

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

Anonymous ACL submission

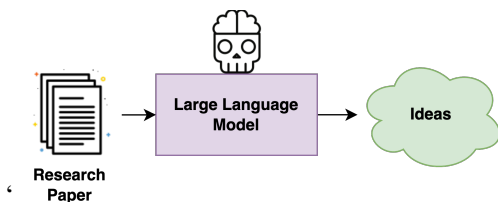


Figure 1: Large language model suggesting future research ideas after reading a research paper

Abstract

“An idea is nothing more nor less than a new combination of old elements” (Young, 2019). 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 explores the capability of LLMs in generating novel research ideas based on information from research papers. We conduct a thorough examination of 4 LLMs in five domains (e.g., Chemistry, Computer, Economics, Medical, and Physics). We found that the future research ideas generated by Claude-2 and GPT-4 are more aligned with the author’s perspective than GPT-3.5 and Gemini. We also found that Claude-2 generates more diverse future research ideas than GPT-4, GPT-3.5, and Gemini 1.0. We further performed a human evaluation of the novelty, relevancy, 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 available¹.

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

¹<https://anonymous.4open.science/r/Future-Idea-Generation-DC42/README.md>

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 be generally understood as a state of focused internally-directed attention involving controlled semantic retrieval (Benedek et al., 2014).

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). 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 spectrum of knowledge, due to their extensive training on vast amounts 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

073 us to ponder:

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

075 Motivated by this, in this paper, we analyze the
076 potential of LLMs in generating future research di-
077 rections/ideas. As LLMs possess knowledge across
078 various domains, we investigate five specific areas,
079 viz. Computer Science, Physics, Chemistry, Eco-
080 nomics, and Medicine. To address this task, we
081 create a dataset of papers published after the year
082 2022 from these five domains. We annotate the
083 papers with future research ideas. To evaluate the
084 novelty and relevance of ideas generated by the
085 LLMs, we propose an Idea Alignment Score (IAS-
086 core). This score reflects how well the generated
087 ideas align with those proposed by the authors. To
088 study the model’s ability to generate diverse ideas,
089 we propose an Idea Distinctness Index. We ana-
090 lyze and discuss the performance and limitations of
091 four LLMs: Gemini (Anil et al., 2023), Claude-2
092 (Anthropic, 2023), GPT-3.5, and GPT-4 (OpenAI,
093 2023). We further conduct a human evaluation of
094 460 generated ideas in computer science to study
095 the novelty, relevance, and feasibility of these re-
096 search ideas. This paper demonstrates that LLMs
097 have the potential to generate relevant, distinct, fea-
098 sible, and novel ideas to some extent.

OUR CONTRIBUTIONS:

- 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 a human evaluation on 460 generated ideas.

099 We hope that this work serves as a foundation
100 for future studies focused on accelerating scien-

tific research by automatically generating research ideas.

2 Related Work

104 Recently, LLMs have shown emergent abilities to
105 perform tasks they were not explicitly trained for
106 (Wei et al., 2022; Bubeck et al., 2023). This in-
107 cludes common sense question answering, code
108 generation, and cross-domain problem solving, en-
109 riching their utility across unforeseen domains
110 (Chen et al., 2021; Sarsa et al., 2022). Their capa-
111 bility extends to advanced scientific domains such
112 as computer science, physics, medicine, and math-
113 ematics (Romera-Paredes et al., 2023; Huang et al.,
114 2023). Technology Semantic Network (TechNet)
115 was proposed to stimulate idea generation in en-
116 gineering design (Sarica et al., 2021). There have
117 been a few works in the discovery of new proteins
118 to accelerate scientific discovery. The prior work
119 reported in (Spangler et al., 2014) involves utilizing
120 published studies to find new protein kinases that
121 phosphorylate the tumor suppressor protein p53.

122 A hypothesis is a hunch, assumption, suspicion,
123 assertion or an idea about a phenomenon, relation-
124 ship or situation, the reality or truth of which you
125 do not know (Kumar, 1996). There have been
126 some works on hypothesis generation. Initial stud-
127 ies on automated hypothesis generation begin by
128 constructing a corpus of distinct concepts. Sub-
129 sequently, they explore the relationships between
130 these concepts using machine learning techniques,
131 such as analyzing the similarities among vectors
132 representing different words (or concepts) (Tshi-
133 toyan et al., 2019), or applying link prediction meth-
134 ods over a graph (where concepts are nodes) (Nad-
135 karni et al., 2021). Recently (Qi et al., 2023; ?)
136 used LLMs and extensive pre-existing knowledge
137 of various scientific fields for hypothesis gener-
138 ation. PaperRobot (Wang et al., 2019) predicts
139 related entities for an input title and writes key
140 elements of a new paper, including the abstract,
141 conclusion, and future work, and predicts a new
142 title.

143 Xu et al. (2023) developed a framework that
144 leverages the concept co-occurrence graphs and
145 a masked language model to explore and verbal-
146 ize academic ideas. Their method involves con-
147 structing evolving concept graphs across various
148 disciplines and utilizing temporal link prediction
149 to identify potential interdisciplinary connections.
150 The framework also incorporates pre-trained lan-

151	guage models to articulate these connections in a		
152	coherent academic context. SciMON (Wang et al.,		
153	2023) showed that LLMs can be guided by seed		
154	terms to generate specific ideas. They applied a		
155	pre-trained sentence classifier to classify sentences		
156	from the title and abstract into categories of Back-		
157	ground, Method, Objective. They considered sen-		
158	tences labeled as problems or motivations as back-		
159	ground, and the remaining were treated as target		
160	output sentences. Additionally, a pre-trained entity		
161	extractor was used to extract salient seed terms.		
162	However, previous works primarily focused on		
163	developing methods (linking and explaining enti-		
164	tities, which may not sufficiently capture the com-		
165	plexity or explain how LLMs can solve real-world		
166	problems) for idea generation, whereas our work		
167	exhaustively focuses on evaluating the capability		
168	of LLMs in generating research ideas. Our goal is		
169	to assess the inherent ability of LLMs to generate		
170	future research ideas/directions.		
171	3 Dataset		
172	Our dataset creation involves three steps: (1)		
173	Dataset Collection, (2) FRI Identification and re-		
174	moval, and (3) FRI generation.		
175	3.1 Dataset Collection		
176	We construct a corpus D from S2ORC collected		
177	100 papers from the domains of Computer Science,		
178	Economics, Physics, Chemistry, Medical from (Lo		
179	et al., 2020). To ensure the quality and relevance		
180	of the data and to utilize the future research ideas		
181	mentioned in a paper, the selected papers must		
182	meet the following requirements: (1) the paper		
183	must contain the full content, and (2) the paper		
184	must include a section on future work.		
185	3.2 FRI Identification and Removal		
186	We first identify and remove any potential research		
187	ideas mentioned in the paper. By doing this, we		
188	ensure that the LLMs have no prior access to these		
189	ideas, which could otherwise affect the objectivity		
190	of the analysis.		
191	3.2.1 Annotation Guidelines		
192	Inspired by Hao et al. (2020), we define a future re-		
193	search idea as a discussion that the authors believe		
194	they will conduct in the future or believe needs to		
195	be investigated in future research. We discuss more		
196	details about the annotation guidelines in Appendix		
197	A.		
	3.2.2 Annotator Training		198
	Given the complexity of the papers and their fre-		199
	quent use of technical terminology, we hired two		200
	doctoral students, each boasting over four years		201
	of experience in scientific research publishing. To		202
	facilitate their training, an expert with more than		203
	ten years of experience in scientific publishing an-		204
	notated 20 random papers from each domain, ad-		205
	hering to our guidelines. After this initial round of		206
	annotation, we reviewed and corrected any misin-		207
	terpretations with the annotators, further refining		208
	their training and enhancing the clarity of our an-		209
	notation guidelines. To assess the effectiveness		210
	of the initial training, we compiled another 20 pa-		211
	pers from each domain. From the second round		212
	onwards, the annotators demonstrated improved		213
	proficiency, accurately identifying at least 95% of		214
	the future research ideas on average.		215
	3.2.3 Annotation Process		216
	We regularly monitored the annotated data, plac-		217
	ing emphasis on identifying and rectifying inconsi-		218
	stencies and cases of confusion. We also imple-		219
	mented an iterative feedback system that contin-		220
	uously aimed to refine and improve the anno-		221
	tation process. In cases of conflict or confusion,		222
	we removed those papers as we wanted only good		223
	quality dataset. Following the annotation phase,		224
	we obtained an average inter-annotator agreement		225
	score of 0.94 using Cohen’s kappa (Cohen, 1960),		226
	indicating a substantial consensus among the anno-		227
	tators.		228
	3.2.4 Annotator’s Pay		229
	We compensated each annotator according to the		230
	standard PhD salaries in India, based on the hours		231
	they worked. The appointment and salaries adhere		232
	to our university’s established practices. Payment		233
	was made per paper since the time required to read		234
	and extract future research ideas from each paper		235
	varies, depending on its complexity, technical ter-		236
	minology, and the annotator’s familiarity with the		237
	subject. Thus, paying based on time spent could		238
	have potentially compromised the quality of the		239
	annotations. To maintain accuracy and prevent fa-		240
	tigue, we imposed a daily limit of six hours for		241
	annotators.		242

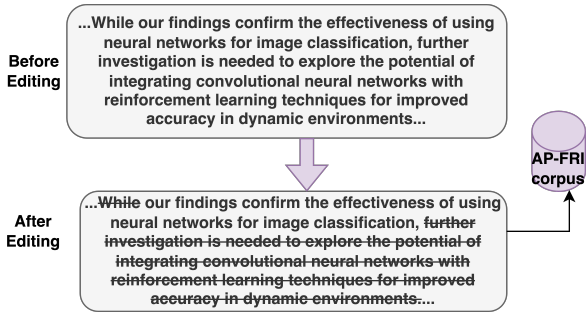


Figure 2: An example for FRI editing; Here the strike through text is removed from the paper text

3.2.5 Future Work Removal

We observed two types of future research ideas (FRIs) ² (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 added 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.

3.3 FRI Generation using LLM

We investigate various prompts and utilize the following prompts to generate FRIs for papers.

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

[paper text]

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.

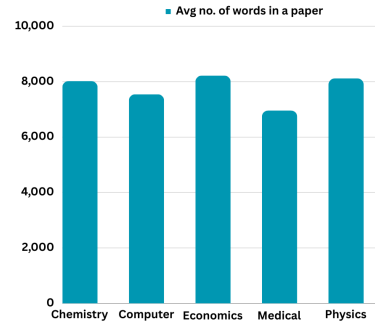


Figure 3: Domain vs Avg. number of words in a paper w/o FWK

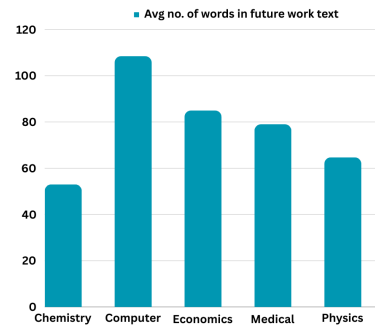


Figure 4: Domain vs Avg. number of words in FWK

3.4 Data Statistics

Figure 3 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 4 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.

4 Experiments

4.1 Challenges

To accurately assess the novelty, relevance, and applicability of ideas generated by LLMs, evaluators

²In this paper, we use the terms ‘ideas,’ ‘research ideas,’ ‘future research ideas,’ and ‘FRI’ interchangeably to frequently refer to future research ideas.

289 must possess a high level of expertise in the specific
 290 domain and a deep understanding of the research
 291 topic to fully grasp the context. Additionally, they
 292 need knowledge of related literature to evaluate the
 293 ideas’ future potential and the broader implications
 294 of their implementation.

295 4.2 Idea Alignment Score (IAScore)

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

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

317 Initially, we compute the average alignment
 318 score AvgScore_j for each paper’s ideas. The
 319 IdeaMatcher model measures the alignment be-
 320 tween the paper’s author *Future Research Ideas*
 321 (AP-FRI_j) and its each generated idea I_{ij} . The
 322 subscript i indexes the i -th idea within the j -th pa-
 323 per, where N_j represents the total number of ideas
 324 proposed in that paper.

$$325 \text{AvgScore}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \text{IM}(\text{AP-FRI}_j, I_{ij}) \quad (1)$$

326 "Here, we refer to IM as 'IdeaMatcher'.

327 LLMs may generate new ideas that even the au-
 328 thor may not have thought of. They can also gen-
 329 erate additional future ideas, which may or may
 330 not be useful. Our goal is for this score is that the
 331 LLMs must have generated at least the author’s
 332 proposed potential future ideas. Therefore, in our
 333 formula of AvgScore_j , the sum of the alignment
 334 scores for a paper’s ideas is divided by the total

number of the author’s proposed ideas, N_j , to nor-
 malize the score.

$$337 \text{IAScore}_{\text{domain}, M} = \frac{1}{P} \sum_{j=1}^P \text{AvgScore}_j \quad (2)$$

338 Subsequently, we aggregate the individual paper
 339 scores to calculate the domain-wise IAScore. This
 340 aggregation, presented in Equation 2, averages the
 341 AvgScore_j values across all P papers within the do-
 342 main. Higher the value of $\text{IAScore}_{\text{domain}}$ signifies
 343 the more alignment of the generated ideas with au-
 344 thor’s perspective of all papers generated by model
 345 M .

346 4.2.1 IdeaMatcher

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

371 We also evaluated the idea-matching capability
 372 of BERTScore (Zhang* et al., 2020), as it utilizes
 373 BERT embeddings for comparison. We discuss the
 374 details in Appendix F. We found that BERTScore
 375 performed better than the entailment technique, re-
 376 sulting in an accuracy of 75.4%. We also tried GPT
 377 by prompting it with various questions and found
 378 that it resulted in 91.8% accuracy when prompted
 379 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 (Natural Language Inference) and similarity measure such as BERTScore. Therefore, we chose GPT for this task³.

4.3 Idea Distinctness 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.

Given a set of generated ideas $I = \{id_1, id_2, \dots, id_n\}$, representing individual ideas, we first encode each idea into a high-dimensional vector space using a pre-trained BERT model (Devlin et al., 2019)⁴ **BERT** : $id_i \mapsto \mathbf{v}_i$, where $\mathbf{v}_i \in R^d$ 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:

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

This measure provides a single scalar value D_I that quantifies the average diverseness of ideas

³We used the OpenAI model GPT-3.5-turbo-0125 using OpenAI API

⁴bert-base-uncased

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 the mean distinctness for that paper. Let $P = \{p_1, p_2, \dots, p_m\}$ 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:

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

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.

4.4 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 and followed the 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-4⁵ models due to their better IAScore and Idea Distinction index. We adopt this approach to avoid author exhaustion and to get an accurate evaluation. 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

⁵We used gpt-4-turbo using OpenAI API for the generation

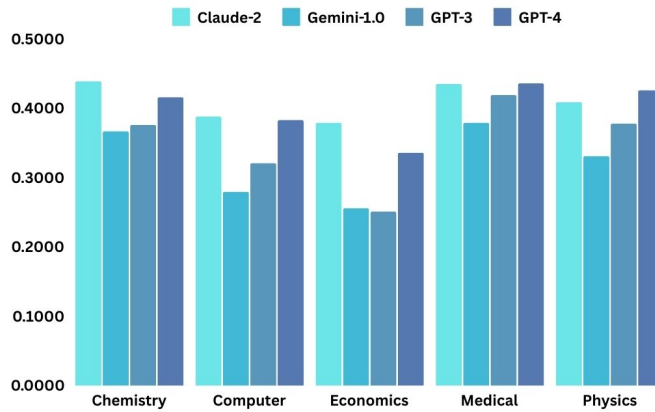


Figure 5: IAScore for each domain and model; a higher value indicates better alignment with the author.

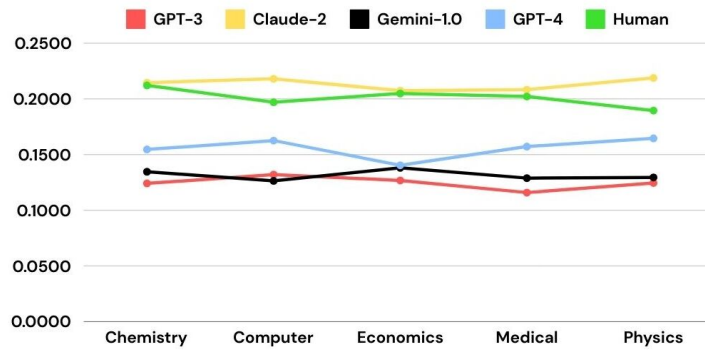


Figure 6: Idea distinctness index analysis; Here human is the authors of the paper

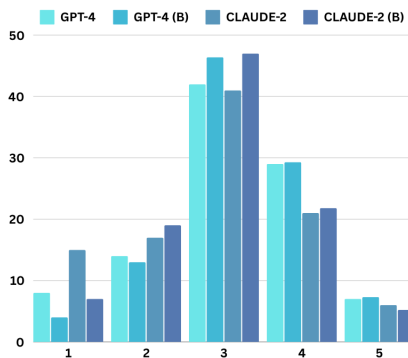


Figure 7: Novelty human evaluation for Computer Science domain; Here (B) means with additional background knowledge

More details about the human evaluation are mentioned in the Appendix B.

5 Results and Discussion

5.1 Alignment Results

Figure 5 provides a comparative overview of the IAScore for four language models⁶ Claude-2, Gemini-1.0, GPT-3, and GPT-4 across five academic domains: Chemistry, Computer Science, Economics, Medical, and Physics.

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 and Gemini have lower scores than both GPT-4 and Claude in ev-

⁶We set maximum token length to 512, and temperature=0 for each models

460 or too vague to be actionable? (Not Possi-
461 ble/Possible)

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

464
465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

ery domain. GPT-3 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 alignment 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). 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 6. The line graph depicts the variation of distinctness between the generated ideas and the human-written ideas (AP-FRIs). GPT-3 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, except in the Computer domain.

The generated ideas of GPT-4 are more distinct than those of Gemini and GPT-3 (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 460 generated ideas for 46 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, thereby confirming the high quality of the annotations of generated research ideas.

Novelty: Figure 7 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. 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.

Relevance and Feasibility: After human evaluation, we found that 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. 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.

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. 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 distinct research ideas than Gemini and GPT 3.5. GPT-4 and Claude aligns 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 evaluations 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 work, we plan to investigate how integrating knowledge from multiple papers can enhance the novelty of ideas generated and prevent the generation of generic and existing ideas.

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 100 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. 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.

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.

Frequently Asked Questions (FAQs)

- **How does our work differ from Scimon?**

⇒ Our paper is fundamentally different from the Scimon paper. We would like to highlight a few major differences. While the focus of Scimon is on developing a framework that generates novel scientific ideas, we clarify that our focus is not on generating ideas but on evaluating the capability of LLMs to generate future research ideas/works. We proposed the novel Idea Alignment Score (IAScore) and the Idea Distinctness Index. Unlike Scimon, we approached authors who are knowledgeable about their paper topics and the broader literature (see Section 4.4). Scimon used only GPT for comparison, while we used GPT-4, GPT-3.5, Claude, and Gemini models. Unlike Scimon, we provide the full paper as input. Scimon used the proposed idea written in the abstract as the target, while we used the future work section written in the full paper as our target. Additionally, they utilized a classifier for this purpose, whereas we employed human evaluators, resulting in fewer chances of error and better evaluation results. Our findings are completely different from those of Scimon. We created a novel annotated dataset for these experiments. While Scimon only experimented with computer science papers from the ACL Anthology, we expanded our experiments to five different domains. Scimon generated ideas guided by seed terms to generate specific ideas. Nonetheless, our goal here is to assess the inherent ability of LLMs to generate future work independently. Introducing external aids or additional context would shift the focus from evaluating the LLM’s standalone capabilities to assessing its performance under enhanced conditions. Such an approach would not align with our objective, which is to understand and measure the raw, unaided generative power of LLMs.

- **Does incorporating extra contextual information alongside individual papers prove counter-productive?**

⇒ A paper encompasses not only its contributions, findings, and methodology, but also includes the related work and introduction sections, which contain significant background information. It is likely that the major recent related papers pertinent to the current work have already been mentioned. Additionally, LLMs possess general knowledge about the many older papers and the paper itself contains some of the most important related papers. However, we also conducted an experiment to understand the effect of adding additional information (using the RAG framework). We discuss the results and details in Appendix E of the paper. Overall, we observed that incorporating additional background knowledge can somewhat help prevent the generation of non-novel or generic ideas. However, further research is needed to enhance the ability of LLMs to generate more novel ideas.

A 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."* Here, how novel research involving magnetogel nanohybrid material will help in future application is written.

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.

B Human Annotation

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 an title and abstract as input to see if LLMs can still retain open-ended generation capabilities. We discovered few cases where GPT-4 still generated novel ideas, such as for a paper (Kumar et al., 2023b) it

generated: "Incorporate explainable AI methods to provide transparency into how the AI model makes its predictions, thereby 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 result (limited information was given as input to the paper). Overall, we found that giving LLMs can still retain open-ended generation because it has past knowledge. But it may not generate many good ideas since it doesn't 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 8. 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.

E Effect of Adding Additional Background Knowledge

We designed our framework based on the Retrieval-Augmented Generation (RAG) model (Lewis et al.,

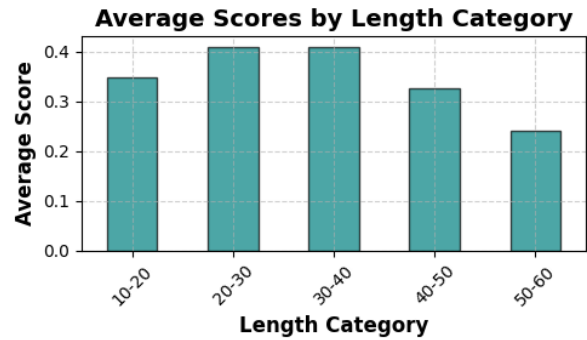


Figure 8: Effect of length on IAScore

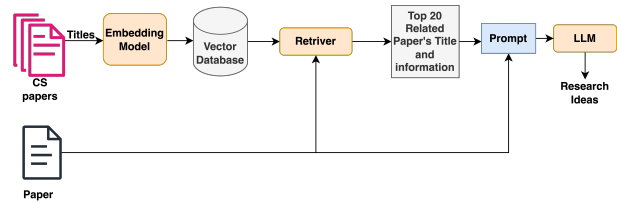


Figure 9: RAG pipeline framework for infusing more background knowledge with the LLMs

2020) to integrate background knowledge into LLMs, as illustrated in Figure 9.

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 this paper title embedding 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 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 LLM⁷ understood its role in extracting 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. After reading the following paper and background knowledge, brainstorm to generate potential top 5 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.

⁷We employed Gemini-Pro model for this task

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 4.4. The results are shown in Figure 7. 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 relevancy 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 threshold⁸ 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 human evaluation using three expert annotators, each with over 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 natural language processing (NLP) models and techniques for contradiction detection.* are generated. These statements are true; however, they are very generic and are common.
- **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., 2023a) 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 2, 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.

References

Gemini Team Google 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 Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anais White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, L. Sifre, Marcello Maggioni, Fred Alcober, Daniel H Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adria Puigdomenech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela de Castro Surita, Jane Labanowski, Abhishek Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravichandra Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin,

⁸We set the threshold 0.68 empirically

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	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Future Work 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Future Work 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Future Work 4	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Future Work 5	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 11: Human Feedback form for novelty

	Relevant	Possible
Future Work 1	<input type="checkbox"/>	<input type="checkbox"/>
Future Work 2	<input type="checkbox"/>	<input type="checkbox"/>
Future Work 3	<input type="checkbox"/>	<input type="checkbox"/>
Future Work 4	<input type="checkbox"/>	<input type="checkbox"/>
Future Work 5	<input type="checkbox"/>	<input type="checkbox"/>

Figure 12: Human Feedback form for relevance, feasibility

998	Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe C. Ashwood,	ter, Priyanka Agrawal, Alex Castro-Ros, George	1035
999	Anton Briukhov, Albert Webson, Sanjay Ganapathy,	van den Driessche, Tao Wang, Fan Yang, Shuo	1036
1000	Smit Sanghavi, Ajay Kannan, Ming-Wei Chang,	yiin Chang, Paul Komarek, Ross McIlroy, Mario	1037
1001	Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur	Luvci'c, Guodong Zhang, Wael Farhan, Michael	1038
1002	Bapna, Matthew Aitchison, Pedram Pejman, Henryk	Sharman, Paul Natsev, Paul Michel, Yong Cheng,	1039
1003	Michalewski, Tianhe Yu, Cindy Wang, J Christo-	Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shak-	1040
1004	pher Love, Junwhan Ahn, Dawn Bloxwich, Kehang	eri, Christina Butterfield, Justin Chung, Paul Kishan	1041
1005	Han, Peter Humphreys, Thibault Sellam, James Brad-	Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar	1042
1006	bury, Varun Godbole, Sina Samangooui, Bogdan	Soparkar, Karel Lenc, Timothy Chung, Aedan Pope,	1043
1007	Damoc, Alex Kaskasoli, S'ebastien M. R. Arnold,	Lorenzo Maggiore, Jackie Kay, Priya Jhakra, Shibo	1044
1008	Vijay Vasudevan, Shubham Agrawal, Jason Riesa,	Wang, Joshua Maynez, Mary Phuong, Taylor Tobin,	1045
1009	Dmitry Lepikhin, Richard Tanburn, Srivatsan Sriniv-	Andrea Tacchetti, Maja Trebacz, Kevin Robinson,	1046
1010	asan, Hyeontaek Lim, Sarah Hodgkinson, Pranav	Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan	1047
1011	Shyam, Johan Ferret, Steven Hand, Ankush Garg,	Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone,	1048
1012	Tom Le Paine, Jian Li, Yujia Li, Minh Giang,	Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gri-	1049
1013	Alexander Neitz, Zaheer Abbas, Sarah York, Machel	bovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music	1050
1014	Reid, Elizabeth Cole, Aakanksha Chowdhery, Di-	Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers,	1051
1015	panjan Das, Dominika Rogozi'nska, Vitaly Niko-	Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed,	1052
1016	laev, Pablo Sprechmann, Zachary Nado, Lukás Zilka,	Tianqi Liu, Richard Powell, Vijay Bolina, Mariko	1053
1017	Flavien Prost, Luheng He, Marianne Monteiro, Gau-	inuma, Polina Zablotskaia, James Besley, Da-Woon	1054
1018	rav Mishra, Christopher A. Welty, Joshua Newlan,	Chung, Timothy Dozat, Ramona Comanescu, Xi-	1055
1019	Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu,	ance Si, Jeremy Greer, Guolong Su, Martin Polacek,	1056
1020	Raoul de Liedekerke, Justin Gilmer, Carl Saroufim,	Raphael Lopez Kaufman, Simon Tokumine, Hex-	1057
1021	Shruti Rijhwani, Shaobo Hou, Disha Shrivastava,	iang Hu, Elena Buchatskaya, Yingjie Miao, Mo-	1058
1022	Anirudh Baddepudi, Alex Goldin, Adnan Ozturel,	hamed Elhawaty, Aditya Siddhant, Nenad Tomaev,	1059
1023	Albin Cassirer, Yunhan Xu, Daniel Sohn, Deven-	Jinwei Xing, Christina Greer, Helen Miller, Shereen	1060
1024	dra Singh Sachan, Reinald Kim Amplayo, Craig	Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Ange-	1061
1025	Swanson, Dessie Petrova, Shashi Narayan, Arthur	los Filos, Milos Besta, Rory Blevins, Ted Klimenko,	1062
1026	Guez, Siddhartha Brahma, Jessica Landon, Miteyan	Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Os-	1063
1027	Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wen-	car Chang, Mantas Pajarskas, Carrie Muir, Vered	1064
1028	hao Jia, Matthew Rahtz, Mai Gim'enez, Legg Yeung,	Cohen, Charline Le Lan, Krishna Haridasan, Amit	1065
1029	Hanzhao Lin, James Keeling, Petko Georgiev, Di-	Marathe, Steven Hansen, Sholto Douglas, Rajku-	1066
1030	ana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro,	mar Samuel, Mingqiu Wang, Sophia Austin, Chang	1067
1031	Kiran Vodrahalli, James Qin, Zeynep Cankara, Ab-	Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo,	1068
1032	hanshu Sharma, Nicholas Fernando, Will Hawkins,	Lars Lowe Sjosund, S'ebastien Cevey, Zach Gle-	1069
1033	Behnam Neyshabur, Solomon Kim, Adrian Hut-	icher, Thi Avrahami, Anudhyan Boral, Hansa Sriniv-	1070
1034		asan, Vittorio Selo, Rhys May, Konstantinos Aiso-	1071

Model	Response
Claude-2	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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. Simulate the impact on infection rates, healthcare burden, and emergence of new variants. 5. Conduct intervention studies to test the effectiveness of community-based participatory approaches in increasing COVID-19 vaccine acceptance among disadvantaged populations. Evaluate culturally-tailored messaging and outreach.
Gemini-1.0	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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. This research could help identify disparities in vaccine distribution and uptake among different population groups. 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. This comparative analysis could provide valuable insights for designing targeted interventions to address global vaccine inequality. 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. By analyzing trends and patterns, researchers could gain a deeper understanding of the evolving dynamics of vaccine distribution and uptake. 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. This study could help identify barriers to vaccination and inform targeted communication strategies to address vaccine hesitancy. 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/>

Model	Response
Claude-2	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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 include 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 1. DComparative 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	<ol style="list-style-type: none"> 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>

1072	pos, L'eonard Hussenot, Livio Baldini Soares, Kate	Romina Datta, Adam Sadovsky, Oskar Bunyan, Do-	1106
1073	Baumli, Michael B. Chang, Adrià Recasens, Ben-	minik Rabiej, Shimu Wu, John Zhang, Gautam Va-	1107
1074	jamin Caine, Alexander Pritzel, Filip Pavetic, Fabio	sudevan, Edouard Leurent, Mahmoud Alnahlawi,	1108
1075	Pardo, Anita Gergely, Justin Frye, Vinay Venkatesh	Ionut-Razvan Georgescu, Nan Wei, Ivy Zheng, Betty	1109
1076	Ramasesh, Dan Horgan, Kartikeya Badola, Nora	Chan, Pam G Rabinovitch, Piotr Stańczyk, Ye Zhang,	1110
1077	Kassner, Subhrajit Roy, Ethan Dyer, V'ictor Cam-	David Steiner, Subhajit Naskar, Michael Azzam,	1111
1078	pos, Alex Tomala, Yunhao Tang, Dalia El Badawy,	Matthew Johnson, Adam Paszke, Chung-Cheng Chiu,	1112
1079	Elspeeth White, Basil Mustafa, Oran Lang, Abhishek	Jaume Sanchez Elias, Afroz Mohiuddin, Faizan	1113
1080	Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles,	Muhammad, Jin Miao, Andrew Lee, Nino Vieil-	1114
1081	Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng,	lard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Ji-	1115
1082	Wojciech Stokowiec, Ce Zheng, Phoebe Thacker,	ageng Zhang, Jeff Stanway, Drew Garmon, Abhi-	1116
1083	cCauglar Unlu, Zhishuai Zhang, Mohammad Saleh,	jit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar,	1117
1084	James Svensson, Maxwell L. Bileschi, Piyush Patil,	Luwei Zhou, Jonathan Evens, William Isaac, Zhe	1118
1085	Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi	Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu,	1119
1086	Vezer, Marco Selvi, Toby Shevlane, Mikel Ro-	Chris Filo Gorgolewski, Peter Grabowski, Yu Mao,	1120
1087	driguez, Tom Kwiatkowski, Samira Daruki, Keran	Alberto Magni, Kaisheng Yao, Javier Snaider, Nor-	1121
1088	Rong, Allan Dafoe, Nicholas FitzGerald, Keren	man Casagrande, Paul Suganthan, Evan Palmer, Ge-	1122
1089	Gu-Lemberg, Mina Khan, Lisa Anne Hendricks,	offrey Irving, Edward Loper, Manaal Faruqi, Isha	1123
1090	Marie Pellat, Vladimir Feinberg, James Cobon-	Arkatkar, Nanxin Chen, Izhak Shafran, Michael	1124
1091	Kerr, Tara N. Sainath, Maribeth Rauh, Sayed Hadi	Fink, Alfonso Castano, Irene Giannoumis, Wooyeol	1125
1092	Hashemi, Richard Ives, Yana Hasson, YaGuang	Kim, Mikolaj Rybi'nski, Ashwin Sreevatsa, Jen-	1126
1093	Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou,	nifer Prendki, David G. Soergel, Adrian Goedeck-	1127
1094	Qingze Wang, Thibault Sottiaux, Michela Paganini,	emeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba,	1128
1095	Jean-Baptiste Lepiau, Alexandre Moufarek, Samer	Jeremy Wiesner, Diana Gage Wright, Yawen Wei,	1129
1096	Hassan, Kaushik Shivakumar, Joost R. van Amers-	Harsha Vashisht, Yana Kulizhskaya, Jay Hoover,	1130
1097	foort, Amol Mandhane, Pratik M. Joshi, Anirudh	Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu,	1131
1098	Goyal, Matthew Tung, Andy Brock, Hannah Sheah-	Kevin Ramirez, A. Ya. Khorlin, Albert Cui, Tian	1132
1099	han, Vedant Misra, Cheng Li, Nemanja Raki'cevi'c,	Lin, Marin Georgiev, Marcus Wu, Ricardo Aguilar,	1133
1100	Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk	Keith Pallo, Abhishek Chakladar, Alena Repina,	1134
1101	Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew	Xihui Wu, Tom van der Weide, Priya Ponnappalli,	1135
1102	Lamm, Nicola De Cao, Charlie Chen, Gamaleldin	Caroline Kaplan, Jiř'ima, Shuangfeng Li, Olivier	1136
1103	Elsayed, Ed Huai hsin Chi, Mahdis Mahdieh, Ian	Dousse, Jeff Piper, Nathan Ie, Minnie Lui, Rama Ku-	1137
1104	Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane,	mar Pasumarthi, Nathan Lintz, Anitha Vijayakumar,	1138
1105	Dylan Scardinaro, Rishub Jain, Jonathan Uesato,	Lam Nguyen Thiet, Daniel Andor, Pedro Valen-	1139

1140	zuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc V. Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel F. Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan S. Barnase, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel Jaymin Mankowitz, Oleksandr Polozov, Nate Kushman, Victoria Krakovna, Sande Minnich Brown, Mohammad-Hossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John F. J. Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Martin Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Kalpesh Krishna, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, P. Narayana, Jing Li, Sabaer Fatehi, J. Michael Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Am'elie H'eliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Daniel Niels Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Milylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Poder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Dehghani Tafti, Raghavendra Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Riviere, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clément Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Østbø Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Víctor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Wei Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durdan, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria E. Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake A. Hechtman, Parker Schuh, Milad Nasr, Mianna Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models . <i>ArXiv</i> , abs/2312.11805.	1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256
1190	Anthropic. 2023. Model card for claude 2 .	1257
1191	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 . <i>NeuroImage</i> , 88:125–133.	1258 1259 1260 1261 1262
1192	Faeze Brahman, Chandra Bhagavatula, Valentina Pyatkin, Jena D Hwang, Xiang Lorraine Li, Hirona J	1263 1264

1265	Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, and Yejin Choi. 2023. Making small language models better procedural knowledge models for (counterfactual) planning . <i>arXiv preprint arXiv:2305.19472</i> .	1322
1266		1323
1267		1324
1268		
1269		
1270	Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, 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 . <i>CoRR</i> , abs/2303.12712.	
1271		
1272		
1273		
1274		
1275		
1276		
1277	Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models . <i>ArXiv</i> , abs/2202.07646.	
1278		
1279		
1280		
1281	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 . <i>CoRR</i> , abs/2107.03374.	
1282		
1283		
1284		
1285		
1286		
1287		
1288		
1289		
1290		
1291		
1292		
1293		
1294		
1295		
1296		
1297		
1298		
1299		
1300		
1301		
1302	Jacob Cohen. 1960. A coefficient of agreement for nominal scales. <i>Educational and Psychological Measurement</i> , 20:37 – 46.	
1303		
1304		
1305	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding . In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)</i> , pages 4171–4186. Association for Computational Linguistics.	
1306		
1307		
1308		
1309		
1310		
1311		
1312		
1313		
1314		
1315	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 <i>JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020</i> , pages 261–269. ACM.	
1316		
1317		
1318		
1319		
1320		
1321		
	Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Benchmarking large language models as ai research agents . <i>ArXiv</i> , abs/2310.03302.	1325
		1326
		1327
		1328
		1329
		1330
		1331
		1332
		1333
		1334
		1335
		1336
		1337
		1338
		1339
		1340
		1341
		1342
	Laura J. Kornish and Jeremy Hutchison-Krupat. 2017. Research on idea generation and selection: Implications for management of technology . <i>Production and Operations Management</i> , 26:633–651.	1343
		1344
		1345
		1346
	Thomas S. Kuhn and David Hawkins. 1963. The structure of scientific revolutions . <i>American Journal of Physics</i> , 31:554–555.	1347
		1348
		1349
	R. Kumar. 1996. Research Methodology: A Step-by-Step Guide for Beginners . Longman Australia.	1350
		1351
	Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023a. When reviewers lock horn: Finding disagreement in scientific peer reviews . <i>arXiv preprint arXiv:2310.18685</i> .	1352
		1353
		1354
		1355
	Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023b. When reviewers lock horns: Finding disagreements in scientific peer reviews . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 16693–16704, Singapore. Association for Computational Linguistics.	1356
		1357
		1358
		1359
		1360
		1361
		1362
	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 <i>Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual</i> .	1363
		1364
		1365
		1366
		1367
		1368
		1369
		1370
		1371
	Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2015. A diversity-promoting objective function for neural conversation models . <i>ArXiv</i> , abs/1510.03055.	1372
		1373
		1374
		1375
	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.	1376
		1377
		1378

1379	Roberta: A robustly optimized bert pretraining approach. <i>arXiv preprint arXiv:1907.11692</i> .	<i>Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014</i> , pages 1877–1886. ACM.	1434
1380			1435
1381	Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 4969–4983, Online. Association for Computational Linguistics.	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 . <i>Nat.</i> , 571(7763):95–98.	1437
1382			1438
1383			1439
1384			1440
1385			1441
1386			1442
1387	J.J. Louviere, T.N. Flynn, and A.A.J. Marley. 2015. <i>Best-Worst Scaling: Theory, Methods and Applications</i> . Cambridge books online. Cambridge University Press.	Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Scimon: Scientific inspiration machines optimized for novelty . <i>arXiv preprint arXiv:2305.14259</i> .	1443
1388			1444
1389			1445
1390			
1391	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 . <i>ArXiv</i> , abs/2106.09700.	Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 1980–1991, Florence, Italy. Association for Computational Linguistics.	1446
1392			1447
1393			1448
1394			1449
1395			1450
1396	OpenAI. 2023. GPT-4 technical report . <i>CoRR</i> , abs/2303.08774.	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 <i>The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022</i> . OpenReview.net.	1451
1397			1452
1398	Plato, G.R.F. Ferrari, and T. Griffith. 2000. <i>Plato: 'The Republic'</i> . Cambridge Texts in the History of Political Thought. Cambridge University Press.	Yi Xu, Shuqian Sheng, Bo Xue, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2023. Exploring and verbalizing academic ideas by concept co-occurrence . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13001–13027, Toronto, Canada. Association for Computational Linguistics.	1453
1399			1454
1400			1455
1401	Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Si-hang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers . <i>ArXiv</i> , abs/2311.05965.	Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models . <i>arXiv preprint arXiv:2210.03629</i> .	1456
1402			1457
1403			1458
1404			1459
1405	Bernardino Romera-Paredes, Mohammadamin Berekatain, 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 . <i>Nature</i> , 625:468 – 475.	J.W. Young. 2019. <i>A Technique for Producing Ideas</i> . Independently Published.	1460
1406			1461
1407			1462
1408			1463
1409			1464
1410			1465
1411			1466
1412			1467
1413			1468
1414	Serhad Sarica, Binyang Song, Jianxi Luo, and Kristin L. Wood. 2021. Idea generation with technology semantic network . <i>Artificial Intelligence for Engineering Design, Analysis and Manufacturing</i> , 35:265 – 283.	Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert . In <i>International Conference on Learning Representations</i> .	1469
1415			1470
1416			1471
1417			1472
1418	Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic generation of programming exercises and code explanations using large language models . In <i>ICER 2022: ACM Conference on International Computing Education Research, Lugano and Virtual Event, Switzerland, August 7 - 11, 2022, Volume 1</i> , pages 27–43. ACM.	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 . <i>ArXiv</i> , abs/2303.18223.	1473
1419			1474
1420			1475
1421			1476
1422			1477
1423			1478
1424			1479
1425	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 <i>The 20th ACM SIGKDD International</i>		1480
1426			1481
1427			1482
1428			1483
1429			
1430			
1431			
1432			
1433			