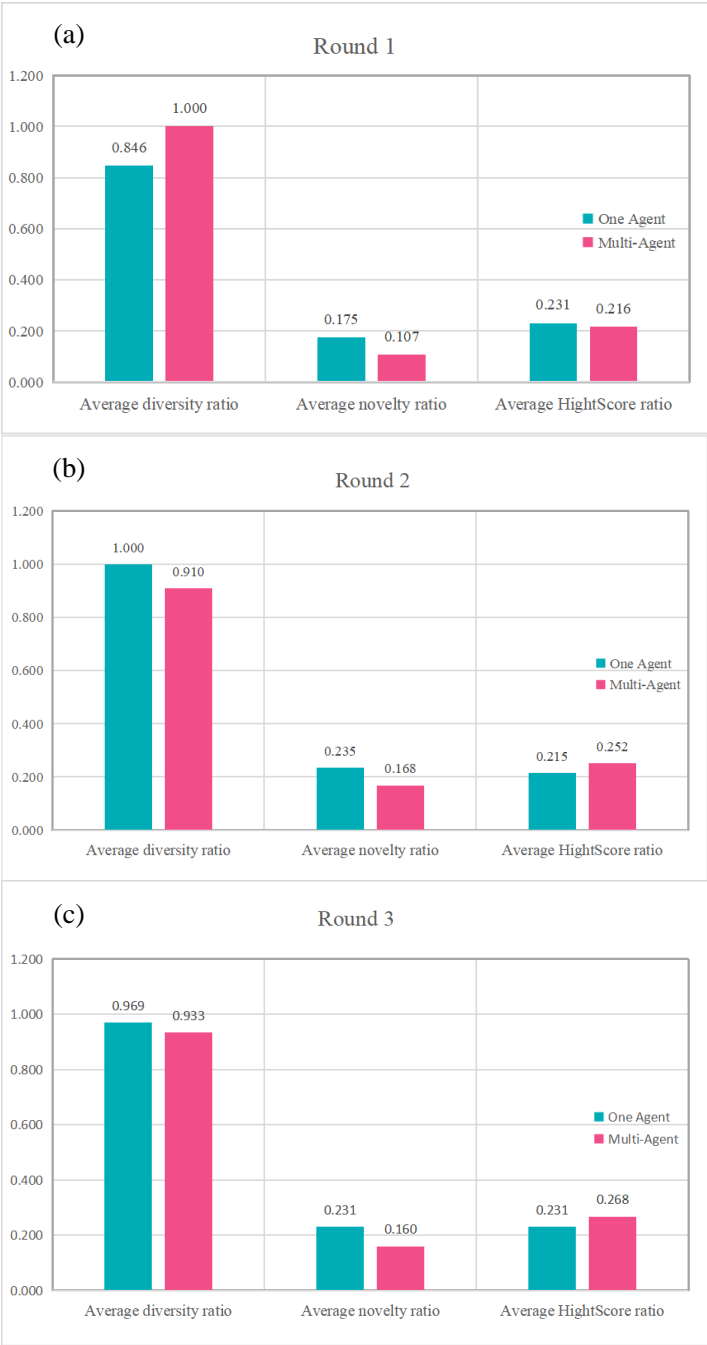


Figure 7. Comparative Performance of Single-Agent vs. Multi-Agent Systems Across Iterations.

We answer RQ2 in this section. The multi-agent iterative planning and search strategy integrates two core modules: knowledge planning and search, and multi-agent generation. A key objective of this study is to determine which module plays

a decisive role in influencing critical metrics. To this end, we set the number of agents to 1, focusing on the impact of a single agent combined with knowledge planning and search on research idea generation. The best performance of the multi-agent iterative planning and search strategy is used as a benchmark for comparison. As shown in Figures 7(a), 7(b), and 7(c), the single-agent approach outperforms in terms of diversity and novelty metrics. This suggests that the knowledge planning and search module positively enhances the generative capabilities of large language models (LLMs), indicating that combinatorial knowledge effectively guides LLMs. However, we also observe a declining trend in the performance of the single-agent system as the experiment progresses, suggesting that it may encounter bottlenecks in the research idea generation process. This finding aligns with the conclusions of Hu et al. (2024) in their study on Nova, where performance similarly plateaued after a certain number of iterations.

In contrast, while the multi-agent system slightly underperforms in diversity and novelty compared to the single-agent approach, it demonstrates significant advantages in the quality of generated ideas. Notably, the multi-agent system exhibits a consistent upward trend across all metrics. This indicates the potential of multi-agent systems and highlights the feasibility of incorporating innovative methodologies. In other words, combinatorial innovation theory and methodological approaches can effectively guide LLMs in the task of generating research ideas.

Discussion

Research Implications

Theoretical Implications

This study applies combinatorial innovation theory to the task of research idea generation, proposing a novel methodological framework aimed at enhancing the diversity and novelty of research ideas generated by large language models (LLMs). Experimental results demonstrate that the proposed method consistently outperforms baseline approaches across key evaluation metrics, including diversity, novelty, and quality scores. This underscores the effectiveness of systematically combining knowledge from diverse domains and employing multi-agent systems to conduct 'brainstorming' sessions. Furthermore, it validates the feasibility of applying combinatorial innovation theory and practical innovation methodologies to the task of research idea generation.

In the ablation study, we compared the individual contributions of the knowledge planning module and the multi-agent system module. The findings reveal that the knowledge planning and search module positively influences the generative capabilities of LLMs, confirming that combinatorial knowledge effectively guides LLMs. Additionally, the multi-agent system, unlike the single-agent approach, maintains an upward performance trend over more iterations, suggesting that practical innovation methodologies can effectively guide LLMs in performing complex reasoning tasks. In summary, combinatorial innovation theory and methodologies are well-suited for the task of research idea generation.

Practical Implications

The multi-agent iterative planning and search strategy significantly enhances the performance of research idea generation tasks in terms of diversity, novelty, and quality. This suggests that adopting a collaborative approach involving multiple modules or agents often yields superior outcomes when conducting research idea generation tasks. Furthermore, leveraging theoretical frameworks to guide the design of each module is crucial. Such an approach not only ensures a more scientific practice but also enhances the interpretability of the results, facilitating a deeper understanding and application of the research findings.

Limitations

In this study, we employ knowledge planning and search alongside a multi-agent system to simulate human innovation processes, aiming to enhance the innovative capabilities of large language models (LLMs) in generating research ideas. Despite promising results, this work has several limitations.

Incomplete Evaluation Metrics: Although the evaluation metrics used in this study encompass novelty, diversity, and quality comparison, they do not account for the value, feasibility, or historical impact of the generated research ideas. This may limit the applicability of the findings.

Lack of Human Expert Evaluation: Although the automated evaluation results indicate that the proposed method in this paper excels in enhancing the novelty and diversity of research ideas generated by large language models, relying solely on automated metrics makes it difficult to comprehensively validate the method's reliability. In subsequent research, we will incorporate human expert evaluation to further verify the practical effectiveness of the method.

Cross-disciplinary applicability remains to be verified: Although this study has demonstrated the effectiveness of multi-agent iterative search strategies for generating research ideas with large language models in the field of natural language processing (NLP), the success in a single discipline is insufficient to prove the universality of this approach. This limitation may hinder its broader application in other subfields of computer science or even more extensive disciplines (such as life sciences, physics, etc.).

Absence of Reward Functions: While the process of generating research ideas with LLMs incorporates combinatorial knowledge from multiple domains, it relies solely on the inherent capabilities of the model without introducing reward mechanisms to guide the generation process. This could potentially impact the quality of the generated ideas.

Conclusion and Future Research Directions

This study introduces combinatorial innovation theory into the task of research idea generation and proposes a multi-agent iterative planning and search strategy that integrates multi-domain knowledge planning and search with a multi-agent system. Experimental results demonstrate that this method outperforms baseline approaches, enhancing the diversity and novelty of generated research ideas. Furthermore, it provides a theoretical explanation, grounded in combinatorial innovation theory and

methodologies, for why the proposed method improves the diversity and novelty of ideas generated by language models, offering new perspectives for future research on idea generation tasks.

Future research efforts will focus on three key directions: (1) Enhancing the research idea generation capability of large language models (LLMs) through fine-grained knowledge entity recombination techniques; (2) Establishing a multi-dimensional evaluation framework to systematically validate the academic value and practical effectiveness of generated content; (3) Constructing domain-specific knowledge graphs to constrain and guide LLM generation processes, thereby effectively mitigating hallucination phenomena.

Ethical Statement

All literature data and author background information used in this study were sourced from publicly available academic databases, and none of the content involves personal privacy or sensitive information. During data processing, we strictly adhered to the terms of use and academic ethics guidelines of each database to ensure no risk of privacy breaches. To protect scholars' personal information security, all author names were anonymized during the analysis.

It is important to emphasize that the system developed in this study is solely intended to assist scientific research. Its design purpose is to provide research support for scholars, not to replace human researchers. Throughout its operation, the system emphasizes the importance of human oversight mechanisms, ensuring the quality of research results through human-machine collaboration.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

- Ajith, A., Xia, M., Chevalier, A., Goyal, T., Chen, D., & Gao, T. (2024). Litsearch: A retrieval benchmark for scientific literature search. *arXiv preprint arXiv:2407.18940*.
- Baek, J., Jauhar, S. K., Cucerzan, S., & Hwang, S. J. (2024). Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.
- Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., & Nyczyk, P. (2024). *Graph of thoughts: Solving elaborate problems with large language models*. Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.
- Boden, M. A. (2004). *The creative mind: Myths and mechanisms*: Routledge.
- Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Gao, D., Li, Z., Pan, X., Kuang, W., Ma, Z., Qian, B., Wei, F., Zhang, W., Xie, Y., & Chen, D. (2024). Agentscope: A flexible yet robust multi-agent platform. *arXiv preprint arXiv:2402.14034*.

- Gu, T., Wang, J., Zhang, Z., & Li, H. (2024). LLMs can realize combinatorial creativity: generating creative ideas via LLMs for scientific research. *arXiv preprint arXiv:2412.14141*.
- Gu, X., & Krenn, M. (2024). Generation and human-expert evaluation of interesting research ideas using knowledge graphs and large language models. *arXiv preprint arXiv:2405.17044*.
- Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Xie, E., Bekiranov, S., & Zhang, A. (2024). IdeaBench: Benchmarking Large Language Models for Research Idea Generation. *arXiv preprint arXiv:2411.02429*.
- Hu, X., Fu, H., Wang, J., Wang, Y., Li, Z., Xu, R., Lu, Y., Jin, Y., Pan, L., & Lan, Z. J. a. p. a. (2024). Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.
- Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., & Cohan, A. (2023). The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199-22213.
- Kumar, S., Ghosal, T., Goyal, V., & Ekbal, A. (2024). Can Large Language Models Unlock Novel Scientific Research Ideas? *arXiv preprint arXiv:2409.06185*.
- Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and technology*, 59(2), 288-296.
- Lee, Y.-N., Walsh, J. P., & Wang, J. J. R. p. (2015). Creativity in scientific teams: Unpacking novelty and impact. 44(3), 684-697.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- Li, R., Jing, L., Han, C., Zhou, J., & Du, X. (2024). Learning to Generate Research Idea with Dynamic Control. *arXiv preprint arXiv:2412.14626*.
- Linstone, H. A., & Turoff, M. (1975). *The delphi method*: Addison-Wesley Reading, MA.
- Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., & Ruan, C. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Liusie, A., Manakul, P., & Gales, M. J. (2023). Zero-shot nlg evaluation through pairware comparisons with llms. *arXiv preprint arXiv:2307.07889*.
- Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.
- Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
- Pu, K., Feng, K., Grossman, T., Hope, T., Mishra, B. D., Latzke, M., Bragg, J., Chang, J. C., & Siangliulue, P. (2024). IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. *arXiv preprint arXiv:2410.04025*.
- Sadler-Smith, E. J. C. r. j. (2015). Wallas' four-stage model of the creative process: More than meets the eye? , 27(4), 342-352.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

- Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., & Barsoum, E. (2025). Agent Laboratory: Using LLM Agents as Research Assistants. *arXiv preprint arXiv:2501.04227*.
- Schumpeter, J. A., & Swedberg, R. (2021). *The theory of economic development*: Routledge.
- Sharpies, M. (2013). An account of writing as creative design *The science of writing* (pp. 127-148): Routledge.
- Shi, F., & Evans, J. J. N. C. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *14*(1), 1641.
- Si, C., Yang, D., & Hashimoto, T. (2024). Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.
- Su, H., Chen, R., Tang, S., Zheng, X., Li, J., Yin, Z., Ouyang, W., & Dong, N. (2024). Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:2410.09403*.
- Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.
- Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. J. S. (2013). Atypical combinations and scientific impact. *342*(6157), 468-472.
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- Wu, L., Wang, D., & Evans, J. A. J. N. (2019). Large teams develop and small teams disrupt science and technology. *566*(7744), 378-382.
- Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D. (2023). *Large language models are diverse role-players for summarization evaluation*. Paper presented at the CCF International Conference on Natural Language Processing and Chinese Computing.
- Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., & Anandkumar, A. (2024). Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Yuan, J., Yan, X., Shi, B., Chen, T., Ouyang, W., Zhang, B., Bai, L., Qiao, Y., & Zhou, B. (2025). Dolphin: Closed-loop Open-ended Auto-research through Thinking, Practice, and Feedback. *arXiv preprint arXiv:2501.03916*.
- Zhang, Z., Zhang, A., Li, M., & Smola, A. J. a. p. a. (2022). Automatic chain of thought prompting in large language models.
- Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., & Le, Q. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zhu, K., Wang, J., Zhao, Q., Xu, R., & Xie, X. (2024). *Dynamic Evaluation of Large Language Models by Meta Probing Agents*. Paper presented at the Forty-first International Conference on Machine Learning.

Table 1. Prompt for initial Research Idea Generation.

<p>System prompt: You are an expert researcher in AI. Your goal is to propose some innovative and valuable research ideas based on the target paper.</p>
<p>Follow these steps to generate innovative research ideas for exploration: Understand the Target Paper and Related Works: Target Paper: This is the core research study you aim to enhance or build upon. It serves as the foundation for identifying and developing new research ideas. Referenced Papers: These are studies cited by the target paper, providing additional context and insights directly relevant to the primary research topic. They are crucial for understanding and expanding upon the target paper. Leverage Scientific Discovery Theories: Select appropriate scientific discovery theories and combine them with insights from the target paper to generate creative, impactful, and feasible research ideas. Explore Scientific Discovery Methodologies: Below are 10 general laws and methodologies of scientific discovery from the philosophy of science. Choose one or more of these methodologies to propose new research ideas for the target paper: {scientific_discovery_theory} Select and Propose New Ideas: Identify the 5 most suitable theories or methods for the target paper and propose 5 new research ideas based on them. Requirements: Output: Provide 5 new research ideas worth exploring. Quality: Aim for ideas with the potential to win best paper awards at top conferences like ACL, NeurIPS, ICLR, and CVPR. Relevance: Skip theories that do not align well with the target paper. Ensure the chosen theories and methods are logical and appropriate for the research context. Input: Target Paper: {target_paper} References: {references} Response Format: Thought: <THOUGHT> IDEA: ```json<JSON> ``` In <THOUGHT>, explain your thought process, including why you chose specific theories and why the proposed ideas have the potential to win best paper awards at top conferences. In <JSON>, provide the new idea with the following fields: Title: A concise title for the idea, suitable for a research paper. Idea: A detailed description of the idea, including its significance and potential impact.</p>

Thinking: A detailed explanation of the thought process, including the scientific discovery theory used.
 Rationale: A detailed justification for why this idea is innovative and feasible.
 Ensure the JSON format is precise, as it will be automatically parsed.

Table 2. An example for initial research idea generation.

Title: Beyond Linguistic Features: Leveraging Cognitive Biases for Machine-Generated Text Detection
Idea: This idea proposes a novel approach to detecting machine-generated text by leveraging cognitive biases inherent in human language processing. While current methods focus on linguistic features, this approach hypothesizes that machine-generated text lacks subtle cognitive biases that humans exhibit, such as confirmation bias, anchoring, and availability heuristic. By training models to detect these biases, we can create a more robust detector that performs well even when linguistic distinctions are minimal. The idea involves collecting a dataset of human and machine-generated texts annotated for cognitive biases, developing a model to detect these biases, and evaluating its performance in diverse scenarios.
Thinking: This idea is inspired by Pierce’s hypothetical deduction method and Kuhn’s paradigm theory. The hypothesis is that machine-generated text lacks cognitive biases, which are subtle but consistent features of human language. By shifting the paradigm from linguistic features to cognitive biases, we can address the anomaly of decreasing linguistic distinctions. The rationale is that while machines can mimic linguistic patterns, they cannot replicate the nuanced cognitive biases that humans naturally exhibit.
Rationale: Current detection methods struggle with out-of-distribution texts and are limited by the decreasing linguistic distinctions between human and machine-generated text. By focusing on cognitive biases, which are deeply ingrained in human language, we can create a more robust detection method that is less reliant on surface-level linguistic features. This approach has the potential to significantly improve detection accuracy in real-world scenarios, making it a strong candidate for a best paper award.

Table 3. Prompt for knowledge planning and search.

System prompt: You are now a senior scholar in the field of artificial intelligence, and you need to develop a search strategy based on a given research idea
<p>Objective: To analyze a given research idea and determine which fields of papers should be retrieved to collect comprehensive information and new knowledge. This will provide a basis for further research and the discovery of new ideas.</p> <p>Input: Research Idea: {idea}</p> <p>Response Format: Thought: <THOUGHT> Search Plans: ```json<JSON> ```</p>

<p>In <THOUGHT>, explain the reasoning behind the search strategy, including why specific fields and keywords were chosen.</p> <p>In <JSON>, provide a detailed search plan with the following fields:</p> <p>Search Plan: A brief description of the overall search strategy.</p> <p>Search Fields: A list of fields relevant to the research idea.</p> <p>Search Keywords: A structured list of keywords for each field, formatted as:</p> <pre>"Search_Keywords": [{"Field": "field_name", "Keywords": ["keyword1", "keyword2", "keyword3"]}]</pre> <p>Ensure the JSON format is precise, as it will be automatically parsed.</p>

Table 4. An example for knowledge planning and search.

Search Plan: "To gather comprehensive information, we will search for papers in cognitive biases, machine-generated text detection, NLP, dataset creation, and model evaluation. Each field will be explored using specific keywords to ensure relevance and depth."
Search Keyword: [{"Field": "Cognitive Biases in Human Language Processing", "Keywords": ["confirmation bias", "anchoring bias", "availability heuristic", "cognitive biases in language", "human language processing"]}, {"Field": "Machine-Generated Text Detection", "Keywords": ["machine-generated text", "AI-generated text", "text detection", "linguistic features", "deepfake text"]}, {"Field": "Natural Language Processing (NLP)", "Keywords": ["text analysis", "language models", "NLP techniques", "text classification", "semantic analysis"]}].
Thought: The research idea revolves around detecting machine-generated text by leveraging cognitive biases in human language processing. To develop a comprehensive search strategy, we need to explore several interconnected fields: (1) Cognitive Biases in Human Language Processing to understand the theoretical foundation of the biases, (2) Machine-Generated Text Detection to review existing methods and their limitations, (3) Natural Language Processing (NLP) for techniques to analyze and model text, (4) Dataset Creation and Annotation for methodologies to collect and label data, and (5) Model Evaluation for metrics and frameworks to assess detector performance. This strategy ensures a holistic understanding of the problem and provides a foundation for developing a novel approach.

Table 5. Scoring rubric.

<p>Novelty: Whether the idea is creative and different from existing works on the topic, and brings fresh insights. You are encouraged to search for related works online. You should consider all papers that appeared online prior to 2024 as existing work when judging the novelty. A rating from 1 to 10. Here are the grading rules:</p> <p>1. Not novel at all - there are many existing ideas that are the same</p>

- 2.
3. Mostly not novel - you can find very similar ideas
- 4.
5. Somewhat novel - there are differences from existing ideas but not enough to turn into a new paper
6. Reasonably novel - there are some notable differences from existing ideas and probably enough to turn into a new paper
- 7.
8. Clearly novel - major differences from all existing ideas
- 9.
10. Very novel - very different from all existing ideas in a very interesting and clever way

Feasible: How feasible it is to implement and execute this idea as a research project? Specifically, how feasible the idea is for a typical CS PhD student to execute within 1-2 months of time. You can assume that we have rich API resources, but only limited hardware resources. A rating from 1 to 10. Here are the grading rules:

1. Impossible: the idea doesn't make sense or the proposed experiments are flawed and cannot be implemented
- 2.
3. Very challenging: there are flaws in the proposed method or experiments, or the experiments require compute/human resources beyond any academic lab
- 4.
5. Moderately feasible: It can probably be executed within the given time frame but would require careful planning, efficient use of APIs or some advanced computational strategies to overcome the limited GPU resources, and would require some modifications to the original proposal to make it work
6. Feasible: Can be executed within the given constraints with some reasonable planning
- 7.
8. Highly Feasible: Straightforward to implement the idea and run all the experiments
- 9.
10. Easy: The whole proposed project can be quickly executed within a few days without requiring advanced technical skills

Excitement: How exciting and impactful this idea would be if executed as a full project. Would the idea change the field and be very influential. A rating from 1 to 10. Here are the grading rules:

1. Poor: You cannot identify the contributions of this idea, or it's not interesting at all and you would fight to have it rejected at any major AI conference
- 2.
3. Mediocre: this idea makes marginal contributions and is very incremental
- 4.
5. Leaning negative: it has interesting bits but overall not exciting enough

6. Learning positive: exciting enough to be accepted at a major AI conference, but still has some weaknesses or somewhat incremental
- 7.
8. Exciting: would deepen the community's understanding or make major progress in this research direction
- 9.
10. Transformative: would change the research field profoundly and worth a best paper award at major AI conferences

Note: Some score values in the scoring rubric lack descriptions. This is because the granularity of the score levels is challenging to articulate in English. For specific details, please refer to the approach used in AI-Researcher¹².

Table 6. Prompt for research idea generation.

System prompt: Your name is Scientist0, you belong to following affiliations ['Westlake University'], you have researched on following topics ['Natural Language Processing Techniques', 'Topic Modeling', 'Multimodal Machine Learning Applications', 'Text Readability and Simplification', 'Speech and dialogue systems'], you have published 24 papers, you have 844 citations.
Task: You are an ambitious scientist who is looking to propose a new idea that will contribute significantly to the field. Improve the seed idea or come up with the next impactful and creative idea for publishing a paper that will contribute significantly to the field by integrating your own knowledge and insights with the information provided.
Here is the seed idea : "{seed_idea}"
Here are the bad reviews of the seed idea, please improve according to the reviews: "{bad reviews}"
<p>You may refer to the following listed references to design a new idea or concept. These references can serve as inspiration, but you are not allowed to directly copy or replicate their content. Ensure that your design is original and addresses a specific problem or meets a unique need. References: {references}</p> <p>"Please respond in the following format:</p> <p>Thought: <THOUGHT></p> <p>New Idea: ``json<JSON>``</p> <p>In <THOUGHT>, briefly discuss your intuitions and motivations for the idea. Justify how this idea differs from existing ones, highlighting its unique aspects.</p> <p>In <JSON>, provide the new idea with the following fields and provide as many details as possible:</p> <ul style="list-style-type: none"> - "Idea": A detailed description of the idea, outlining its significance and potential impact. - "Title": A title for the idea, will be used for the paper writing.

¹² <https://github.com/NoviSci/AI-Researcher>

<p>- "Experiment": An outline of the implementation process. Describe your high-level design plan, including necessary design steps and the ideal outcomes of the experiments.</p> <p>- "Excitement": {excitement scoring rules} - "Excitement Rationale": {prompt for rationale}</p> <p>- "Feasibility": {feasibility scoring rules} - "Feasibility Rationale": {prompt for rationale}</p> <p>- "Novelty": {novelty scoring rules} - "Novelty Rationale": {prompt for rationale}</p> <p>Be cautious and realistic on your ratings. This JSON will be automatically parsed, so ensure the format is precise, and the content should be longer than 600 words. You only need to output one idea.</p>

Table 7 (continued). An example for research idea generation.

Title: "Multimodal Few-Shot Learning for Cognitive Bias-Based Detection of Machine-Generated Text",
Idea: "This idea proposes a multimodal, few-shot learning framework for detecting machine-generated text by leveraging cognitive biases and contextual cues across text, speech, and visual modalities. Unlike existing methods that focus solely on linguistic features, this approach integrates psychological patterns (e.g., confirmation bias, anchoring) with multimodal data to create a more robust and adaptive detection system. The framework involves: (1) collecting a diverse dataset of human and machine-generated content annotated for cognitive biases and contextual features, (2) developing a multimodal deep learning model that combines text, speech, and visual embeddings, and (3) employing few-shot learning techniques to enable the model to generalize across new domains and languages. The system will be evaluated in real-time scenarios, such as social media moderation and customer service chatbots, to assess its effectiveness in distinguishing machine-generated content from human-authored text. This approach has the potential to significantly improve the accuracy and adaptability of machine-generated text detection, particularly in dynamic and multilingual environments."
Experiment : "1. Dataset Collection: Curate a multimodal dataset containing text, speech, and visual data from both human and machine-generated sources. Annotate the dataset for cognitive biases and contextual features. 2. Model Development: Design a multimodal deep learning model that integrates text, speech, and visual embeddings using transformer architectures. Incorporate few-shot learning techniques to enable domain adaptation. 3. Training and Evaluation: Train the model on the annotated dataset and evaluate its performance in detecting machine-generated text across diverse scenarios, including social media moderation and customer service chatbots. 4. Real-Time Testing: Deploy the model in real-time applications to assess its effectiveness and adaptability. Measure accuracy, recall, and F1-score as key performance metrics. 5. Comparative Analysis: Compare the

proposed model with state-of-the-art methods to demonstrate its superiority in terms of accuracy and adaptability."
Excitement: 9, Excitement Rationale: "This idea is highly exciting because it addresses a critical challenge in NLP and multimodal learning by integrating cognitive biases with advanced machine learning techniques. It has the potential to significantly improve the detection of machine-generated text in real-time applications, which is increasingly important in the era of large language models and AI-generated content. The combination of multimodal data and few-shot learning makes this approach highly adaptable and generalizable, setting it apart from existing methods."
Feasibility: 7, Feasibility Rationale: "The idea is feasible but requires careful planning and efficient use of resources. While the multimodal dataset collection and model development are achievable within 1-2 months, the real-time testing and deployment may require additional computational resources. The use of APIs and pre-trained models can help mitigate some of these challenges, but the integration of few-shot learning techniques may require advanced expertise."

Table 8. Prompt for comparison of research ideas.

System prompt: You are a reviewer specialized in Natural Language Processing. You are given two project summaries. One of them is accepted by a top AI conference (like ICLR or ACL) and the other one is rejected. Your task is to identify the one that has been accepted.
Now decide which one is the accepted idea. give a meta-review to each paper. <pre> '''json{ Decision: <DECISION> ReviewForPaper1: <ReviewForPaper1> ReviewForPaper2: <ReviewForPaper2> } ''' </pre> In <DECISION>, return a number 1 or 2 and nothing else In <ReviewForPaper1>, write the review for paper 1, reasons for acceptance or rejection. In <ReviewForPaper2>, write the review for paper 2, reasons for acceptance or rejection. This JSON will be automatically parsed, so ensure the format is precise.

Table 9. Prompt for research idea abstract generation.

System prompt: You are now a senior scholar in the field of artificial intelligence..
You are an ambitious scientist who will generate a summary based on given research idea and experimental steps. Requirements: The content of the abstract should cover: research questions and objectives, research methods, expected research results, and conclusions. Do not exceed 300 words.

Here is the research idea: "{idea}"

Here is the experimental steps: "{experiment}"

"Please respond in the following format:

Thought: <THOUGHT>

Abstract: ``json<JSON>``

In <THOUGHT>, please briefly describe your thinking.

In <JSON>, provide the abstract with the following fields:

- "Title": A title for the abstract.
- "Abstract": abstract.

Be cautious and realistic on your ratings. This JSON will be automatically parsed, so ensure the format is precise.

Table 10. An example for research idea abstract generation.

Thought: The research idea focuses on developing a robust framework for detecting AI-generated content across multiple modalities by leveraging cognitive biases and advanced fusion techniques. The experimental steps outline a systematic approach to dataset collection, model development, fusion mechanism implementation, domain adaptation, evaluation, and deployment. The abstract should succinctly capture the research objectives, methods, expected results, and conclusions while maintaining clarity and precision.

Abstract:

{"Title": "Contrastive Meta-Style Adversarial Fusion Network for Robust Detection of AI-Generated Content",

"Abstract": "This research proposes the Contrastive Meta-Style Adversarial Fusion Network (CMSAF-Net), a novel framework designed to detect AI-generated content across text, audio, and visual modalities by leveraging cognitive biases. The framework integrates modality-specific bias detection models, a contrastive hierarchical fusion mechanism, and a meta-style adversarial domain adaptation module. The methodology involves dataset collection, training of modality-specific models using advanced NLP, speech analysis, and computer vision techniques, and the development of a hierarchical fusion algorithm enhanced with contrastive learning. Additionally, meta-style adversarial training is employed to ensure generalization across diverse datasets. The framework is evaluated on a held-out dataset, with performance metrics including accuracy, precision, recall, and F1-score. Expected results indicate that CMSAF-Net will outperform existing unimodal and multimodal detection methods, providing a robust and generalizable solution for detecting synthetic media. The research concludes with the deployment of a user-friendly interface, aiming to mitigate the challenges posed by AI-generated content in misinformation campaigns."}