NOVA: An Iterative Planning Framework for Enhancing Scientific Innovation with Large Language Models

Xiang Hu[4,*], **Hongyu Fu**[5,*], **Jinge Wang**[1], **Yifeng Wang**[6], **Zhikun Li**[7]
Renjun Xu[2], **Yu Lu**[1], **Yaochu Jin**[1], **Lili Pan**[†,3], **Zhenzhong Lan**[†,1]
Westlake University[1] Zhejiang University[2]
University of Electronic Science and Technology of China[3]
China Life R&D Center[4] Carnegie Mellon University[5]
Southeast University[6] University of Oxford[7]
huxiang2022@e-chinalife.com hongyuf@andrew.cmu.edu
lilipan@uestc.edu.cn lanzhenzhong@westlake.edu.cn

Abstract

Scientific innovation is pivotal for humanity, and harnessing large language models (LLMs) to generate research ideas could transform discovery. However, existing LLMs often produce simplistic and repetitive suggestions due to their limited ability in acquiring external knowledge for innovation. To address this problem, we introduce an enhanced planning and search methodology designed to boost the creative potential of LLM-based systems. Our approach involves an iterative process to purposely plan the retrieval of external knowledge, progressively enriching the idea generation with broader and deeper insights. Validation through automated and human assessments demonstrates that our framework substantially elevates the quality of generated ideas, particularly in novelty and diversity. The number of unique novel ideas produced by our framework is 3.4 times higher than without it. Moreover, our method outperforms the current state-of-the-art, generating at least 2.5 times more top-rated ideas based on 170 seed papers in a Swiss Tournament evaluation. Our code is available at https://github.com/hflyzju/Nova

1 Introduction

In recent years, LLMs have demonstrated remarkable progress across various challenging tasks, including solving mathematical problems (Romera-Paredes et al., 2024), proving mathematical theory (Wang et al., 2024a), and generating code to solve analytical or computational tasks (Huang et al., 2024). These advances have opened up new possibilities to utilize LLMs to accelerate research (Wang et al., 2023a), including generating novel research ideas (Si et al., 2024; Wang et al., 2024b; Baek et al., 2024; Li et al., 2024a; Pu et al., 2024; Kumar et al., 2024; Guo et al., 2024).

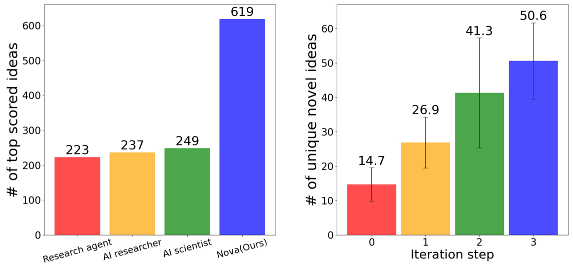

Figure 1: **Nova's Performance comparison. Left:** Performance comparison against baseline systems across Swiss Tournament scores (higher is better). **Right:** Number of unique novel ideas generated at each iteration step, showing 3.4× improvement over baseline.

Our work addresses the fundamental challenge of enabling LLMs to generate high-caliber research ideas by introducing a systematic framework that combines theoretical principles from scientific discovery with dynamic knowledge exploration. Unlike existing approaches that rely primarily on retrieval or prompt engineering, we propose a structured method for expanding both the breadth and depth of idea generation. Existing studies (Wang et al., 2024b; Si et al., 2024) have tackled this challenge by integrating additional knowledge into the idea generation process. Baek et al. (2024) enrich the process by incorporating co-occurrence entities with existing knowledge, prompting LLMs to generate ideas based on these entities. Si et al. (2024) suggest an iterative approach to retrieve topic-relevant papers through the Semantic Scholar API, utilizing retrieval-augmented generation (RAG) for idea generation. They find that *"LLM-generated ideas are judged as more novel ($p < 0.05$) than human experts"*. However, they also show that *"LLMs lack diversity in idea generation"*. We argue that this limitation stems from the constrained scope and lack of direction in knowledge acquisition within these methods.

[*] Equal contribution.

[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21330–21359
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Implementation	RAG	SI Theory Guided	Iteration	Planning Guided
SciMon	✓		✓	
ResearchAgent			✓	
AI-Researcher	✓			
AI-Scientist			✓	
Nova	✓	✓	✓	✓

Table 1: Detailed comparison with SciMon (Wang et al., 2024b), ResearchAgent (Baek et al., 2024), AI-Scientist (Lu et al., 2024) and AI-Researcher (Si et al., 2024). Here, "SI Theory Guided" means generating seed ideas under the guidance of Scientific Innovation Theory as mentioned in Section 3.1, and "Planning Guided" refers to planning guided RAG, as detailed in Section 3.2.

Broadening the scope of the search, both in terms of breadth and depth, presents a significant challenge. The crux of the issue lies in determining what knowledge to retrieve. Traditional methods of entity and keyword retrieval are not goal-oriented and frequently yield knowledge that is not conducive to fostering innovation.

In order to address the above problem, we introduce an iterative planning framework for LLM-based idea generation that specifically targets the enhancement of the novelty and diversity of the ideas produced. Starting with seed ideas that are generated using scientific innovation theories and related knowledge, our framework undergoes multiple iterations of planning and searching. In each iteration, the model is tasked with devising a search plan aimed at identifying papers from multiple fields that will enhance the novelty and diversity of the current set of ideas.

As depicted in Figure 1, the proposed iterative planning framework significantly enhances the quality of ideas generated from recent 170 LLM-related papers (from top conferences like ACL, ICLR, and CVPR). The number of high-quality ideas (as measured by the Swiss Tournament Score (Si et al., 2024)) is at least 2.5 times greater than those produced by other state-of-the-art methods. Moreover, the number of unique novel ideas generated by our iterative planning framework is 3.4 times higher compared to approaches that do not incorporate such a framework.

2 Related work

2.1 LLM-based Scientific Innovation

In the past year, several studies on LLM-based scientific innovation (Yang et al., 2024; Baek et al., 2024; Lu et al., 2024; Wang et al., 2024b; Gu and Krenn, 2024; Li et al., 2024b) have been proposed, garnering significant attention from the LLM community. Among these studies, Baek et al. (2024) introduced a research agent that utilizes an external knowledge graph for co-occurrence entity search and incorporated retrieved entities into the idea generation process. To avoid generating similar ideas, Lu et al. (2024) treat past generated ideas as negative examples and instruct the LLM on what constitutes a negative example. To explore more external knowledge for innovation, some other works (Wang et al., 2024b; Gu and Krenn, 2024) propose prompting the LLM to generate ideas integrated with external knowledge, such as retrieved external entities or problem-solution pairs.

Concurrent with our research, Si et al. (2024) introduce AI-Researcher, which, for the first time, demonstrates that LLMs can generate ideas deemed more novel than those written by human experts. In addition, they point out that using LLMs to directly evaluate different dimensions of scientific ideas is unreliable and propose an idea ranking method based on pairwise comparison, achieving an accuracy of 71.4% in distinguishing accepted and rejected submissions on real ICLR 2024 data.

Table 1 provides a detailed comparison with these methods. Although effective, the above approach often generates repetitive ideas (Si et al., 2024) due to the lack of direction in acquiring new knowledge. In contrast, our method provides a plan for searching for new knowledge and suffers less from the repetitive problem.

2.2 Reasoning and Planning

Reasoning has been proven to be an effective technique for enhancing the problem-solving capabilities of LLMs (Wei et al., 2022; Wang et al., 2023c; Yao et al., 2023a). Among them, Wei et al. (2022) propose chain of thought (CoT), which involves guiding LLMs to solve complex problems by generating a step-by-step reasoning process. Wang et al. (2023c) improve CoT by sampling and comparing diverse reasoning pathways to enhance the consistency of the reasoning process. To perform deliberate decision making, Yao et al. (2023a) propose tree of thought, enabling LLMs to explore multiple reasoning paths and conduct self-evaluations when determining the next action. To further improve exploration, Xie et al. (2024) enhance the reasoning capabilities of LLMs by introducing Monte Carlo Tree Search (MCTS) with iterative preference learning.

These methods significantly enhance the reasoning capabilities of LLMs, but rarely take into ac-

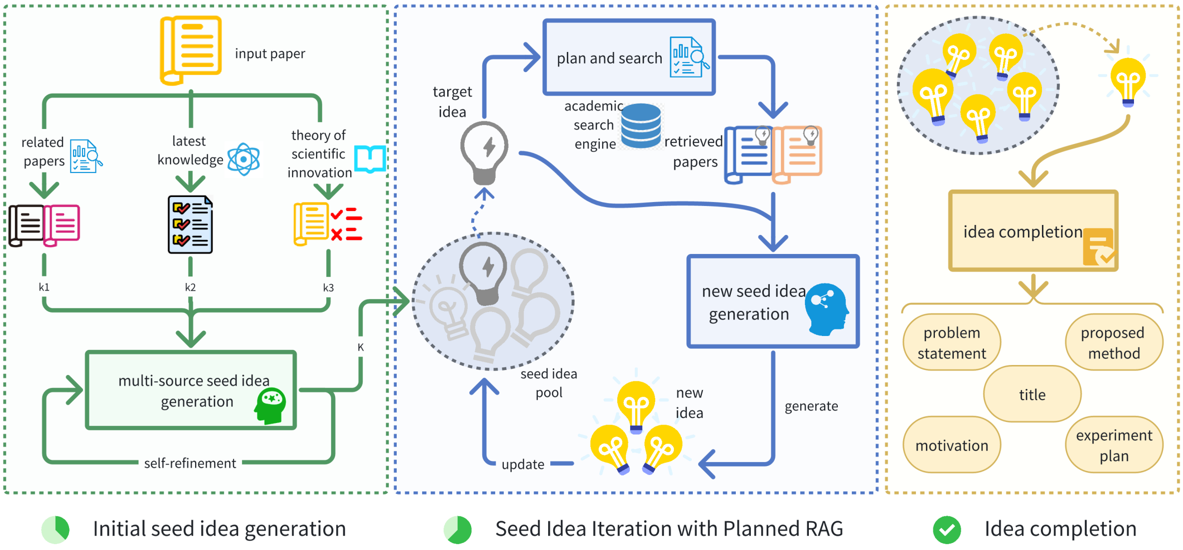

Figure 2: **Nova Pipeline.** The idea generation process begins with multi-source seed idea generation, incorporating theoretical principles, latest research trends, and domain expertise. These initial ideas undergo iterative refinement through our novel planning mechanism, which identifies and integrates knowledge from multiple distinct fields per iteration. Finally, the system expands promising ideas through structured elaboration and validation.

count the interaction with the external environment. To address this limitation, Trivedi et al. (2023) integrate CoT with knowledge retrieval, interleaving reasoning with searching to acquire additional external knowledge for knowledge-intensive question answering. Yao et al. (2023b) propose ReAct, combining reasoning and acting for solving language reasoning and decision-making tasks.

The reasoning capabilities of LLMs can also be applied to planning, such as generating plausible goal-driven action plans that can be enacted in interactive, embodied environments (Relex et al., 2022). Furthermore, plan-to-solve prompting (Wang et al., 2023b) can generate a plan that divides complex reasoning tasks into subtasks, allowing LLMs to execute each subtask according to the outlined plan. Our work marks the inaugural integration of planning methodologies into the complex domain of research tasks.

3 Nova Pipeline

Nova implements a three-stage pipeline (Figure 2) that systematically integrates scientific innovation theories with dynamic knowledge exploration. (1) The process begins with multi-source seed idea generation, incorporating theoretical principles, latest research trends, and domain expertise. (2) These initial ideas undergo iterative refinement through our novel planning mechanism, which identifies and integrates knowledge from multiple distinct fields per iteration. (3) Finally, the system expands promising ideas through structured elaboration and validation.

3.1 Initial Seed Idea Generation

To generate high-quality initial seed ideas, we propose a multi-source seed idea generation module. Upon receiving an input paper, the module activates the LLM to generate initial seed ideas, drawing on scientific innovation theories, recent research trends, and related works. The initial seed idea represents a potential direction for exploration based on the input paper, with "initial" signifying its nascent stage, subject to further refinement. In addition to prompting the LLM to generate ideas, we also ask it to outline the corresponding thinking/reasoning process behind each proposed idea, as shown in Table 2.

To incorporate the theoretical principles, we introduce ten scientific innovation theories proposed by Kuhn and other scientists (Kuhn, 1997; Doppelt, 1986; Simon, 2013; Popper, 1963; Lakatos et al., 1978; Feyerabend, 2020; Laudan et al., 1986; Resnik, 1994) to guide the identification of new research problems during the initial seed idea generation. Specifically, we input the title and abstract of the input paper, along with the theories, as context and require the LLM to identify the most appro-

Input Paper: ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models (Baek et al., 2024)

Scientific innovation theory—defining new scientific problems
Method: Identifying anomalies in existing theories, exploring theoretical boundaries and scope of application, integrating interdisciplinary knowledge and discovering new problems, and re-examining neglected historical problems.
Theoretical basis: Kuhn's paradigm theory(Kuhn, 1997), Laudan's problem-solving model(Doppelt, 1986), Nichols's problem-generation theory(Nichols and Knobe, 2007).
Thinking: By defining new scientific problems, we can explore the theoretical boundaries of ResearchAgent models. One potential area is the integration of domain-specific ontologies to enhance the contextual understanding of generated research ideas. This can lead to more accurate and relevant idea generation.
Initial seed idea: Integrate domain-specific ontologies into ResearchAgent to enhance the contextual understanding and relevance of generated research ideas.
...
Final seed idea: Dynamic Ontology Updating with LLMs to keep the ontology up-to-date with the latest research, improving the relevance of generated ideas.

Scientific innovation theory—exploring the shortcomings of current methods
Method: Critically analyzing existing methods, finding deviations between theoretical predictions and experimental results, exploring the performance of methods under extreme conditions, and interdisciplinary comparative methodology.
Theoretical basis: Popper's falsificationism (Popper, 1963), Lakatos's research program methodology (Lakatos et al., 1978), Feyerabend's methodological anarchism (Feyerabend, 2020).
Thinking: Exploring the shortcomings of current methods can reveal areas for improvement. By critically analyzing the scalability of ResearchAgent, we can investigate the use of distributed computing and parallel processing to handle larger volumes of scientific literature more efficiently.
Initial seed idea: Investigate the use of distributed computing and parallel processing techniques to enhance the scalability of ResearchAgent for handling larger volumes of scientific literature.
...
Final seed idea: Hybrid predictive analytics and real-time data streams for load balancing in ResearchAgent.

Table 2: Generated Seed Ideas based on Scientific Innovation Theory (Critical Elements Highlighted in Pink)

priate theories for the initial seed idea generation. Some examples of generated seed ideas paired with the corresponding innovation theories are shown in Table 2. See Appendix A.1 for the Prompt.

To incorporate the latest research trends, we design a knowledge tracking module to monitor recent papers on HuggingFace[†], GitHub[†], and academic preprint servers, summarizing the latest research trends for initial seed idea generation. The most influential recent papers, identified based on user engagement metrics such as likes, comments, and reposts, are collected. We leverage LLMs to summarize prevailing research trends from these papers and generate initial seed ideas that integrate these trends. See Appendix A.6 for the prompt and Appendix A.7 for a trend report example.

To incorporate the domain expertise, we prompt the LLM to assess the shortcomings of the input paper and related works, thereby proposing potential scientific problems to be addressed. The related works are retrieved from our academic search engine, and the top three relevant papers are selected.

It is worth mentioning that we employ self-refinement (Madaan et al., 2024) to avoid hallucination and improve the logicality of the initial seed ideas generated. These methods partly guarantee that the ideas generated from the multiple source are logical and reasonable.

3.2 Seed Idea Iteration with Planned RAG

Once an initial seed idea pool has been generated, we start to iteratively planning and search new knowledge according to the initial seed idea and generate new seed ideas using the acquired new knowledge.

3.2.1 Planned RAG

Unlike traditional RAG approaches that rely solely on similarity-based retrieval, our planned RAG mechanism employs: (1) LLM-guided knowledge exploration planning that identifies promising research directions and required knowledge domains, and (2) targeted retrieval that combines semantic similarity with planned exploration paths. This approach enables systematic knowledge acquisition that aligns with research objectives while maintaining exploration breadth.

In the planning phase, we guide LLM to identify key fields for comprehensive and novel knowledge acquisition to enhance further research and idea generation based on the given seed idea. The planning process and a detailed search plan example are illustrated in Figure 3 and Table 3.

[†] https://huggingface.co/papers

[†] https://github.com/dair-ai/ML-Papers-of-the-Week

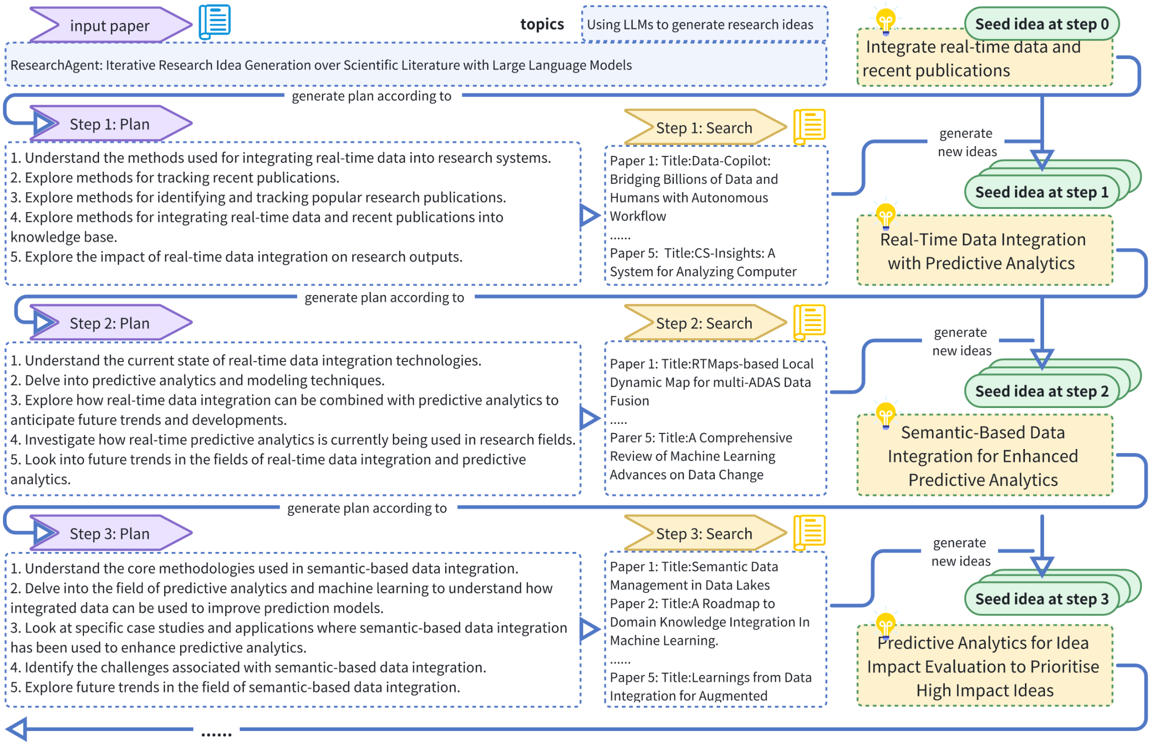
input paper
topics
Using LLMs to generate research ideas
ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models
Seed idea at step 0
Integrate real-time data and recent publications
generate plan according to
Step 1: Plan
1. Understand the methods used for integrating real-time data into research systems.
2. Explore methods for tracking recent publications.
3. Explore methods for identifying and tracking popular research publications.
4. Explore methods for integrating real-time data and recent publications into knowledge base.
5. Explore the impact of real-time data integration on research outputs.
Step 1: Search
Paper 1: Title:Data-Copilot: Bridging Billions of Data and Humans with Autonomous Workflow
......
Paper 5: Title:CS-Insights: A System for Analyzing Computer
generate new ideas
Seed idea at step 1
Real-Time Data Integration with Predictive Analytics
generate plan according to
Step 2: Plan
1. Understand the current state of real-time data integration technologies.
2. Delve into predictive analytics and modeling techniques.
3. Explore how real-time data integration can be combined with predictive analytics to anticipate future trends and developments.
4. Investigate how real-time predictive analytics is currently being used in research fields.
5. Look into future trends in the fields of real-time data integration and predictive analytics.
Step 2: Search
Paper 1: Title:RTMaps-based Local Dynamic Map for multi-ADAS Data Fusion
.....
Parer 5: Title:A Comprehensive Review of Machine Learning Advances on Data Change
generate new ideas
Seed idea at step 2
Semantic-Based Data Integration for Enhanced Predictive Analytics
generate plan according to
Step 3: Plan
1. Understand the core methodologies used in semantic-based data integration.
2. Delve into the field of predictive analytics and machine learning to understand how integrated data can be used to improve prediction models.
3. Look at specific case studies and applications where semantic-based data integration has been used to enhance predictive analytics.
4. Identify the challenges associated with semantic-based data integration.
5. Explore future trends in the field of semantic-based data integration.
Step 3: Search
Paper 1: Title:Semantic Data Management in Data Lakes
Paper 2: Title:A Roadmap to Domain Knowledge Integration In Machine Learning.
......
Paper 5: Title:Learnings from Data Integration for Augmented
generate new ideas
Seed idea at step 3
Predictive Analytics for Idea Impact Evaluation to Prioritise High Impact Ideas
......

Figure 3: **Example of Planning-Driven Seed Idea Iteration.** Starting from an initial seed idea, a human-like literature search plan is formulated for a comprehensive relevant literature search, together with iteration to help generate seed ideas with broader and deeper insights.

This approach, demonstrated through an in-context learning example, leverages the LLM's internal knowledge to determine useful knowledge for new ideas, surpassing traditional entity or keyword-based retrieval methods.

In the retrieval phase, we construct an academic search engine based on Facebook AI Similarity Search (FAISS)[†]. For each item in the plan, we use the search engine to identify the most relevant literature based on keywords representing the fields for new knowledge acquisition.

In the generation phase, upon acquiring new knowledge, a new seed idea is generated based on the retrieved papers, the initial seed idea, and the input paper. To improve the quality of the generated ideas, we also incorporate self-refinement.

3.2.2 Seed Idea Iteration

In each iteration, the old seed ideas are replaced with the newly generated seed ideas. For each previous seed idea, three times as many new seed ideas are generated based on the retrieved relevant papers. This allows our agent to dive deeper, largely expanding the scope of the search.

3.3 Proposal Generation

After completing the seed idea iteration, we obtain a final seed idea pool. Each seed ideas in the pool is then expanded into the final proposal. Specifically, we first prompt the LLM to generate the initial proposal based on the final seed ideas, the retrieved papers, and the input paper. Next, we instruct the LLM to decompose the methods section of the initial proposal into several sub-modules, and have the LLM design these sub-modules in more detail. Finally, the final proposal is generated based on the initial proposal, the papers, and the detailed sub-modules. The templates for the initial and final proposals are provided in Appendix C.

4 Experiment

To validate our proposed iterative planning framework, we perform comprehensive comparisons with state-of-the-art research idea generation methods and conduct an ablation study.

4.1 Experimental Setup

Data. We construct our evaluation dataset from high-quality papers published at CVPR 2024, ACL

[†] https://github.com/facebookresearch/faiss

Seed idea: Integrate **real-time data feeds and recent publications into the ResearchAgent's** knowledge base to ensure that the generated ideas are aligned with the latest scientific advancements (at step 1 in Fig. 3).

Search Plan:
1. Understanding the methods and technologies used for integrating real-time data into research systems. This involves exploring data integration techniques, real-time data processing, and the infrastructure required to support such systems.
2. Exploring methods for tracking recent publications and integrating them into research databases. This includes studying publication databases, indexing services, and algorithms for categorizing new research papers.
3. Exploring methods for identifying and tracking popular research publications. This involves studying citation analysis, impact metrics, and algorithms for determining the popularity and influence of research papers.
4. Exploring methods for integrating real-time data and recent publications into a knowledge base. This includes studying knowledge base management, data integration frameworks, and algorithms for updating and maintaining the knowledge base.
5. Explorethe impact of real-time data integration on research outputs. This involves evaluating the effectiveness of real-time data integration, and understanding how it influences the quality and relevance of research ideas.

Table 3: Example Search Plan (Promising Research Directions and Required Knowledge Highlighted in Pink).

2024, and ICLR 2024, following a systematic filtering process. Starting with an initial corpus of 7,805 papers, we apply LLM-related keyword filtering to identify approximately 2,000 relevant papers. We further refine this set using citation thresholds ($\geq$20 for ICLR 2024, $\geq$10 for CVPR/ACL 2024), resulting in 153 papers. The final dataset comprises 170 papers after incorporating additional 17 highly-rated papers from HuggingFace Daily Papers.

To evaluate our proposed approach against state-of-the-art methods, we select three leading approaches as baselines: AI-Researcher (Si et al., 2024), AI-Scientist (Lu et al., 2024), and Research-Agent (Baek et al., 2024). For AI-Researcher, we execute their original code on collected data. The implementation details for the other baseline methods remain consistent with those in the original work. Across different baselines, the input format is adjusted from topic/code to paper, and the output is the final proposal.

Implementations Details. We implement all experiments using GPT-4o (2024-02-15-preview) with consistent hyperparameter settings across all phases (temperature=0.3, top$_p$=1.0). Our academic search engine is built on FAISS, with AI-related papers covering the period from 2022 to 2024. For embedding generation and similarity computation, we utilize the all-MiniLM-L6-v2[†] model, which provides a balance between efficiency and performance for academic paper representation.

In the automatic evaluation experiment, each method generates 100 ideas from an input paper. For Nova, the iteration step is set to 3, with an initial pool of 15 seed ideas. By the final iteration, 405 seed ideas are generated, up from 105 in the previous iteration. For each seed idea, three times as many new seed ideas are created using the same retrieved papers. The seed ideas are then clustered into 100 groups using k-means, and one idea is randomly selected from each cluster to form the final set of 100 ideas.

Automatic Evaluation. Automatic evaluation mainly focuses on overall quality evaluation and also concern with the novelty and diversity.

1. Quality. Following (Si et al., 2024), we employ the Swiss System Tournament[†] with Claude-3.5-Sonnet zero-shot ranker to evaluate the quality of ideas. Each idea undergoes five rounds of pairwise comparisons, with wins scored as 1 point. This quality assessment method has been shown to be better than direct comparison (Lu et al., 2024). It has demonstrated 71.4% accuracy in distinguishing accepted from rejected submissions on real ICLR 2024 data.

2. Novelty. Following Si et al. (2024), we assess novelty through a two-step process: (1) retrieving the top-10 most relevant papers using our academic search engine, and (2) employing LLM-based analysis to determine if the generated idea presents novel contributions beyond existing work. An idea is considered novel if no retrieved paper contains substantially similar concepts or methodologies. The same academic search engine as mentioned in plannedRAG (Section 4.1) is used in relevant paper retrieval. The prompt can be found in Appendix A.5.

3. Diversity. Similar to (Si et al., 2024), we use the proportion of unique ideas to measure generation diversity. To be specific, diversity is measured by the number of different ideas generated within every 100 ideas, cosine distance between proposal

[†] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[†] https://en.wikipedia.org/wiki/Swiss-system_tournament

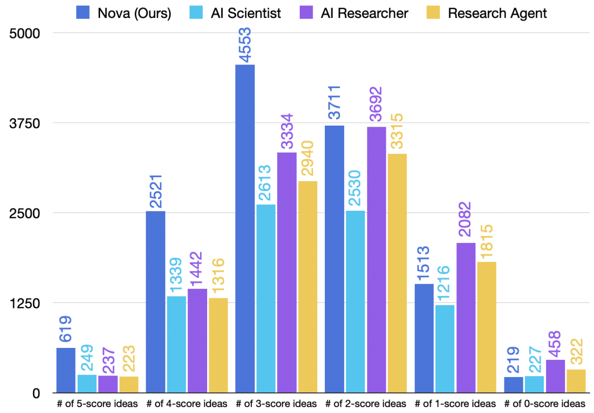

Figure 4: The score distribution of different methods in the Swiss Tournament. This score distribution reveals that Nova not only generates more unique ideas but also produces a higher proportion of high-quality ideas. Among the 619 and 2,521 ideas generated by Nova, those with scores of 4 and 5 significantly outperform the baseline methods.

	Idea Reviewer ($N = 10$)				
Metric	Mean	Median	Min	Max	SD
papers	27	18	2	94	25
citations	1967	802	5	11637	3314
h-index	10	9	2	26	6
i10-index	12	8	0	40	12

Table 4: Research profile metrics of the idea reviewer. Data are from Google Scholar.

embeddings is used to measure similarity and the duplication threshold is set to be 0.8.

Human Evaluation. To validate the effectiveness of our automatic evaluation, we have an additional human evaluation. Our goal is to assess how well our automatic evaluation aligns with human expert evaluations. We recruit a panel of 10 experts, all with a PhD or professorship (4 in NLP, 3 in ML, and 3 in CV), doing research in LLM-related fields. Table 4 records the average citation and average h-index of the 10 experts. These experts evaluated ideas based on novelty and overall quality (including feasibility and effectiveness). We select five ideas generated by each agent based on the same input paper. These ideas correspond to the 1^{st}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles of the automatic evaluation, resulting in a total of 20 ideas per topic. This process is repeated for 20 times (7 from ACL, 7 from CVPR, and 6 from ICLR). Each expert reviews four groups, ensuring that at least two independent experts evaluate each idea. Experts evaluat topics matching their expertise, with some interdisciplinary experts assigned multiple topics across their fields of knowledge. The final score for each idea is averaged across all ratings from different experts.

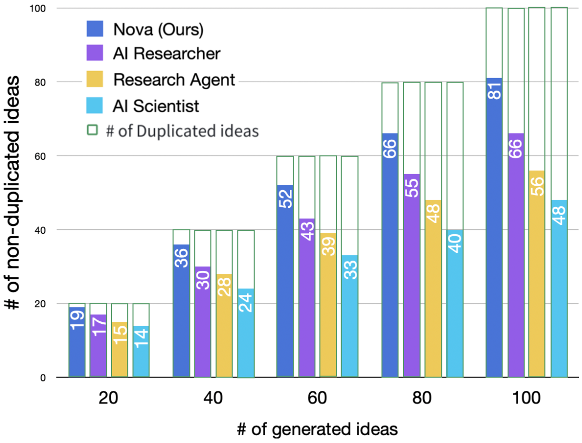

Figure 5: Non-Duplicate Percentage Comparison. Nova continuously generate new ideas through iterative planning and search. In Non-Duplicate Percentage, it significantly outperforms others, with over 80% of the ideas being unique, compared to AI-Scientist's 48%.

We compare the distributions of expert evaluations against those in automatic evaluation. Specifically, we track which methods produce the top 20 percent of ideas, as ranked by the experts. This helps determine which methods outperform the others. Moreover, this approach reveals whether the model evaluation aligns with human evaluation.

4.2 Results

4.2.1 Automatic Evaluation Results

The Swiss Tournament score comparison are shown in Figure 4. The novelty and diversity comparison are shown in Figure 5.

Clearly, Nova achieves a significantly higher Swiss score. 619 and 2521 of the ideas generated by Nova are scored at 4 and 5, significantly surpassing the performance of other agents. By incorporating iterative planning and search for external knowledge retrieval, Nova engages in more effective exploration for innovation. This may significantly enhance the novelty of the generated ideas. Since novelty is often the most important factor in evaluating idea quality, Nova is consistently better than other state-of-the-art methods.

Figure 5 shows that Nova generates significantly more diverse ideas. As the number of generated ideas increases, Nova can continuously generate new ideas through iterative planning and search. In Non-Duplicate Percentage, Nova significantly outperforms others, with over 80% of the ideas being unique.

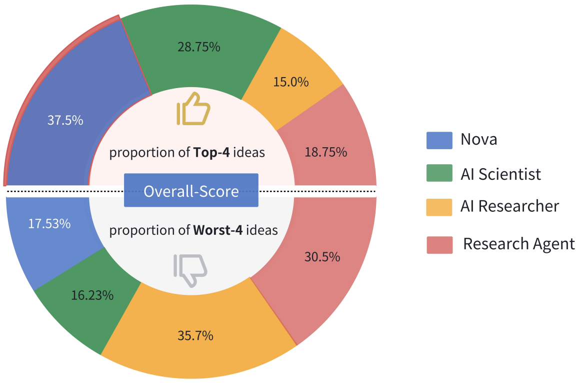

Figure 6: Human Evaluation for Overall Quality.

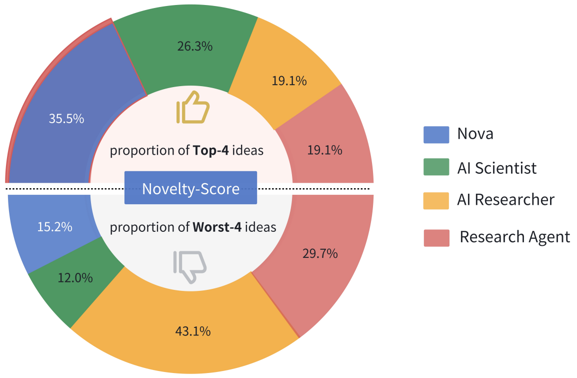

Figure 7: Human Evaluation for Novelty.

4.2.2 Human Evaluation Results

In human evaluation, Nova achieves the highest scores for both overall quality and novelty. As shown in Figure 6, Nova contributes 37.5% of the top 4 ideas, the highest among the four methods. Additionally, Nova has a notably low percentage of the worst 4 ideas, accounting for only 17.53% in terms of overall quality. In Figure 7, a similar pattern is observed in novelty evaluation.

To ensure the reliability of human evaluations, we also assess the consistency among the expert panel. Considering the length of each proposal and the total number of proposals to review, this constituted a challenging task even for experts. Each proposal is evaluated by two distinct experts, and we use the Pearson correlation coefficient to test the consistency between the scores on the same idea rated by the two experts. The Pearson correlation coefficient of the ratings is approximately 0.118 with a p-value of 0.02. This indicates a weak but statistically significant positive relationship between the expert ratings, meaning the weak correlation is unlikely to have occurred by chance.

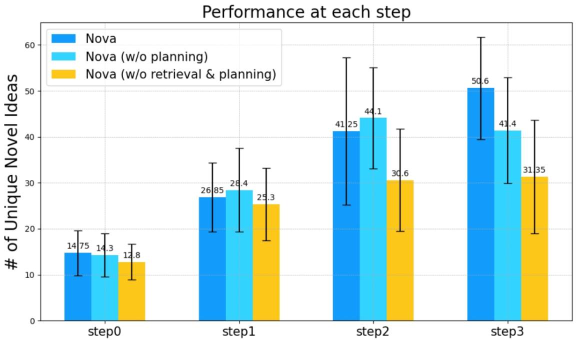

Figure 8: Ablation studies for Nova. We can find that both retrieval and planning significantly enhance the generation of unique novel ideas.

By comparing the distribution of top-rated ideas in both human and automated evaluations (Fig. 4 and 6), it is evident that human reviewers and the LLM evaluate the performance of the four methods in a similar pattern. In both human and automatic evaluations, our method generates the highest proportion of top-rated ideas, followed by AI-Scientist, ResearchAgent, and finally AI-Researcher. This indicates that our automatic review mechanism effectively captures human reviewers' true preferences.

4.3 Ablation Study

To assess the effectiveness of planning and search in Nova, we conduct comparisons by gradually removing the planning and retrieval components. All methods retrieve the same number of papers, specifically $K = 5$. Both retrieval and planning are found to significantly enhance the generation of unique and novel ideas. When planning is excluded, the number of unique ideas at step 3 (44.1) no longer increases compared to step 2 (42.4). This suggests that without planning, relying solely on retrieval based on seed ideas limits access to valuable external knowledge for innovation. This limitation may arise from the restricted scope of search when planning is absent. Obviously, when planning and retrieval are both removed, the number of unique novel ideas increases slightly at step 2 (from 25.3 to 30.6) and stagnates at step 3 (from 30.6 to 31.35), due to no external knowledge being introduced.

5 Conclusion

In this paper, we propose an LLM-based scientific innovation method, Nova, which introduces iterative planning and search to retrieve external knowledge for innovation. Nova leverages the internal knowledge of LLMs to generate search plans for external knowledge retrieval, significantly enhancing the effectiveness of the retrieval process. The ablation study demonstrates the effect of the iterative planning and search framework on promoting the novelty of generating ideas. The automatic and human evaluations show that Nova significantly and consistently outperforms state-of-the-art scientific innovation methods. Future work could explore: (1) incorporating multi-modal knowledge sources beyond textual academic literature, (2) developing reinforced and adaptive planning strategies that dynamically adjust the exploration depth based on idea potential, and (3) investigating methods for maintaining idea diversity while scaling to larger knowledge bases.

6 Limitations

While experimental results demonstrate Nova's effectiveness in enhancing research ideation, we identify several promising directions for future work:

Knowledge Integration Scope. Current implementation focuses primarily on computer science literature, opening opportunities for exploring cross-domain knowledge integration to potentially enable even more innovative interdisciplinary insights.

Computational Efficiency. As the system explores multiple knowledge fields iteratively, there are opportunities to optimize the search and integration processes while maintaining idea quality. Future work could investigate more efficient knowledge retrieval and filtering mechanisms.

Planning Strategies. While our current planning mechanism significantly improves idea diversity, future research could explore incorporating reward functions or learning-based approaches to further enhance planning effectiveness.

These directions suggest rich opportunities for extending Nova's capabilities while building on its demonstrated strengths in enhancing research ideation.

7 Ethics Statement

Publication Policy. The increasing use of AI to generate research ideas poses significant challenges to academic integrity. The growing accessibility of LLMs and the rising usefulness of LLMs in research may lead to deterioration in the overall quality of scholarly content, as individuals may rely on AI for both creativity and submission reviews. Therefore, there is a legitimate concern that students or researchers would exploit these technologies and present low-quality research proposals. To mitigate these risks, it is crucial to hold accountability for outputs generated through AI tools in scientific submissions.

Intellectual Credit. Generative AI in the research cycle poses great concerns about intellectual credit on the submitted works. While traditional frameworks were more like a tool for human researchers, LLMs are more potent in a way that plays a more significant role in the scientific research process if used. It is still unclear how intellectual credit should be distributed in the case of AI-driven research. To better attribute credit to AI-supported research, researchers should adopt transparent documentation about their research process, including the extent of AI involvement in generating ideas and developing experiments.

Potential for Misuse. AI-generated research ideas, particularly those introducing novel concepts, possess the potential for misuse. This could lead to harmful outcomes. Ideation agents may be exploited to develop adversarial attack strategies or other unethical applications. Therefore, it is important to develop anti-jailbreak mechanisms or safety checks on AI-generated content and the use of generative AI in research.

Idea Homogenization. If AI was widely used in scientific research, this would raise concerns about the potential idea of homogenization. The wide adoption of LLMs in research could reflect a narrower set of perspectives or systematic biases compared to human researchers not using AI assistance. Therefore, it is important to recognize the limitations of current LLM-generated ideas, and future work should focus more on enhancing the generation diversity either by improving the models themselves or by refining the ideation process.

Impact on Human Researchers. The challenge posed by AI's integration into research should be well recognized because research is fundamentally and historically a community-driven and collaborative effort. It is still unclear on the negative consequences of the introduction of AI in the research process. People should be cautious and aware of the potential decline in human thought and a reduction in opportunities for human collaboration after the introduction of AI in research. Future works should explore other methods of human-AI collaboration. Understanding how LLM should be integrated into the research process will be an ongoing problem.

8 Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported by the Scientific Research Project of Westlake University (Grant No. WU2024B003).

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. *Preprint*, arXiv:2404.07738.

Gerald Doppelt. 1986. Relativism and The Reticulational Model of Scientific Rationality. *Synthese*, pages 225–252.

Paul Feyerabend. 2020. *Against Method: Outline Of An Anarchistic Theory Of Knowledge*. Verso Books.

Xuemei Gu and Mario Krenn. 2024. Generation and Human-expert Evaluation of Interesting Research Ideas Using Knowledge Graphs and Large Language Models. *Preprint*, arXiv:2405.17044.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. 2024. Ideabench: Benchmarking large language models for research idea generation. *Preprint*, arXiv:2411.02429.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. In *ICML 2024*.

Thomas S Kuhn. 1997. *The Structure of Scientific Revolutions*, volume 962. University of Chicago press Chicago.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2024. Can large language models unlock novel scientific research ideas? *Preprint*, arXiv:2409.06185.

Imre Lakatos et al. 1978. The Methodology of Scientific Research Programmes.

Larry Laudan, Arthur Donovan, Rachel Laudan, Peter Barker, Harold Brown, Jarrett Leplin, Paul Thagard, and Steve Wykstra. 1986. Scientific Change: Philosophical Models and Historical Research. *Synthese*, 69:141–223.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024a. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *Preprint*, arXiv:2410.13185.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024b. Mlr-copilot: Autonomous Machine Learning Research Based on Large Language Models Agents. *Preprint*, arXiv:2408.14033.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *Preprint*, arXiv:2408.06292.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative Refinement with Self-Feedback. *Advances in Neural Information Processing Systems*, 36.

Shaun Nichols and Joshua Knobe. 2007. Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous*, 41(4):663–685.

Karl R Popper. 1963. Science as Falsification. *Conjectures and refutations*, 1(1963):33–39.

Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. *Preprint*, arXiv:2410.04025.

Relex, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In *ICML 2022*.

David B Resnik. 1994. Hacking's Experimental Realism. *Canadian Journal of Philosophy*, 24(3):395–411.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical Discoveries from Program Search with Large Language Models. *Nature*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When Large Language Models Meet Personalization. *Preprint*, arXiv:2304.11406.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *Preprint*, arXiv:2409.04109.

Herbert A Simon. 2013. Understanding the Processes of Science: The Psychology of Scientific Discovery. In *Progress in Science and Its Social Conditions: Nobel Symposium 58 Held at Lidingö, Sweden, 15–19 August 1983*, page 159. Elsevier.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *ACL 2023*.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Lin Li, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, et al. 2024a. Lego-Prover: Neural Theorem Proving with Growing Libraries. In *ICLR 2024*.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023a. Scientific Discovery in The Age of Artificial Intelligence. *Nature*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *ACL 2023*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024b. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *In ACL 2024*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *ICLR 2023*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS 2022*.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning. *Preprint*, arXiv:2405.00451.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large Language Models for Automated Open-domain Scientific Hypotheses Discovery. *Preprint*, arXiv:2309.02726.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS 2023*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR, 2023*.

A Prompts and Examples

A.1 Prompt for Generating Seed Ideas Based on Scientific Innovation Theory

Role: You are an expert researcher in AI. You are familiar with the Theory of Science Discovery,, and you can use this theory to propose some innovative and valuable research idea based on the information provided by users.

Skill: Follow these steps to generate seed ideas for exploration.

(1) **Understanding of the target paper, related papers is essential:**

  (a) The target paper is the primary research study you aim to enhance or build upon through future research, serving as the central source and focus for identifying and developing the specific research idea.

  (b) The related papers are studies that are related to the primary research topic you are focusing on and provide additional context and ideas that are essential for understanding and extending the target paper.

(2) **Understanding of the theory of science discovery is essential:** You need to select appropriate theories and combine the information provided by the current paper to come up with creative, influential and feasible ideas.

(3) Here are 10 general laws and methodologies of scientific discovery from the perspective of philosophy of science. You can choose one or more of these methodologies and propose new scientific research ideas for the target paper: *self.theoryOfScientificDiscovery*.

(4) Select *k* most appropriate theories and methods that are most suitable for target paper and put forward *k* new idea.

Requirements:

(1) Output about *k* new idea worth exploring.

(2) You should aim for new research ideas that can potentially win best paper awards at top conferences like ACL and NeurIPS and ICLR and CVPR..

(3) Skip the theory of scientific discovery may not fit well with the target paper, the theory and method you used should make sense and be reasonable for the target paper.

(4) Thinking is your thought process. Please explain which theory you used in your thinking process.

(5) The exist idea set is some ideas that have been proposed before, you need to propose some different one.

(6) Please output your thought process.

(7) please thinking step by step.

Output Format: *self.scientificIdeaJsonFormat*.

Example: Here are some examples for the json format result, you can follow the format only, and need to propose new idea according to the input bellow. *self.scientificIdeaExample*.

Input:

(1) related paper titles:*relatedPaperTitles*

(2) related papers abstracts:*relatedPaperAbstracts*.

(3) target paper title:*targetPaperTitle*.

(4) target paper abstract:*targetPaperAbstract*.

(5) exist idea set:*existIdeaSet*.

SelfRefinement: if *self.selfRefinement* is True: In the thinking step, you can first think of about 10 new ideas and analyze the advantages and disadvantages of each of them. Your final *K* idea can absorb their advantages and discard their disadvantages.

Output:

(1) **Thinking:** output your thinking process here, explain why you choose these theory to discover new idea and why it should have change to win the best paper awards at top conferences.

(2) **IdeaJsonList:**<JSON>.

A.2 Prompt for Generating Search Plan about Seed Ideas

Role: You are an expert researcher in AI. You can think out of the box, develop a detailed paper search plan for a given research idea base on target paper.

Skill: Here is a research idea for you, please analyze the sequence of different fields which you should search for relevant papers. This way, you can gather comprehensive information and new knowledge, further expanding your research perspective and finding new ideas.

Example: You can follow these examples to get a sense of how the plan should be formatted (but don't borrow the plan themselves):
self.fewShotExample

Format: In <JSON>, provide the search plan list in JSON format, every plan with the following fields:

(1) **title:** Indicate the field and direction in which you want to search for literature.

(2) **thinking:** The thought process behind proposing the plan.

(3) **keywords:** keywords that can help search in google scholar then you can find related paper.

(4) **rationale:** The reason for generating this plan, explain why this should help you gather comprehensive information.

Requirements:

(1) The search plan needs to be developed around the given research idea.

(2) The plan also needs to focus on how to optimize Target paper.

(3) Please provide the thought process and search keywords.

(4) Please thinking step by step.

Input:

1. **Research idea:** *ideaInfo*
2. **Target paper title:** targetPaperTitle

Output:

(1) **Thinking:** output your thinking process here.

(2) **Search plan output:**<JSON>.

To enhance retrieval precision, search keyword generation occurs concurrently with the development of plan descriptions. When generating search plans from seed ideas, the prompt should specify distinct requirements for both the descriptive planning component and the keyword extraction process, as list below:

In <JSON>, provide the search plan list in JSON format, along with the following contents:

a. Plan Description: Indicate the field and direction in which you want to search for the literature.

b. Thinking: The thought process behind proposing of the plan. As mentioned in CoT(Wei et al., 2022), thinking mainly refers to the step by step thinking process.

c. Keywords: keywords that can help search in Google Scholar, then you can find related papers.

d. Rationale: The reason for generating this plan is to explain why this should help you gather comprehensive information. As mentioned in (Madaan et al., 2024), the rationale is about why output the current results.

During knowledge retrieval, the generated keywords are integrated with the search plan description to create a comprehensive search query. This dual-component approach is designed to improve the overall quality and effectiveness of the generated search plan.

A.3 Prompt for Generating Initial Proposal Base on Seed Idea

Role: You are an expert researcher in Large Language Models. Now I want you to help me brainstorm the detail research project proposal base on the idea: *idea*.

Background: You should generate detail proposal base on the given knowledge. Try to be creative.

- **Target Paper:** the giving idea are driving from targetPaperTitle:*targetPaperTitle*, targetPaperAbstract:*targetPaperAbstract*, this just for your background knowledge.
- **Related Papers:** Here are some relevant papers on this idea just for your background knowledge: *planThenRetrievalKnowledge*.

Example: if self.useFewShotExample is True: You can follow these examples to get a sense of how the proposal should be formatted (but don't borrow the proposal themselves):
self.fewShotExample

Format: The poposal should be described as:

(1) **Problem:** State the problem statement, which should be closely related to the idea description and something that large language models cannot solve well yet.

(2) **Existing Methods:** Mention some existing benchmarks and baseline methods if there are any.

(3) **Motivation:** Explain the inspiration of the proposed method and why it would work well.

(4) **Proposed Method:** Propose your new method and describe it in detail. The proposed method should be maximally different from all existing work and baselines, and be more advanced and effective than the baselines. You should be as creative as possible in proposing new methods, we love unhinged ideas that sound crazy. This should be the most detailed section of the proposal.

(5) **Experiment Plan:** Specify the experiment steps, baselines, and evaluation metrics.

Requirements:

(1) You should generate detail proposal base on the given knowledge. Try to be creative.

(2) The above papers are only for inspiration and you should not cite them and just make some incremental modifications. Instead, you should make sure your proposal are novel and distinct from the prior literature.

(3) You should aim for projects that can potentially win best paper awards at top conferences like ACL and NeurIPS.

(4) Please thinking step by step.

Attention:

(1) You should make sure to come up with your own novel proposal for the specified idea: *idea*, You should try to tackle important problems that are well recognized in the field and considered challenging for current models. For example, think of novel solutions for problems with existing benchmarks and baselines. In rare cases, you can propose to tackle a new problem, but you will have to justify why it is important and how to set up proper evaluation.

(2) proposal should base on the idea: *idea*.

(3) topic should follow the targetPaperTitle:*targetPaperTitle*, targetPaperabstract:**targetPaperabstract**

(4) Please write down thinking and your final proposal. Output the final proposal in json format as a dictionary, where you should generate a short proposal name (e.g., N̈on-Linear Story Understanding̈, or M̈ulti-Agent Negotiation)̈ as the key and the actual proposal description as the value (following the above format). "

Output:

(1) **Thinking:** output your thinking process here.

(2) **Proposal:**<JSON>.

A.4 Prompt for Generating Final Proposal Base on Initial Proposal

Role: You are an expert researcher in AI and your job is to expand a brief project idea into a full project proposal with detailed methodology and experiment plans so that your students can follow the steps and execute the full project.

Input: You should generate detail proposal base on the given knowledge. Try to be creative.

- **Initial Proposal:** *initialProposal*
- **Target Paper:** The target paper is the primary research study you aim to enhance or build upon through future research, serving as the central source and focus for identifying and developing the specific research idea. targetPaperTitle: *targetPaperTitle*, targetPaperAbstract: *targetPaperAbstract*.

Example: if self.useDemoExample is True: Note that we only provide examples related to prompt work. Please refer to this format in other fields, You can follow these examples to get a sense of how the idea should be formatted (but don't bore the ideas), Below are a few examples of how the full experiment plans should look like:
self.demoExamples

Format: Now you should come up with the full proposal covering:

(1) **Title:** A concise statement of the main research question to be used as the paper title.

(2) **Problem Statement:** Clearly define the problem your research intends to address. Explain clearly why this problem is interesting and important.

(3) **Motivation:** Explain why existing methods (both classic ones and recent ones) are not good enough to solve the problem, and explain the inspiration behind the new proposed method. You should also motivate why the proposed method would work better than existing baselines on the problem.

(4) **Proposed Method:** Explain how the proposed method works, describe all the steps. Make sure every step is clearly described and feasible to implement.

(5) **Step-by-Step Experiment Plan:** Break down every single step of the experiments, make sure every step is executable. Cover all essential details such as the datasets, models, and metrics to be used. If the project involves prompting, give example prompts for each step.

Requirements:

(1) The experiment plan should not include any background introduction (you can skip the literature review, paper writing tips, and ethical discussion). Just give instructions on the experiments

(2) Be consistent in your methodology and experiment design, for example, if you will use black-box LLM APIs such as GPT and Claude for your experiments, then you shouldn't propose any experiments that require white-box model weights or data access and you should edit them accordingly to follow the black-box assumptions.

(3) Consider novelty, significance, correctness, and reproducibility to ensure the high quality of the final proposal.

(4) You should aim for final proposal that can potentially win best paper awards at top conferences like ACL and NeurIPS and ICLR and CVPR.

(5) first give an overview of the proposed method, then give a detailed design.

(6) Please output your thought process.

(7) please thinking step by step.

Output: Now please write down your final proposal in JSON format (keys should be the section names, just like the above examples). Make sure to be as detailed as possible so that a student can directly follow the plan to implement the project.

(1) **Thinking:** output your thinking process here.

(2) **FinalProposal:**<JSON>.

A.5 Prompt for Novelty Evaluation

Role: You are a professor specialized in AI.

Task: You have a research proposal. Your job is to determine whether the given paper is directly relevant to the proposal.

Requirements:

(1) The research proposal and paper abstract are considered to be matched if both the research problems and the approaches are the same. Note that the method details do not matter, you should only focus on the high-level concepts and judge whether they are directly relevant.

(2) You should first specify what is the proposed research problem and method. Then, response with a binary judgment, saying either "Yes" or "No". If answering yes, your explanation should be the one-sentence sumerizing both the abstract and the proposal and their similarity. If answering no, give the short summarization of the abstract and the proposal separately, then highlight their differences.

Input:

1. **The research proposal is:** *proposal*
2. **The paper is:** *related paper*

Output:

(1) **Thinking:** *output your thinking process here.*

(2) **Result:** *output your final analysis results here.*

A.6 Prompt for Summarizing Current Research Trends

Role: You are an AI expert researcher. You can summarise the current hot research trends from the list of recent AI papers.

Skill: You will analyze the research trending based on the recent popular papers, provide us with the research trending report.

Requirements:

(1) Provide a comprehensive analysis, including the hot research directions, the highlights of the technologies and methods, and discuss whether these technologies can be used in other fields.

Input:

(1) **Popular Paper List:** *popular paper list*

Output:

(1) **Result:** *output your final report here.*

A.7 A Current Research Trend Example

Long-Context Language Models (LLMs) and Retrieval-Augmented Generation (RAG)

Core Papers:

(1) "RAG in the Era of Long-Context LLMs", "LongCite", "MemLong", "Improved RAG with Self-Reasoning", "LongWriter", "EfficientRAG", "Enhanced RAG with Long-Context LLMs", "GraphReader"

Highlights:

(1) Addressing the challenge of maintaining focus and relevance in long-context LLMs. Combining RAG mechanisms with long-context capabilities to improve performance in tasks like question answering and citation generation. Innovations such as order-preserving RAG, external retrievers, and graph-based systems to enhance context handling.

Applications:

(1) These advancements can be applied in fields requiring extensive document analysis, such as legal research, academic literature review, and medical records analysis.

B Scientific Innovation Theories

1. Defining new scientific problems:
Theoretical basis: Kuhn’s paradigm theory, Laudan’s problem-solving model, Nichols’s problem-generation theory.
Method: Identifying anomalies in existing theories, exploring theoretical boundaries and scope of application, integrating interdisciplinary knowledge and discovering new problems, and re-examining neglected historical problems.
2. Proposing new hypotheses:
Theoretical basis: Pierce’s hypothetical deduction method, Weber’s theory of accidental discovery, and Simon’s scientific discovery as problem-solving.
Method: Analogical reasoning, thought experiment, intuition and creative leaps, and reductio ad absurdum thinking.
3. Exploring the limitations and shortcomings of current methods:
Theoretical basis: Popper’s falsificationism, Lakatos’s research program methodology, Feyerabend’s methodological anarchism.
Method: Critically analyzing existing methods, finding deviations between theoretical predictions and experimental results, exploring the performance of methods under extreme conditions, and interdisciplinary comparative methodology.
4. Improving existing methods:
Theoretical basis: Laudan’s methodological improvement model, Ziemann’s creative extension theory, and Hacking’s experimental system theory.
Method: Integrating new technologies and tools, improving experimental design and control, improving measurement accuracy and resolution, and developing new data analysis methods.
5. Concluding the general laws behind multiple related studies:
Theoretical basis: Whewell’s conceptual synthesis theory, Carnap’s inductive logic, and Glaser and Strauss’s grounded theory.
Method: Comparative analysis of multiple case studies, identifying common patterns and structures, constructing conceptual frameworks and theoretical models, and Formal and mathematical descriptions.
6. Constructing and modifying theoretical models:
Theoretical basis: Quine’s holism, Lakoff’s conceptual metaphor theory, and Kitcher’s unified theory of science.
Method: Forming a balance between reductionism and emergence, developing an interdisciplinary theoretical framework, mathematical modeling and computer simulation, and theoretical simplification and unification.
7. Designing pivotal Experiments:
Theoretical basis: Duhem-Quine thesis, Bayesian experimental design theory, and Mayo’s theory of experimental reasoning.
Method: Designing experiments to distinguish between competing theories, exploring extreme conditions and boundary cases, developing novel observation and measurement techniques, and designing natural and quasi-experiments.
8. Explaining and integrating anomalous findings:
Theoretical basis: ansen’s theory of anomalous findings, Sutton’s model of scientific serendipity, and Kuhn’s theory of crises and scientific revolutions.
Method: Revisiting foundational assumptions, developing auxiliary hypotheses, exploring new explanatory frameworks, and integrating multidisciplinary perspectives.
9. Evaluating and selecting competing theories:
Theoretical basis: Reichenbach’s confirmation theory, Sober’s theory selection criteria, and Laudan’s problem-solving progress assessment.
Method: Comparing theories for explanatory power and predictive power, evaluating the simplicity and elegance of theories, considering the heuristics and research agenda of theories, and weighing the empirical adequacy and conceptual coherence of theories.

10. Changing scientific paradigm:
Theoretical basis: Kuhn's theory of scientific revolutions, Toulmin's model of conceptual evolution, and Hall's dynamic system theory.
Method: Identifying accumulated anomalies and crises, developing new conceptual frameworks, reinterpreting and organizing known facts, and establishing new research traditions and practices.

C Proposal Template

Follow (Si et al., 2024), here we set the template of the initial proposal and the final proposal. As mentioned above, we evaluate diversity based on the initial proposal and evaluate quality and novelty based on the final proposal.

C.1 Initial proposal

1. Problem: State the problem statement, which should be closely related to the idea description and something that large language models cannot solve well yet.
2. Existing Methods: Mention some existing benchmarks and baseline methods if there are any.
3. Motivation: Explain the inspiration of the proposed method and why it would work well.
4. Experiment Plan: Specify the experiment steps, baselines, and evaluation metrics.

C.2 Final Proposal Template

1. Title: A concise statement of the main research question to be used as the paper title.
2. Problem Statement: Clearly define the problem your research intends to address. Explain clearly why this problem is interesting and important.
3. Motivation: Explain why existing methods are not good enough to solve the problem, and explain the inspiration behind the new proposed method. You should also motivate why the proposed method would work better than existing baselines on the problem.
4. Proposed Method: Explain how the proposed method works, describe all the essential steps.
5. Step-by-Step Experiment Plan: Break down every single step of the experiments, make sure every step is executable. Cover all essential details such as the datasets, models, and metrics to be used. If the project involves prompting, give some example prompts for each step.

D Human Annotation

This section provides details about human annotation in our human evaluation experiment.

D.1 Annotation Instructions

The complete annotation instruction for the idea reviewer is given below:

Please evaluate the given twenty ideas based on four criteria (Novelty, Feasibility, Effectiveness, and Overall), identify the best four and the worst four ideas, and rank them accordingly. The principles include:

Annotation Dimensions: See Table 5 for detailed dimension instructions.

Criteria	**Definition**
Novelty	Novelty refers to the originality and innovativeness of the idea. It assesses how new and unique the idea is compared to existing work in the field.
Overall	Overall evaluates the general quality and potential of the idea, taking into account all other criteria (Novelty, Feasibility, and Effectiveness). It provides a holistic assessment of the idea's value.
Feasibility	Feasibility assesses the practicality and implementability of the idea. It considers whether the idea can be realistically executed with available resources and within a reasonable timeframe.
Effectiveness	Effectiveness evaluates the expected impact and success of the idea in achieving its intended goals. It considers how well the idea is likely to perform in practice.

Table 5: Evaluation Criteria and Definitions for Online Idea Assessment Based on Novelty, Feasibility, Effectiveness, and Overall Quality

Annotation Method: Annotate the best 4 and the worst 4 ideas in each of the 4 dimensions. For the best ideas, mark them as 1, 2, 3, and 4 respectively, while 1 refers to the best idea. For the worst ideas, mark them as 17, 18, 19, and 20, while 20 refers to the worst idea. No need to annotate ideas other than the best 4 and the worst 4. For an example, see Table 6.

Rank	Label
Best	1
Second Best	2
Third Best	3
Fourth Best	4
...	...
...	...
Fourth Worst	17
Third Worst	18
Second Worst	19
Worst	20

Table 6: An example of Ranking Labels for Annotating the Best and Worst Ideas Across Four Evaluation Dimensions

D.2 Data Description

We manually selected 20 different papers from ACL2024, CVPR2024, and ICLR2024. Each paper is carefully selected, and the 20 papers are from various research fields, including Natural Language Processing, Computer Vision, and Large Language Models in general, and have varied academic significance measured by citations to represent a broad scope of research papers. The online pilot study gives the human evaluators a form of twenty rows and five columns, along with a hyperlink to the original paper the ideas are generated. Each row is of an idea generated by one of the four different methods, Nova, AI-Scientist, AI-Researcher, or ResearchAgent. Still, human experts have no information on which method to generate that particular idea. The four columns are summary, novelty, feasibility, effectiveness, and overall. The summary is the research plan generated from one of the four methods, and the remaining four columns are entries for human experts to input their rankings. The four best ideas are labeled as 1, 2, 3, or 4, and the four worst ideas are labeled as 17, 18, 19, or 20. The rest of the entries are left blank.

D.3 Risk Statement

1. Physical Risk. This study does not involve any activities that may cause physical harm or discomfort.

2. Psychological Risk. This study does not involve sensitive topics or psychological experiments.

3. Social Risk. This study does not involve activities that could affect participants' social relationships or reputations. No personal information will be disclosed.

4. Economic Risk. This study will not result in any economic loss for the participants.

5. Privacy and Data Security Risk. All annotation data will be randomly assigned to anonymous experts. No personally sensitive information will be collected.

E Case Study

E.1 Case Study: Predictive Analytics for Idea Impact Evaluation to Prioritise High Impact Ideas

Predictive Analytics for Idea Impact Evaluation to Prioritise High Impact Ideas

Input Paper: ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models.

1. Problem Statement: Large language models (LLMs) are capable of generating numerous research ideas, but not all ideas have the same potential for impact and novelty. Evaluating and prioritizing these ideas manually is time-consuming and subjective, leading to inefficiencies and missed opportunities.

2. Motivation: Current methods for idea evaluation often rely on qualitative assessments by experts or simple quantitative metrics such as citation counts. These methods are limited by human biases, scalability issues, and the inability to predict future impact accurately. Inspired by the need for a more efficient and objective way to evaluate the potential impact of research ideas, we propose using predictive analytics to assess and prioritize ideas generated by LLMs. By leveraging data-driven prediction techniques, we can create a system that identifies high-impact ideas with greater accuracy and efficiency.

3. Proposed Method: Our method involves the following steps: **(1) Data Collection:** Gather a comprehensive dataset of past research ideas, their associated features (e.g., keywords, abstract, authors), and their impact metrics (e.g., citation counts, publication venues). **(2) Feature Engineering:** Extract relevant features from the dataset, including textual features (e.g., TF-IDF, embeddings), author features (e.g., h-index, collaboration network), and contextual features (e.g., research trends, funding availability). **(3) Model Training:** Train a predictive model using machine learning algorithms (e.g., random forests, gradient boosting, neural networks) to predict the impact of research ideas based on the extracted features. **(4) Idea Generation:** Use an LLM to generate new research ideas, incorporating the core paper and relevant publications as input. **(5) Impact Prediction:** Apply the trained predictive model to evaluate the potential impact of the generated ideas. **(6) Iterative Refinement:** Use feedback from the predictive model to iteratively refine and improve the generated ideas, focusing on those with the highest predicted impact.

4. Step-by-Step Experiment Plan:

(1) **Collect and Preprocess the Dataset:** Gather a comprehensive dataset of past research ideas from various scientific publications across multiple disciplines. Preprocess the dataset to extract features such as keywords, abstracts, authors, citation counts, and publication venues.

(2) **Extract Features:** Extract relevant features from the dataset, including textual features (e.g., TF-IDF, embeddings), author features (e.g., h-index, collaboration network), and contextual features (e.g., research trends, funding availability).

(3) **Train the Predictive Model:** Train a predictive model using machine learning algorithms (e.g., random forests, gradient boosting, neural networks) to predict the impact of research ideas based on the extracted features. Split the dataset into training and validation sets to evaluate the model's performance.

(4) **Generate Research Ideas Using an LLM:** Use an LLM (e.g., GPT-4) to generate new research ideas. Provide the LLM with the core paper and relevant publications as input to guide the idea generation process. Example prompt: ’Based on the core paper [title] and the following relevant publications [list], generate new research ideas that build on these works.

(5) **Evaluate the Impact of the Generated Ideas:** Apply the trained predictive model to evaluate the potential impact of the generated ideas. Rank the ideas based on their predicted impact scores.

(6) **Refine the Ideas Iteratively:** Use feedback from the predictive model to iteratively refine and improve the generated ideas. Focus on enhancing the ideas with the highest predicted impact. Example prompt for refinement: ’Given the feedback that the idea [description] has high potential impact, refine this idea to make it more novel and feasible.

(7) **Compare Results with Baseline Methods:** Compare the performance of the predictive model with baseline methods (e.g., citation count prediction, expert evaluation) using metrics such as precision, recall, and F1 score. Additionally, evaluate the quality and novelty of the generated ideas through human and model-based assessments.

(8) **Analyze Results:** Analyze the results to determine whether the proposed method improves the efficiency and accuracy of research idea evaluation compared to baseline methods. Document the findings and identify areas for further improvement.

In Section E.1, we present an intriguing case study based on the input paper–**ResearchAgent**(Baek et al., 2024): Iterative Research Idea Generation over Scientific Literature with Large Language Models. The focus of this study is the **utilisation of LLM for the generation of research ideas.** The motivation for this approach is to address the limitations of manual evaluation and prioritization of ideas, which are both time-consuming and subjective, leading to inefficiencies. This study offers a promising avenue for further improving the LLM based idea generation. The idea employs **predictive analytics to assess and priorities ideas generated by LLMs**. The use of data-driven prediction techniques is also a reasonable.

E.2 Case Study: An Ethical and Privacy-Preserving Framework for Personalized AI Systems

An Ethical and Privacy-Preserving Framework for Personalized AI Systems

Input Paper: LaMP: When Large Language Models Meet Personalization.

1. Problem Statement: Current personalization techniques often compromise user privacy and lack transparency, leading to ethical concerns and reduced user trust. There is a need for a framework that ensures responsible AI usage in personalization, protecting user privacy while providing relevant outputs.
2. Motivation: Existing methods like differential privacy, federated learning, and explainable AI are often used in isolation and do not fully address the multifaceted ethical concerns in personalization. Combining federated learning, explainable AI, and user-centric consent management can create a robust framework that addresses privacy, transparency, and user control simultaneously. Federated learning ensures data privacy, explainable AI builds user trust, and consent management empowers users.
3. Proposed Method: Our method involves the following steps: **(1) Federated Learning Engine:** Purpose: To train personalization models on user devices, ensuring data remains local and leveraging collaborative learning to improve model performance. Implementation: Local Model Training: Personalization models are trained locally on user devices using local data. Model Aggregation: Aggregated model updates are sent to a central server without sharing raw data, ensuring privacy. Privacy-Preserving Techniques: Techniques like differential privacy and secure multi-party computation are used to enhance privacy. **(2) Explainable AI Component:** Purpose: To generate clear and understandable explanations for personalized outputs, increasing transparency and helping users understand how their data is used. Implementation: Explanation Generation: Mechanisms to generate explanations for model decisions, such as feature importance and decision trees. User Interface: A user-friendly interface to present explanations in an understandable manner. Feedback Loop: A feedback mechanism to improve explanations based on user input. **(3) User-Centric Consent Management:** Purpose: To provide users with granular control over their data, allowing them to opt-in or opt-out of personalization features, and aligning with regulatory requirements. Implementation: Consent Interface: A user-friendly interface for managing consent preferences. Granular Control: Options for users to control specific aspects of data usage and personalization features. Regulatory Compliance: Ensuring compliance with regulations like GDPR and CCPA. **(4) Contextual Integrity Module:** Purpose: To ensure that data usage aligns with user expectations and social norms by adapting to various contexts through dynamic consent management and context-aware explanations. Implementation: Dynamic Consent Management: Mechanisms to dynamically adjust consent preferences based on context. Context-Aware Explanations: Generating explanations that consider the context in which data is used. User Feedback Integration: Incorporating user feedback to continuously improve context-aware mechanisms.
4. Step-by-Step Experiment Plan:

(1) **Gather Datasets:** We will use datasets relevant to personalization tasks, including recommendation systems and personalized content delivery. Example datasets include MovieLens for recommendations and personalized news datasets.

(2) **Implement Federated Learning Engine:** Local Model Training: Train personalization models locally on user devices using local data from the gathered datasets.Model Aggregation: Implement a central server to aggregate model updates without sharing raw data. Privacy-Preserving Techniques

(3) **Implement Explainable AI Component:** Explanation Generation: Develop mechanisms to generate explanations for model decisions, such as feature importance and decision trees. User Interface: Design a user-friendly interface to present explanations in an understandable manner. Feedback Loop: Create a feedback mechanism to

(4) **Implement User-Centric Consent Management:** Consent Interface: Develop a user-friendly interface for managing consent preferences. Granular Control: Provide options for users to control specific aspects of data usage and personalization features. Regulatory Compliance: Ensure compliance with regulations like GDPR and CCPA.

(5) **Implement Contextual Integrity Module:** Dynamic Consent Management: Develop mechanisms to dynamically adjust consent preferences based on context. Context-Aware Explanations: Generate explanations that consider the context in which data is used. User Feedback Integration:

(6) **Evaluate Privacy:** Measure the privacy guarantees provided by federated learning using metrics like differential privacy.

(7) **Evaluate Transparency:** Assess the quality and clarity of explanations generated by the explainable AI component through user studies.

(8) **Evaluate User Control:** Conduct surveys and user studies to evaluate user satisfaction and trust.

(9) **Evaluate Personalization Quality:** Compare the performance of the EFPF with traditional personalization methods using metrics like precision, recall, and user engagement. Baselines will include standard federated learning, traditional centralized personalization, and existing explainable AI techniques.

In Section E.2, we present another case study based on the input paper—**LaMP**(Salemi et al., 2024): When Large Language Models Meet Personalization. The focus of this study is the **personalization of large language models** The proposed method emphasizes the integration of different methods including federated learning, explainable AI, and user-centric consent management to address challenges in personalization. The idea proposes to use explainable AI, dynamic consent management, and context-aware mechanisms to ensure that personalization adapts to different scenarios. Through a detailed experimental plan, the idea generated demonstrates viability as well as novelty.

F Qualitative Comparison of Nova with Other Methods

Towards Responsible Personalization: Integrating Federated Learning, Explainability, and Contextual Consent

Input Paper: LaMP: When Large Language Models Meet Personalization

Idea Generated by Nova, Rated 1 (Top Quality) by Our Experts:
1. Problem Statement: Current personalization techniques often compromise user privacy and lack transparency, leading to ethical concerns and reduced user trust. There is a need for a framework that ensures responsible AI usage in personalization, protecting user privacy while providing relevant outputs. **2. Motivation:** Existing methods like differential privacy, federated learning, and explainable AI are often used in isolation and do not fully address the multifaceted ethical concerns in personalization. Combining federated learning, explainable AI, and user-centric consent management can create a robust framework that addresses privacy, transparency, and user control simultaneously. Federated learning ensures data privacy, explainable AI builds user trust, and consent management empowers users. **3. Proposed Method:** Our method involves the following steps: (1) Federated Learning Engine: Purpose: To train personalization models on user devices, ensuring data remains local and leveraging collaborative learning to improve model performance. Implementation: Local Model Training: Personalization models are trained locally on user devices using local data. Model Aggregation: Aggregated model updates are sent to a central server without sharing raw data, ensuring privacy. Privacy-Preserving Techniques: Techniques like differential privacy and secure multi-party computation are used to enhance privacy. (2) Explainable AI Component: Purpose: To generate clear and understandable explanations for personalized outputs, increasing transparency and helping users understand how their data is used. Implementation: Explanation Generation: Mechanisms to generate explanations for model decisions, such as feature importance and decision trees. User Interface: A user-friendly interface to present explanations in an understandable manner. Feedback Loop: A feedback mechanism to improve explanations based on user input. (3) User-Centric Consent Management: Purpose: To provide users with granular control over their data, allowing them to opt-in or opt-out of personalization features, and aligning with regulatory requirements. Implementation: Consent Interface: A user-friendly interface for managing consent preferences. Granular Control: Options for users to control specific aspects of data usage and personalization features. Regulatory Compliance: Ensuring compliance with regulations like GDPR and CCPA. (4) Contextual Integrity Module: Purpose: To ensure that data usage aligns with user expectations and social norms by adapting to various contexts through dynamic consent management and context-aware explanations. Implementation: Dynamic Consent Management: Mechanisms to dynamically adjust consent preferences based on context. Context-Aware Explanations: Generating explanations that consider the context in which data is used. User Feedback Integration: Incorporating user feedback to continuously improve context-aware mechanisms. **4. Step-by-Step Experiment Plan:** (1) Gather Datasets: We will use datasets relevant to personalization tasks, including recommendation systems and personalized content delivery. Example datasets include MovieLens for recommendations and personalized news datasets. (2) Implement Federated Learning Engine: Local Model Training: Train personalization models locally on user devices using local data from the gathered datasets.Model Aggregation: Implement a central server to aggregate model updates without sharing raw data. Privacy-Preserving Techniques (3) Implement Explainable AI Component: Explanation Generation: Develop mechanisms to generate explanations for model decisions, such as feature importance and decision trees. User Interface: Design a user-friendly interface to present explanations in an understandable manner.

ForgetMeNot: Incremental Learning Framework with Memory-Augmented Networks for LLMs

Input Paper: LaMP: When Large Language Models Meet Personalization

Idea Generated by ResearchAgent, Rated 20 (Lowest Quality) by Our Experts:
1. Problem Statement: Large Language Models (LLMs) face significant challenges in continuously learning from new data without forgetting previously learned information, a phenomenon known as catastrophic forgetting. This issue hampers the sustained relevance and accuracy of LLMs over time. **2. Motivation:** Existing methods, including LaMP, do not adequately address the challenge of incremental learning and the risk of catastrophic forgetting. Incremental learning allows models to adapt to new information while retaining previously learned knowledge, making them more robust and effective over time. Our proposed method aims to enhance the ability of LLMs to learn incrementally by leveraging memory-augmented networks. **3. Proposed Method:** We propose an incremental learning framework that integrates online learning algorithms with memory-augmented networks. The memory networks will store and retrieve past interactions to inform future outputs, ensuring the model adapts to new data without forgetting previous information. The framework consists of the following steps: (1) Data Collection: Continuously collect interaction data from participants. (2) Memory Network Design: Implement a memory-augmented network that can store and retrieve past interactions. (3) Online Learning Algorithm: Develop an online learning algorithm that updates the model incrementally using new data while consulting the memory network to retain past knowledge. (4) Integration: Integrate the memory network and online learning algorithm into the LLM. (5) Evaluation: Evaluate the performance of the framework using metrics like relevance, accuracy, and user satisfaction. **4. Step-by-Step Experiment Plan:** (1) Gather Datasets: Recruit participants and continuously collect interaction data. Use datasets like Personalization Tasks from LaMP, which include diverse language tasks and multiple entries for each user profile. (2) Design Memory Networks: Implement memory-augmented networks using architectures like Neural Turing Machines (NTMs) or Differentiable Neural Computers (DNCs). Ensure the memory network can store and retrieve past interactions effectively. (3) Develop Online Learning Algorithm: Create an online learning algorithm that updates the model incrementally. Use techniques like Elastic Weight Consolidation (EWC) to mitigate catastrophic forgetting.

In Section F, we present an qualitative comparison based on the input paper–**LaMP**(Salemi et al., 2024): When Large Language Models Meet Personalization.

Given the presented proposals generated by Nova and ResearchAgent, one could more thoroughly analyze the difference in novelty and details. Nova generated a more thorough, detailed, and innovative proposal, while ResearchAgent generated a proposal that were too general and opaque. This discrepancy is reflected in the ratings provided by human experts, with Nova receiving a rating of 1 (top quality idea) and ResearchAgent receiving a rating of 20.