# ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models

**Anonymous ACL submission**


## Abstract

The pace of scientific research, vital for improving human life, is complex, slow, and needs specialized expertise. Meanwhile, novel, impactful research often stems from both a deep understanding of prior work, and a cross-pollination of ideas across domains and fields. To enhance the productivity of researchers, we propose ResearchAgent, which leverages the encyclopedic knowledge and linguistic reasoning capabilities of Large Language Models (LLMs) to assist them in their work. This system automatically defines novel problems, proposes methods and designs experiments, while iteratively refining them based on the feedback from collaborative LLM-powered reviewing agents. Specifically, starting with a core scientific paper, ResearchAgent is augmented not only with relevant publications by connecting information over an academic graph but also entities retrieved from a knowledge store derived from shared underlying concepts mined across numerous papers. Then, mimicking a scientific approach to improving ideas with peer discussions, we leverage multiple LLM-based ReviewingAgents that provide reviews and feedback via iterative revision processes. These reviewing agents are instantiated with human preference-aligned LLMs whose criteria for evaluation are elicited from actual human judgments via LLM prompting. We experimentally validate our ResearchAgent on scientific publications across multiple disciplines, showing its effectiveness in generating novel, clear, and valid ideas based on both human and model-based evaluation results. Our initial foray into AI-mediated scientific research has important implications for the development of future systems aimed at supporting researchers in their ideation and operationalization of novel work.


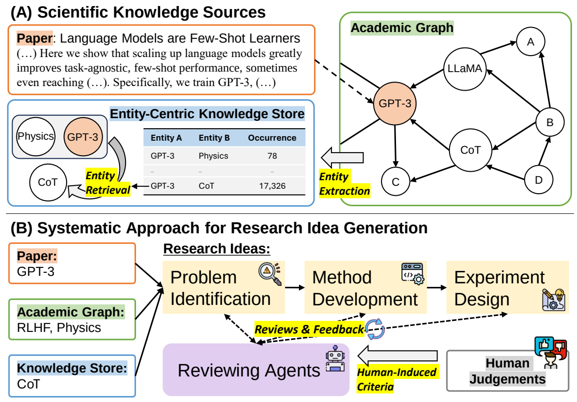


Figure 1: (A) The scientific knowledge used for research idea generation consists of a paper, its relationships over an academic graph, and entities within a knowledge store extracted from numerous papers. (B) Given them, the proposed research idea generation process involves problem identification, method development, and experiment design. Those are also iteratively refined by reviews and feedback from reviewing agents, aligned with criteria induced from human judgements.

## 1 Introduction

Scientific research plays a crucial role in driving innovation, advancing knowledge, solving problems, expanding our understanding of the world, and ultimately improving the lives of people in tangible ways. This process usually consists of two key components: the formulation of new research ideas and the validation of these ideas through well-crafted experiments, which are typically conducted by human researchers (Hope et al., 2023; Wang et al., 2023a; Huang et al., 2023). However, this is a slow, effort-intensive process, which requires reading and synthesizing overwhelming amounts of knowledge over the vast corpus of rapidly growing scientific literature to formulate research ideas, as well as design and perform experimental validations of those ideas. For example, the number of academic papers published per year is more than 7 million (Fire and Guestrin, 2019). Similarly, the process of testing a new pharmaceutical drug requires deep expertise, and is massively expensive and labor-intensive, often taking several years (Vamathevan et al., 2019).

In the meantime, recent Large Language Models (LLMs) (Touvron et al., 2023; OpenAI, 2023; Anil et al., 2023) have shown impressive capabilities in processing and generating text with remarkable accuracy, even outperforming human experts across diverse specialized domains including math, physics, history, law, medicine, and ethics. They are able to process and analyze large volumes of data at speeds and scales far exceeding human capabilities, have internalized large swaths of human knowledge from being trained on virtually the entire web, and can identify patterns, trends, and correlations that may not be immediately apparent to human researchers (such as the usage of quantum mechanics in medical imaging or applying psychological insights in AI). This renders them ideally poised to become foundational tools to accelerate the two phases of the scientific research process: ideation of novel research opportunities, and scientific validation of those research hypotheses.

A few recent papers in the domain of LLM-augmented scientific discovery have focused on the second phase. Specifically, they attempt (Huang et al., 2023; AI4Science and Quantum, 2023; Bran et al., 2023) to mainly accelerate the experimental validation process, by writing code for machine-learning models, facilitating the exploration of chemical spaces, or advancing the simulation of molecular dynamics. Thus, in this paper, we leverage LLMs in the first phase of scientific research – specifically idea generation, whose key focus is conceptualizing novel research questions, methodologies, and experiments. To the best of our knowledge, our work is the first to leverage and evaluate the capabilities of LLMs to act as mediators in scientific idea generation in an open-ended setting.

Given our goal to build an LLM-powered ResearchAgent, we draw inspiration from how human researchers position themselves to come up with novel research ideas. We draw distinctions between three key components of their workflow: a broad and deep understanding of related scientific literature, an encyclopedic view of concepts and how they relate to one another both within and across domains, and a community of colleagues on which to rely for feedback and constructive criticism.

We model each of these three aspects in our ResearchAgent. Specifically, in order to imbibe related work, the system begins with a core scientific paper and then explores a range of related papers through references and citation relationships. Further, to develop an encyclopedic view of related concepts, we build and then augment ResearchAgent with an entity-centric knowledge store derived from co-occurrences of key concepts in the scientific literature. This repository is aimed at capturing novel underlying relationships within and across domains, thereby increasing the chances of a cross-pollination of ideas (Wahle et al., 2023). Finally, to simulate robust feedback mechanisms, we instantiate a number of LLM-powered ReviewingAgents that help the ResearchAgent to iterate on research idea generation with constructive critiques. Crucially, these ReviewingAgents are prompted with evaluation criteria that are induced from real researchers' judgements, thus aligning them with actual scientific preferential standards. An illustration of our system is provided in Figure 1.

We validate the effectiveness of ResearchAgent for research idea generation based on scientific literature across multiple disciplines. Then, on a battery of tests conducted with both human- and model-based evaluations, we demonstrate that ResearchAgent outperforms strong LLM-powered baselines by large margins, generating more clear, relevant, and significant ideas that are especially novel. Furthermore, analyses show the efficacy of our comprehensive approach to modeling ResearchAgent: the entity-centric knowledge store and the iterative idea refinement steps help the system generate meaningfully better ideas compared with an instantiation that is purely based on prior related work.

These findings indicate the huge promise of AI-mediated research assistants, and our initial novel foray into scientific idea generation has important implications for future work that seeks to explore and improve upon the work we have proposed here. These include better support and operationalization to experimentally validate scientific ideas, and the design and evaluation of the utility of these systems to end users, applications, and industries.

## 2 Related Work

**Large Language Models** LLMs have shown impressive performances across a wide range of tasks (OpenAI, 2023; Anil et al., 2023), including ones in scientific fields such as mathematics, physics, medicine, and computer science (Portenoy et al., 2021; Romera-Paredes et al., 2023; Bran et al., 2023; Huang et al., 2023; Liu et al., 2024). A study on GPT-4 shows that it is capable of understanding DNA sequences, designing biomolecules, predicting the behavior of molecular systems, and solving Partial Differential Equation (PDE) problems (AI4Science and Quantum, 2023). However, LLMs have mainly been used for accelerating the experimental validation of already identified research ideas, but not for identifying new problems.

**Hypothesis Generation** The principle of hypothesis generation is based on literature-based discovery (Swanson, 1986), which aims to discover relationships between concepts (Henry and McInnes, 2017). For instance, these concepts could be a specific disease and a compound not yet considered as a treatment for it. Early works on automatic hypothesis generation first build a corpus of discrete concepts, and then identify their relationships with machine learning approaches, e.g., using similarities between word (concept) vectors (Tshitoyan et al., 2019) or applying link prediction methods over a graph (where concepts are nodes) (Sybrandt et al., 2020; Nadkarni et al., 2021). Recent approaches are further powered by LLMs (Wang et al., 2023b; Qi et al., 2023; Yang et al., 2023), leveraging their prior knowledge about scientific disciplines. However, all these approaches perform idea generation in a localized manner and are designed to identify potential relationships between two variables or to generate textual descriptions about them, which may be sub-optimal to capture the complexity and multifaceted nature of real-world problems (e.g., urban planning involve numerous interacting variables). Meanwhile, we do not artificially restrict the generated research idea to be a predictive single variable or simple binary link, instead allowing the model to generate ideas in an open-ended fashion.

**Knowledge-Augmented LLMs** The approach to augment LLMs with external knowledge enhances their utility, making them more accurate and relevant to specific target contexts. Much prior work aims at improving the factuality of LLM responses to given queries by retrieving the relevant documents and then injecting them into the input of LLMs (Lazaridou et al., 2022; Ram et al., 2023; Shi et al., 2023). In addition, given that entities or facts are atomic units for representing knowledge, recent studies further augment LLMs with them (Baek et al., 2023; Wu et al., 2023). In contrast to these efforts which use knowledge units piecemeal, we instead jointly leverage accumulated knowledge over massive troves of scientific papers. More recently, Baek et al. (2024) proposes to use accumulated entities (extracted from various web search contexts) for query suggestion, which – while similar – has the entirely different objective of narrowing the focus of LLMs to entities already present in an LLM's context.

**Iterative Refinements with LLMs** Similar to humans, LLMs do not always generate optimal outputs on their first attempt. Drawing inpiration from humans who can iteratively refine their thoughts based on critiques from themselves and their peers, many recent studies (including some hypothesis generation work) have investigated the potential of LLMs to correct and refine their outputs, demonstrating that they indeed possess those capabilities (Welleck et al., 2023; Madaan et al., 2023; Shridhar et al., 2023; Ganguli et al., 2023; Wang et al., 2023b; Qi et al., 2023; Yang et al., 2023).

## 3 Method

We present ResearchAgent, a system that automatically proposes research ideas with LLMs.

### 3.1 LLM-Powered Research Idea Generation

We begin by formally introducing the new problem of research idea generation, followed by an explanation of how LLMs are utilized to tackle it.

**Research Idea Generation** The goal of the research idea generation task is to formulate new and valid research ideas, to enhance the overall efficiency of the first phase of scientific discovery. While we acknowledge that the real process by which humans conduct research is varied and complex to an extent well beyond the scope of this scientific study, we attempt to model simulacra in three systematic steps that would likely be maximally beneficial to a researcher seeking assistance from an AI system. These are namely, identifying novel research ideas, proposing methods to validate these ideas, and designing experiments to measure the success of these methods in relation to the ideas.

To accomplish the aforementioned steps, we utilize the existing literature (e.g., academic publications) as a primary source, which provides insights about existing knowledge along with gaps and unanswered questions[1]. Formally, let $\mathcal{L}$ be the literature, and $\boldsymbol{o}$ be the ideas that consist of the problem $\boldsymbol{p}$, method $\boldsymbol{m}$, and experiment design $\boldsymbol{d}$, as follows: $\boldsymbol{o} = [\boldsymbol{p}, \boldsymbol{m}, \boldsymbol{d}]$ where each item consists of a sequence of tokens and $[\cdot]$ denotes a concatenation operation. Then, the idea generation model $f$ can be represented as follows: $\boldsymbol{o} = f(\mathcal{L})$, which is further decomposed into three submodular steps: $\boldsymbol{p} = f(\mathcal{L})$ for identifying problems, $\boldsymbol{m} = f(\boldsymbol{p}, \mathcal{L})$ for developing methods, and $\boldsymbol{d} = f(\boldsymbol{p}, \boldsymbol{m}, \mathcal{L})$ for designing experiments. In this work, we opera-

[1] We focus on the existing literature-based idea generation by following the paradigm that a *new idea* is more often than not just a new combination of old elements (Young, 2003).

tionalize $f$ with LLMs, leveraging their capability to understand and generate academic text.

**Large Language Models** Before describing the LLM in the context of our problem setup, let us first provide its general definition, which takes an input sequence of tokens $\boldsymbol{x}$ and generates an output sequence of tokens $\boldsymbol{y}$, as follows: $\boldsymbol{y} = \texttt{LLM}_\theta(\mathcal{T}(\boldsymbol{x}))$. Here, the model parameters $\theta$ are typically fixed after training, due to the high costs of further fine-tuning. In addition, the prompt template $\mathcal{T}$ serves as a structured format that outlines the context (including the task descriptions and instructions) to direct the model in generating the desired outputs.

### 3.2 Knowledge-Augmented LLMs for Research Idea Generation

We now turn to our primary focus of automatically generating research ideas with LLMs. Recall that we aim to produce a complete idea consisting of the problem, method, and experiment design ($\boldsymbol{o} = [\boldsymbol{p}, \boldsymbol{m}, \boldsymbol{d}]$), while using the existing literature $\mathcal{L}$ as a primary source of information. We operationalize this with LLMs by instantiating the aforementioned research idea generation function $f$ with $\texttt{LLM}$ coupled with the task-specific template. Formally, $\boldsymbol{p} = \texttt{LLM}(\mathcal{T}_p(\mathcal{L}))$ indicates the problem identification step, followed by $\boldsymbol{m} = \texttt{LLM}(\mathcal{T}_m(\boldsymbol{p}, \mathcal{L}))$ for method development and $\boldsymbol{d} = \texttt{LLM}(\mathcal{T}_e(\boldsymbol{p}, \boldsymbol{m}, \mathcal{L}))$ for experiment design, which constitutes the full idea: $\boldsymbol{o} = [\boldsymbol{p}, \boldsymbol{m}, \boldsymbol{d}]$.

Following this general formulation, the important question to answer is how the body of scientific literature is leveraged for actually generating research ideas with LLMs. Here, we outline three key desiderata that contribute to the success of human researchers ideating novel research ideas: a broad and deep understanding of related work, an encyclopedic perspecitve on the interconnectedness of concepts within and across scientific domains, and a community of peers who help iteratively improve ideas through constructive critiques. We describe our operationalization of these three desiderata using the prior literature and LLMs in what follows.

**Citation Graph based Literature Survey** Due to the constraints on their input lengths and their reasoning abilities, particularly over very long contexts (Liu et al., 2023a), it is not possible to incorporate all the existing publications from the literature $\mathcal{L}$ into the LLM input. Instead, we need to find a meaningful subset relevant to the problem at hand. To achieve this, we mirror the process followed by human researchers, who expand their knowledge of a paper by perusing other papers that either cite or are cited by it. Concretely, for the $\texttt{LLM}$, we initiate its literature review process by providing a core paper $l_0$ from $\mathcal{L}$ and then selectively incorporating subsequent papers $\{l_1, ..., l_n\}$ that are directly connected based on a citation graph. This procedure makes the $\texttt{LLM}$ input for idea generation more manageable and coherent. In addition, we operationalize the selection process of the core paper and its relevant citations with two design choices: 1) the core paper is selected based on its citation count (e.g., exceeding 100 over 3 months) typically indicating high impact; 2) its relevant papers (which may be potentially numerous) are further narrow-downed based on their similarities of abstracts with the core paper, ensuring a more focused and relevant set of related work.

**Entity-Centric Knowledge Augmentation** In order to model an encyclopedic view of interconnected concepts, we must effectively design a framework to extract, store and effectively leverage the vast amount of knowledge in scientific literature $\mathcal{L}$. In this work, we view entities as the atomic units of knowledge, which allows for ease of representation and accumulation over papers in a unified manner across different disciplines. For example, we can easily extract the term "database" whenever it appears in any paper, using existing off-the-shelf entity linking methods and then aggregate their linked occurrences into a knowledge store. Then, if the term "database" is prevalent within the realm of medical science but less so in hematology (which is a subdomain of medical science), the constructed knowledge store can capture the affinity between those two domains based on overlapping entities. This representational paradigm can then be used to suggest the term "database" when formulating the ideas about hematology. In other words, this approach enables providing novel and interdisciplinary insights by leveraging the interconnectedness of entities across various fields.

Formally, we design the knowledge store as a two-dimensional matrix $\mathcal{K} \in \mathcal{R}^{m \times m}$ where $m$ is the total number of unique entities identified and $\mathcal{K}$ is implemented in a sparse format. This knowledge store is constructed by extracting entities over all the available scientific articles in literature $\mathcal{L}$[2], which not only counts the co-occurrences between

[2]As extracting entities on all articles is computationally infeasible, we target papers appearing after May 01, 2023.

entity pairs within individual papers but also quantifies the count for each entity. Our approach is versatile, thus, we can use any entity linker (Wu et al., 2020). Also, despite the lack of entity linkers customized for the scientific domain, the off-the-shelf system proved capable of extracting key scientific entities, as shown in Table 16. Specifically, this linker tags and canonicalizes entities in a paper $l$ from $\mathcal{L}$, formalized as follows: $\mathcal{E}_l = \texttt{EL}(l)$ where $\mathcal{E}_l$ denotes a multiset of entities (allowing for repetitions) appearing in $l$[3]. Upon extracting entities $\mathcal{E}$, to store them into the knowledge store $\mathcal{K}$, we consider all possible pairs of $\mathcal{E}$ represented as follows: $\{e_i, e_j\}_{(i,j)\in\mathcal{C}(|\mathcal{E}|,2)}$ where $e \in \mathcal{E}$.

Given this knowledge store $\mathcal{K}$, our next goal is to enhance the previous vanilla research idea generation process implemented based on a group of interconnected papers, denoted as follows: $\boldsymbol{o} = \texttt{LLM}(\mathcal{T}(\{l_0, l_1, ..., l_n\}))$. We do this by augmenting the LLM with the relevant entities from $\mathcal{K}$, which can expand the contextual knowledge – what LLMs can consume – by offering additional knowledge. In other words, this knowledge is not seen in the current group of papers but is relevant to it, identified based on entity (co-)occurrence information stored in $\mathcal{K}$. Formally, let us define entities extracted from the group of interconnected papers, as follows: $\mathcal{E}_{\{l_0,...,l_n\}} = \bigcup_{i=0}^{n} \texttt{EL}(l_i)$. Then, the probabilistic form of retrieving the top-$k$ relevant external entities can be represented as follows:

$$\texttt{Ret}(\{l_0, ..., l_n\}; \mathcal{K}) = \underset{I\subset[m]:|I|=k}{\arg\max} \prod P(e_i|\mathcal{E}_{\{l_0,...,l_n\}}), \quad (1)$$

where $[m] = \{1, ..., m\}$ and $e_i \notin \mathcal{E}_{\{l_0,...,l_n\}}$. Also, for simplicity, by applying Bayes' rule and assuming that entities are independent, the retrieval operation (Equation 1) can be approximated as follows:

$$\underset{I\subset[m]:|I|=k}{\arg\max} \prod ( \prod_{e_j\in\mathcal{E}_{\{l_0,...,l_n\}}} P(e_j|e_i)) \times P(e_i), \quad (2)$$

where $P(e_j|e_i)$ and $P(e_i)$ can be derived from values in the two-dimensional matrix $\mathcal{K}$, suitably normalized. We note that the formulation in Equation 2 is only one instance of operationalizing retrieval; this could be replaced with other retrieval strategies – for example, the embedding-based retrieval (discussions and results are provided in Appendix B.4). Hereafter, the instantiation of research proposal generation augmented with relevant entity-centric knowledge is formalized as follows: $\boldsymbol{o} = \texttt{LLM}(\mathcal{T}(\{l_0, ..., l_n\}, \texttt{Ret}(\{l_0, ..., l_n\}; \mathcal{K})))$[4]. We call this knowledge-augmented LLM-powered idea generation approach ResearchAgent, and provide the templates to instantiate it in Tables 6, 7, and 8.

**Iterative Research Idea Refinements** Finally, in order to model a community of peers for idea improvement, we propose a set of LLM-powered reviewing agents (called ReviewingAgents). These agents provide the ResearchAgent with reviews and feedback according to specific criteria in order to help it iteratively improve idea generation.

Specifically, similar to our approach to instantiate ResearchAgent with an LLM (LLM) and template ($\mathcal{T}$), ReviewingAgents are instantiated similarly but with different templates (See Tables 9, 10, and 11). Then, with ReviewingAgents, each of the generated research ideas (problem, method, and experiment design) is separately evaluated according to its own specific five criteria[5], which are provided in labels of Figure 2 and detailed in Table 12. Based on the reviews and feedback from ReviewingAgents, the ResearchAgent iteratively updates and refines its generation of research ideas.

Despite the proficiency of LLMs in the evaluation of machine-generated texts (Zheng et al., 2023; Fu et al., 2023), their judgments on research ideas may not be aligned with the judgments of real human researchers. However, there are no ground truth reference judgments available, and collecting them to align LLMs is expensive and often infeasible. Ideally, the judgments made by LLMs should be similar to the ones made by humans, and we aim to ensure this by automatically generating human preference-aligned evaluation criteria (used for automatic evaluations) with a few human annotations. Specifically, to obtain these human-aligned evaluation criteria, we first collect 10 pairs of research ideas and their associated scores (on a 5-point Likert scale annotated by human researchers having at least 3 papers) on every evaluation criterion. Then, we prompt the LLM with these human-annotated pairs to induce detailed descriptions for evaluation criteria (Lin et al., 2024) (See Tables 13, 14, and 15) that reflect the human preferences[6], which are then used as evaluation criteria by the ReviewingAgents.

[3]Due to the extensive length of scientific publications, the target of entity extraction is restricted to titles and abstracts.

[4]There may be additional knowledge sources (beyond the existing literature and entities) for research idea generation, and we leave exploring them as future work.

[5]We select the top five criteria which we consider as the most important, and leave exploring others as future work.

[6]We additionally ask five human annotators, who evaluate research ideas, to judge the quality of the induced criteria, and two of them strongly agree and three of them agree with them.

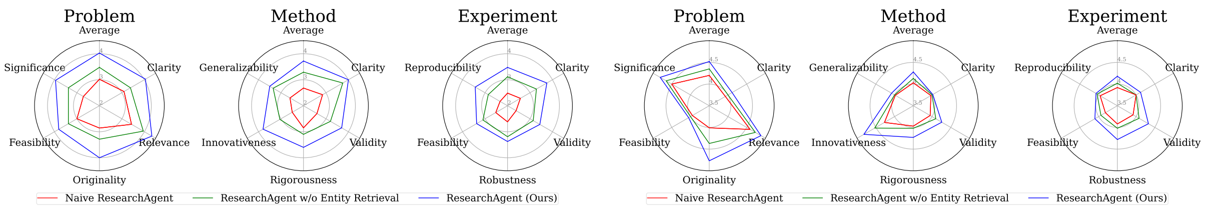


Figure 2: Main results on our research idea generation task with human- (left) and model-based (right) evaluations, where we report the score of each idea (problem, method, or experiment design) based on its own five criteria and their average score.

# 4 Experimental Setups

In this section, we describe the datasets, models, evaluation setup, and implementation details.

## 4.1 Data

The main source to generate research ideas is scientific literature $\mathcal{L}$, which we obtain from Semantic Scholar Academic Graph API[7]. From this, we select papers appearing after May 01, 2023, because LLMs that we use in our experiments are trained on data from the open web available before this point. This follows the procedure of existing literature-based hypothesis generation work (Qi et al., 2023). Then, we select high-impact papers (that have more than 20 citations) as core papers, mirroring human researchers' tendency to leverage influential work, to ensure the high quality of the generated ideas. The resulting data is still very large; thus, we further randomly sample a subset of 300 papers as core papers to obtain a reasonably sized benchmark dataset. The average number of reference papers for each core paper is 87; the abstract of each paper has 2.17 entities on average. The distribution of disciplines for all papers is provided in Figure 7.

## 4.2 Baselines and Our Model

As we target the novel task of research idea generation involving the generation of problems, methods, and experimental designs (whose setup differs from existing hypothesis generation works that identify relationships between two variables), there are no existing baselines that would serve as direct comparison[8]. Thus, we mainly compare our full ResearchAgent model, which utilizes both references and entities, against its ablated variants as follows: 1. **Naive ResearchAgent** – which uses only a core paper to generate research ideas. 2. **ResearchAgent w/o Entity Retrieval** – which uses the core paper and its relevant references without considering entities. 3. **ResearchAgent** – which is our full model that uses the relevant references and entities along with the core paper, to augment LLMs.

[7]https://www.semanticscholar.org/product/api

[8]The comparison results of ResearchAgent against hypothesis generation approaches are discussed in Appendix B.3.

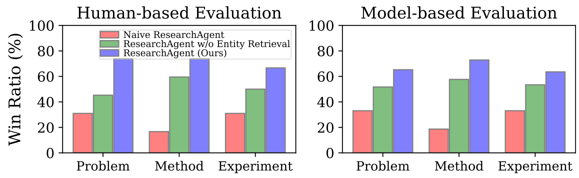


Figure 3: Results of pairwise comparisons between ideas from two of any different approaches, where we report the win ratio.

## 4.3 Evaluation Setup

Given our formulation of idea generation (Sec 3.1), there are no ground-truth answers to measure the quality of the generated ideas. Meanwhile, exhaustively listing pairs of core papers and reference research ideas is suboptimal, since there may exist a large number of valid research ideas for each core paper, and this process requires much time, effort and expertise on the part of human researchers. Thus, we turn to model-based automatic evaluation as well as manual human evaluation to validate different models on our experimental benchmark.

**Model-based Evaluation** Following the recent trends in using LLMs to judge the quality of output texts (especially in the setting of reference-free evaluations) (Zheng et al., 2023; Fu et al., 2023; Liu et al., 2023b), we use GPT-4 to judge the quality of research ideas. Note that each of the problem, method, and experiment design is evaluated with five different criteria (See labels of Figure 2 for criteria and see Table 12 for their detailed descriptions). We ask the LLM-based evaluation model to either rate the generated idea on a 5-point Likert scale for each criterion or perform pairwise comparisons between two ideas from different models. We provide the prompts for evaluations in Appendix A.

**Human Evaluation** Similar to model-based evaluations, we perform human evaluations that involve assigning a score for each criterion and conducting pairwise comparisons between two ideas. As the generated ideas are knowledge-intensive, we carefully select annotators who are well-versed in the field and provide them with ideas that are highly relevant to their field of expertise. Specifically, we choose ten expert researchers who have authored at least three papers and ask them to judge only the ideas that are generated based on their own papers.

Table 1: Results of agreements between two human annotation results and between human and model evaluation results.

| Categories | Metrics | Problem | Method | Experiment |
|---|---|---|---|---|
| **Human and Human** | Scoring | 0.83 | 0.76 | 0.67 |
| | Pairwise | 0.62 | 0.62 | 0.41 |
| **Human and Model** | Scoring | 0.64 | 0.58 | 0.49 |
| | Pairwise | 0.71 | 0.62 | 0.52 |

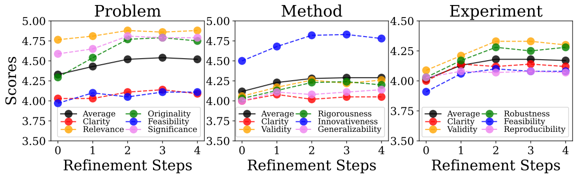


Figure 4: Results with varying the number of refinement steps.

### 4.4 Implementation Details

We mainly use the GPT-4 (OpenAI, 2023) release from Nov 06 as the basis for all models, which is, notably, reported to be trained with data up to Apr 2023 (meanwhile, the papers used for idea generation appear after May 2023). To extract entities and build the entity-centric knowledge store, we use the off-the-shelf BLINK entity linker (Wu et al., 2020), with papers from May 01, 2023, to Dec 31, 2023 (available from Semantic Scholar API) along with their references, which number 50,091 in total. We provide detailed prompts used to elicit responses for research idea generation in Appendix A.3.

## 5 Experimental Results and Analyses

We present experimental results and various analyses, showing the effectiveness of ResearchAgent.

**Main Results** Our main results on scoring with human and model-based evaluations are provided in Figure 2. These demonstrate that our full ResearchAgent outperforms all baselines by large margins on all metrics across all the problems, methods, and experiment designs generated (constituting the complete research ideas). Particularly, the full ResearchAgent augmented with relevant entities exhibits strong gains on metrics related to creativity (such as Originality for problems and Innovativeness for methods) since entities may offer novel concepts and views that may not be observable in the group of papers (core paper and its references) used for generating ideas. In addition, the results of pairwise comparisons between two of any models with human and model-based evaluations are reported in Figure 3, on which the full ResearchAgent shows the highest win ratio over its baselines.

**Analysis on Inter-Annotator Agreements** To validate the quality and reliability of human annotations, we measure the inter-annotator agreements, where 20% of the generated ideas are evaluated by two humans, and report the results in Table 1. Specifically, for the scoring, we first rank scores from each annotator and measure Spearman's correlation coefficient (Pirie, 2006) between the ranked scores of two annotators. For the pairwise comparison between two judges, we measure Cohen's kappa coefficient (Cohen, 1960). As shown in Table 1, we observe that inter-annotator agreement is high, confirming the reliability of our assessments about the quality of generated research ideas.

**Analysis on Human-Model Agreements** Similar to what we did for the aforementioned inter-annotator agreements, we measure agreements between human-based and model-based evaluations, to ensure the reliability of model-based evaluations. As shown in Table 1, we further confirm that agreements between humans and models are high, indicating that model-based evaluations are a reasonable alternative to judge research idea generation.

**Analysis of Refinement Steps** To see the effectiveness of iterative refinements of research ideas with ReviewingAgents, in Figure 4, we report the averaged scores on the generated ideas as a function of refinement steps. We first observe initial improvements in the quality of generated ideas with increased refinement steps. Yet, the performance becomes saturated after three iterations, which may indicate diminishing returns for subsequent iteration steps, which aligns with the pattern observed in agent-based refinement work (Du et al., 2023).

**Ablation on Knowledge Sources** Recall that the full ResearchAgent is augmented with two different knowledge sources, namely relevant references and entities. To see their individual contribution, we perform an ablation study by either excluding one of the knowledge sources or replacing it with random elements. As shown in Table 2, each knowledge source appears to contribute to performance improvement, and the relevant references are especially helpful. We also note that providing random elements is more helpful than providing no elements at all; we hypothesize that this may be due to the LLM's capability to filter out noise while still gaining incidental value from random inputs.

**Analysis on Human Alignment for Evaluation** Recall that to align judgments from model-based evaluations with actual human preferences, we generated the evaluation criteria based on human evaluation results and used them as the criteria for

Table 2: Results of ablation study on references and entities.

| Methods | Problem | Method | Experiment |
|---|---|---|---|
| ResearchAgent | **4.52** | **4.28** | **4.18** |
| - w/o Entities | 4.35 | 4.13 | 4.02 |
| - w/ Random Entities | 4.41 | 4.19 | 4.13 |
| - w/o References | 4.26 | 4.08 | 3.97 |
| - w/ Random References | 4.35 | 4.16 | 4.02 |
| - w/o Entities & References | 4.20 | 4.03 | 3.92 |

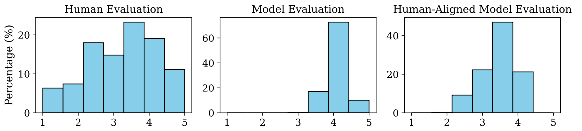

Figure 5: Distributions of model-based evaluation results with and without the human-induced score criteria alignment (middle and right), as well as human evaluation results (left).

model-based evaluations. Figure 5 demonstrates the efficacy of this strategy, presenting the score distribution of human evaluation compared with the distributions of model-based evaluations with and without human alignment. We find that the score distribution of model-based evaluations without alignment is skewed and different from the score distribution of human judgments. Meanwhile, after aligning the model-based evaluations with human-induced score criteria, the calibrated distribution more closely resembles the distribution of humans.

**Correlation on Citation Counts** We further investigate whether a high-impact paper (when used as a core paper) leads to high-quality research ideas. To measure this, we categorize papers by their citation count (as a proxy for impact), and visualize the average score of each bucket (with model-based evaluations) in Figure 6. We find that ideas from high-impact papers tend to be of higher quality, likely due to their ability to identify research gaps, propose feasible methods, and connect with other works. Additionally, based on the paper distribution (See Figure 7) and for the ease of manual quality check, evaluation criteria for model-based evaluations are induced mainly with computer science papers. To see whether those criteria are applicable to diverse fields, we also compare a correlation between scores of computer science papers and all papers in Figure 6. From this, we observe that the scores increase when the citation increases for both domains, which may support the generalizability of human-preference-induced evaluation criteria.

**Analysis using Different LLMs** To see how the performance of ResearchAgent changes if an LLM other than the GPT-4 is used, we conduct an auxiliary analysis instantiating the ResearachAgent with different LLMs, such as Llama3, Mixtral, Qwen1.5, and GPT-3.5 (Bai et al., 2023; Jiang et al., 2024),

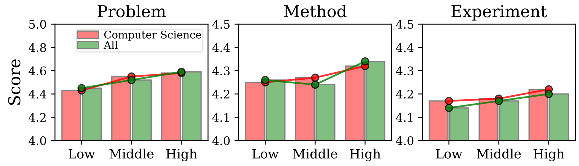

Figure 6: Results with bucketing papers based on citations.

Table 3: Results with different, open and proprietary LLMs.

| LLMs | Models | Problem | Method | Experiment |
|---|---|---|---|---|
| **GPT-4.0** | Naive ResearchAgent | 4.20 | 4.03 | 3.92 |
| | ResearchAgent (Ours) | 4.52 | 4.28 | 4.18 |
| **GPT-3.5** | Naive ResearchAgent | 3.56 | 3.56 | 3.63 |
| | ResearchAgent (Ours) | 3.58 | 3.58 | 3.60 |
| **Llama3 (8B)** | Naive ResearchAgent | 3.76 | 3.69 | 3.54 |
| | ResearchAgent (Ours) | 4.18 | 4.03 | 3.95 |
| **Mixtral (8x7B)** | Naive ResearchAgent | 3.31 | 3.27 | 3.20 |
| | ResearchAgent (Ours) | 3.28 | 3.35 | 3.31 |
| **Qwen1.5 (32B)** | Naive ResearchAgent | 3.64 | 3.74 | 3.66 |
| | ResearchAgent (Ours) | 4.02 | 3.97 | 3.94 |

and present the model-based evaluation results in Table 3. We then find that the performance with less capable models (other than GPT-4) drops significantly. Moreover, the performance differences between the Naive ResearchAgent without knowledge augmentation and the full ResearchAgent become marginal, for Mixtral and GPT-3.5, which indicates that they might not be capable of capturing complex concepts across different scientific papers. This is unsurprising if taken in the context of the emergent abilities of LLMs for complex reasoning (but not in smaller LMs) (Wei et al., 2022).

## 6 Conclusion

In this work, we presented ResearchAgent – a system that aims to assist researchers in their workflow by automatically generating research ideas, which consists of novel problem identification, method development, and experiment design. Drawing inspiration from the human process of research ideation, we developed an approach that simultaneously conducts a broad and deep review of relevant literature, leverages encyclopedic knowledge through interconnected concepts across domains to help cross-pollination of ideas, and leverages a community of reviewing agents to provide constructive critiques for iteratively refining the research ideas. Through human and model-based evaluations, we showed that ResearchAgent generates ideas that are more creative, valid, and clear than ones from baselines. While we envision ResearchAgent as a collaborative partner for scientists, this initial foray has only demonstrated early signs of the promise of AI-mediated research assistants. There are multiple important avenues of future research to pursue, including improving and building upon ResearchAgent, operationalizing experimental validation of its research hypotheses, and measuring the real-world value it brings to researchers and their productivity.

## Limitations

ResearchAgent has some limitations that we hope to address in future work, discussed as follows:

First, recall that we built the entity-centric knowledge store to propose beneficial entities during idea generation; however this store is constructed by extracting entities from the titles and abstracts of a limited number of publications (due to the costs of processing them) thereby precluding a large number of other entities and their interconnectedness.

In addition, the number of entities that we obtain from the BLINK entity linker (Wu et al., 2020) amounts to 3 per paper on average, indicating limited coverage (it is an open-domain linker after all), although it exhibits the generally strong understanding of scientific contexts, as demonstrated by the improvement achieved by the inclusion of its predictions (See Figures 2 and 3, and Table 16).

Lastly, since our ResearchAgent is powered by LLMs, similar to any other approaches based on LLMs, it may hallucinate the generated research ideas. While our proposed ResearchAgent can partially mitigate this problem by augmenting LLMs with additional elements, such as references to the target paper and greater entity-centric knowledge, which help ground the generation process in more accurate and relevant information, validating these generated research ideas with experiments is essential to truly accelerate scientific research.

## Ethics Statement

We are aware that the ResearchAgent may have the potential to be misused for nefarious purposes, such as generating research ideas about new explosives, malicious software, and invasive surveillance tools. Notably, this vulnerability is not unique to our approach but a common challenge faced by existing LLMs that possess significant creative and reasoning capabilities, occasionally generating content that may be deemed undesirable. Consequently, it underscores the necessity to enhance the robustness and safety of LLMs more broadly.

Also, we recognize the risks of unintentional plagiarism associated with using ResearchAgent, where the system might generate ideas that closely mirror existing research due to the recitation of training data. While mitigation strategies, such as integrating access to a comprehensive knowledge base to inform users of prior work, can be employed, we understand that building and maintaining such a resource is inherently complex and may not fully eliminate the risk. To further reduce the possibility of plagiarism, recording and tracking all generated ideas could help identify similarities and guide the model to avoid repetition, though this approach would necessitate explicit user consent.

## References


Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106, Toronto, Canada. Association for Computational Linguistics.

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen Herring, and Sujay Kumar Jauhar. 2024. Knowledge-augmented large language models for personalized contextual query suggestion. *WWW*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023.

Chemcrow: Augmenting large-language models with chemistry tools.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

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

Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: Observing goodhart's law in action. *GigaScience*, 8.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamile Lukosiute, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Sam Henry and Bridget T. McInnes. 2017. Literature based discovery: Models, methods, and trends. *Journal of biomedical informatics*, 74:20–32.

Tom Hope, Doug Downey, Daniel S. Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. *Commun. ACM*, 66(8):62–73.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Benchmarking large language models as AI research agents. *arXiv preprint arXiv:2310.03302*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv: 2401.04088*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Ying-Chun Lin, Jennifer Neville, Jack W Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. Interpretable user satisfaction estimation for conversational systems with large language models. *arXiv preprint arXiv:2403.12388*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How AI processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 17:1–17:25. ACM.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

R.K. Nadkarni, David Wadden, Iz Beltagy, Noah A. Smith, Hannaneh Hajishirzi, and Tom Hope. 2021. Scientific language models for biomedical knowledge base completion: An empirical study. *ArXiv*, abs/2106.09700.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

W. Pirie. 2006. *Spearman Rank Correlation Coefficient*, volume 8.

Jason Portenoy, Marissa Radensky, Jevin West, Eric Horvitz, Daniel S. Weld, and Tom Hope. 2021. Bridger: Toward bursting scientific filter bubbles and boosting innovation via novel author discovery. *arXiv preprint arXiv:2108.05669*.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, Alhussein Fawzi, Josh Grochow, Andrea Lodi, Jean-Baptiste Mouret, Talia Ringer, and Tao Yu. 2023. Mathematical discoveries from program search with large language models. *Nature*, 625:468 – 475.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ram Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. 2023. The ART of LLM refinement: Ask, refine, and trust. *arXiv preprint arXiv:2311.07961*.

Don R. Swanson. 1986. Undiscovered public knowledge. *The Library Quarterly*, 56:103–118.

Justin Sybrandt, Ilya Tyagin, M. Shtutman, and Ilya Safro. 2020. Agatha: Automatic graph mining and transformer based hypothesis generation approach. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alex Dunn, Ziqin Rong, Olga Vitalievna Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95 – 98.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. 2019. Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery*, 18(6):463—477.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif M. Mohammad. 2023. We are who we cite: Bridges of influence between natural language processing and other academic fields. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12896–12913. Association for Computational Linguistics.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora S. Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. 2023a. Scientific discovery in the age of artificial intelligence. *Nat.*, 620(7972):47–60.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Yike Wu, Nan Hu, Sheng Bi, Guilin Qi, J. Ren, Anhuan Xie, and Wei Song. 2023. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. *arXiv preprint arXiv:2309.11206*.

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

J. Young. 2003. *A Technique for Producing Ideas*. McGraw Hill LLC.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

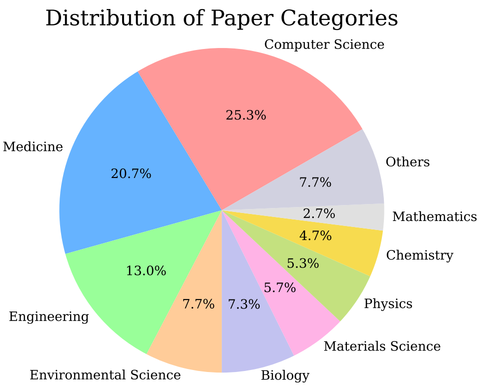


Figure 7: Visualization of the distribution of disciplines for all core papers, selected for research idea generation.

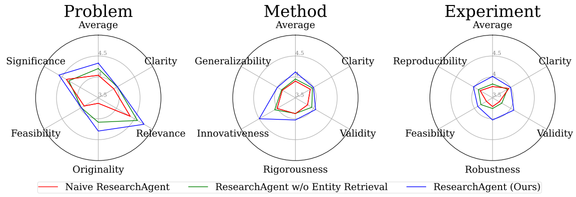


Figure 8: Results on our research idea generation task with model-based evaluation, where we exclude refinement steps.

# A Additional Experimental Details

In this section, we provide additional details on experiments, including datasets, human evaluation setups, prompts (used for research idea generation and validation), and human-induced criteria.

## A.1 Data Statistics

We visualize a distribution of core paper categories used for idea generation in Figure 7, where the categories are obtained from Semantic Scholar API[9]. From this, we find that the top 3 categories are computer science, medicine, and engineering.

## A.2 Details on Human Evaluation

To conduct evaluations with human judges, we recruited 10 researchers from the United States and South Korea, majoring in computer science, medicine, and biology, each with a minimum of 3 published papers. For annotation, they were provided with a 6-page guideline document, which includes the task instruction and annotation examples. After reading this document, the annotators access the Label Studio platform, on which they first read the title and abstract of the target paper, and then review and evaluate the generated research ideas from different models. During the evaluation process, they are allowed to use any external tools, such as web searches. We note that they were compensated at a rate of $22.20 per hour. Also, on average, for one hour, they evaluated 3 sets of research ideas (that are generated from their own papers), with each set comprising three sub-ideas (the problem, method, and experiment design) from three different approaches (i.e., a total of 9 ideas for one hour). We perform three rounds of human evaluations with refinements in between, and, due to the cost associated with human annotations, we are able to fully evaluate a total of 150 ideas.

[9]https://www.semanticscholar.org/product/api

## A.3 Prompts for Ideas Generation

We provide the prompts used to elicit the idea generations from our full ResearchAgent, specifically for instantiating problem identification, method development, and experiment design in Table 6, Table 7, and Table 8, respectively.

## A.4 Prompts for Idea Validation

We provide the prompts used to elicit the idea validation from our ReviewingAgents as well as the model-based evaluations, specifically for instantiating problem validation, method validation, and experiment design validation in Table 9, Table 10, and Table 11, respectively. In addition, we provide the criteria used, which are induced by human judgments in the next subsection (Appendix A.5).

## A.5 Criteria Induced by Human Judgements

Recall that, to align model-based evaluations with human preferences, we induce the criteria (used for automatic evaluations) with actual human judgments. We note that this is done by prompting GPT-4 with 10 pairs of generated ideas and (randomly selected) human judgments. We provide the resulting criteria for validations of problems, methods, and experiment designs in Table 13, Table 14, and Table 15, respectively.

# B Additional Experimental Results

We provide additional experimental results, including comparisons without refinements and examples of the generated research ideas.

## B.1 Results without Refinement Steps

To see whether the proposed ResearchAgent is consistently effective even without ReviewingAgents, we show the model-based evaluation results without any refinement steps in Figure 8. From this, we clearly observe that the full ResearchAgent outperforms its variants, demonstrating its effectiveness.

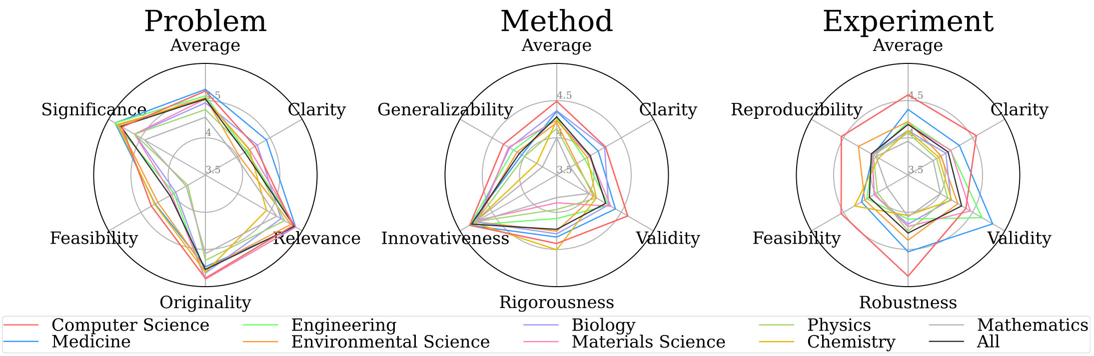


Figure 9: Breakdown results of the research ideas generated from our full ResearchAgent across different domains.

Table 4: Performance comparisons of ResearchAgent against prior hypothesis generation baselines, such as SciMon (Wang et al., 2023b) and Hypothesis Proposer (Yang et al., 2023)

| Methods | Clarity | Relevance | Originality | Feasibility | Significance |
|---|---|---|---|---|---|
| SciMON | 4.04 | 4.37 | 4.56 | 3.98 | 4.15 |
| Hypothesis Proposer | 3.97 | 4.14 | 4.07 | 4.01 | 4.11 |
| ResearchAgent | **4.11** | **4.88** | **4.77** | **4.05** | **4.81** |

Table 5: Results with different entity retrieval strategies.

| Methods | Problem | Method | Experiment |
|---|---|---|---|
| ResearchAgent | | | |
| - w/ Co-occurrence-based Retrieval | **4.52** | 4.28 | **4.18** |
| - w/ Embedding-based Retrieval | 4.49 | **4.34** | 4.16 |
| - w/o Entity Retrieval | 4.35 | 4.13 | 4.02 |

## B.2 Results on Generated Ideas by Domain

To see the quality of the generated research ideas across different domains, we breakdown the performance of ResearchAgent according to the categories of core papers in Figure 7, and present the results in Figure 9. From this, we observe that the generated research ideas on the high-resource domains (such as Computer Science, Medicine, and Engineering where there is a greater volume of existing literature as shown in Figure 7) are superior to those generated from the low-resource domain papers (such as Physics, Chemistry, and Mathematics). This disparity might be attributed to the fact that the underlying LLMs used to generate research ideas are likely trained on data predominantly sourced from high-resource domains, which leads to enhancing their ability to comprehend scientific concepts and produce relevant research ideas in these high-resource fields.

## B.3 Comparisons to Hypothesis Generation

Recall that, as explained in Section 2, the existing approaches for hypothesis generation (Wang et al., 2023b; Qi et al., 2023; Yang et al., 2023) is based on the principle of literature-based discovery (Swanson, 1986); thus, they have a different problem setup, which either predicts a link between two variables or generates one conditioned on another. In contrast, our approach is designed to generate open-ended research ideas (including problems, methods, and experimental setups) given the set of relevant literature from the academic graph and key concepts from the entity-centric knowledge store, without any restrictions on the inputs and outputs regarding variables and their relationships. Therefore, not only the target outputs generated from prior works differ from our work but also the elements used for generating research ideas, which makes comparisons between them non-standard.

Nevertheless, to understand how the quality of the generated research ideas from prior works differs from the ones by our ResearchAgent, we additionally conduct experiments comparing our ResearchAgent against existing works (Wang et al., 2023b; Yang et al., 2023) (that do not require manual annotations of input-output pairs), by generating research ideas with them and then measuring their quality according to Clarity, Relevance, Originality, Feasibility, and Significance via LLM-based evaluations. As shown in Table 4, we observe that our proposed ResearchAgent is capable of generating superior research ideas compared to them.

## B.4 Analysis with Different Entity Retrieval

To see the effectiveness of different entity retrieval strategies, we perform additional experiments, replacing the co-occurrence-based entity retrieval in Equation 2 to the contextual embedding-based retrieval. Notably, this contextual embedding-based retrieval approach uses the entities that have the highest similarity to the entities appearing in the current literature (i.e., core paper and its references) used for idea generation, where the similarities are calculated based on embedding-level similarities between entities over the latent space represented by the entity linker (Wu et al., 2020). Therefore, unlike the previous co-occurrence-based entity retrieval that may retrieve entities that have opposite concepts to the main idea of the current core paper (since we often mention limitations of previous work along with the proposed ideas), this embedding-based approach may provide the ResearchAgent with mostly the entities having similar concepts to the core paper. Nevertheless, as shown in Table 5, the results with the strategy of entity co-occurrence-based retrieval are comparable to the results with the new embedding-based contextual retrieval. These results might confirm that there is not much difference in the quality of entity retrieval among those two strategies, i.e., most entities retrieved from the co-occurrence-based retrieval are contextually relevant for generating research ideas.

### B.5 Examples

We provide examples of generated research ideas (including problems, methods, and experiment designs) in Table 16.

Table 6: The prompt used in the full instantiation of ResearchAgent for problem identification.

| **Types** | **Texts** |
|---|---|
| **System Message** | You are an AI assistant whose primary goal is to identify promising, new, and key scientific problems based on existing scientific literature, in order to aid researchers in discovering novel and significant research opportunities that can advance the field. |
| **User Message** | You are going to generate a research problem that should be original, clear, feasible, relevant, and significant to its field. This will be based on the title and abstract of the target paper, those of {len(references)} related papers in the existing literature, and {len(entities)} entities potentially connected to the research area.<br><br>Understanding of the target paper, related papers, and entities is essential:<br>- 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 problem.<br>- The related papers are studies that have cited the target paper, indicating their direct relevance and connection to the primary research topic you are focusing on, and providing additional context and insights that are essential for understanding and expanding upon the target paper.<br>- The entities can include topics, keywords, individuals, events, or any subjects with possible direct or indirect connections to the target paper or the related studies, serving as auxiliary sources of inspiration or information that may be instrumental in formulating the research problem.<br><br>Your approach should be systematic:<br>- Start by thoroughly reading the title and abstract of the target paper to understand its core focus.<br>- Next, proceed to read the titles and abstracts of the related papers to gain a broader perspective and insights relevant to the primary research topic.<br>- Finally, explore the entities to further broaden your perspective, drawing upon a diverse pool of inspiration and information, while keeping in mind that not all may be relevant.<br><br>I am going to provide the target paper, related papers, and entities, as follows:<br>Target paper title: {paper[’title’]}<br>Target paper abstract: {paper[’abstract’]}<br>Related paper titles: {relatedPaper[’titles’]}<br>Related paper abstracts: {relatedPaper[’abstracts’]}<br>Entities: {Entities}<br><br>With the provided target paper, related papers, and entities, your objective now is to formulate a research problem that not only builds upon these existing studies but also strives to be original, clear, feasible, relevant, and significant. Before crafting the research problem, revisit the title and abstract of the target paper, to ensure it remains the focal point of your research problem identification process.<br><br>Target paper title: {paper[’title’]}<br>Target paper abstract: {paper[’abstract’]}<br><br>Then, following your review of the above content, please proceed to generate one research problem with the rationale, in the format of<br>Problem:<br>Rationale: |

Table 7: The prompt used in the full instantiation of ResearchAgent for method development.

| Types | Texts |
|---|---|
| **System Message** | You are an AI assistant whose primary goal is to propose innovative, rigorous, and valid methodologies to solve newly identified scientific problems derived from existing scientific literature, in order to empower researchers to pioneer groundbreaking solutions that catalyze breakthroughs in their fields. |
| **User Message** | You are going to propose a scientific method to address a specific research problem. Your method should be clear, innovative, rigorous, valid, and generalizable. This will be based on a deep understanding of the research problem, its rationale, existing studies, and various entities.<br><br>Understanding of the research problem, existing studies, and entities is essential:<br>- The research problem has been formulated based on an in-depth review of existing studies and a potential exploration of relevant entities, which should be the cornerstone of your method development.<br>- The existing studies refer to the target paper that has been pivotal in identifying the problem, as well as the related papers that have been additionally referenced in the problem discovery phase, all serving as foundational material for developing the method.<br>- The entities can include topics, keywords, individuals, events, or any subjects with possible direct or indirect connections to the existing studies, serving as auxiliary sources of inspiration or information that may be instrumental in method development.<br><br>Your approach should be systematic:<br>- Start by thoroughly reading the research problem and its rationale, to understand your primary focus.<br>- Next, proceed to review the titles and abstracts of existing studies, to gain a broader perspective and insights relevant to the primary research topic.<br>- Finally, explore the entities to further broaden your perspective, drawing upon a diverse pool of inspiration and information, while keeping in mind that not all may be relevant.<br><br>I am going to provide the research problem, existing studies (target paper & related papers), and entities, as follows:<br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br>Target paper title: {paper['title']}<br>Target paper abstract: {paper['abstract']}<br>Related paper titles: {relatedPaper['titles']}<br>Related paper abstracts: {relatedPaper['abstracts']}<br>Entities: {Entities}<br><br>With the provided research problem, existing studies, and entities, your objective now is to formulate a method that not only leverages these resources but also strives to be clear, innovative, rigorous, valid, and generalizable. Before crafting the method, revisit the research problem, to ensure it remains the focal point of your method development process.<br><br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br><br>Then, following your review of the above content, please proceed to propose your method with its rationale, in the format of<br>Method:<br>Rationale: |

Table 8: The prompt used in the full instantiation of ResearchAgent for experiment design.

| Types | Texts |
|---|---|
| **System Message** | You are an AI assistant whose primary goal is to design robust, feasible, and impactful experiments based on identified scientific problems and proposed methodologies from existing scientific literature, in order to enable researchers to systematically test hypotheses and validate groundbreaking discoveries that can transform their respective fields. |
| **User Message** | You are going to design an experiment, aimed at validating a proposed method to address a specific research problem. Your experiment design should be clear, robust, reproducible, valid, and feasible. This will be based on a deep understanding of the research problem, scientific method, existing studies, and various entities.<br><br>Understanding of the research problem, scientific method, existing studies, and entities is essential:<br>- The research problem has been formulated based on an in-depth review of existing studies and a potential exploration of relevant entities.<br>- The scientific method has been proposed to tackle the research problem, which has been informed by insights gained from existing studies and relevant entities.<br>- The existing studies refer to the target paper that has been pivotal in identifying the problem and method, as well as the related papers that have been additionally referenced in the discovery phase of the problem and method, all serving as foundational material for designing the experiment.<br>- The entities can include topics, keywords, individuals, events, or any subjects with possible direct or indirect connections to the existing studies, serving as auxiliary sources of inspiration or information that may be instrumental in your experiment design.<br><br>Your approach should be systematic:<br>- Start by thoroughly reading the research problem and its rationale followed by the proposed method and its rationale, to pinpoint your primary focus.<br>- Next, proceed to review the titles and abstracts of existing studies, to gain a broader perspective and insights relevant to the primary research topic.<br>- Finally, explore the entities to further broaden your perspective, drawing upon a diverse pool of inspiration and information, while keeping in mind that not all may be relevant.<br><br>I am going to provide the research problem, scientific method, existing studies (target paper & related papers), and entities, as follows:<br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br>Scientific method: {scientificMethod}<br>Rationale: {scientificMethodRationale}<br>Target paper title: {paper['title']}<br>Target paper abstract: {paper['abstract']}<br>Related paper titles: {relatedPaper['titles']}<br>Related paper abstracts: {relatedPaper['abstracts']}<br>Entities: {Entities}<br><br>With the provided research problem, scientific method, existing studies, and entities, your objective now is to design an experiment that not only leverages these resources but also strives to be clear, robust, reproducible, valid, and feasible. Before crafting the experiment design, revisit the research problem and proposed method, to ensure they remain at the center of your experiment design process.<br><br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br>Scientific method: {scientificMethod}<br>Rationale: {scientificMethodRationale}<br><br>Then, following your review of the above content, please proceed to outline your experiment with its rationale, in the format of<br>Experiment:<br>Rationale: |

Table 9: The prompt used in the full instantiation of ReviewingAgent for problem validation.

| Types | Texts |
|---|---|
| **System Message** | You are an AI assistant whose primary goal is to assess the quality and validity of scientific problems across diverse dimensions, in order to aid researchers in refining their problems based on your evaluations and feedback, thereby enhancing the impact and reach of their work. |
| **User Message** | You are going to evaluate a research problem for its {metric}, focusing on how well it is defined in a clear, precise, and understandable manner.<br><br>As part of your evaluation, you can refer to the existing studies that may be related to the problem, which will help in understanding the context of the problem for a more comprehensive assessment.<br>- The existing studies refer to the target paper that has been pivotal in identifying the problem, as well as the related papers that have been additionally referenced in the discovery phase of the problem.<br><br>The existing studies (target paper & related papers) are as follows:<br>Target paper title: {paper['title']}<br>Target paper abstract: {paper['abstract']}<br>Related paper titles: {relatedPaper['titles']}<br>Related paper abstracts: {relatedPaper['abstracts']}<br><br>Now, proceed with your {metric} evaluation approach that should be systematic:<br>- Start by thoroughly reading the research problem and its rationale, keeping in mind the context provided by the existing studies mentioned above.<br>- Next, generate a review and feedback that should be constructive, helpful, and concise, focusing on the {metric} of the problem.<br>- Finally, provide a score on a 5-point Likert scale, with 1 being the lowest, please ensuring a discerning and critical evaluation to avoid a tendency towards uniformly high ratings (4-5) unless fully justified:<br>{criteria}<br><br>I am going to provide the research problem with its rationale, as follows:<br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br><br>After your evaluation of the above content, please provide your review, feedback, and rating, in the format of<br>Review:<br>Feedback:<br>Rating (1-5): |

Table 10: The prompt used in the full instantiation of ReviewingAgent for method validation.

| Types | Texts |
|---|---|
| **System Message** | You are an AI assistant whose primary goal is to assess the quality and soundness of scientific methods across diverse dimensions, in order to aid researchers in refining their methods based on your evaluations and feedback, thereby enhancing the impact and reach of their work. |
| **User Message** | You are going to evaluate a scientific method for its {metric} in addressing a research problem, focusing on how well it is described in a clear, precise, and understandable manner that allows for replication and comprehension of the approach.<br><br>As part of your evaluation, you can refer to the research problem, and existing studies, which will help in understanding the context of the proposed method for a more comprehensive assessment.<br>- The research problem has been used as the cornerstone of the method development, formulated based on an in-depth review of existing studies and a potential exploration of relevant entities.<br>- The existing studies refer to the target paper that has been pivotal in identifying the problem and method, as well as the related papers that have been additionally referenced in the discovery phase of the problem and method.<br><br>The research problem and existing studies (target paper & related papers) are as follows:<br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br>Target paper title: {paper['title']}<br>Target paper abstract: {paper['abstract']}<br>Related paper titles: {relatedPaper['titles']}<br>Related paper abstracts: {relatedPaper['abstracts']}<br><br>Now, proceed with your {metric} evaluation approach that should be systematic:<br>- Start by thoroughly reading the proposed method and its rationale, keeping in mind the context provided by the research problem, and existing studies mentioned above.<br>- Next, generate a review and feedback that should be constructive, helpful, and concise, focusing on the {metric} of the method.<br>- Finally, provide a score on a 5-point Likert scale, with 1 being the lowest, please ensuring a discerning and critical evaluation to avoid a tendency towards uniformly high ratings (4-5) unless fully justified:<br>{criteria}<br><br>I am going to provide the proposed method with its rationale, as follows:<br>Scientific method: {scientificMethod}<br>Rationale: {scientificMethodRationale}<br><br>After your evaluation of the above content, please provide your review, feedback, and rating, in the format of<br>Review:<br>Feedback:<br>Rating (1-5): |

Table 11: The prompt used in the full instantiation of ReviewingAgent for experiment design validation.

| Types | Texts |
|---|---|
| **System Message** | You are an AI assistant whose primary goal is to meticulously evaluate the experimental designs of scientific papers across diverse dimensions, in order to aid researchers in refining their experimental approaches based on your evaluations and feedback, thereby amplifying the quality and impact of their scientific contributions. |
| **User Message** | You are going to evaluate an experiment design for its {metric} in validating a scientific method to address a research problem, focusing on how well it is described in a clear, precise, and understandable manner, enabling others to grasp the setup, procedure, and expected outcomes.<br><br>As part of your evaluation, you can refer to the research problem, scientific method, and existing studies, which will help in understanding the context of the designed experiment for a more comprehensive assessment.<br>- The research problem has been formulated based on an in-depth review of existing studies and a potential exploration of relevant entities.<br>- The scientific method has been proposed to tackle the research problem, which has been informed by insights gained from existing studies and relevant entities.<br>- The existing studies refer to the target paper that has been pivotal in identifying the problem, method, and experiment, as well as the related papers that have been additionally referenced in their discovery phases.<br><br>The research problem, scientific method, and existing studies (target paper & related papers) are as follows:<br>Research problem: {researchProblem}<br>Rationale: {researchProblemRationale}<br>Scientific method: {scientificMethod}<br>Rationale: {scientificMethodRationale}<br>Target paper title: {paper['title']}<br>Target paper abstract: {paper['abstract']}<br>Related paper titles: {relatedPaper['titles']}<br>Related paper abstracts: {relatedPaper['abstracts']}<br><br>Now, proceed with your {metric} evaluation approach that should be systematic:<br>- Start by thoroughly reading the experiment design and its rationale, keeping in mind the context provided by the research problem, scientific method, and existing studies mentioned above.<br>- Next, generate a review and feedback that should be constructive, helpful, and concise, focusing on the {metric} of the experiment.<br>- Finally, provide a score on a 5-point Likert scale, with 1 being the lowest, please ensuring a discerning and critical evaluation to avoid a tendency towards uniformly high ratings (4-5) unless fully justified:<br>{criteria}<br><br>I am going to provide the designed experiment with its rationale, as follows:<br>Experiment design: {experimentDesign}<br>Rationale: {experimentDesignRationale}<br><br>After your evaluation of the above content, please provide your review, feedback, and rating, in the format of<br>Review:<br>Feedback:<br>Rating (1-5): |

Table 12: The criteria used for evaluating research ideas: problems, methods, and experiment designs.

| Types | Criteria | Texts |
|---|---|---|
| **Problem** | **Clarity** | It assesses whether the problem is defined in a clear, precise, and understandable manner. |
| | **Relevance** | It measures whether the problem is pertinent and applicable to the current field or context of study. |
| | **Originality** | It evaluates whether the problem presents a novel challenge or unique perspective that has not been extensively explored before. |
| | **Feasibility** | It examines whether the problem can realistically be investigated or solved with the available resources and within reasonable constraints. |
| | **Significance** | It assesses the importance and potential impact of solving the problem, including its contribution to the field or its broader implications. |
| **Method** | **Clarity** | It assesses whether the method is described in a clear, precise, and understandable manner that allows for replication and comprehension of the approach. |
| | **Validity** | It measures the accuracy, relevance, and soundness of the method in addressing the research problem, ensuring that it is appropriate and directly relevant to the objectives of the study. |
| | **Rigorousness** | It examines the thoroughness, precision, and consistency of the method, ensuring that the approach is systematic, well-structured, and adheres to high standards of research quality. |
| | **Innovativeness** | It evaluates whether the method introduces new techniques, approaches, or perspectives to the research field that differ from standard research practices and advance them in the field. |
| | **Generalizability** | It assesses the extent to which the method can be applied to or is relevant for other contexts, populations, or settings beyond the scope of the study. |
| **Experiment** | **Clarity** | It determines whether the experiment design is described in a clear, precise, and understandable manner, enabling others to grasp the setup, procedure, and expected outcomes. |
| | **Validity** | It measures the appropriateness and soundness of the experimental design in accurately addressing the research questions or effectively validating the proposed methods, ensuring that the design effectively tests what it is intended to examine. |
| | **Robustness** | It evaluates the durability of the experimental design across a wide range of conditions and variables, ensuring that the outcomes are not reliant on a few specific cases and remain consistent across a broad spectrum of scenarios. |
| | **Feasibility** | It evaluates whether the experiment design can realistically be implemented with the available resources, time, and technological or methodological constraints, ensuring that the experiment is practical and achievable. |
| | **Reproducibility** | It examines whether the information provided is sufficient and detailed enough for other researchers to reproduce the experiment using the same methodology and conditions, ensuring the reliability of the findings. |

Table 13: The criteria induced from human judgments for validating the identified problems, which are used to align model-based evaluations with actual human preferences.

| Types | Criteria | Texts |
|---|---|---|
| Problem | Clarity | 1. The problem is presented in a highly ambiguous manner, lacking clear definition and leaving significant room for interpretation or confusion.<br>2. The problem is somewhat defined but suffers from vague terms and insufficient detail, making it challenging to grasp the full scope or objective.<br>3. The problem is stated in a straightforward manner, but lacks the depth or specificity needed to fully convey the nuances and boundaries of the research scope.<br>4. The problem is clearly articulated with precise terminology and sufficient detail, providing a solid understanding of the scope and objectives with minimal ambiguity.<br>5. The problem is exceptionally clear, concise, and specific, with every term and aspect well-defined, leaving no room for misinterpretation and fully encapsulating the research scope and aims. |
| | Relevance | 1. The problem shows almost no relevance to the current field, failing to connect with the established context or build upon existing work.<br>2. The problem has minimal relevance, with only superficial connections to the field and a lack of meaningful integration with prior studies.<br>3. The problem is somewhat relevant, making a moderate attempt to align with the field but lacking significant innovation or depth.<br>4. The problem is relevant and well-connected to the field, demonstrating a good understanding of existing work and offering promising contributions.<br>5. The problem is highly relevant, deeply integrated with the current context, and represents a significant advancement in the field. |
| | Originality | 1. The problem exhibits no discernible originality, closely mirroring existing studies without introducing any novel perspectives or challenges.<br>2. The problem shows minimal originality, with slight variations from known studies, lacking significant new insights or innovative approaches.<br>3. The problem demonstrates moderate originality, offering some new insights or angles, but these are not sufficiently groundbreaking or distinct from existing work.<br>4. The problem is notably original, presenting a unique challenge or perspective that is well-differentiated from existing studies, contributing valuable new understanding to the field.<br>5. The problem is highly original, introducing a pioneering challenge or perspective that has not been explored before, setting a new direction for future research. |
| | Feasibility | 1. The problem is fundamentally infeasible due to insurmountable resource constraints, lack of foundational research, or critical methodological flaws.<br>2. The problem faces significant feasibility challenges related to resource availability, existing knowledge gaps, or technical limitations, making progress unlikely.<br>3. The problem is feasible to some extent but faces notable obstacles in resources, existing research support, or technical implementation, which could hinder significant advancements.<br>4. The problem is mostly feasible with manageable challenges in resources, supported by adequate existing research, and has a clear, achievable methodology, though minor issues may persist.<br>5. The problem is highly feasible with minimal barriers, well-supported by existing research, ample resources, and a robust, clear methodology, promising significant advancements. |
| | Significance | 1. The problem shows minimal to no significance, lacking relevance or potential impact in advancing the field or contributing to practical applications.<br>2. The problem has limited significance, with a narrow scope of impact and minor contributions to the field, offering little to no practical implications.<br>3. The problem demonstrates average significance, with some contributions to the field and potential practical implications, but lacks innovation or broader impact.<br>4. The problem is significant, offering notable contributions to the field and valuable practical implications, with evidence of potential for broader impact and advancement.<br>5. The problem presents exceptional significance, with groundbreaking contributions to the field, broad and transformative potential impacts, and substantial practical applications across diverse domains. |

Table 14: The criteria induced from human judgments for validating the developed methods, which used to align model-based evaluations with actual human preferences.

| Types | Criteria | Texts |
|---|---|---|
| Method | Clarity | 1. The method is explained in an extremely vague or ambiguous manner, making it impossible to understand or replicate the approach without additional information or clarification.<br>2. The method is described with some detail, but significant gaps in explanation or logic leave the reader with considerable confusion and uncertainty about how to apply or replicate the approach.<br>3. The method is described with sufficient detail to understand the basic approach, but lacks the precision or specificity needed to fully replicate or grasp the nuances of the methodology without further guidance.<br>4. The method is clearly and precisely described, with most details provided to allow for replication and comprehension, though minor areas may benefit from further clarification or elaboration.<br>5. The method is articulated in an exceptionally clear, precise, and detailed manner, enabling straightforward replication and thorough understanding of the approach with no ambiguities. |
| | Validity | 1. The method shows a fundamental misunderstanding of the research problem and lacks any credible alignment with established scientific principles or relevant studies.<br>2. The method partially addresses the research problem but exhibits significant flaws in its scientific underpinning, making its validity questionable despite some alignment with existing literature.<br>3. The method adequately addresses the research problem but with some limitations in its scientific validity, showing a mix of strengths and weaknesses in its alignment with related studies.<br>4. The method effectively addresses the research problem, demonstrating a strong scientific basis and sound alignment with existing literature, albeit with minor areas for improvement.<br>5. The method exemplifies an exceptional understanding of the research problem, grounded in a robust scientific foundation, and shows exemplary integration and advancement of existing studies' findings. |
| | Rigorousness | 1. The method demonstrates a fundamental lack of systematic approach, with significant inconsistencies and inaccuracies in addressing the research problem, showing a disregard for established research standards.<br>2. The method shows a minimal level of systematic effort but is marred by notable inaccuracies, lack of precision, and inconsistencies that undermine the rigorousness of the method in tackling the research problem.<br>3. The method exhibits an average level of systematic structure and adherence to research standards but lacks the thoroughness, precision, and consistency required for a rigorous scientific inquiry.<br>4. The method is well-structured and systematic, with a good level of precision and consistency, indicating a strong adherence to research standards, though it falls short of exemplifying the highest level of rigorousness.<br>5. The method exemplifies exceptional rigorousness, with outstanding thoroughness, precision, and consistency in its systematic approach, setting a benchmark for high standards in scientific research quality. |
| | Innovativeness | 1. The method introduces no novel elements, fully relying on existing techniques without any attempt to modify or adapt them for the specific research problem, showing a lack of innovativeness.<br>2. The method shows minimal innovation, with only slight modifications to existing techniques that do not substantially change or improve the approach to the research problem.<br>3. The method demonstrates moderate innovativeness, incorporating known techniques with some new elements or combinations that offer a somewhat fresh approach to the research problem but fall short of a significant breakthrough.<br>4. The method is highly innovative, introducing new techniques or novel combinations of existing methods that significantly differ from standard practices, offering a new perspective or solution to the research problem.<br>5. The method represents a groundbreaking innovation, fundamentally transforming the approach to the research problem with novel techniques or methodologies that redefine the field's standard practices. |
| | Generalizability | 1. The method shows no adaptability, failing to extend its applicability beyond its original context or dataset, showing a complete lack of generalizability.<br>2. The method demonstrates minimal adaptability, with limited evidence of potential applicability to contexts slightly different from the original.<br>3. The method exhibits some level of adaptability, suggesting it could be applicable to related contexts or datasets with modifications.<br>4. The method is adaptable and shows evidence of applicability to a variety of contexts or datasets beyond the original.<br>5. The method is highly adaptable, demonstrating clear evidence of broad applicability across diverse contexts, populations, and settings. |

Table 15: The criteria induced from human judgments for validating the experiment designs, which are used to align model-based evaluations with actual human preferences.

| Types | Criteria | Texts |
|---|---|---|
| Experiment | Clarity | 1. The experiment design is extremely unclear, with critical details missing or ambiguous, making it nearly impossible for others to understand the setup, procedure, or expected outcomes.<br>2. The experiment design lacks significant clarity, with many important aspects poorly explained or omitted, challenging others to grasp the essential elements of the setup, procedure, or expected outcomes.<br>3. The experiment design is moderately clear, but some aspects are not detailed enough, leaving room for interpretation or confusion about the setup, procedure, or expected outcomes.<br>4. The experiment design is mostly clear, with most aspects well-described, allowing others to understand the setup, procedure, and expected outcomes with minimal ambiguity.<br>5. The experiment design is exceptionally clear, precise, and detailed, enabling easy understanding of the setup, procedure, and expected outcomes, with no ambiguity or need for further clarification. |
| | Validity | 1. The experiment design demonstrates a fundamental misunderstanding of the research problem, lacks alignment with scientific methods, and shows no evidence of validity in addressing the research questions or testing the proposed methods.<br>2. The experiment design has significant flaws in its approach to the research problem and scientific method, with minimal or questionable evidence of validity, making it largely ineffective in addressing the research questions or testing the proposed methods.<br>3. The experiment design is generally aligned with the research problem and scientific method but has some limitations in its validity, offering moderate evidence that it can somewhat effectively address the research questions or test the proposed methods.<br>4. The experiment design is well-aligned with the research problem and scientific method, providing strong evidence of validity and effectively addressing the research questions and testing the proposed methods, despite minor limitations.<br>5. The experiment design excellently aligns with the research problem and scientific method, demonstrating robust evidence of validity and outstandingly addressing the research questions and testing the proposed methods without significant limitations. |
| | Robustness | 1. The experiment design demonstrates a fundamental lack of understanding of the scientific method, with no evidence of durability or adaptability across varying conditions, leading to highly unreliable and non-replicable results.<br>2. The experiment design shows minimal consideration for robustness, with significant oversights in addressing variability and ensuring consistency across different scenarios, resulting in largely unreliable outcomes.<br>3. The experiment design adequately addresses some aspects of robustness but lacks comprehensive measures to ensure durability and consistency across a wide range of conditions, leading to moderate reliability.<br>4. The experiment design incorporates a solid understanding of robustness, with clear efforts to ensure the experiment's durability and consistency across diverse conditions, though minor improvements are still possible for optimal reliability.<br>5. The experiment design exemplifies an exceptional commitment to robustness, with meticulous attention to durability and adaptability across all possible conditions, ensuring highly reliable and universally applicable results. |
| | Feasibility | 1. The experiment design is fundamentally unfeasible, with insurmountable resource, time, or technological constraints that make implementation virtually impossible within the proposed framework.<br>2. The experiment design faces significant feasibility challenges, including major resource, time, or technological limitations, that heavily compromise its practical execution and likelihood of success.<br>3. The experiment design is somewhat feasible, with moderate constraints on resources, time, or technology that could be addressed with adjustments, though these may not guarantee success.<br>4. The experiment design is largely feasible, with minor resource, time, or technological limitations that can be effectively managed or mitigated, ensuring a high probability of successful implementation.<br>5. The experiment design is highly feasible, with no significant constraints on resources, time, or technology, indicating that it can be implemented smoothly and successfully within the proposed framework. |
| | Reproducibility | 1. The experiment design lacks critical details, making it virtually impossible for other researchers to replicate the study under the same conditions or methodologies.<br>2. The experiment provides some essential information but omits significant details needed for replication, leading to considerable ambiguity in methodology or conditions.<br>3. The experiment design includes sufficient details for replication, but lacks clarity or completeness in certain areas, posing challenges for seamless reproducibility.<br>4. The experiment is well-documented with clear, detailed instructions and methodologies that allow for consistent replication, albeit with minor areas for improvement.<br>5. The experiment design is exemplary in its clarity, detail, and comprehensiveness, ensuring that other researchers can precisely and effortlessly replicate the study under identical conditions and methodologies. |

Table 16: The examples of research idea generation results from the proposed full ResearchAgent.

| Index | Types | Texts |
|---|---|---|
| 1 | Input | **Title**: Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering<br>**Abstract**: Large Language Models (LLMs) are capable of performing zero-shot closed-book question answering tasks, based on their internal knowledge stored in parameters during pre-training. However, such internalized knowledge might be insufficient and incorrect, which could lead LLMs to generate factually wrong answers. Furthermore, fine-tuning LLMs to update their knowledge is expensive. To this end, we propose to augment the knowledge directly in the input of LLMs. Specifically, we first retrieve the relevant facts to the input question from the knowledge graph based on semantic similarities between the question and its associated facts. After that, we prepend the retrieved facts to the input question in the form of the prompt, which is then forwarded to LLMs to generate the answer. Our framework, Knowledge-Augmented language model PromptING (KAPING), requires no model training, thus completely zero-shot. We validate the performance of our KAPING framework on the knowledge graph question answering task, that aims to answer the user's question based on facts over a knowledge graph, on which ours outperforms relevant zero-shot baselines by up to 48% in average, across multiple LLMs of various sizes.<br>**Entities**: ['Natural language', 'Learning management system', 'Question answering', 'Wikipedia', 'Artificial intelligence', 'Named-entity recognition', 'Quality assurance', 'Long short-term memory', 'English language', 'Google Books', 'Natural-language programming', 'LMS color space', 'GitHub', 'Logic learning machine', 'Machine translation', 'DBpedia', 'Integrated library system', 'Spanish language', 'Neural machine translation', 'ArXiv', 'Artificial general intelligence', 'Natural-language understanding', 'Nonlinear programming', 'Codex', 'LoRa', 'Adhesive', 'Common Crawl', 'ROUGE (metric)', 'Transformer', 'Tomas Mikolov'] |
| | Problem | Developing a Multimodal Knowledge-Aware Prompting System for Multilingual Zero-Shot Question Answering across Structured and Unstructured Data Sources |
| | Method | The further refined and enhanced method for developing a multimodal knowledge-aware prompting system for multilingual zero-shot question answering (MMKP-ZSQA) incorporates advanced refinements in the following components to address the requirements for innovativeness, clarity, and generalizability:<br><br>1. Interconnected Multimodal Knowledge Orchestration (IMKO):<br>- IMKO is an advanced evolution of SMKI, incorporating state-of-the-art language and vision models to handle data verbalization with context-aware disambiguation methods.<br>- Techniques like attention mechanisms are employed to weigh the relevance of different data pieces when integrated, revealing how textual and visual data augment each other in a synthesized representation.<br><br>2. Linguistically Inclusive Retrieval Engine (LIRE):<br>- LIRE extends EMKA with an emphasis on semantic understanding, using transformer-based models trained on diverse datasets, including idiomatic and cultural nuances across languages.<br>- Specific algorithms to handle linguistic phenomena such as code-switching and transliteration are included, enhancing the application to a broader set of languages and dialects.<br><br>3. Prompt Learning and Optimization Nexus (PLON):<br>- Building on AMPL, PLON focuses on creating a library of optimized prompts categorized by linguistic features and data modalities, using Bayesian optimization algorithms.<br>- It includes domain adaptation techniques and a wider array of meta-learning strategies with case studies for high and low-resource languages, and outlines theoretical frameworks for their implementation.<br><br>4. Cross-Modal Integrative Learning System (C-MILS):<br>- C-MILS advances SCAS by detailing the use of multi-head attention across modalities for effective and scalable reasoning, with exemplifications on how each modality can enhance comprehension synergistically.<br>- The component now incorporates a layer of abstraction to distill knowledge into a modality-agnostic format, aiding reasoning and facilitating interpretability across languages and data types.<br><br>5. User-Centric Adaptation and Privacy Framework (UCAPF):<br>- UCAPF enriches ALCUM with a user-centric design, focusing on interaction protocols that describe user engagement cycles, feedback loops, and privacy-preserving active learning paradigms, with practical workflows and pseudocode.<br>- Aligns with GDPR and other privacy frameworks to fortify trust and address the ethical use of data in dynamic learning environments.<br><br>6. Global Evaluation and Reflective Testing System (GERTS):<br>- GERTS expands CEB by presenting a multi-tiered validation and reflection methodology to adjust system components based on a diverse set of metrics, including fairness, bias, interpretability, and computational efficiency.<br>- Offers a structured approach to cross-cultural evaluation, including the use of demographic and regional diversity in forming test cohorts. |

*Continued on the next page*

**Table 16 – Continued from the previous page**

| Index | Types | Texts |
|---|---|---|
| | **Experiment** | The experiment, named "Refined Experiment for Multimodal Knowledge-Aware Prompting System for Multilingual Zero-shot Question Answering (RE-MKP-ZSQA)", aims to methodically develop and validate an advanced AI system. The experiment is streamlined to address feasibility, clarity, and reproducibility concerns while upholding robustness and validity by adhering to the following refined phases:<br><br>1. Detailed System Implementation Plan:<br>- Provide a publicly accessible project roadmap with specific milestones, resource allocation, and timelines.<br><br>2. Dataset Curation with Clear Guidelines:<br>- Publish precise annotation guidelines with strategies to prevent bias.<br>- Document the dataset assembly process, including source selection and data processing procedures.<br><br>3. Transparent System Training:<br>- Offer a detailed training protocol with hyperparameters, optimization strategies, and Bayesian optimization processes used in PLON.<br><br>4. Structured Zero-Shot Evaluation:<br>- Outline evaluation metrics derived from GERTS with benchmark datasets to test zero-shot capabilities.<br><br>5. Clearer Interdisciplinary Evaluation Protocol:<br>- Specify the composition of the evaluation committee, criteria for assessments, and methods for integrating the feedback.<br><br>6. Iterative Improvement with Validation Metrics:<br>- Describe statistical methods for reflective assessment and continuous improvement, aligned with multi-tiered GERTS methodology.<br><br>7. User-Centric Design and Privacy Compliance Evaluation:<br>- Structure user studies with targeted data points to assess usability and cultural adaptability.<br>- Outline privacy compliance protocols to adhere to international standards.<br><br>8. Detailed Global Scalability Evaluation Method:<br>- Define evaluation metrics for scalability tests and describe diverse infrastructural setups.<br><br>9. Enhanced Reporting for Reproducibility:<br>- Commit to creating a comprehensive report with precise specifications, configurations, and instructions for replication purposes.<br>- Utilize GitHub for version-controlled deposition of code and datasets, and arXiv for openly accessible experiment protocols and findings. |
| 2 | **Input** | **Title**: Test-Time Self-Adaptive Small Language Models for Question Answering<br><br>**Abstract**: Recent instruction-finetuned large language models (LMs) have achieved notable performances in various tasks, such as question-answering (QA). However, despite their ability to memorize a vast amount of general knowledge across diverse tasks, they might be suboptimal on specific tasks due to their limited capacity to transfer and adapt knowledge to target tasks. Moreover, further finetuning LMs with labeled datasets is often infeasible due to their absence, but it is also questionable if we can transfer smaller LMs having limited knowledge only with unlabeled test data. In this work, we show and investigate the capabilities of smaller self-adaptive LMs, only with unlabeled test data. In particular, we first stochastically generate multiple answers, and then ensemble them while filtering out low-quality samples to mitigate noise from inaccurate labels. Our proposed self-adaption strategy demonstrates significant performance improvements on benchmark QA datasets with higher robustness across diverse prompts, enabling LMs to stay stable.<br><br>**Entities**: ['Codex', 'Natural language', 'English language', 'United States', 'Question answering', 'Natural-language programming', 'GTRI Information and Communications Laboratory', 'Artificial intelligence', 'LoRa', 'Llama', 'Python (programming language)', 'Learning management system', 'Natural language processing', 'Reinforcement learning', 'LMS color space', 'Wikipedia', 'GitHub', 'Natural-language understanding', 'London, Midland and Scottish Railway', 'Integrated library system', 'Language model', 'Chinese language', 'Lumen (unit)', 'Spanish language', 'English Wikipedia', 'Logic learning machine', 'Gradient descent', 'Alternative public offering', 'Technology transfer', 'Dialogue system'] |
| | **Problem** | Developing a Scalable, Domain-Adaptive Test-Time Training Protocol for Low-Resource Language QA Using Small Language Models |

*Continued on the next page*

**Table 16 – Continued from the previous page**

| Index | Types | Texts |
|---|---|---|
| | **Method** | 1. Selection of Scalable Compact Language Models (CLMs): Identify and evaluate existing CLMs suitable for adaptation, emphasizing models with minimal computational requirements.<br><br>2. Creation of a Multilingual Test-Time Training (TTT) Framework: Develop a TTT protocol that enables CLMs to adapt to new domains and languages during the inference phase, leveraging unsupervised learning techniques and pseudo-label generation.<br><br>3. Synthetic and Unsupervised Data Generation: Utilize a combination of unsupervised and synthetic data generation methods to produce multilingual QA pairs, employing techniques such as back-translation and context-based question synthesis.<br><br>4. Domain-Adaptive Mechanisms: Introduce domain-adaptive components, including feature adaptation layers and meta-learning algorithms, which tailor the model's behavior to new contexts and languages at test time.<br><br>5. Incremental Language Addition and Dominance Assessment: Start with a subset of linguistically diverse, low-resource languages. Evaluate domain adaptability for each language via an iterative process, ensuring models learn to prioritize resource efficiency.<br><br>6. Model Robustness and Generalization: Perform robustness tuning (RT) to prepare models for unforeseen linguistic variations and conduct thorough evaluations across multiple domains to ensure models can generalize their learning effectively.<br><br>7. Human-In-The-Loop Evaluation: Conduct evaluations with native speakers and domain experts to validate the relevance and accuracy of the QA outputs, incorporating feedback into the iterative training process.<br><br>8. Open-Sourcing and Community Collaboration: Make the TTT protocol, trained models, and evaluation benchmarks publicly available for the research community, fostering collaboration and further innovation. |
| | **Experiment** | 1. Selection and Preparation:<br>- Identify potential compact language models (CLMs) suitable for domain adaptation and test-time training, focusing on those with minimal computational requirement and the ability to be fine-tuned or adapted in an unsupervised manner.<br>- Prepare a diverse set of low-resource languages and corresponding text corpora, ensuring linguistic diversity and sociocultural significance. Select benchmark datasets for these languages if available.<br><br>2. Training and Adaptation Procedure:<br>- Create a Test-Time Training (TTT) framework that allows selected CLMs to adapt to various domains in the selected low-resource languages during the inference phase.<br>- Implement unsupervised learning techniques and pseudo-label generation to produce QA pairs, utilizing back-translation and context-based question synthesis to generate synthetic datasets for languages with limited or no available QA datasets.<br>- Integrate domain-adaptive components and meta-learning algorithms into the CLMs to enable domain-specific adaptations at test time.<br><br>3. Iterative Evaluation and Refinement:<br>- Begin adaptation and training with a single low-resource language and gradually add additional languages, monitoring the domain adaptability and model performance metrics after each addition.<br>- Perform robustness tuning and cross-domain evaluations for each CLM and language adaptation to ensure generalizability and prevent overfitting.<br><br>4. Human-In-The-Loop Assessment:<br>- Enlist native speakers and domain experts to evaluate the relevance and accuracy of the model's QA outputs for each language.<br>- Incorporate feedback into the iterative training process, refining and re-adapting the models accordingly.<br><br>5. Open-Sourcing and Community Feedback:<br>- Make the TTT protocol, adaptive CLMs, evaluation benchmarks, and any synthetic datasets publicly available for the research community.<br><br>6. Experiment Monitoring and Documentation:<br>- Record all the parameters, datasets, model configurations, and evaluation metrics meticulously to ensure robustness and reproducibility.<br>- Document any challenges faced, unexpected results, or adaptions made during the experiment for open-sourcing purposes.<br><br>7. Data Analysis and Reporting:<br>- Analyze the collected performance data quantitatively, using appropriate statistical methods to compare with non-adaptive baselines.<br>- Report qualitative findings from human-in-the-loop evaluations, interpreting the implications for language model performance in low-resource language domains. |

*Continued on the next page*

**Table 16 – Continued from the previous page**

| Index | Types | Texts |
|---|---|---|
| **3** | **Input** | **Title**: Whole-brain annotation and multi-connectome cell typing quantifies circuit stereotypy in Drosophila<br><br>**Abstract**: The fruit fly Drosophila melanogaster combines surprisingly sophisticated behaviour with a highly tractable nervous system. A large part of the fly's success as a model organism in modern neuroscience stems from the concentration of collaboratively generated molecular genetic and digital resources. As presented in our FlyWire companion paper1, this now includes the first full brain connectome of an adult animal. Here we report the systematic and hierarchical annotation of this 130,000-neuron connectome including neuronal classes, cell types and developmental units (hemilineages). This enables any researcher to navigate this huge dataset and find systems and neurons of interest, linked to the literature through the Virtual Fly Brain database2. Crucially, this resource includes 4,552 cell types. 3,094 are rigorous consensus validations of cell types previously proposed in the "hemibrain" connectome3. In addition, we propose 1,458 new cell types, arising mostly from the fact that the FlyWire connectome spans the whole brain, whereas the hemibrain derives from a subvolume. Comparison of FlyWire and the hemibrain showed that cell type counts and strong connections were largely stable, but connection weights were surprisingly variable within and across animals. Further analysis defined simple heuristics for connectome interpretation: connections stronger than 10 unitary synapses or providing >1% of the input to a target cell are highly conserved. Some cell types showed increased variability across connectomes: the most common cell type in the mushroom body, required for learning and memory, is almost twice as numerous in FlyWire as the hemibrain. We find evidence for functional homeostasis through adjustments of the absolute amount of excitatory input while maintaining the excitation-inhibition ratio. Finally, and surprisingly, about one third of the cell types proposed in the hemibrain connectome could not yet be reliably identified in the FlyWire connectome. We therefore suggest that cell types should be defined to be robust to inter-individual variation, namely as groups of cells that are quantitatively more similar to cells in a different brain than to any other cell in the same brain. Joint analysis of the FlyWire and hemibrain connectomes demonstrates the viability and utility of this new definition. Our work defines a consensus cell type atlas for the fly brain and provides both an intellectual framework and open source toolchain for brain-scale comparative connectomics.<br><br>**Entities**: ['Virtual Fly Brain', 'Central nervous system', 'Transposable element', 'SUMO protein', 'Kenyon cell', 'Romani people', 'Induced stem cells', 'Ventral nerve cord', 'FlyBase', "Parkinson's disease", 'Virtual Network Computing', 'P element', 'Piwi-interacting RNA', 'Drosophila Genetic Reference Panel', 'Bateson–Dobzhansky–Muller model', 'J. B. S. Haldane', 'ATG7', "Haldane's rule", 'Oxford Nanopore Technologies', 'Drosophila mauritiana', 'Germline', 'PINK1', 'Migratory locust', 'CRISPR', 'Heliconius', 'GINS (protein complex)', 'Parkin (ligase)', 'Lepidoptera', 'Illumina, Inc.', 'Drosophila'] |
| | **Problem** | Investigating the Functional Implications of Connectome Variability in Drosophila's Learning and Memory Circuits Across Different Environmental and Genetic Contexts |
| | **Method** | The proposed method involves a multi-tiered approach that integrates connectomics, behavioral assays, genetic manipulation, and computational modeling to investigate the functional implications of connectome variability in Drosophila's learning and memory circuits. The method consists of the following steps:<br><br>1. Connectome Mapping and Variability Analysis:<br>a. Utilize the Virtual Fly Brain database to identify and compare individual connectomes, focusing on the mushroom body.<br>b. Quantify the variability in connection weights and cell type counts using statistical methods and machine learning algorithms to identify patterns of variability.<br><br>2. Behavioral Assays:<br>a. Design a series of learning and memory tasks for Drosophila, such as olfactory conditioning or visual pattern recognition.<br>b. Test groups of flies with known connectome profiles under controlled environmental conditions to establish baseline behavioral data.<br><br>3. Environmental and Genetic Perturbations:<br>a. Expose different groups of flies to varied learning paradigms and sensory inputs to create environmental perturbations.<br>b. Use CRISPR-Cas9 technology to introduce targeted mutations in genes like PINK1 or Parkin, creating genetic perturbations.<br>c. Assess the impact of these perturbations on connectome structure using high-resolution imaging and reconstruction techniques.<br><br>4. Transcriptomic and Spatial Analysis:<br>a. Apply single-cell RNA sequencing and spatial transcriptomics to profile gene expression changes in response to environmental and genetic perturbations.<br>b. Correlate transcriptomic data with connectome changes to identify molecular pathways associated with structural and functional plasticity.<br><br>5. Computational Modeling and Network Analysis:<br>a. Develop computational models to simulate the effects of connectome variability on neural circuit function.<br>b. Use network analysis tools to explore information flow and circuit dynamics, integrating data from related papers on olfactory projection neurons and information flow in the olfactory system.<br><br>6. Synthesis and Validation:<br>a. Integrate findings from behavioral assays, connectome mapping, transcriptomic analysis, and computational modeling.<br>b. Validate the proposed models and hypotheses through iterative experimentation and refinement. |

*Continued on the next page*

**Table 16 – Continued from the previous page**

| Index | Types | Texts |
|---|---|---|
| | **Experiment** | 1. Pre-experimental Setup:<br>a. Establish a Drosophila melanogaster breeding program to ensure a consistent supply of genetically similar flies for the experiments.<br>b. Develop a standardized protocol for rearing flies under controlled conditions to minimize pre-experimental variability.<br>c. Select and validate CRISPR-Cas9 constructs for targeted gene editing of PINK1, Parkin, and other genes of interest.<br>d. Train personnel in the use of the Virtual Fly Brain database and relevant computational tools for connectome analysis.<br><br>2. Connectome Mapping and Variability Analysis:<br>a. Randomly assign individual flies to either a control group or various treatment groups (environmental and genetic perturbations).<br>b. Utilize high-resolution imaging techniques to map the connectomes of flies from each group, with a focus on the mushroom body.<br>c. Apply statistical and machine learning algorithms to quantify and compare the variability in connection weights and cell type counts across groups.<br><br>3. Behavioral Assays:<br>a. Design and validate a series of learning and memory tasks, such as olfactory conditioning and visual pattern recognition, ensuring tasks are sensitive to subtle differences in performance.<br>b. Test flies from each group in the behavioral tasks and record performance metrics.<br>c. Analyze behavioral data to establish correlations with connectome profiles.<br><br>4. Environmental and Genetic Perturbations:<br>a. Expose flies to different learning paradigms and sensory inputs to induce environmental perturbations.<br>b. Perform gene editing using CRISPR-Cas9 to create genetic perturbations in the treatment groups.<br>c. Re-map connectomes post-perturbation to assess structural changes.<br><br>5. Transcriptomic and Spatial Analysis:<br>a. Collect brain tissue from flies post-behavioral assays and perform single-cell RNA sequencing and spatial transcriptomics.<br>b. Analyze transcriptomic data to identify gene expression changes and correlate these with observed connectome and behavioral variations.<br><br>6. Computational Modeling and Network Analysis:<br>a. Develop computational models to simulate the impact of observed connectome variability on neural circuit function.<br>b. Use network analysis to integrate behavioral, connectomic, and transcriptomic data, focusing on information flow and circuit dynamics.<br><br>7. Synthesis and Validation:<br>a. Integrate findings across all experimental components to formulate a cohesive understanding of the functional implications of connectome variability.<br>b. Validate models and refine hypotheses through additional targeted experiments, informed by initial findings. |