Can Large Language Models *Unlock* Novel Scientific Research Ideas?

Sandeep Kumar†, Tirthankar Ghosal‡, Vinayak Goyal†, Asif Ekbal†
†Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
‡National Center for Computational Sciences, Oak Ridge National Laboratory, USA
†(sandeep_2121cs29,2201ai52_vinayak,asif)@iitp.ac.in
‡ghosalt@ornl.gov

Abstract

The widespread adoption of Large Language Models (LLMs) and publicly available ChatGPT have marked a significant turning point in the integration of Artificial Intelligence (AI) into people's everyday lives. This study examines the ability of Large Language Models (LLMs) to generate future research ideas from scientific papers. Unlike tasks such as summarization or translation, idea generation lacks a clearly defined reference set or structure, making manual evaluation the default standard. However, human evaluation in this setting is extremely challenging — it requires substantial domain expertise, contextual understanding of the paper, and awareness of the current research landscape. This makes it time-consuming, costly, and fundamentally non-scalable, particularly as new LLMs are being released at a rapid pace. Currently, there is no automated evaluation metric specifically designed for this task. To address this gap, we propose two automated evaluation metrics: Idea Alignment Score (IAScore) and Idea Distinctness Index. We further conducted human evaluation to assess the novelty, relevance, and feasibility of the generated future research ideas. This investigation offers insights into the evolving role of LLMs in idea generation, highlighting both its capability and limitations. Our work contributes to the ongoing efforts in evaluating and utilizing language models for generating future research ideas. We make our datasets and codes publicly available[1].

"Innovation is seeing what everybody has seen and thinking what nobody has thought" —Dr. Albert Szent-Györgyi

1 Introduction

An *idea* can be defined as a thought or suggestion aimed at solving a problem or considering a possibility. This concept is central to fields ranging from philosophy to science and economics. According to (Plato et al., 2000), ideas are archetypal forms that represent the most accurate reality. In the context of scientific research, (Kuhn and Hawkins, 1963) in "The Structure of Scientific Revolutions" describes an idea as a realization or hypothesis that can challenge and shift paradigms within a scientific community. Therefore, an idea can be understood as a cognitive construct that arises from the human mind's ability to process information, reflect, and imagine, serving as a cornerstone for creativity, problem-solving, and innovation. Idea generation can generally be understood as a state of focused internally-directed attention involving controlled semantic retrieval (Benedek et al., 2014).

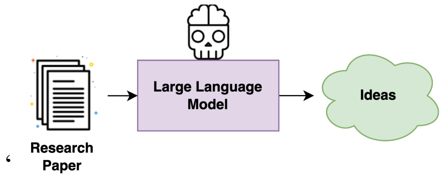


Figure 1: Large language model suggesting future research ideas after reading a research paper

As technology improves, new capabilities emerge. Ever since the Turing Test was proposed in the 1950s, humans have explored the mastering of language intelligence by machine (Zhao et al., 2023). Technological advancements serve two key functions in innovation. Firstly, they influence the goals of generating and selecting ideas. Secondly, they impact the methodology of how ideas are generated and chosen (Kornish and Hutchison-Krupat, 2017). Large Language Models (LLMs) have exhibited unparalleled mastery of natural language processing (NLP). Since, these have become increasingly powerful, researchers have begun to investigate their reasoning ability in problem-solving tasks (Yao et al., 2022; Brahman et al., 2023). The concept of an idea is essentially a new combination of old elements. LLMs have access to a broad spec-

[1]https://github.com/sandeep82945/Future-Idea-Generation.git

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33552–33576
November 4-9, 2025

trum of knowledge, due to their extensive training on vast amount of text data. However, understanding how information extracted from a research paper can give rise to new ideas, which have not yet been explored much. This leads us to ponder:

> Can Large Language Models read a scientific paper and suggest new research ideas or directions?

Motivated by this, in this paper, we analyze the potential of LLMs in generating future research directions or ideas. As LLMs possess knowledge across various domains, we investigate five specific areas, *viz.* Computer Science, Physics, Chemistry, Economics, and Medicine. To address this task, we create a dataset of papers published post 2022 from these five domains. To the best of our knowledge, there is no existing automated evaluation metric specifically designed for assessing the quality of future research ideas (FRIs) generated by LLMs. Unlike tasks such as summarization or translation, idea generation lacks a clearly defined reference set or structure, making manual evaluation the default standard. However, human evaluation in this context is particularly challenging—it demands substantial domain expertise, a deep understanding of the source paper, and awareness of the current research landscape. As a result, it becomes time-consuming, costly, and fundamentally non-scalable, especially given the rapid pace at which new LLMs are being released. To evaluate the novelty and relevance of ideas generated by LLMs, we propose an Idea Alignment Score (IAScore). This score reflects how well the generated ideas align with those proposed by the authors. Our proposed IAScore provides a scalable, automated, and interpretable metric that captures how well a model can generate author-aligned research directions. While IAScore does not capture all novel ideas, it serves as a lower-bound indicator of a model's competence. We annotate research papers with future research ideas to create a benchmark for evaluation. To assess the models' ability to generate diverse ideas, we introduce the Idea Distinctness Index. We further conduct a human evaluation of 660 generated ideas in the field of computer science to assess their novelty, relevance, and feasibility. Additionally, we analyze and discuss the performance and limitations of four LLMs: Gemini (Anil et al., 2023), Claude-2 (Anthropic, 2023), GPT-3.5, and GPT-4 (OpenAI, 2023). Our findings demonstrate that LLMs have the potential to generate research ideas that are relevant, distinct, feasible, and novel—to some extent.

To summarize, our main contributions in this paper are as follows:

- We contribute to the ongoing exploration of LLMs' capabilities in generating future research ideas.
- To address the task, we create a novel dataset of recent papers of five domains (Computer science, Economics, Chemistry, Physics, Medical).
- To assess the quality of generated ideas from LLMs, we propose Idea Alignment Score and Idea Distinctness Index to evaluate the idea generation capability.
- We discuss the challenges associated with human evaluation and conduct an in-depth analysis on the generated ideas.

We hope that this work serves as a foundation for future studies focused on accelerating scientific research by automatically generating research ideas.

2 Related Work

Recently, LLMs have shown tremendous abilities to perform tasks they were not explicitly trained for (Wei et al., 2022; Bubeck et al., 2023). This includes common sense question answering, code generation, and cross-domain problem solving, enriching their utility across unforeseen domains (Chen et al., 2021; Sarsa et al., 2022). Their capability extends to advanced scientific domains such as computer science, physics, medicine, and mathematics (Romera-Paredes et al., 2023; Huang et al., 2023). Technology Semantic Network (TechNet) was proposed to stimulate idea generation in engineering design (Sarica et al., 2021). There have been a few works in the discovery of new proteins to accelerate scientific discovery. The prior work reported in (Spangler et al., 2014) involves utilizing published studies to find new protein kinases that phosphorylate the tumor suppressor protein p53.

A hypothesis is a hunch, assumption, suspicion, assertion or an idea about a phenomenon, relationship or situation, the reality or truth of which you do not know (Kumar, 1996). There have been some works on hypothesis generation. Initial studies on automated hypothesis generation begin by constructing a corpus of distinct concepts. Subsequently, they explore the relationships between these concepts using machine learning techniques, such as analyzing the similarities among vectors representing different words (or concepts) (Tshitoyan et al., 2019), or applying link prediction methods over a graph (where concepts are nodes) (Nadkarni et al., 2021). Recently, (Qi et al., 2023) used LLMs and extensive pre-existing knowledge of various scientific fields for hypothesis generation. PaperRobot (Wang et al., 2019) predicts related entities for an input title and writes key elements of a new paper, including the abstract, conclusion, and future work, and predicts a new title.

Xu et al. (2023) developed a framework that leverages the concept co-occurrence graphs and a masked language model to explore and verbalize academic ideas. Their method involves constructing evolving concept graphs across various disciplines and utilizing temporal link prediction to identify potential interdisciplinary connections. The framework also incorporates pre-trained language models to articulate these connections in a coherent academic context. SciMON (Wang et al., 2023) showed that LLMs can be guided by seed terms to generate specific ideas.

However, previous works primarily focused on developing methods (linking and explaining entities, which may not sufficiently capture the complexity or explain how LLMs can solve real-world problems) for idea generation, whereas our work exhaustively focuses on evaluating the capability of LLMs in generating research ideas. Our goal is to assess the inherent ability of LLMs to generate future research ideas/directions.

3 Dataset

Our dataset creation involves three steps: (1) Dataset Collection, (2) Future Research Ideas (FRI) Identification and removal, and (3) FRI generation.

3.1 Dataset Collection

We construct a corpus D by collecting 1,250 papers from S2ORC[2] from the domains of Computer Science, Economics, Physics, Chemistry, Medical from (Lo et al., 2020)[3]. We prompted the LLM with paper text after removing the future work for generating future ideas.

To ensure the quality and relevance of the data and to utilize the future research ideas mentioned in a paper, the selected papers must meet the following requirements: (1) the paper must contain the full content, and (2) the paper must include a section on future work.

3.2 FRI Identification and Removal

We first identify and remove any potential research ideas mentioned throughout the paper. By doing this, we ensure that LLMs have no prior access to these ideas, which could otherwise affect the objectivity of the analysis.

3.2.1 Annotation Process

Inspired by Hao et al. (2020), we define a future research idea as a discussion that the authors believe they will conduct in the future or believe needs to be investigated in future research. We discuss more details about the annotation guidelines, annotation training, annotation process, and annotator's pay in Appendix A.

3.2.2 Future Work Removal

We observed two types of future research ideas (FRIs)[4] (Direct FRI and Mixed FRI). We discuss them in details in Appendix H.

AP-FRI Corpus: We removed the sentence from the paper's input text if it pertains to Direct FRI. However, in the case of Mixed FRI, we did not entirely remove the sentences; instead, we eliminated only parts of sentences or markers indicating future research ideas. We add the removed future ideas to a corpus, which we refer to as the AP-FRI (Author Perspective Future Research Idea Corpus). This corpus contains the future research ideas proposed by the authors of the paper. Also, before adding to the AP-FRI corpus, we merged the sentences about the same topic into a single group. We illustrate this by an example in Figure 8, mentioned in the Appendix of the paper.

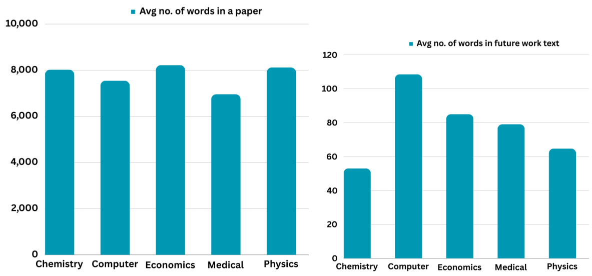
3.3 Data Statistics

Figures 2a and 2b present a domain-wise analysis of average word counts in academic papers and their future work (FWK) sections, revealing significant variations across disciplines. Notably,

[2] 250 papers from each domain.

[3] They used Science Parse and Grobid tool to extract data from PDF. We used the plain text without figure and tables and references for our experiment.

[4] In this paper, we use the terms 'ideas,' 'research ideas,' 'future research ideas,' and 'FRI' interchangeably to frequently refer to future research ideas.



(a) Domain vs Avg. number of words in a paper w/o FWK

(b) Domain vs Avg. number of words in FWK

Figure 2: Comparison of average word counts in papers with and without FWK across domains

Computer Science emphasizes future research extensively, while Chemistry adopts a more concise approach. Refer to the appendix Section A.5 for more details.

4 Experiments

4.1 FRI Generation using LLM

We investigate various prompts and utilize the following prompts to generate FRIs from the paper text after removing the FRIs mentioned in the paper as discussed in Section 3.

> **System**: You are a research scientist.
> **User**: Imagine you are a research scientist. 1) Read the full paper and understand it. 2) Find out the related works in this direction 3) Brainstorm and follow a step-by-step reasoning approach to generate potential future research ideas:
>
> [paper text]
>
> Make sure the future research ideas are very distinct from the related papers. Potential future research ideas from the paper in bullet points are:

Here, '[paper text]' contains the full content of the paper after removal of future work sections. In case the text was larger than the context size of the particular LLM we divided the paper into parts and finally combined the ideas.

4.2 Challenges

To accurately assess the novelty, relevance, and applicability of ideas generated by LLMs, evaluators must possess a high level of expertise in the specific domain and a deep understanding of the research topic to fully grasp the context. Additionally, they need knowledge of related literature to evaluate the ideas' future potential and the broader implications of their implementation.

4.3 Idea Alignment Score (IAScore)

With the above challenges, the evaluation of ideas generated by LLMs is a challenging process that demands a high number of domain-specific experts. We, therefore, proposed an *Idea Alignment Score* (IAScore), which reflects how well the generated ideas align with those proposed by the author. The underlying idea for this score is that authors of accepted papers can be regarded as experts in their respective subjects. The reason being that they possess thorough background knowledge and have conducted deep analyses of the research topic before getting the paper accepted. Consequently, they are well-acquainted with the pertinent challenges which also may have been discussed by expert reviewers. Therefore, we propose that future ideas mentioned by the authors in the paper could be utilized as good quality of potential FRIs.

The IAScore quantifies the alignment of newly generated ideas with author's perspectives within a specific domain, and is computed via a two-step process, detailed in Equations 1 and 2.

Initially, we compute the average alignment score AvgScore_j for each paper's ideas. The `IdeaMatcher` model measures the alignment between the paper's author *Future Research Ideas* (AP-FRI_j) and its each generated idea I_{ij}. The subscript i indexes the i-th idea within the j-th paper, where N_j represents the total number of ideas proposed in that paper.

$$\text{AvgScore}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \texttt{IM}\left(\text{AP-FRI}_j, \text{I}_{ij}\right) \tag{1}$$

LLMs may generate new ideas that even the author may not have thought of. They can also generate additional future ideas, which may or may not be useful. Our goal with this score is to ensure that the LLMs generate potential future research ideas originally proposed by the authors[5]. Therefore, in our formula of AvgScore_j, the sum of the alignment scores for a paper's ideas is divided by the total number of the author's proposed ideas, N_j, to normalize the score.

$$\text{IAScore}_{\text{domain, M}} = \frac{1}{P} \sum_{j=1}^{P} \text{AvgScore}_j \tag{2}$$

Subsequently, we aggregate the individual paper scores to calculate the domain-wise IAScore. This

[5]We intentionally avoided constraining the LLMs by not specifying a fixed number of ideas, allowing them to freely generate additional ideas beyond those proposed by the authors.

aggregation, presented in Equation 2, averages the AvgScore_j values across all P papers within the domain. Higher the value of $\text{IAScore}_{\text{domain}}$ signifies the more alignment of the generated ideas with author's perspective of all papers generated by model M.

Let's define:

- $I_{\text{valid}} \supseteq I_{\text{auth}}$: the set of all valid FRIs for a paper, including all author-defined FRIs.
- The probability that a generated idea $I \in I_{\text{gen}}$ is useful can be written as:

$$P(I \in I_{\text{valid}}) = P(I \in I_{\text{auth}}) + P(I \in I_{\text{valid}} \setminus I_{\text{auth}}).$$

Now assume we cannot directly measure I_{valid}. The only known subset is I_{auth}, written by experts. So:

- IAScore serves as a *lower bound estimate* for $P(I \in I_{\text{valid}})$
- A model that scores poorly on the IAScore is unlikely to generate $I \in I_{\text{auth}} \subset I_{\text{valid}}$

Hence, if a model can not even match the known subset I_{auth}, we should not expect it to robustly hit the full space I_{valid}. IAScore is not intended to penalize novelty but to ensure a baseline level of credibility and expert alignment. It acts as a precision-driven benchmark: if a model can not even regenerate well-grounded ideas by domain experts, we cannot trust it to propose higher-quality novel ones. Hence, IAScore is a necessary but not sufficient component of idea evaluation and should complement—not replace—novelty-based costly human evaluation.

4.3.1 IdeaMatcher

To select an effective IdeaMatcher, we create a small annotated corpus. Our dataset was divided using the 30:70 ratio for validation and test sets, respectively. Since our study involves comparing two ideas using a pre-trained model, we did not require a separate training set. We first manually searched for matching pairs of ideas from generated ideas and AP-FRI of the paper. After obtaining 61 matching pairs, we searched for non-matching pairs of ideas, which is straightforward as only one generated idea will match or would not match with another one from AP-FRI while others would not match, so we picked an equal number of non-matching pairs. Then, we experimented with the idea-matching task by considering it similar to the Natural Language Inference (NLI) task. In particular, we considered the generated FRIs to be hypotheses and their corresponding AP-FRIs of the paper to be premises. If the idea matches, the hypothesis should be entailed by the premise. In particular, we used a pre-trained RoBERTa MNLI model (Liu et al., 2019) for this task. We found that this technique produces many false negative cases, resulting in an accuracy of 65.5%.

We also evaluated the idea-matching capability of BERTScore (Zhang* et al., 2020), as it utilizes BERT embeddings for comparison. We discuss the details in Appendix F. We found that BERTScore performed better than the entailment technique, resulting in an accuracy of 75.4%. We also tried GPT by prompting it with various questions and found that it resulted in 91.8% accuracy when prompted with a specific question prompt below:-

> **Prompt:** Your task is to examine whether a particular idea is incorporated within a set of ideas and to what degree.
> **Collection of ideas:** {API-FRIs}
> **Single idea:** {A generated Idea}
> Is the single idea contained within the collection of ideas?
> If yes, quantify its degree of presence or relevance of the single idea in the collection of ideas on a scale from 0 to 1.

We found that GPT performs better than the existing NLI and similarity measure, such as BERTScore. Therefore, we chose GPT for this task[6]

4.4 IdeaDistinctness Index

Distinct-N (Li et al., 2015), is a metric that measures the diversity of a sentence. It focuses on the number of distinct n-grams of a sentence, and thus penalizes sentences with a lot of repeated words. However, comparing two ideas need semantic comparisons rather than just syntactic differences. So, we introduce a method to semantically evaluate the distinctness of the generated ideas. This method in particular leverages semantic embedding to capture the essence of each idea and computes their distinctness based on semantic similarity measures.

[6] We used the OpenAI model GPT-3.5-turbo-0125 using OpenAI API. However, given the rapid advancements in language models, we recognize that newly released and upcoming models could further enhance the potential of IdeaMatcher.

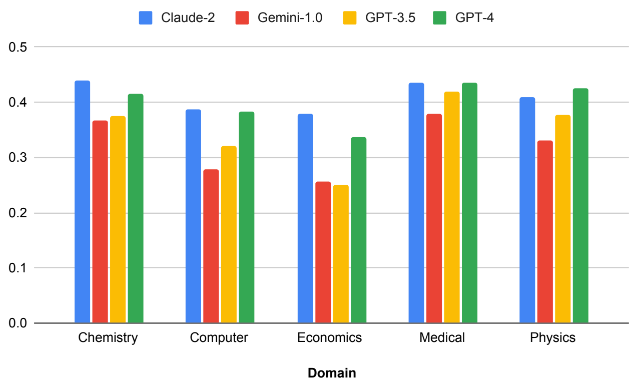


Figure 3: IAScore for each domain and model; a higher value indicates better alignment with the author.

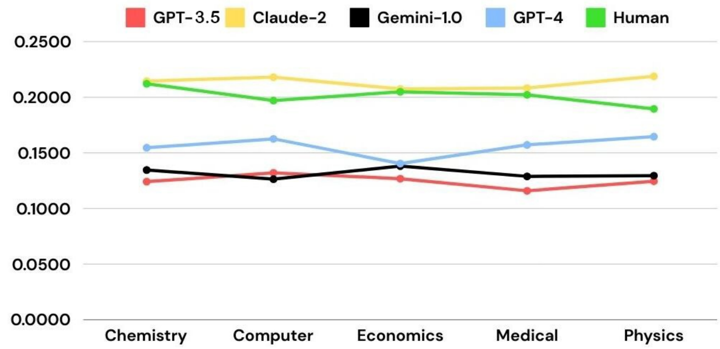


Figure 4: IdeaDistinctness index analysis; Here human is the authors of the paper

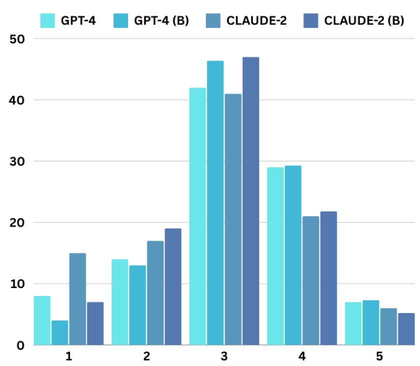


Figure 5: Novelty human evaluation for Computer Science domain. Here, (B) means with additional background knowledge; The x-axis represents the scale of novelty annotated by humans.

Given a set of generated ideas $I = \{\mathrm{id}_1, \mathrm{id}_2, \ldots, \mathrm{id}_n\}$, representing individual ideas, we first encode each idea into a high-dimensional vector space using a pre-trained GPT embedding (OpenAI, 2022)[7] **GPT** : $id_i \mapsto \mathbf{v}_i$, where $\mathbf{v}_i \in \mathbb{R}^d$ is the embedding of idea id_i and d is the dimensionality of the embedding space.

To quantify the distinctness between pairs of ideas, we compute the cosine similarity between their embeddings, $sim(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|}$, for each pair of ideas (id_i, id_j) in I. The distinctness D_{ij} between two ideas i and j is then inversely related to their similarity: $D_{ij} = 1 - sim(\mathbf{v}_i, \mathbf{v}_j)$.

The overall distinctness of the set I is calculated as the mean of all pairwise distinctness scores:

$$D_I = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} D_{ij} \tag{3}$$

This measure provides a single scalar value D_I that quantifies the average diverseness of ideas within a corpus of ideas, with higher values indicating a greater degree of diverseness among the ideas.

Subsequently, we aggregated the distinctness scores across all ideas in each paper to compute

[7] text-embedding-ada-002

the mean distinctness for that paper. Let $P = \{p_1, p_2, \ldots, p_m\}$ represent the set of papers in a domain, where m is the number of papers in the domain. Finally, for a comprehensive assessment of model performance within a domain, we averaged the mean distinctness scores of all papers generated by model M as follows:

$$D_{\text{domain},M} = \frac{1}{m} \sum_{p=1}^{m} D_{I_{pM}} \tag{4}$$

The resultant metric, $D_{\text{domain},M}$, represents the average idea distinctness for model M in a given domain, indicating the model's ability to generate diverse ideas.

We compute IA Score and Distinctness Index separately per domain to capture how LLM performance varies across disciplines. This design avoids conflating unrelated domains (e.g., chemistry vs. economics) and enables more granular, interpretable evaluation of LLM capabilities.

4.5 Adding additional background knowledge

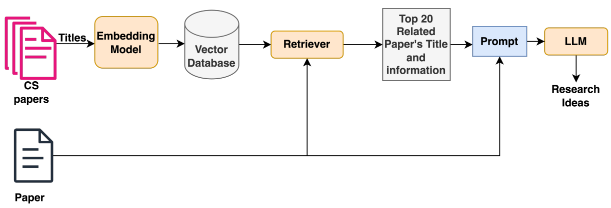


Figure 6: RAG Pipeline framework for infusing infusing more background knowledge with the LLMs

We designed our framework based on the Retrieval-Augmented Generation (RAG) model (Lewis et al., 2020) to integrate background knowledge into LLMs, as illustrated in Figure 6. A detailed explanation is available in the Appendix E, but we summarize the key components here. 1) **Vector Database**: We collected 190K CS paper titles and abstracts via Semantic Scholar API (Kinney et al., 2023) and stored BERT-based embeddings in a vector database. 2) **Retriever**: Computed cosine similarity with target title to retrieve top 20 related papers; used LLM prompts to extract their key contributions from abstracts. 3) **Generator**: Using the retrieved background knowledge, we prompted an LLM to generate distinct and novel future research ideas.

4.6 Human Evaluation

The evaluation of generated future ideas necessitates familiarity with both previous works related to the subject and the work being evaluated. Specifically, the evaluator must be an expert in the domain and topic. Given the complexity of human evaluation, we approached authors (as the authors have the knowledge of their paper and they also have knowledge of the literature) who have published papers in reputable venues, possess over 5 years of experience in scientific publishing, and have authored more than 5 scientific papers. We collected their accepted papers (published between 2023 and 2024) and followed the process of dataset preparation as we discussed in Section 3, and generated FRIs. We modify the prompt slightly to specifically generate only the top five results (see Appendix B). We selected the outputs from Claude and GPT-4[8] models due to their better IAScore and Idea Distinction index. We adopt this approach to avoid author exhaustion and to get an accurate evaluation. More details about the human evaluation are mentioned in the Appendix B.

5 Results and Discussion

5.1 Alignment Results

Figure 3 provides a comparative overview of the IAScore for four language models[9] Claude-2, Gemini-1.0, GPT-3.5, and GPT-4 across five academic domains: Chemistry, Computer Science, Economics, Medical, and Physics[10].

In the Chemistry and Economics domains, Claude has the highest IAScore, indicating strong alignment with the authors' future research ideas. Claude and GPT-4 have almost similar values for the Computer, Medical, and Physics domains (with GPT-4 slightly higher). GPT-3.5 and Gemini have lower scores than both GPT-4 and Claude in every domain. GPT-3.5 has almost the same score as Gemini in the Chemistry and Economics domains. However, it scores higher than Gemini in the Computer, Medical, and Physics domains. The results underscore the advancements in language model capabilities, with each model showcasing domain-specific strengths in idea generation. This align-

[8] We used gpt-4-turbo using OpenAI API for the generation

[9] We set maximum token length to 512, and temperature=0 for each model

[10] We used GPT-3.5-turbo-0125 for GPT-3.5 (cutoff: December 2023), gpt-4-0125-preview for GPT-4 (cutoff: December 2023), anthropic.claude-v2 for Claude 2 (cutoff: August 2023), and gemini-1.0-pro-002 (cutoff: February 2023).

ment of LLMs shows that LLMs are able to generate relevant and novel ideas to some extent. We also studied the effect of length of future work on IAScore (See Appendix D). Our overall analysis shows that ideas of moderate length (20-40 words) achieve the highest Impact Assessment Scores (IAScores), balancing detail and clarity, while shorter (<20 words) and longer (40-60 words) ideas tend to score lower due to insufficient detail. We also conducted a human analysis to understand the quality of research ideas generated when the IAScore is low (see Appendix G).

5.2 Distinctness Results

We show the comparative evaluation of idea distinctness scores in Figure 4. The line graph depicts the variation of distinctness between the generated ideas and the human-written ideas (AP-FRIs). GPT-3.5 shows the least distinctness among the generated ideas, except in the Computer domain, where it is slightly more distinct than Gemini. As shown in the graph, the distinctness of Gemini is also quite low; however, it is slightly better than GPT-3.5, except in the Computer domain.

The generated ideas of GPT-4 are more distinct than those of Gemini and GPT-3.5 (except for economics, whereas the distinctness of GPT-4 is the same as Gemini). However, it is lower than both Claude and Human. The Idea Distinctness Index of the generated ideas from Claude are almost the same as those of humans for Chemistry, Economics, and Medical domains. However, they are higher than even human scores in the Computer and Physics domains, which shows that it generates very distinct FRIs.

5.3 Human Evaluation Results

We conducted a human evaluation on 660 generated ideas for 66 papers in the computer science domain. To validate the quality of human annotation, we measure the inter-annotator agreement ratio where 20% of the generated ideas are evaluated by two different authors of the same paper. We measured Cohen's kappa coefficient (Cohen, 1960), which was 0.83, confirming the high-quality annotation of the research ideas generated.

Novelty: Figure 5 displays the results of the human evaluation. We observed that Claude generates 14.78% of non-novel and 16.52% generic FRIs, 41.73% moderately novel, 20.86% very novel, and 16.52% extremely novel FRIs. GPT generates 7.83% not-novel, 13.91% generic, 42.61% moderately novel, 28.70% very novel, and 6.96% extremely novel ideas. Claude generates more non-novel and generic ideas than GPT-4, while GPT-4 produces more very novel ideas and nearly the same number of excellent ideas.

We conducted an additional human evaluation in the physics domain, analyzing 630 generated ideas across 63 papers. We observed that Claude generates 15.28% non-novel, 16.02% generic, 31.22% moderately novel, 20.36% very novel, and 17.12% extremely novel FRIs. GPT produces 7.33% not-novel, 14.53% generic, 42.22% moderately novel, 28.0% very novel, and 7.95% extremely novel ideas, similar to previous observations in computer science. This demonstrates that although LLMs also generate generic or already explored ideas, they are capable of producing novel ideas that have either not been explored or have been minimally explored[11].

Relevance and Feasibility: After human evaluation, for Computer Science domain, we found that 76.67% of the ideas generated by Claude and 93.34% by GPT-4 are relevant. Furthermore, 83.34% of Claude's generated ideas and 96.64% of GPT-4's ideas were judged to be practically feasible and factually correct. Similarly, for the Physics domain, after human evaluation, we found that 78.45% of the ideas generated by Claude and 91.56% by GPT-4 are relevant. Furthermore, 85.67% of Claude's generated ideas and 94.78% of GPT-4's ideas were judged to be practically feasible and factually correct.

These results highlight that Claude and GPT-4 can generate relevant and feasible research ideas. However, the reason Claude generates more impractical and irrelevant research ideas may be that Claude attempts to generate more distinct research ideas than GPT-4, as we evaluated and discussed in Section 5.2.

5.4 Open-ended generation:

We tested whether LLMs could retain open-ended generation capabilities by providing only a title and abstract as input. Our findings showed that, overall, LLMs can still generate open-ended content due to their past knowledge. However, they may not produce many high-quality ideas, as they lack access

[11]Comprehensive human evaluations across all domains require significant expertise and logistical resources, making it a challenge for the present study. However, the observed results strongly suggest that similar trends are likely to hold across other domains.

to recent publications and methodological insights relevant to the current paper. We discuss this in details in the Appendix C.

5.5 Impact on Adding Background Knowledge

We found that adding background knowledge reduced the generation of generic or non-novel ideas and improved relevance and factual accuracy. However, further research is needed to boost the novelty of generated ideas. We discuss this in details in the Appendix E.

5.6 Analysis on New Models

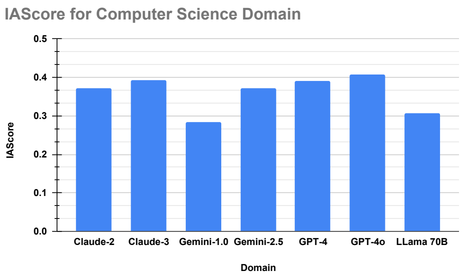


Figure 7: IAScore result for CS domain

As several new models were released after our initial experiments, we extended our analysis to include their results as well, as shown in Figure 7. GPT-4o achieves the highest IAScore in the Computer Science domain (≈ 0.41), marginally surpassing GPT-4 and Claude-3 (both ≈ 0.39). Claude-2 and Gemini-2.5 follow at $\approx$ 0.37, while Llama-70B and Gemini-1.0 trail at $\approx$ 0.31 and $\approx$ 0.28, respectively, illustrating the incremental gains delivered by successive model releases. We discuss more details on Appendix K.

6 Conclusion and Future Work

In conclusion, we present the first attempt to evaluate the potential of LLMs in generating future research ideas across five domains: Computer Science, Economics, Chemistry, Physics, and Medicine. Here, we presented the first systematic attempt to evaluate the potential of LLMs for generating future research ideas across five diverse domains. By introducing the Idea Alignment Score (IAScore) and the Idea Distinctness Index, we offered scalable, interpretable metrics that complement costly human evaluation and provide a structured way to assess the quality of generated ideas. Our results and analysis show that LLMs possess domain-specific strengths in idea generation. Furthermore, the results from the Idea Distinctness Index indicate that LLMs, such as Claude and GPT-4, generate more distinct research ideas than Gemini and GPT 3.5. GPT-4 and Claude align better with authors written future research ideas than Gemini and GPT-4. The alignment of LLMs with the authors of generated ideas, and our human evaluation on relevance, novelty, and feasibility, reveal that although LLMs often produce non-novel and generic ideas, they have the potential to generate relevant and novel and diverse ideas to a significant extent. We hope that the findings and experiments of this work will unlock the potential of LLMs in idea generation and will foster new advancements in automated scientific innovation.

In future, we plan to automate the extraction and annotation pipeline to enable scalable, automatic evaluation of our metrics for newly released models. Also, beyond relying on authors' stated plans, future datasets could integrate ideas from related work, citing papers, and community discussions to reduce bias and capture broader research directions.

7 Limitations

7.1 Limitations of Data Collection

We extracted papers using the Semantic Scholar Academic Graph API from January 2023 to February 2024. The number of papers available is limited by the scope of our data extraction from the Semantic Scholar Academic Graph. We excluded papers that are not in English, as well as those whose abstracts could not be correctly parsed from the PDFs. Not all of these papers include sections on future work; therefore, we annotated only those that contained sections outlining future research directions. So due to such limitations, we collected 250 papers from each domain for analysis.

7.2 Memorization

(Carlini et al., 2022) highlight that LLMs are prone to memorizing portions of their training data, a significant concern in the evaluation of contemporary LLMs. Despite this, the data used for pre-training and post-training includes "a small amount" of more recent data (Wang et al., 2023). Therefore, we gathered recent papers from 2023 and 2024. By focusing our evaluation on papers published in these years, the likelihood of test papers appearing in the pre-training corpora for the models is substantially reduced. In addition, we conducted a manual review of these papers to assess memorization. This involved asking various questions related to the papers, such as their titles, publishing venues, author names, etc., to see if the models could supply the missing information. Our findings showed no evidence of such memorization occurring. A similar approach is also followed by (Wang et al., 2023) (discussed in Section 6.4) and even they did not find any evidence of this occurring.

7.3 Limitation of IAScore and Idea Distinctness Index

IAScore provides a systematic and interpretable method for measuring the alignment between ideas generated by LLMs and those identified by domain experts. While effective as a benchmark for assessing credibility and relevance, IAScore primarily reflects alignment with author-specified future research directions. Consequently, it may not fully capture innovative ideas proposed by LLMs that authors themselves have not explicitly mentioned. IAScore's performance is sensitivity to accuracy and comprehensiveness of the author-identified future directions. Moreover, IAScore emphasizes alignment rather than explicitly evaluating the depth, complexity, or practical feasibility of generated ideas.

As there is currently no reliable automatic metric for evaluating idea generation, human evaluation remains the primary approach. However, it is costly, resource-intensive, and not scalable. Our findings show that IAScore strongly aligns with expert judgments, but for more accurate and nuanced evaluation, human assessment is still preferred. Also, Idea Distinctness Index, which relies on cosine similarity, might miss subtle differences between ideas that appear semantically similar but actually represent distinct research directions.

Ethics Statement

We have utilized the open source dataset for our work. Our aim for this work is to assess the potential of language models in generating ideas. Our Institutional Review Board (IRB) evaluated and approved this study. We do not encourage the use of LLMs to generate AI generated research papers (by generating new ideas) or misuse it for harmful idea generation. LLMs can process and synthesize vast amount of literature faster than humans, potentially identifying new patterns or gaps in research that might not be obvious, thus accelerating scientific discovery. However, since LLMs can generate content that may be similar to existing materials, this raises concerns about intellectual property rights and the originality of ideas. LLMs utilized for generating ideas might be misapplied to produce harmful materials such as plans for schemes for designs for destructive devices, explosive devices, ideas for spamming. Notably, it is a common challenge among existing LLMs with strong creative and reasoning abilities. So, we emphasize the responsible use of LLMs for idea generation and the need to broadly improve the safety of LLMs.

Acknowledgement

Sandeep Kumar acknowledges the Prime Minister Research Fellowship (PMRF) program of the Govt of India for its support. We acknowledge Google for the "Gemma Academic Program GCP Credit Award", which provided Cloud credits to support this research. This research used the resources of the Oak Ridge Leadership Compute Facility (OLCF) at the Oak Ridge National Laboratory (ORNL), which is supported by the United States Department of Energy (DOE)'s Office of Science under Contract No. DE-AC05-00OR22725.

References

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. *CoRR*, abs/2312.11805.

Anthropic. 2023. Model card for claude 2.

Mathias Benedek, Emanuel Jauk, Andreas Fink, Karl Koschutnig, Gernot Reishofer, Franz Ebner, and Aljoscha C. Neubauer. 2014. To create or to recall? neural mechanisms underlying the generation of creative new ideas. *NeuroImage*, 88:125–133.

Faeze Brahman, Chandra Bhagavatula, Valentina Pyatkin, Jena D Hwang, Xiang Lorraine Li, Hirona J Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, and Yejin Choi. 2023. : Making small language models better procedural knowledge models for (counterfactual) planning. *arXiv preprint arXiv:2305.19472*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *ArXiv*, abs/2202.07646.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

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

Wenke Hao, Zhicheng Li, Yuchen Qian, Yuzhuo Wang, and Chengzhi Zhang. 2020. The ACL FWS-RC: A dataset for recognition and classification of sentence about future works. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, pages 261–269. ACM.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Benchmarking large language models as ai research agents. *ArXiv*, abs/2310.03302.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Laura J. Kornish and Jeremy Hutchison-Krupat. 2017. Research on idea generation and selection: Implications for management of technology. *Production and Operations Management*, 26:633–651.

Thomas S. Kuhn and David Hawkins. 1963. The structure of scientific revolutions. *American Journal of Physics*, 31:554–555.

R. Kumar. 1996. *Research Methodology: A Step-by Step Guide for Beginners*. Longman Australia.

Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16693–16704, Singapore. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2015. A diversity-promoting objective function for neural conversation models. *ArXiv*, abs/1510.03055.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

J.J. Louviere, T.N. Flynn, and A.A.J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge books online. Cambridge University Press.

Rahul 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. 2022. text-embedding-ada-002. `https://platform.openai.com/docs/guides/embeddings`. OpenAI embedding model used via API.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Plato, G.R.F. Ferrari, and T. Griffith. 2000. *Plato: 'The Republic'*. Cambridge Texts in the History of Political Thought. Cambridge University Press.

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*, abs/2311.05965.

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.

Serhad Sarica, Binyang Song, Jianxi Luo, and Kristin L. Wood. 2021. Idea generation with technology semantic network. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 35:265 – 283.

Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic generation of programming exercises and code explanations using large language models. In *ICER 2022: ACM Conference on International Computing Education Research, Lugano and Virtual Event, Switzerland, August 7 - 11, 2022, Volume 1*, pages 27–43. ACM.

W. Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter J. Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence A. Donehower, Ying Chen, and Olivier Lichtarge. 2014. Automated hypothesis generation based on mining scientific literature. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1877–1886. ACM.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nat.*, 571(7763):95–98.

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

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yi Xu, Shuqian Sheng, Bo Xue, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2023. Exploring and verbalizing academic ideas by concept co-occurrence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13001–13027, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

A Dataset Annotation

A.1 Dataset Annotation Guidelines

Recognizing future research idea in a paper involves analyzing the portion of text containing directions for future research. The following steps can be followed:

Step 1: Begin by reading the Title and Abstract of the paper to gain an understanding of its subject matter. It is important to read these sections multiple times to grasp the paper's main points, such as its motivation, contributions, and other relevant aspects. If necessary, refer to the paper itself or read related material to enhance your understanding.

Step 2: Identify Key Sections for Analysis Focus primarily on the Discussion and Conclusion sections of the paper, as these areas often contain explicit mentions of future research directions. Scan the Methodology section as well, as sometimes suggestions for improving future studies or addressing current study limitations are mentioned here.

Step 3: Distinguish Future Research Ideas from General Statements: Differentiate explicit future research suggestions from general discussion. Future research directions usually involve specific recommendations, plans, or identified gaps that require further exploration. These are often phrased using terms like "future studies should," "further research is needed," or "additional work will." Avoid confusing these with broader statements of potential relevance or applicability, which do not provide direct guidance on future work.

We offer multiple examples of papers with its future research ideas to assist and direct the annotators. We found a few text which looks like future work but is on contrary the motivation of the work. As an example, consider the following: *"The goal of this work was to direct attention to emerging and novel research involving "magnetogel nanohybrid materials" that might be relevant in future applications for the treatment of wastewater, as well as in other fields.*

The second example is: *"Our data could be useful for designing high-quality trials in the future to define the exact role of hemoadsorption in ARDS."*.

This sentence describes the future applications of existing research on magnetogel nanohybrid materials, not a specific direction for future research.

Also another example is: *"The goal of this work was to direct attention to emerging and novel research involving magnetogel nanohybrid materials that might be relevant in future applications for the treatment of wastewater, as well as in other fields."* This is the application in future, and not the future work.

Step 4: Separate Future Research from Limitations: Carefully examine any limitations mentioned in the paper to determine if they are explicitly linked to future research. Only consider a limitation as future work if the authors clearly indicate a direct intention to address it in subsequent studies. This helps avoid assuming that all limitations naturally lead to future research directions.

There is also very thin line between limitation and future work, where a limitation can or cannot be a future work. There were few cases where limitations were mentioned *"One limitation of this paper is the absence of a coordinated attention structure to capture cross-channel information."*. As limitations can or cannot be a future work, we only take those limitations which is explicitly mentioned by the author to be a future work. Hence, we only considered the explicit mention of the future work by the author in their paper.

A.2 Annotator Training

Given the complexity of the papers and their frequent use of technical terminology, we hired two doctoral students, each boasting over four years of experience in scientific research publishing. To facilitate their training, an expert with more than ten years of experience in scientific publishing annotated 20 random papers from each domain, adhering to our guidelines. After this initial round of annotation, we reviewed and corrected any misinterpretations with the annotators, further refining their training and enhancing the clarity of our annotation guidelines. To assess the effectiveness of the initial training, we compiled another 20 papers from each domain. From the second round onward, the annotators consistently identified 95% or more of the future research ideas correctly, based on comparison with expert-annotated labels.

A.3 Annotation Process

We regularly monitored the annotated data, placing emphasis on identifying and rectifying inconsistencies and cases of confusion. We also implemented an iterative feedback system that continuously aimed to refine and improve the annotation process. In cases of conflict or confusion, we removed those papers as we wanted only good quality dataset. Following the annotation phase, we obtained an average inter-annotator agreement score of 0.94 using Cohen's kappa (Cohen, 1960), indicating a substantial consensus among the annotators.

A.4 Annotator's Pay

We compensated each annotator according to the standard PhD salaries in India, based on the hours they worked. The appointment and salaries adhere to our university's established practices. Payment was made per paper since the time required to read and extract future research ideas from each paper varies, depending on its complexity, technical terminology, and the annotator's familiarity with the subject. Thus, paying based on time spent could

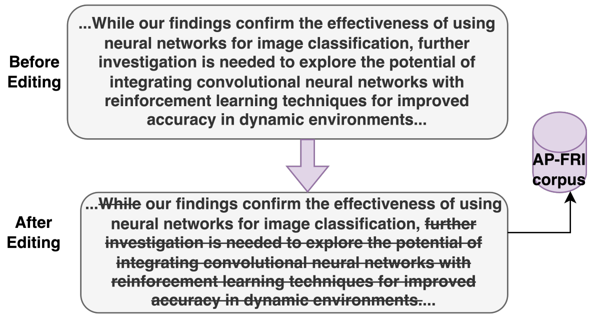


Figure 8: An example for FRI editing; Here the strike through text is removed from the paper text

have potentially compromised the quality of the annotations. During the pilot phase, we observed that hourly-based payments encouraged speed over accuracy, potentially compromising annotation quality. In contrast, paying per paper ensured that annotators could dedicate the necessary effort and time required by each individual paper without worrying about time constraints or inefficiencies associated with clock-based metrics. In cases where papers were especially complex or ambiguous, we excluded them from the dataset rather than risk subpar annotation quality. To maintain accuracy and prevent fatigue, we imposed a daily limit of six hours for annotators.

A.5 Detailed Dataset Statistics

Figure 2a provides a domain-wise distribution of the average word count in academic papers, excluding discussions on future work (FWK). It can be observed that the length of papers across all fields falls within a range of 7,000 to 8,000 words. Additionally, we calculated the average word count of extracted future work within each domain, providing comparative insights into how different fields prioritize discussions of future research directions. Figure 2b compares the average word count of future work text across six distinct scholarly domains. We observed that the literature in Computer Science notably prioritizes extensive discourse on future research, with an average word count significantly higher than that of other disciplines. In contrast, the literature in Chemistry demonstrates a more concise approach to discussions of future research, as evidenced by its lower average word count.

B Human Annotation

We ask the following questions from each human evaluator:-

- Q1: Is the idea relevant with the research topic of the paper. (Relevant/Not relevant)
- Q2: Assess the originality/novelty of the research idea (5 scale)
- Q3: Review the research idea for factual correctness and feasibility. Is the idea impractical or too vague to be actionable? (Not Possible/Possible)

For Q2, we used Best-Worst Scaling (Louviere et al., 2015) on a 5-point scale.

We prepared a Google Form for each paper and provided the links to the annotators. We also specified instructions for them at the beginning of the form. We have added an example of the form for a paper in Figure 10, Figure 11, and Figure 12.

Here is the little modified from for human evaluation that generates only top 5 research ideas:-

System: You are a research scientist.
User: Imagine you are a research scientist. After reading the following paper, brainstorm to generate potential top 5 future research ideas:

[paper text]

Potential top 5 future research ideas from the paper in bullet points are:

Here, '[paper text]' contains the full content of the paper after removal of future work sections.

C Effect of giving only Title and Abstract as Input

We found a few cases where we provided only a title and abstract as input to see if LLMs can still retain open-ended generation capabilities. We discovered a few cases where GPT-4 still generated novel ideas, such as for a paper (Kumar et al., 2023) it generated: *"Incorporate explainable AI methods to provide transparency into how the AI model makes its predictions, thus making the outcomes more interpretable and acceptable to human editors."*. This kind of analysis has not been done yet and could be helpful. After providing full paper content to the model we found that same idea was again generated.

There were also cases where GPT-4 generated a novel idea of solving the problem using transformers for a task (The task was mostly solved using techniques like RNN), which had not been done before. However, after providing the full paper content, the model understood that this transformer has already been implemented for this task, so further suggested to add more contextual information to it to boost the performance (limited information was given as input to the paper). Overall, we found that LLMs can still retain open-ended generation because it has past knowledge. But it may not generate many good ideas since it does not have access to recently published papers or other methodological findings related to the current paper.

D Effect of Length of Idea on IAScore

In our analysis, we explore the relationship between the length of ideas and their corresponding Impact Assessment Score (IAScore), specifically focusing on computer science papers and outputs generated by GPT-4. This relationship is visually represented in the bar chart found in Appendix Figure 9. The data reveal that shorter ideas, typically under 20 words, tend to receive lower IAScores. This could be attributed to their lack of detailed information, which might be essential for a comprehensive understanding and assessment. Conversely, we observe that ideas spanning 40-60 words also tend to score lower. This may result from their verbosity; excessive information can dilute the core message, making it challenging to discern the main points. Interestingly, ideas with a moderate length, ranging from 20 to 40 words, achieve the highest IAScores. This length seems optimal as it allows for sufficient detail without overwhelming the reader, striking a balance that facilitates clearer understanding.

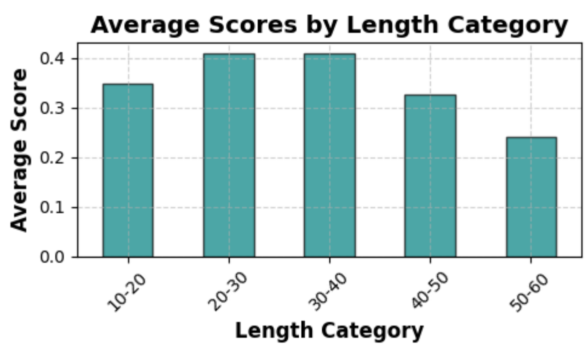


Figure 9: Effect of length on IAScore

E Effect of Adding Additional Background Knowledge

Below we describe our pipeline for retrieving background knowledge and generating future research ideas using LLMs in detail :-

E.1 Vector Database

We utilized the Semantic Scholar API (Kinney et al., 2023) to collect the titles and abstracts of approximately 1.9 lakh existing computer science research papers. We employed BERT embeddings to create vector representations for the titles of these papers, which were then stored in a vector database.

E.2 Retriever

To retrieve relevant papers, we created embeddings for the title of the paper for which we have to generate ideas. We computed the cosine similarity between the embedding of the title of this paper and those from our vector database. We then retrieved the top 20 research papers that exhibited the highest similarity to our target paper title. Finally, we extracted the contributions from these papers to gather relevant data from their abstracts.

We used the following prompt to instruct LLM to extract useful information from abstract of the paper:

> **System**: You are a helpful research agent that generates background knowledge or related works given abstracts of papers.
> **User**: You are given abstracts of research papers and your task is to extract contributions or findings or methods proposed in the paper. You are not allowed to make any changes to data given to you. Return the response as it is and return response for all 20 papers in passage. Return title of paper followed by its contributions or findings or methods in less than 100 words. If no contributions or findings or methods are found, return NONE.
>
> PASSAGE: '{relevant_passage}'
>
> Potential top 5 future research ideas from the paper in bullet points are:

We designed the above query prompt to ensure that the LLM[12] understood its role in extracting

[12]We employed Gemini-Pro model for this task

relevant information without altering the provided information.

E.3 Generator

Next we produced the ideas using a prompt that includes the prompt using the paper and the retrieved background knowledge.

Specifically, we used the below prompt for our task:-

> **System**: You are a research scientist.
> **User**: Imagine you are a research scientist. 1) Read the full paper and understand it. 2) Find out the related works in this direction 3) Brainstorm and follow a step-by-step reasoning approach to generate to potential future research ideas:
>
> [paper text] [background knowledge]
>
> Make sure the future research ideas are very distinct from the background knowledge provided. Potential top 5 future research ideas from the paper in bullet points are:

Here, '[paper text]' contains the full content of the paper after removal of future work sections. '[background knowledge]' contains the background knowledge retrieved. An example of background knowledge is shown in Appendix Table 6.

We performed this experiment on the same set of papers and conducted human evaluations for novelty following the same methodology as we discussed in Section 4.6. The results are shown in Figure 5. Initially, we observed that adding background knowledge affected the LLM's performance; it primarily generated ideas that already existed, merely creating new combinations from the background knowledge. Subsequently, we modified the prompt to instruct the model not to repeat ideas that were mentioned in the background knowledge.

We found that adding background slightly improved the task. The results show that the improvements for GPT-4 and Claude were 50% and 53.33%, respectively, in reducing the generation of non-novel ideas. Also, it resulted in the improvement of 7.14% and 11.76% not generating generic ideas of GPT-4 and Claude. We observed that GPT-4 generated 9.52% and 14.63% more moderately novel ideas. However, we noted only a very slight improvement in the generation of highly novel or extremely novel ideas.

The analysis revealed that 73.71% of the ideas generated by Claude and 93.34% by GPT-4 were relevant. We observed that the relevance score for Claude decreased by 2.96%, and GPT-4 increased by a slight 0.77%. Furthermore, 83.14% of Claude's generated ideas and 96.98% of GPT-4's ideas were judged to be practically feasible and factually correct. The score for Claude decreased by 0.20%, and the score for GPT-4 increased by 0.34%. It seems that additional information negatively impacts Claude's performance by generating ideas that are irrelevant, non-novel, and infeasible. However, for GPT-4, we observed that incorporating additional background knowledge helps prevent the generation of non-novel or generic ideas and slightly improves the relevance and factual correctness of the generated ideas. However, further research is needed to enhance the ability of LLMs to generate more novel ideas.

F BERTScore Implementation Details

The motivation to use BERT embeddings is that the generated and the original ideas often do not use the same words, so we need to understand the contextual meanings of the ideas in order to compare them. We used the default setting of the BERTScore metric, which employs a 24-layer RoBERTa-large model and utilizes the 17th layer for embedding. We determined the threshold[13] using the validation set. If the similarity exceeds that threshold, we classify those pairs of ideas as similar, and vice versa.

G Error Analysis:

We conducted a human evaluation using three expert annotators, each with more than five years of experience in this field. They reviewed 15 papers. We assigned papers to each reviewer based on their familiarity with the subject matter of the papers. We identified two major reasons for the low IAS score:

- Generic Ideas: Few ideas such as *"Explore different explainability methods like LIME, SHAP to generate model explanations instead of just rationales. Compare their effectiveness.", Building on the baseline model, future research could explore more advanced NLP*

[13]We set the threshold 0.68 empirically

models and techniques for contradiction detection. are generated. These statements are true; however, they are very generic and are common.

- Not Feasible Ideas: We observed that LLMs occasionally propose ideas that are not feasible, even in principle. For instance, in one case, the model suggested "*collecting real-time, large-scale annotated data on private peer review discussions between authors and reviewers to study the evolution of scientific ideas.*" Such data is not only unavailable, but also extremely unlikely to ever be accessible due to strict confidentiality and privacy concerns in the peer review process. Human researchers, on the other hand, are typically more attuned to such ethical and practical constraints and tend to propose ideas that are realistically actionable.

- Author Miss: Due to page limits or more novel ideas, the author fails to mention a few ideas in a paper. For example, for a paper (Kumar et al., 2023) GPT-4 generated idea: *"Exploring the Impact of Contradictions on Review Outcomes: An interesting area for future research would be to study the impact of reviewer contradictions on the outcomes of the peer review process. This could involve analyzing the correlation between the presence and nature of contradictions and the final decisions made by editors (acceptance, rejection, major/minor revisions). Such studies could provide valuable insights into how contradictions influence the decision-making process and how they might be effectively managed to improve the fairness and quality of peer review."*. This represents a strong, novel research problem not mentioned by the authors, which warrants future investigation

H Direct FRI and Mixed FRI

- **Direct FRI**: When the sentences that mention future research idea only contains future research idea. For example *"In future work, we plan to extend our approach to other code-mixed languages and evaluate its performance on more NLP tasks."*

- **Mixed FRI**: We found that sometimes research papers articulate future research ideas along with other essential information of the paper in a single sentence. For example in Figure 8, this sentence not only summarizes the current research findings but also clearly outlines a direction for future work.

I Output Examples

Our LLM generated future research output can be found in Table 1, Table 2, Table 3, Table 4 and Table 5.

J Model Training Cutoff

Regarding model training cutoffs, we used GPT-3.5-turbo-0125 for GPT-3.5 (cutoff: December 2023), gpt-4-0125-preview for GPT-4 (cutoff: December 2023), anthropic.claude-v2 for Claude 2 (cutoff: August 2023), and gemini-1.0-pro-002 (cutoff: February 2023), Llama 3.3 70B Instruct (cutoff: December 2023), gpt-4o-2024-08-06 for GPT4o (cutoff: October 2023), claude-3-opus-20240229 for Claude 3 (cutoff: August 2023), gemini-2.5-pro-preview-05-06 for Gemini-2.5 (cutoff: December 2025).

K Evaluation on Newly Released Models

To evaluate the performance of recently released language models, we constructed a supplementary Computer Science dataset comprising 200 papers published between January and March 2025. These papers were sourced from S2ORC using the same English-language, PDF-availability, and *Future Work* section filters as described in Section 3. Annotation was carried out following the same protocol as for the main dataset (Appendix A), ensuring consistency in labeling. This new dataset forms the basis for the IAScore results reported for GPT-4o, Claude-3, Gemini-2.5, and Llama-70B (Figure 7). We plan to automate the process to automatically evaluate the IAScore for newly released models.

Please read the research idea carefully and select the appropriate response.

We explain each questions below:-

Q1: Is the idea relevant with the research topic of the paper. (Relevant/Not relevant)

- **Not Relevant**: If the idea seems unrelated to the research topic of the paper.
- **Relevant**: If the idea aligns with the research topic of the paper.

Q2: Assess the originality/novelty of the research idea. (Not Novel/Novel)

- **Not Novel**: If you believe the idea is generic or already exists.
- **Slightly novel**: The idea is somewhat generic
- **Moderately novel**: This is a promising research direction but has already been explored by a significant number of researchers.
- **Very novel:** The idea represents a good research direction and has been pursued by only a few researchers.
- **Extremely novel**: The idea is highly innovative and, to my knowledge, has not been explored by anyone.

Q3: Review the research idea for factual correctness and feasibility. Is the idea impractical or too vague to be actionable? (Not Possible/Possible)

- **Not Possible**: If you feel that the idea is impractical/doesn't make any sense/ not realistic
- **Possible:** If you believe the idea is practical and realistic.

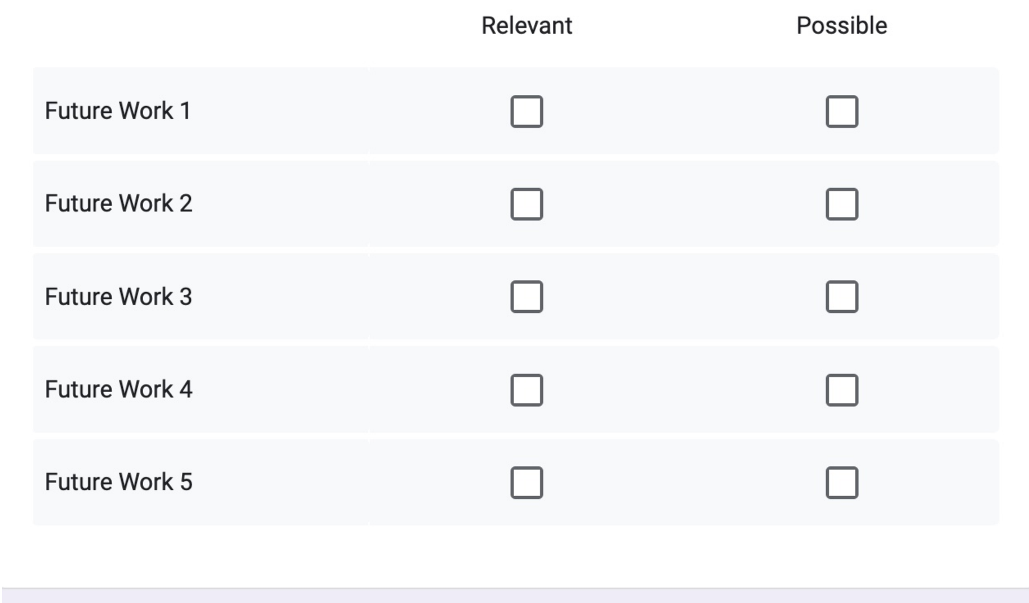
Figure 10: Instructions on Human Feedback form

Future Work 4: Add lexical constraints during decoding to prevent minimal word changes. This could enforce more semantic changes for style transfer.

Future Work 5: Evaluate the model on other styles like formality, gender, political slant etc. This could test the generalizability of the approach to other text style transfer tasks.

	Not novel	Slightly novel: The idea is somewhat generic	Moderately novel: This is a promising research direction but has already been explored by a significant number of researchers.	Very novel: The idea represents a good research direction and has been pursued by only a few researchers.	Extremely novel: The idea is highly innovative and, to my knowledge, has not been explored by anyone.
Future Work 1	○	○	○	○	○
Future Work 2	○	○	○	○	○
Future Work 3	○	○	○	○	○
Future Work 4	○	○	○	○	○
Future Work 5	○	○	○	○	○

Figure 11: Human Feedback form for novelty



	Relevant	Possible
Future Work 1	☐	☐
Future Work 2	☐	☐
Future Work 3	☐	☐
Future Work 4	☐	☐
Future Work 5	☐	☐

Figure 12: Human Feedback form for relevance, feasibility

Model	Response
Claude-2	1. Investigate the antioxidant and oil-stabilizing effects of peptides from the soy protein hydrolysates. The paper showed the pepsin hydrolysate had better activity, likely due to its peptide composition. Specific peptides could be isolated and tested for their antioxidant effects in oil. 2. Study the effects of the soy protein hydrolysates on the quality and shelf life of the fried food. This paper focused only on the oil stability. It would be useful to also assess how the hydrolysates affect lipid oxidation, texture, color etc. in the fried banana chips. 3. Evaluate the antioxidant effects of the soy protein hydrolysates in other vegetable oils such as sunflower, canola, corn oil etc. during frying. Since oil composition affects stability, the effects may differ in other oils high in polyunsaturated fatty acids. 4. Examine the impacts of frying parameters such as temperature, food type, frying time etc. on the efficacy of the soy protein hydrolysates. Varying these factors could provide insights into the optimal conditions for the hydrolysates to exert antioxidant effects. 5. Study the sensory qualities of fried foods with the soy protein hydrolysate additions. An analysis of taste, aroma, color acceptability by consumers would provide important information for potential commercial use of the hydrolysates as natural antioxidants.
Gemini-1.0	1. Investigate the effects of different hydrolysis conditions on the antioxidant activity of soy protein hydrolysates. This could involve varying the enzyme concentration, temperature, and pH of the hydrolysis reaction. 2. Identify the specific peptides in the pepsin hydrolysate that are responsible for its antioxidant activity. This could be done using mass spectrometry and other analytical techniques. 3. Test the antioxidant activity of soy protein hydrolysates in other food systems. This could include using them to stabilize other types of oils or to protect food products from oxidation. 4. Investigate the potential health benefits of soy protein hydrolysates. This could include studying their effects on oxidative stress, inflammation, and other chronic diseases. 5. Develop new food products that incorporate soy protein hydrolysates as antioxidants. This could include developing new types of oil-based products, such as salad dressings and mayonnaise, as well as new types of food products, such as fortified cereals and snacks.
GPT-4	1. Exploration of Peptide Sequences and Amino Acid Compositions for Enhanced Oxidative Stability: Investigate the specific peptide sequences and amino acid compositions derived from different proteases besides trypsin and pepsin to understand their contribution to the oxidative stability of oils. This could involve using a variety of proteolytic enzymes to hydrolyze soy proteins and other plant-based proteins, followed by an analysis of their peptide profiles and antioxidant activities in oil and food systems. 2. Application of Soy Protein Hydrolysates in Various Food Systems: Extend the application of soy protein hydrolysates to other food systems beyond frying oils, such as baked goods, dairy products, and meat alternatives, to assess their potential in improving oxidative stability, shelf-life, and nutritional profiles. This could involve comparative studies of hydrolysate effectiveness in different food matrices and processing conditions. 3. Mechanistic Studies on the Antioxidant Action of Hydrolysates: Conduct in-depth mechanistic studies to understand how soy protein hydrolysates interact with lipid molecules and radical species at the molecular level. This could involve computational modeling and in vitro experiments to explore the dynamics of peptide-lipid interactions and the role of specific amino acids in scavenging free radicals and metal ions. 4. Development of Environmentally Friendly and Health-focused Antioxidants: Given the interest in replacing synthetic antioxidants with natural alternatives, further research could focus on optimizing the production, purification, and application of soy protein hydrolysates as natural antioxidants. This could include exploring sustainable extraction methods, evaluating the safety and efficacy of hydrolysates in food systems, and assessing consumer acceptance. 5. Comparative Analysis of Hydrolysates from Different Oilseed Proteins: Since the study focused on soy proteins, future research could explore the antioxidative potential of hydrolysates derived from other oilseed proteins, such as sunflower, canola, or hemp. This would involve a comparative analysis of their antioxidative activities, peptide compositions, and effectiveness in enhancing the oxidative stability of different types of oils and fats used in food processing and storage.
GPT-3.5	1. Investigate the impact of different proteases on the antioxidant properties of soy protein hydrolysates in stabilizing oils during frying cycles. 2. Study the influence of peptide size and amino acid composition on the oxidative stability of oils when treated with protein hydrolysates. 3. Explore the potential of specific amino acids, such as tryptophan, histidine, and methionine, in enhancing the antioxidant activity of hydrolyzed soy proteins in frying oils. 4. Examine the effects of pH adjustments during enzymatic hydrolysis on the antioxidant performance of soy protein hydrolysates. 5. Evaluate the long-term stability and effectiveness of pepsin and trypsin hydrolyzed soy protein in preventing oil oxidation in different frying conditions over extended periods.

Table 1: Example of generated response for chemistry domain. Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10528889/pdf/foods-12-03498.pdf

Model	Response
Claude-2	1. Investigate other machine learning models like recurrent neural networks for multi-round transfer learning in lung cancer detection. The paper focused on convolutional neural networks, but RNNs may capture temporal patterns in medical imaging data. 2. Explore different loss functions and regularization techniques to further minimize negative transfer. The paper proposed custom loss functions, but there is room to experiment with other formulations. 3. Evaluate the multi-round transfer learning approach on diverse medical imaging datasets beyond lung cancer, such as brain MRI scans. This could demonstrate the generalizability of the method. 4. Develop more advanced generative models like CycleGAN to synthesize intermediate domains between source and target. This could create higher quality bridging distributions. 5. Perform multi-round transfer learning in a low-data regime by leveraging data augmentation and semi-supervised learning techniques. This could show the approach is viable when target data is extremely scarce.
Gemini-1.0	1. Investigate the impact of different prioritization algorithms on the performance of MTL-MGAN.The paper uses a hybrid approach that merges modified 2D dynamic warping (M2DW) and Silhouette coefficient to prioritize source domains. Future research could explore other prioritization algorithms, such as mutual information, transfer entropy, or graph-based methods, to determine if they can further improve the performance of MTL-MGAN. 2. Develop new loss functions to minimize negative transfer.** The paper proposes loss functions that are designed to minimize negative transfer in the aspects of domains, instances, and features. Future research could investigate other loss functions that can further reduce negative transfer, such as adversarial loss, maximum mean discrepancy, or knowledge distillation. 3. Explore different architectures for the MGAN.** The paper uses a modified GAN (MGAN) to create two intermediate domains as bridges between the source domain and the target domain. Future research could explore other GAN architectures, such as deep convolutional GAN (DCGAN), Wasserstein GAN (WGAN), or StyleGAN, to determine if they can further improve the performance of MTL-MGAN. 4. Apply MTL-MGAN to other medical diagnosis tasks.** The paper evaluates the performance of MTL-MGAN on 10 benchmark datasets related to lung cancer detection. Future research could apply MTL-MGAN to other medical diagnosis tasks, such as breast cancer detection, skin cancer detection, or diabetic retinopathy detection, to determine its generalizability. 5. Investigate the use of MTL-MGAN for other types of data.** The paper applies MTL-MGAN to medical images. Future research could investigate the use of MTL-MGAN for other types of data, such as text, audio, or time series data, to determine its applicability to a wider range of problems.
GPT-4	1. Extension to Other Forms of Cancer and Diseases: Investigating the applicability and effectiveness of the MTL-MGAN approach for the detection and diagnosis of other types of cancer, such as breast, prostate, or skin cancer, and potentially other diseases that have distinct imaging characteristics. This would involve adapting the transfer learning and MGAN components to handle different imaging modalities and disease signatures. 2. Real-Time Diagnostic Systems Development: Developing real-time, AI-driven diagnostic systems using the MTL-MGAN framework that can be integrated into clinical workflows. This could significantly reduce diagnosis times and improve patient outcomes by providing immediate, accurate diagnostic information during patient examinations or surgeries. 3. Integration with Electronic Nose Technology for Early Detection: Combining the MTL-MGAN approach with emerging technologies like electronic nose devices, which can detect volatile organic compounds associated with lung cancer from breath samples. Research could focus on creating a hybrid detection system that leverages imaging data and breath analysis for earlier detection of lung cancer. 4. Exploration of Unsupervised and Semi-supervised Learning Techniques: Given the challenge of obtaining large, annotated medical imaging datasets, exploring unsupervised and semi-supervised learning methods within the MTL-MGAN framework could be valuable. This might involve developing algorithms that can learn from unlabelled data or very small labelled datasets, potentially increasing the model's applicability in low-resource settings. 5. Improvement of Transfer Learning Techniques for Heterogeneous Data Sources: Investigating advanced transfer learning techniques that can effectively handle more significant variations between source and target domains, particularly when dealing with heterogeneous data sources. This research could include the development of novel algorithms for better domain adaptation and intermediate domain creation to facilitate the transfer of knowledge across diverse medical datasets and imaging modalities.
GPT-3.5	1. Exploring Multiround Transfer Learning in other Healthcare Applications: Investigate the applicability of multiround transfer learning in various healthcare domains beyond lung cancer detection, such as cardiovascular disease diagnosis, diabetic retinopathy detection, or infectious disease prediction. 2. Advanced Loss Function Design for Negative Transfer Avoidance:Develop novel loss functions to further minimize the impact of negative transfer in transfer learning models by incorporating more sophisticated domain, instance, and feature-based penalties. 3. Enhanced Generative Adversarial Networks for Bridging Domain Gaps:Investigate the development of advanced generative adversarial networks that can create more effective intermediate domains to bridge the gap between diverse source and target domains in transfer learning scenarios. 4. Automated Prioritization Algorithm for Source Domains:Develop an automated algorithm that can dynamically prioritize source domains based on their similarities to the target domain, thereby enhancing the efficiency and accuracy of the transfer learning process. 5. Evaluation of MTL-MGAN in Real-World Clinical Settings:Conduct a comprehensive evaluation of the proposed multiround transfer learning and modified generative adversarial network algorithm in real-world clinical settings to assess its performance, scalability, and generalizability in diverse healthcare environments.

Table 2: Example of generated response for computer domain. Paper: https://www.hindawi.com/journals/ijis/2023/6376275/

Model	Response
Claude-2	1. Conduct qualitative research through interviews or focus groups to understand the specific barriers and facilitators to COVID-19 vaccination uptake in different socioeconomic groups within a country. This could provide insights to inform targeted interventions. 2. Examine the impact of COVID-19 vaccination policies and public health messaging on vaccination uptake across socioeconomic groups. For example, study how vaccination requirements or incentives affect different populations. 3. Investigate strategies to improve global collaboration and coordination in vaccine distribution and access. Research equitable pricing models, technology transfers, and policies to balance local needs with global supply. 4. Develop mathematical models to forecast COVID-19 vaccination coverage over time under different policy scenarios. 5. Conduct intervention studies to test the effectiveness of community-based participatory approaches in increasing COVID-19 vaccine acceptance among disadvantaged populations.
Gemini-1.0	1. Investigate the impact of socioeconomic factors on vaccine hesitancy and refusal.This could be done through qualitative studies to understand the underlying reasons for vaccine hesitancy and refusal, and quantitative studies to examine the association between socioeconomic factors and vaccine uptake. 2. Examine the role of social networks and community engagement in promoting vaccination.This could involve studying the influence of social networks on vaccination decisions, and exploring the effectiveness of community-based interventions in increasing vaccination rates. 3. Assess the effectiveness of different strategies for addressing vaccine inequality.This could include evaluating the impact of financial incentives, targeted outreach programs, and mobile vaccination clinics on vaccination rates among underserved populations. 4. Develop and evaluate interventions to improve the equitable distribution of vaccines.** This could involve working with governments and international organizations to develop policies and programs that ensure that vaccines are distributed fairly and equitably to all populations. 5. Conduct longitudinal studies to track the long-term impact of socioeconomic factors on vaccination rates.This could help to identify trends and patterns in vaccination uptake over time, and inform the development of policies and interventions to address persistent disparities.
GPT-4	1. Impact of Vaccine Acceptance and Hesitancy on Socioeconomic Disparities: Investigate how vaccine acceptance and hesitancy, influenced by cultural, historical, and information dissemination factors, contribute to the observed socioeconomic disparities in vaccination rates. This research could further dissect how misinformation and trust in healthcare systems intersect with socioeconomic status to affect vaccine uptake. 2. Longitudinal Studies on the Evolution of Vaccine Inequality: Conduct longitudinal studies to track changes in vaccine inequality over time, particularly as new variants emerge and booster vaccinations become more common. This research would provide insights into how socioeconomic disparities in vaccination rates evolve with the pandemic's progression and the introduction of new vaccine technologies or strategies. 3. Role of International and Public-Private Partnerships in Addressing Vaccine Inequity: Explore the effectiveness of international collaborations and public-private partnerships in improving vaccine distribution equity. This study could examine case studies of successful collaborations and identify best practices for leveraging these partnerships to reduce disparities in vaccine access. 4. Comparative Analysis of Vaccination Strategies and Socioeconomic Impact: Compare the vaccination strategies of different countries or regions with similar socioeconomic profiles to identify what strategies are most effective in mitigating inequalities in vaccination rates. This research could include the role of mobile vaccination units, community engagement initiatives, and targeted communication strategies in improving vaccine uptake among disadvantaged populations. 5. Impact of Socioeconomic Factors on Post-Vaccination Outcomes: Investigate how socioeconomic factors influence outcomes after vaccination, such as breakthrough infection rates, severity of breakthrough cases, and long-term immunity. This research could help identify whether socioeconomic disparities persist even after vaccination and how these disparities can be addressed through booster doses, tailored public health strategies, and improved healthcare access.
GPT-3.5	1. Investigating the Impact of Socioeconomic Factors on Access and Uptake of COVID-19 Booster Vaccines: A study could focus on analyzing how socioeconomic indicators, such as GDP, human development index, and corruption perception index, influence access to and acceptance of booster doses of COVID-19 vaccines. 2. Comparative Analysis of Inequality in COVID-19 Vaccination Across Continents: Future research could compare and contrast the levels of inequality in COVID-19 vaccination across different continents, especially focusing on the factors contributing to variations in vaccination rates and disparities in vaccine distribution. 3. Longitudinal Study on the Impact of Socioeconomic Factors on COVID-19 Vaccination: A longitudinal study could be conducted to track changes in vaccination rates over time and assess how socioeconomic factors impact the progress of COVID-19 vaccination campaigns. 4. Exploring Public Perception and Vaccine Hesitancy Among Different Socioeconomic Groups: Future research could delve into public perceptions and attitudes towards COVID-19 vaccination among diverse socioeconomic groups, investigating factors such as education level, income, and access to healthcare. 5. Evaluation of Intervention Strategies to Reduce Inequality in COVID-19 Vaccination: Researchers could conduct a study to evaluate the effectiveness of various intervention strategies aimed at reducing inequality in COVID-19 vaccination. By implementing targeted initiatives and measuring their impact on vaccine distribution and uptake, this research could provide evidence-based recommendations for policy interventions to promote equitable access to vaccines.

Table 3: Example of generated response for economics domain. Paper: https://www.mdpi.com/2076-393X/11/7/1245

Model	Response
Claude-2	1. Further optimize the minimal active H2 relaxin analog B9-31 through additional amino acid substitutions and modifications to improve binding affinity and activity. 2. Explore alternative methods to induce helicity and stabilize the minimal H2 relaxin analogs, such as cyclization or incorporation of other helix-inducing amino acids. 3. Develop longer-acting analogs by modifying B9-31 with polyethylene glycol, albumin binding moieties, or other half-life extending strategies. 4. Test the in vivo efficacy of optimized minimal analogs in animal models of fibrosis, heart failure, and other conditions to assess therapeutic potential. 5. Design biased analogs to selectively activate certain signaling pathways downstream of RXFP1 to optimize beneficial effects and reduce side effects.
Gemini-1.0	1. Investigate the potential benefits of using ECMO in combination with other therapies for severe respiratory failure. For example, ECMO could be used in combination with lung transplantation, mechanical ventilation, or extracorporeal membrane oxygenation (ECMO). This could help to improve patient outcomes by providing more comprehensive support for the lungs and other organs. 2. Develop new strategies for managing complications associated with ECMO. Common complications of ECMO include bleeding, infection, and thrombosis. New strategies for managing these complications could help to improve patient safety and outcomes. 3. Study the long-term effects of ECMO on patients.There is limited data on the long-term effects of ECMO on patients. Studies are needed to investigate the potential risks and benefits of ECMO in the long term. 4. Develop new technologies to improve the performance of ECMO.New technologies could help to make ECMO more effective, safer, and easier to use. For example, new oxygenators could be developed that are more efficient and less likely to cause complications. 5. Conduct research to better understand the mechanisms of respiratory failure.This could help to identify new targets for therapy and improve the overall management of respiratory failure.
GPT-4	1. Comparative Effectiveness of ECMO Configurations: Conduct randomized controlled trials comparing different ECMO cannulation strategies (e.g., dual-lumen cannula via the right internal jugular vein vs. traditional multi-cannula approaches) to assess their impact on patient outcomes, including survival rates, complication rates, and the ability to mobilize and rehabilitate during ECMO support. This research could also explore the use of image guidance for cannula placement and its impact on patient outcomes. 2. ECMO in the Management of COVID-19-Induced ARDS: Given the limited randomized controlled trial data on the efficacy of ECMO for COVID-19-associated respiratory failure, further research is needed to define the role of ECMO in this context. Future studies could focus on identifying which patients benefit most from ECMO, the timing of initiation, and how ECMO is best integrated with other supportive therapies (e.g., antivirals, anti-inflammatory agents). 3. Long-term Outcomes and Quality of Life Post-ECMO: Investigate the long-term outcomes of patients who have undergone ECMO support for severe respiratory failure, including physical, cognitive, and psychological recovery. This research could also assess the quality-adjusted life years (QALYs) gained through ECMO support and identify factors that predict better long-term outcomes. 4. Optimization of Anticoagulation Strategies: Develop and test new anticoagulation protocols or agents that minimize the risks of bleeding and thrombotic complications in ECMO patients. This could include comparative studies of direct thrombin inhibitors versus unfractionated heparin, as well as the exploration of novel anticoagulants and biocompatible circuit coatings that may reduce the need for systemic anticoagulation. 5. Integration of ECMO with Rehabilitation: Explore the feasibility, safety, and benefits of early rehabilitation protocols in ECMO-supported patients, including those on "awake ECMO." This research could assess the impact of rehabilitation on outcomes such as muscle strength, functional status at discharge, duration of hospital stay, and overall mortality. Studies could also investigate the best practices for physical therapy, nutritional support, and psychological support during ECMO.
GPT-3.5	1. Investigating the long-term outcomes and quality of life of patients who have undergone VV-ECMO support for respiratory failure, including factors influencing long-term recovery and rehabilitation post-ECMO. 2. Exploring new technologies and techniques to improve the efficiency and safety of VV-ECMO, such as developing novel anticoagulation strategies or improving cannulation methods to reduce complications and enhance patient outcomes. 3. Conducting comparative studies to evaluate the effectiveness of different ventilation strategies in conjunction with VV-ECMO support, aiming to optimize patient management and potentially improve survival rates in severe respiratory failure cases. 4. Examining the impact of ECMO utilization during disease outbreaks, such as COVID-19, on healthcare resource allocation, patient outcomes, and healthcare system preparedness, to inform future pandemic response strategies. 5. Investigating the role of personalized medicine in VV-ECMO patient selection and management, including the potential use of biomarkers, genetic profiling, or other advanced techniques to tailor ECMO therapy to individual patient characteristics and optimize outcomes.

Table 4: Example of generated response for medical domain. Paper: https://link.springer.com/article/10.1007/s41030-023-00214-2

Model	Response
Claude-2	1. Investigate the antioxidant and oil-stabilizing effects of specific peptide sequences identified in the pepsin and trypsin hydrolysates. The paper showed the pepsin hydrolysate had higher activity but did not identify the key active peptides. Bioinformatics could be used to predict promising peptides. 2. Test the soy protein hydrolysates in stabilizing other vegetable oils like canola, sunflower, etc. during deep frying. The effects may differ for oils with varying fatty acid compositions. 3. Evaluate the sensory qualities of foods fried in palm olein stabilized with the soy protein hydrolysates compared to unstabilized oil. This could include assessing odor, taste, and textural changes. 4. Study the effects of the hydrolysates on the nutritional quality of the palm olein during repeated deep frying cycles. Parameters could i+nclude vitamin E content, formation of trans fats, and retention of essential fatty acids. 5. Investigate combining the soy protein hydrolysates with other antioxidants like rosemary extract or green tea to evaluate synergistic effects on stabilizing frying oils. Mixtures may provide greater protection than individual additives.
Gemini-1.0	1. Investigate the effect of different hydrolysis conditions on the antioxidant activity of soy protein hydrolysates.This could involve varying the enzyme concentration, temperature, pH, and hydrolysis time to determine the optimal conditions for producing hydrolysates with the highest antioxidant activity. 2. Evaluate the antioxidant activity of soy protein hydrolysates in different food systems. This could involve adding hydrolysates to edible oils, fried foods, or other food products and measuring the oxidative stability of the products over time. This would help to determine the potential applications of soy protein hydrolysates as natural antioxidants in food products. 3. Study the mechanism by which soy protein hydrolysates exert their antioxidant activity.** This could involve using in vitro and in vivo assays to investigate the ability of hydrolysates to scavenge free radicals, chelate metal ions, or inhibit lipid peroxidation. This information would help to elucidate the molecular mechanisms underlying the antioxidant activity of soy protein hydrolysates. 4. Explore the potential health benefits of soy protein hydrolysates. This could involve conducting animal or human studies to investigate the effects of hydrolysates on oxidative stress, inflammation, and other health markers. This would help to determine the potential role of soy protein hydrolysates in promoting human health. 5. Develop novel food products containing soy protein hydrolysates. This could involve incorporating hydrolysates into functional foods, such as beverages, snacks, or supplements, to improve their nutritional value and antioxidant content. This would help to increase the consumption of soy protein hydrolysates and promote their potential health benefits.
GPT-4	1. Comparative Analysis of Different Plant-Based Proteins on Oil Stability: Extending the research to other plant-based proteins, such as pea, hemp, and rice proteins, to compare their effectiveness in improving oil oxidative stability during frying cycles. This could provide a broader understanding of the potential of plant-based proteins in enhancing the nutritional and safety profiles of fried foods. 2. Mechanism of Action at Molecular Level: Delving deeper into the molecular mechanisms behind the antioxidant activities of different hydrolysates. This could involve investigating the interaction between specific peptide sequences and oxidative compounds within the oil, using computational modeling and in vitro experiments to elucidate the pathways for their antioxidant actions. 3. Impact of Hydrolysate Concentration and Frying Duration: Exploring the effect of varying concentrations of the most effective hydrolysates (e.g., from pepsin digestion) over extended frying cycles beyond 12, to determine the optimal concentration for maximum oxidative stability without affecting the sensory qualities of the oil and the fried product. 4. Application in Industrial Frying Processes: Conducting pilot-scale studies in industrial settings to assess the feasibility and economic viability of incorporating hydrolysates into commercial frying oils. This research could include the development of guidelines for the use of hydrolysates in different types of frying operations and their impact on the quality and shelf-life of fried foods. 5. Development of Natural Antioxidant Blends: Formulating blends of hydrolysates with other natural antioxidants (e.g., extracts from spices, herbs, or other food by-products rich in phenolic compounds) to synergistically enhance the oxidative stability of frying oils. This research could also investigate the sensory impact of these blends on fried products and consumer acceptance.
GPT-3.5	1. Investigating the impact of different proteases on the generation of antioxidant peptides from soy protein hydrolysates and their effectiveness in stabilizing oils during frying cycles. 2. Exploring the structural features of peptides derived from soy protein hydrolysates to determine the key amino acids responsible for their antioxidant properties. 3. Comparing the oxidative stability of different types of oils (e.g., sesame oil, canola oil, corn oil) when stabilized with soy protein hydrolysates under varied frying conditions. 4. Examining the influence of hydrolyzed protein residues on the formation of secondary oxidation products in oils during frying cycles and their impact on food quality and safety. 5. Utilizing bioinformatics tools to predict and select specific peptide sequences from soy protein hydrolysates that exhibit the highest antioxidant capacity and stability-enhancing properties in fried oils..

Table 5: Example of generated response for physics domain. Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10528889/pdf/foods-12-03498.pdf

1. Test-time Adaptation of Discriminative Models via Diffusion Generative Feedback Adapts pre-trained discriminative models to each unlabelled example in the test set using generative feedback from a diffusion model. 2. Adaptive Discriminative Generative Model for Object Tracking Formulates a novel discriminative generative framework that generalizes the conventional Fisher Linear Discriminant algorithm with a generative model and renders a proper probabilistic interpretation. 3. Classification with Hybrid Generative/Discriminative Models Describes a hybrid model in which a high-dimensional subset of the parameters are trained to maximize generative likelihood, and another, small, subset of parameters are discriminatively trained to maximize conditional likelihood. 4. Discriminative Level Set for Contour Tracking Integrates discriminative methods into a level set framework when constructing the level set energy function. 5. ManiFPT Defining and Analyzing Fingerprints of Generative Models Formalizes the definition of artifact and fingerprint in generative models, proposes an algorithm for computing them in practice, and finally study its effectiveness in distinguishing a large array of different generative models. 6. Generative Models for 3D Point Clouds Experiments with transformer encoders, latent-space flow models, and autoregressive decoders to improve the performance of point cloud latent-space generative models. 7. Models and Modeling 8. Do text-free diffusion models learn discriminative visual representations? Explores the possibility of a unified representation learner, a diffusion model, which addresses both generative and discriminative tasks simultaneously. 9. Fine-Tuning Generative Models as an Inference Method for Robotic Tasks Investigates how to quickly adapt the sample generation of neural network models to observations in robotic tasks. 10. Discriminative locally document embedding Learning a smooth affine map by approximation of the probabilistic generative structure of subspace 11. Working with Deep Generative Models and Tabular Data Imputation Provides a fair comparison of proposed methods for imputing missing values in tabular data using deep generative models. 12. Robust Discriminative Principal Component Analysis 13. Generative Second Language Acquisition 14. Nonlinear Models 15. Understanding how Differentially Private Generative Models Spend their Privacy Budget Analyzes how DP generative models distribute privacy budgets across rows and columns of tabular data. 16. Online multiple object tracking by hierarchical association of detection responses Presents a framework for multi-pedestrian tracking using a hierarchical association of detection responses, learning both discriminative and generative appearance models online. 17. Two-Stage Generative Learning Objects 18. Generative design games activity 19. First vs second quantization 20. Non-discrimination Criteria for Generative Language Models Studies how to uncover and quantify the presence of gender biases in generative language models, deriving generative AI analogues of three well-known non-discrimination criteria from classification.

Table 6: Example of background knowledge of https://ieeexplore.ieee.org/document/10191295

Domain	**Best Performing LLM**	**Worst Performing LLM**
Chemistry	Claude-2	Gemini-1.0
Economics	Claude-2	Gemini-1.0
Computer Science	GPT-4 (IAScore), Claude-2 (Distinctness)	Gemini-1.0
Medical	GPT-4	Gemini-1.0
Physics	GPT-4 (IAScore), Claude-2 (Distinctness)	Gemini-1.0

Table 7: Summary of best and worst performing LLMs across domains based on alignment (IAScore) and diversity (Distinctness).

Model	**Novelty**	**Relevance**	**Feasibility**
Claude-2	Moderate	Moderate	High
GPT-4	High	Very High	Very High
Claude-2 (+BG)	Higher	Slightly Lower	Slightly Lower
GPT-4 (+BG)	Highest	Highest	Highest

Table 8: Human evaluation results in Computer Science domain. (+BG) indicates use of background knowledge.