IdeaBench: Benchmarking Large Language Models for Research Idea Generation

Sikun Guo
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
qkm6sq@virginia.edu

Amir Hassan Shariatmadari
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
ahs5ce@virginia.edu

Guangzhi Xiong
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
hhu4zu@virginia.edu

Albert Huang
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
kfa7fg@virginia.edu

Myles Kim
School of Medicine
University of Virginia
Charlottesville, Virginia, USA
mbt8hz@virginia.edu

Corey M. Williams
School of Medicine
University of Virginia
Charlottesville, Virginia, USA
cmw6pa@virginia.edu

Stefan Bekiranov
School of Medicine
University of Virginia
Charlottesville, Virginia, USA
sb3de@virginia.edu

Aidong Zhang
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
aidong@virginia.edu

Abstract

Large Language Models (LLMs) have revolutionized interactions between human and artificial intelligence (AI) systems, demonstrating state-of-the-art performance across various domains, including scientific discovery and hypothesis generation. However, the absence of a comprehensive and systematic evaluation framework for LLM-driven research idea generation hinders a rigorous understanding of their strengths and limitations. To address this gap, we propose IdeaBench, a benchmark system that provides a structured dataset and evaluation framework for standardizing the assessment of research idea generation by LLMs. Our dataset comprises titles and abstracts from 2,374 influential papers across eight research domains, along with their 29,408 referenced works, creating a context-rich environment that mirrors human researchers' ideation processes. By profiling LLMs as domain-specific researchers and grounding them in similar contextual constraints, we directly leverage the models' knowledge learned from the pre-training stage to generate new research ideas. To systematically evaluate LLMs' research ideation capability and approximate human assessment, we propose a reference-based metric that aligns with human judgment to quantify idea quality with the assistance of LLMs. Through this evaluation, we find that while LLMs excel at generating novel ideas, they may struggle with generating feasible ideas. IdeaBench serves as a critical resource for benchmarking and comparing LLMs, ultimately advancing research on AI's role in automating scientific discovery. We provide our code and dataset in the following repository: https://github.com/amir-hassan25/IdeaBench.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737419

CCS Concepts

• **Applied computing** → *Genetics*; *Sociology*; **Engineering**; *Chemistry*; **Life and medical sciences**.

Keywords

Hypothesis Generation, Large Language Models, AI for Science

ACM Reference Format:
Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. 2025. IdeaBench: Benchmarking Large Language Models for Research Idea Generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3737419

1 Introduction

Large Language Models (LLMs) have emerged as transformative tools in artificial intelligence (AI), demonstrating remarkable capabilities across diverse domains, including natural language understanding, content generation, and complex problem-solving. Advances in architectures such as GPT-4 [17] and the LLaMA series [24] have further enhanced these models, enabling them to generalize across a wide range of tasks through in-context learning. Unlike traditional machine learning models that require extensive task-specific training, LLMs can adapt dynamically to new problems by conditioning on contextual information provided within a prompt [4]. This adaptability has led to state-of-the-art performance in numerous applications, from conversational AI to creative writing, translation, and even the synthesis of scientific knowledge [5].

Beyond consumer applications, the potential of LLMs to facilitate scientific discovery has sparked increasing interest in academia. Researchers have begun to explore their use in hypothesis generation, literature review automation, and research idea formulation [1, 2, 9, 10, 19, 27–29, 33]. These studies highlight the promise of LLMs as intelligent assistants capable of accelerating scientific ideation. However, despite the growing enthusiasm, a major challenge remains: the lack of a systematic framework to evaluate the quality of research ideas generated by LLMs. Without standardized benchmarks and evaluation metrics, it is difficult to rigorously assess their generative strengths and limitations, hindering progress in understanding their role in scientific discovery.

To address this gap, we propose *IdeaBench*, a benchmark system designed to systematically evaluate the capabilities of LLMs in research idea generation. *IdeaBench* is built on a structured evaluation framework that mirrors the human research ideation process. The intuition behind our framework is grounded in the typical workflow that human researchers follow when generating new scientific ideas:

1. Identifying a specific research topic of interest.
2. Reviewing relevant literature, particularly recent findings and methodologies.
3. Recognizing gaps in existing knowledge or methodologies.
4. Proposing new research ideas to address these gaps.

To mimic this process, we construct a benchmark dataset that consists of 2,374 influential research papers across eight domains, each paired with their referenced work. These referenced papers serve as the contextual basis for idea generation, allowing LLMs to simulate the research ideation process in a structured manner. By profiling LLMs as domain-specific researchers in our prompts and grounding them within relevant contextual constraints, we aim to probe their knowledge learned from the pre-training stage to generate meaningful and insightful research ideas.

To assess the quality of generated research ideas, we introduce a two-stage evaluation framework. First, we employ a ranking mechanism that uses LLMs to assess ideas based on user-defined quality indicators such as novelty and feasibility. This scalable approach allows personalization and adaptation to different research domains. Second, we introduce a reference-based scoring metric, termed "Insight Score," which quantifies the quality of the generated ideas with user-specified quality indicators. This evaluation framework provides a rigorous and adaptable means of assessing LLM-generated research ideas that align with human judgment, facilitating more informed comparisons between different models.

Through extensive experimentation with recent high-capacity LLMs, we find that while these models are adept at generating novel research ideas, they often struggle to ensure feasibility. This observation underscores the need for further refinement in the way LLMs conceptualize and validate scientific hypotheses. We hope that *IdeaBench* will serve as a critical resource for the research community, enabling systematic benchmarking and inspiring future advancements in AI-driven scientific ideation.

To summarize, our main contributions are as follows:

- We construct *IdeaBench*, a benchmark dataset comprising 2,374 influential research articles in eight research domains along with their 29,408 reference articles, providing a structured context to evaluate LLM's capabilities in the generation of research ideas.
- We propose an evaluation framework that introduces a scalable and versatile metric, the "Insight Score," which enables the quantification of various quality indicators such as novelty and feasibility, as defined by human researchers.
- We conduct extensive experiments to evaluate the ability of multiple LLMs to generate research ideas, providing empirical insights that while LLMs excel at generating novel ideas, there is still room for improvement in generating feasible ones.

By providing a systematic and rigorous evaluation framework, our aim is to contribute to the evolving role of LLMs in scientific discovery, fostering a deeper understanding of their potential and limitations in automating research ideation.

2 Related Works

2.1 Machine Learning for Hypothesis Generation

Most existing research on hypothesis generation has focused on literature-based discovery (LBD), which predicts relationships between discrete scientific concepts [26]. While effective, LBD approaches assume all concepts are known and primarily identify implicit connections from literature snapshots. Recent studies have explored large language models (LLMs) for hypothesis generation [22]. SciMON [27] uses retrieval-based LLMs to ground hypotheses in past literature, optimizing for novelty. MOOSE [29] leverages multi-level self-feedback to facilitate hypothesis discovery, particularly in social sciences, while ResearchAgent [2] integrates entity-centric knowledge graphs to refine research problems and methodologies. Other approaches enhance LLM-based hypothesis generation using prompting strategies [33], search mechanisms [11], and reinforcement learning for controllable generation [14]. Beyond idea generation, Agent Laboratory [21] introduces an LLM-driven research assistant capable of autonomously conducting literature review, experimentation, and report writing, highlighting the potential of automating the research process.

2.2 Evaluation for Open-ended Text Generation

Although human judgment is still considered the golden standard for evaluating open-ended text generation, the Natural Language Processing community has tried to develop different approaches to approximate human evaluation in a scalable way. Traditional metrics like BLEU [18] and ROUGE [15] measure the lexical overlap between model generated content and ground-truth reference. Later on, several efforts use pre-trained language models to measure distributional similarity [31, 32] or token probabilities [23, 30]. With the increasing popularity and impressive performance of Large Language Models, recent endeavors employ LLMs as autoraters for open-ended text generation [3, 5–7, 16, 25], the effectiveness of using LLMs as autoraters is often reflected by its correlation with human-ratings, making autoraters a promising alternative to human evaluators for large-scale evaluation.

3 Benchmark Design

To establish a systematic framework for evaluating LLM-driven research ideation, we design *IdeaBench* around three core components: (1) a structured dataset capturing the relationship between published research ideas and their supporting literature, (2) a standardized research idea generation protocol that emulates human scientific ideation, and (3) a robust evaluation framework that quantitatively assesses the quality of generated ideas. Each of these

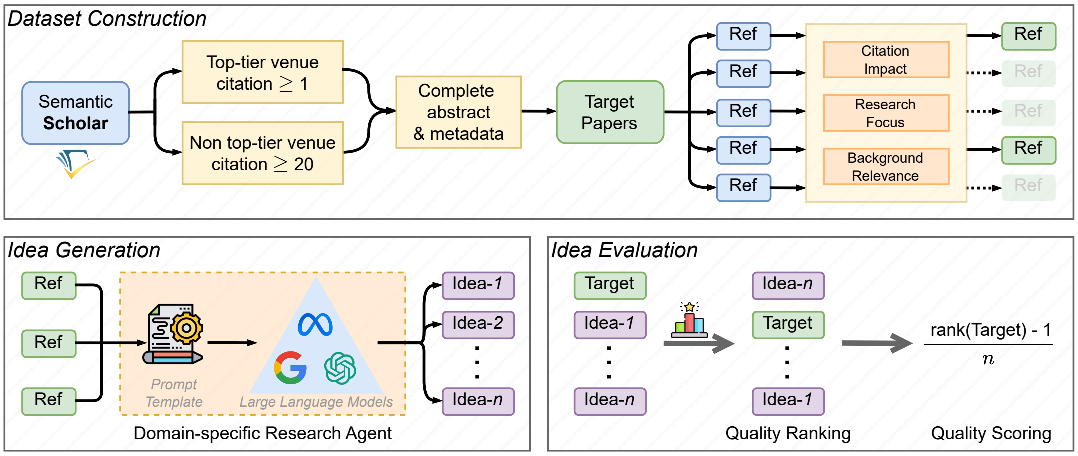


Figure 1: Benchmark Overview. *Dataset Construction* depicts the process of identifying research papers of interest (target papers), extracting their reference papers and filtering them based on relevance and significance. *Idea Generation* shows how an LLM utilizes a target paper's references to generate domain-specific research ideas. *Idea Evaluation* illustrates our proposed evaluation framework, which ranks the research idea from a target paper alongside LLM-generated research ideas, derived from the target paper's references, based on a personalized quality indicator.

components is carefully constructed to ensure the validity, relevance, and fairness of the benchmark. We outline our benchmark dataset construction, idea generation protocol, and research idea evaluation framework in Figure 1. Below, we provide a detailed design principle and methodology underpinning each component.

3.1 Dataset Construction

A fundamental requirement for evaluating research idea generation is a well-structured dataset that mirrors the real-world process of scientific discovery. To this end, we construct a dataset comprising two interconnected elements: (1) **target papers**, which encapsulate ground-truth research ideas, and (2) **reference papers**, which provide the contextual background that inspired these ideas. This dataset enables LLMs to simulate the ideation process by leveraging relevant prior work.

3.1.1 Curation of Target Papers To ensure that *IdeaBench* reflects state-of-the-art scientific contributions and to ensure the target papers do not appear in the training data of the models, we select a set of high-quality research articles published in 2024. Our selection criteria prioritize rigorous peer-reviewed work that has gained early recognition in the scientific community. Specifically, we retrieve papers from (1) top-tier venues as ranked by Google Scholar [8] with at least one citation and (2) other venues where papers have received at least 20 citations. This dual criterion ensures that our dataset includes both highly regarded publications and emerging influential research.

To preserve the integrity of our evaluation, we exclude papers with missing abstracts or incomplete metadata. The resulting set of **target papers** serves as the benchmark against which LLM-generated research ideas are compared.

3.1.2 Contextual Reference Papers Scientific ideation does not occur in isolation; researchers build on existing knowledge to formulate new ideas. To replicate this dynamic, we retrieve the **reference papers** cited in each target paper using the Semantic Scholar API [12]. These references provide the foundational context that human researchers leveraged when developing their research ideas. By grounding LLMs in the same contextual information, we ensure a realistic and meaningful evaluation of their ideation capabilities.

3.1.3 Filtering for Relevance and Significance Not all references cited in a paper are equally relevant to its central research idea. Some citations serve peripheral roles, such as general background or methodological comparisons, rather than directly informing the core hypothesis. To optimize the dataset while maintaining efficiency in computational resource allocation, we implement a filtering process to retain only the most pertinent references.

The filtering process follows three key principles:

(1) **Citation Impact:** We exclude reference papers with fewer than five citations, ensuring that retained references are recognized and influential within the field.
(2) **Primary Research Focus:** We remove non-primary research sources (e.g., review articles, editorials, books) as classified by Semantic Scholar, since these do not directly contribute to novel hypothesis generation.
(3) **Background Section Relevance:** We prioritize reference papers cited within the background or introduction section of the target paper, as these are most likely to have shaped the research idea.

This filtering process ensures that LLMs receive the most relevant context when generating new research ideas. As an additional baseline, we compare our approach to a random reference selection strategy, with ablation studies presented in later sections.

3.1.4 Benchmark Dataset Details Based on the construction method we discussed in this section, we curated our dataset consisting of 2,374 target articles across eight domains and their corresponding 29,408 reference articles. The eight domains are: (1) Health & Medicine, (2) Genetics & Molecular Biology, (3) Environmental Sciences, (4) Neuroscience & Cognitive Sciences, (5) Technology

& Engineering, (6) Social & Behavioral Sciences, (7) Materials Science, and (8) Public Health & Policy. After applying our reference filtering process, we retain 23,460 high-quality reference papers. This dataset forms the foundation for the evaluation of LLM-driven research ideas by providing the contextual background necessary to generate and assess research ideas. Note that our dataset construction method can be applied to update the paper selection or customize datasets for any domain of interest.

Table 1: Total counts of the dataset's target papers and references.

Description	Count
Total number of target papers	2,374
Total number of reference papers (with filtering)	23,460
Total number of reference papers (w/o filtering)	29,408

Table 2: Descriptive statistics of the number of target papers for covered research domains in the dataset.

Research Domain	Count
Health & Medicine	764
Genetics & Molecular Biology	760
Environmental Sciences	123
Neuroscience & Cognitive Sciences	248
Technology & Engineering	114
Social & Behavioral Sciences	181
Materials Science	55
Public Health & Policy	129

Table 3: Descriptive statistics of the number of references per target paper.

Statistic	With Filtering	w/o Filtering
Mean	9.882	12.388
Standard Deviation	6.521	7.946
Minimum	3	3
25% Percentile	5	6
50% Percentile (Median)	8	10
75% Percentile	13	16
Maximum	51	62

Table 1 shows the total number of target papers in our dataset along with the number of filtered and unfiltered references. Table 2 shows the descriptive statistics of the number of target papers for eight covered research domains in the dataset. Table 3 shows the descriptive statistics of the filtered and unfiltered reference papers in our dataset.

3.2 Research Idea Generation

To systematically assess the research ideation capabilities of LLMs, we design a controlled generation protocol that closely mirrors human scientific thought processes. In academic research, new ideas typically emerge from analyzing existing literature and identifying unexplored opportunities. We emulate this process by prompting LLMs with the same contextual references that human researchers had access to when formulating their hypotheses.

A core challenge in using LLMs for the generation of research ideas is to ensure that the models generate ideas that are *inspired* by prior work rather than simply regurgitating existing concepts. To address this, we construct a structured prompt template that explicitly instructs LLMs to synthesize novel insights from reference papers while avoiding direct replication.

Figure 2 illustrates our prompt structure, in which the model is assigned the role of a **domain-specific researcher** tasked with proposing a hypothesis based on a curated set of reference abstracts. This approach leverages the knowledge models learned from the pretraining stage while grounding it in domain-relevant context. By systematically guiding LLMs through a structured ideation process, our goal is to approximate human-like research idea generation.

Our approach establishes a simple method for evaluating LLMs in research ideation, providing a fair baseline to assess their basic capability in generating ideas. LLMs may struggle with conceptual depth or practical feasibility, highlighting areas for future improvement. Therefore, we encourage the research community to explore more advanced methods for enhancing LLM-driven ideation, such as retrieval-augmented generation, in-context learning optimizations, or fine-tuning strategies tailored to scientific discovery. By iteratively improving research idea generation techniques, we can work toward unlocking the full potential of AI-assisted scientific creativity. Future advanced techniques can also be assessed using the evaluation method introduced in Section 3.3.

Prompt template for generating research ideas

You are a `{domain_specific}` researcher. You are tasked with creating a hypothesis or research idea given some background knowledge. The background knowledge is provided by abstracts from other papers.

Here are the abstracts:
Abstract 1:`{reference_paper_1_abstract}`
Abstract 2:`{reference_paper_2_abstract}`
......
Abstract n:`{reference_paper_n_abstract}`

Using these abstracts, reason over them and come up with a novel hypothesis. Please avoid copying ideas directly, rather use the insights to inspire a novel hypothesis in the form of a brief and concise paragraph.

Figure 2: Prompt template used to generate research ideas.

3.3 Evaluation of Generated Research Ideas

Evaluating the generation of research ideas is inherently challenging, as the quality of an idea cannot be determined solely by conventional metrics such as accuracy or similarity. Unlike tasks with clearly defined ground-truth labels, research ideation requires an assessment of more nuanced qualities, such as novelty, feasibility, and overall scientific value. To address this, we design a two-stage evaluation framework that consists of (1) **personalized quality ranking**, where generated research ideas are ranked based on a user-specified quality indicator, and (2) **relative quality scoring**, which quantifies how well an LLM performs relative to human-generated research ideas using a newly proposed metric called the *Insight Score.*

3.3.1 Personalized Quality Ranking for Generated Research Ideas Evaluating research ideas is inherently subjective, as different researchers may prioritize different aspects such as novelty and feasibility, etc. To accommodate diverse evaluation preferences, we develop a flexible ranking mechanism that allows users to specify a quality indicator of interest. Specifically, we employ LLMs to rank a set of generated research ideas alongside the research idea from the target paper, ensuring a comparative assessment against real human-produced contributions.

The evaluation process begins by constructing an **idea set** for each target paper, consisting of multiple LLM-generated research ideas and the extracted research idea from the target paper, which serves as a reference. Further details on the extraction process are provided in the Appendix. To ensure consistency and scalability, we employ a structured prompt template, as illustrated in Figure 3. The prompt dynamically incorporates the specified quality indicator (e.g., novelty, feasibility) and instructs the evaluator, an LLM, to rank the research ideas accordingly.

Leveraging LLMs' in-context learning ability [13], our approach allows users to refine quality definitions based on their preferences. For instance, "novelty" might be interpreted as the degree to which the work introduces original and significant contributions that depart from or extend existing knowledge in a specific field, and "feasibility" might be regarded as the extent to which the proposed study can be practically and successfully executed, given the available resources, time, expertise, and constraints. This adaptability ensures that the evaluation framework aligns with different research perspectives.

To prevent bias, the LLM evaluator is not informed which idea originates from the target paper, ensuring a fair comparison between LLM-generated and human-generated ideas. This design enables a standardized yet customizable approach to assessing the quality of research ideas, facilitating reproducible benchmarking across different LLMs. Finally, the ranked list based on the user-specified quality indicator for each target paper will be passed to quality scoring to obtain the Insight Score.

3.3.2 Quality Scoring: The Insight Score Building upon the personalized quality ranking, we introduce the **Insight Score**, a metric designed to quantitatively assess an LLM's ability to generate high-quality research ideas. This metric captures how often LLM-generated ideas surpass the target paper's idea in a ranked list based on a specified quality indicator. For the ith target paper and the user-specified quality indicator q, we use $\rho_i(q)$ to denote the rank of the target paper's research idea within the ranked list. When n ideas are generated, the quality of the generated ideas can be reflected by $\rho_i(q)$. We use $r_i(q)$ to denote the relative position of $\rho_i(q)$ given the quality indicator q, and it can be calculated via:

$$r_i(q) = \frac{\rho_i(q) - 1}{n} \tag{1}$$

where:

- $r_i(q)$ takes values in the range $[0, 1]$.
- If $r_i(q) = 0$, the target paper's idea ranks first, meaning no generated idea surpasses its quality.
- If $r_i(q) = 1$, the target paper's idea ranks last, indicating that all generated ideas outperform it.

Prompt template used to rank research ideas based on user-specified quality indicators

```
You are a reviewer tasked with ranking the quality of a set
of research ideas based on their {quality_indicator}.
The idea with the highest {quality_indicator} should
be ranked first.

Please rank the following hypotheses in the format:
1. Hypothesis (insert number):(insert brief rationale)
2. Hypothesis (insert number):(insert brief rationale)
......
n. Hypothesis (insert number):(insert brief rationale)

Please rank the following hypotheses:
Hypothesis 1: {target_paper_idea}
Hypothesis 2: {generated_idea_1}
......
Hypothesis n: {generated_idea_n}
```

Figure 3: Prompt template used to rank research ideas based on user-specified quality indicators.

With $r_i(q)$, we can formally define the **Insight Score** $I(LLM, q)$ as the average $r_i(q)$ over the dataset, which measures the given LLM's capability in generating research ideas with respect to the user-specified quality indicator:

$$I(LLM, q) = \frac{1}{m} \sum_{i=1}^{m} r_i(q) \tag{2}$$

where m is the number of target papers in the dataset.

Formulas (1) and (2) together define a reference-based metric which provides a relative quality assessment of LLM-generated research ideas for different quality indicators. Unlike absolute similarity-based metrics, the **Insight Score** offers a direct comparative measure of how well LLMs generate ideas that are competitive with or superior to those in high-impact research papers.

To ensure fair comparisons across different models, it is essential to generate a consistent number of ideas n per query. Our analysis indicates that the rank position of the target paper's idea varies with n, directly influencing the computed **Insight Score**.

By integrating personalized ranking with a quality scoring system, our evaluation framework offers a scalable and flexible approach for assessing LLM-generated research ideas. This methodology not only enables standardized comparisons across different models but also accommodates domain-specific research priorities through customizable quality indicators.

4 Experiments

In this section, we evaluate the ability of Large Language Models (LLMs) to generate research ideas by benchmarking them with our proposed *IdeaBench* framework. We investigate their capacity to produce novel and feasible ideas, and the impact of resource constraints on performance. Furthermore, we demonstrate the effectiveness of our proposed *Insight Score* in capturing quality dimensions that traditional similarity metrics overlook and show that it aligns well with human evaluation for novelty but not as well for feasibility.

4.1 Experimental Setup

4.1.1 Evaluated Models To assess LLMs' capability in generating research ideas, we evaluate several state-of-the-art commercial and open-source models, spanning different architectures and parameter sizes. Specifically, we test the Meta LLaMA [24], Google Gemini [20], and OpenAI GPT [17] series. All models have a training data cutoff prior to January 1, 2024, ensuring that the target papers included in our dataset—published in 2024—were not part of their pretraining corpus, thus eliminating data leakage and ensuring a fair comparison.

4.1.2 Evaluation Metrics To comprehensively evaluate LLM-generated research ideas, we compare models using both traditional similarity metrics and our proposed *Insight Score*:

- **Semantic Similarity:** We use BERTScore (F1 score) [31] to measure the semantic similarity between the generated research idea and the target paper's research idea.
- **Idea Overlap:** Inspired by the positive correlation between GPT models and human judgments in rating-based evaluations across various benchmarks [16], we compute an LLM similarity rating using GPT-4o. This rating quantifies the conceptual overlap between the generated research idea and the target paper's abstract on a scale from 1 to 10. The prompt template used to obtain this rating is provided in the Appendix (Figure 8).
- **Insight Score:** Our proposed metric evaluates the relative quality of generated research ideas based on user-defined indicators, such as novelty and feasibility.

To ensure a fair comparison, we use the highest-scoring generated idea (out of $n = 3$ per query) for semantic similarity and idea overlap measurements.

4.1.3 Resource Constraints: Low vs. High Resource Scenarios To assess the impact of input constraints on research idea generation, we evaluate models in two scenarios:

- **Low Resource:** The LLM is provided with only the top five references, as selected by our reference filtering method.
- **High Resource:** The LLM is provided with all unfiltered references. For GPT-3.5 Turbo, which has a limited context window, we input as many references as the model allows.

By comparing these scenarios, we analyze how access to more contextual information affects LLM-generated research ideas.

4.2 Main Results

For a concise illustration, we further categorize the eight research domains into three categories. We put Health & Medicine, Genetics & Molecular Biology, Neuroscience & Cognitive Sciences, and Public Health & Policy into Life & Health Sciences. We put Environmental Sciences and Material Sciences into Physical & Environmental Sciences. We put Technology & Engineering and Social & Behavioral Sciences into Technonology, Society & Engineering. The primary benchmark results are presented in Table 4. We present more detailed benchmark results for the eight research domains in our GitHub repository. Below, we analyze the key findings.

4.2.1 How well do LLM Research Ideas Align with Human-Authored Research Ideas? LLMs generate research ideas that semantically align with human-authored ideas, although the actual degree of overlap varies by model and resource scenario. As shown in Table 4, models such as GPT-3.5 Turbo and GPT-4o Mini (High Resource) consistently achieve the highest semantic similarity scores and produce the greatest idea overlap, typically ranging between 4 and 6. In all domains tested, high resource scenarios lead to better alignment compared to low resource settings.

Moreover, as demonstrated in Table 5, increasing the number of reference papers improves semantic similarity and idea overlap. Providing more references gives the LLM additional context, mirroring the cues human authors use to craft their research ideas.

This trend holds across multiple domains, indicating that LLMs can generate domain-specific research ideas without further fine-tuning when provided with proper context, thanks to the knowledge they learned from their extensive and diverse training data.

4.2.2 Do LLMs Generate Novel Research Ideas? As shown in Table 4, most LLMs generate research ideas that are at least as novel as those found in human-authored papers. Notably, GPT-4o (high resource) achieves the highest novelty insight scores across all domains, indicating that its ideas are often ranked even higher in novelty than those proposed by human researchers. These results highlight the potential of LLMs to push the boundaries of scientific discovery by proposing unconventional and exploratory hypotheses.

The high novelty of LLM-generated ideas can be attributed to their ability to synthesize vast amounts of knowledge across disciplines and form unexpected connections. Unlike human researchers, who are often influenced by limited background knowledge, cognitive biases, or feasibility constraints, LLMs operate without such limitations. This allows them to explore unorthodox perspectives, re-purpose insights from seemingly unrelated fields, and generate hypotheses that may be neglected by human researchers.

This insight suggests that LLM-assisted ideation could complement human creativity, particularly in early-stage research where generating bold, unconventional ideas is crucial. By leveraging their capacity for knowledge recombination, LLMs could serve as valuable collaborators in brainstorming sessions, pushing researchers to consider possibilities they might not have initially explored. However, effective use of LLM-generated ideas will require human expertise to refine, validate, and adapt these ideas into actionable research directions.

4.2.3 Do LLMs Generate Feasible Research Ideas? LLMs struggle with feasibility. As shown in Table 4, feasibility Insight Scores are consistently lower than novelty scores, with no model exceeding 0.408. This suggests that while LLMs are adept at generating innovative ideas, they face significant challenges in producing ideas that are practically viable within real-world research constraints.

Generating a feasible research idea requires more than just creativity—it demands a deep understanding of domain-specific constraints, such as resource availability, technical expertise, experimental feasibility, funding, and time limitations. Although LLMs are prompted to adopt a researcher's perspective, they tend to prioritize creative associations over pragmatic considerations, often

Table 4: Main benchmark results. The table displays semantic similarity (mean) and the idea overlap (mean) between generated research ideas and target papers. It also shows the evaluation of novelty and feasibility using our proposed Insight Scores for various LLMs in high and low resource settings. Bold scores represent the highest score of a given metric.

Model	Resource Scenario	Semantic Similarity ↑	Idea Overlap ↑	Novelty Insight Score ↑	Feasibility Insight Score ↑
Life & Health Sciences					
Llama 3.1 70B-Instruct	low	0.586	4.805	0.626	0.141
Llama 3.1 70B-Instruct	high	0.600	5.544	0.599	0.132
Llama 3.1 405B-Instruct	low	0.561	5.206	0.648	0.119
Llama 3.1 405B-Instruct	high	0.584	5.648	0.689	0.122
Gemini 1.5 Flash	low	0.603	4.600	0.440	0.234
Gemini 1.5 Flash	high	0.608	5.381	0.581	0.290
Gemini 1.5 Pro	low	0.620	4.304	0.508	0.286
Gemini 1.5 Pro	high	0.630	5.205	0.652	0.291
GPT-3.5 Turbo	low	0.640	5.145	0.409	0.175
GPT-3.5 Turbo	high	0.646	5.912	0.211	**0.308**
GPT-4o Mini	low	0.637	4.943	0.450	0.148
GPT-4o Mini	high	**0.649**	**6.069**	0.535	0.193
GPT-4o	low	0.619	5.051	0.623	0.140
GPT-4o	high	0.627	5.846	**0.777**	0.150
Physical & Environmental Sciences					
Llama 3.1 70B-Instruct	low	0.578	4.523	0.602	0.166
Llama 3.1 70B-Instruct	high	0.584	5.338	0.575	0.215
Llama 3.1 405B-Instruct	low	0.547	4.944	0.621	0.181
Llama 3.1 405B-Instruct	high	0.566	5.185	0.655	0.161
Gemini 1.5 Flash	low	0.595	4.343	0.470	0.251
Gemini 1.5 Flash	high	0.598	5.079	0.558	0.346
Gemini 1.5 Pro	low	0.612	4.000	0.522	**0.408**
Gemini 1.5 Pro	high	0.623	5.017	0.634	0.343
GPT-3.5 Turbo	low	0.631	4.787	0.377	0.228
GPT-3.5 Turbo	high	0.638	**5.742**	0.092	0.380
GPT-4o Mini	low	0.632	4.680	0.420	0.205
GPT-4o Mini	high	**0.640**	5.713	0.494	0.238
GPT-4o	low	0.611	4.961	0.582	0.161
GPT-4o	high	0.615	5.354	**0.697**	0.222
Technology, Society & Engineering					
Llama 3.1 70B-Instruct	low	0.585	4.434	0.619	0.196
Llama 3.1 70B-Instruct	high	0.593	5.016	0.640	0.212
Llama 3.1 405B-Instruct	low	0.547	4.569	0.657	0.171
Llama 3.1 405B-Instruct	high	0.564	5.102	0.606	0.179
Gemini 1.5 Flash	low	0.591	4.384	0.345	0.285
Gemini 1.5 Flash	high	0.594	4.885	0.489	0.355
Gemini 1.5 Pro	low	0.603	3.775	0.512	0.367
Gemini 1.5 Pro	high	0.613	4.588	0.614	**0.372**
GPT-3.5 Turbo	low	0.628	4.698	0.363	0.261
GPT-3.5 Turbo	high	**0.636**	**5.627**	0.198	0.244
GPT-4o Mini	low	0.627	4.644	0.431	0.204
GPT-4o Mini	high	0.634	5.508	0.502	0.254
GPT-4o	low	0.606	4.831	0.575	0.155
GPT-4o	high	0.609	5.329	**0.735**	0.238

proposing ideas that sound promising in theory but lack a realistic pathway to implementation. The consistently lower feasibility scores suggest that while LLMs can recombine existing knowledge in novel ways, they struggle to assess whether those ideas can be executed with available methods and resources.

A key distinction between human-generated and LLM-generated ideas lies in verification. Human researchers are trained to generate ideas that are feasible because feasibility is an implicit prerequisite for publication—if an idea were not viable, it would not progress through peer review or be recognized as a meaningful scientific contribution. In contrast, LLM-generated ideas remain unverified, requiring further validation through experimentation and expert assessment. Without empirical grounding, LLM-generated hypotheses risk being speculative rather than actionable.

Additionally, the models tested in this study are general-purpose LLMs, not models specifically fine-tuned for scientific research. Unlike domain-trained systems that integrate structured scientific knowledge or empirical constraints, general-purpose LLMs lack an inherent mechanism to evaluate whether a proposed idea is experimentally achievable. This limitation is particularly pronounced in resource-intensive disciplines, where feasibility depends on factors beyond theoretical soundness, such as equipment availability, regulatory constraints, or interdisciplinary collaboration.

Ultimately, while LLMs show promise in assisting ideation, there is still much room for them to improve their capability to autonomously generate research ideas ready for implementation. Currently, their role may be best suited to expanding the conceptual space of inquiry, providing inspiration for researchers who can then refine, validate, and adapt these ideas to practical constraints.

Table 5: Comparison of semantic similarity and idea overlap scores for research ideas generated by GPT-4o Mini with filtered and unfiltered references. Underlined scores are higher when compared to their filtered/unfiltered counterpart. Bold values are the highest for each measure.

Num Ref	Similarity (Filtered)	Similarity (Unfiltered)	Idea Overlap (Filtered)	Idea Overlap (Unfiltered)
1	0.619	0.616	2.946	2.640
3	0.631	0.630	4.636	4.410
5	0.637	0.635	5.089	4.913
7	0.639	0.638	5.349	5.304
10	0.641	0.642	5.720	5.508
13	0.642	0.643	5.694	5.685
15	0.643	0.642	**6.002**	5.748
All	**0.644**	**0.648**	5.915	**6.302**

4.2.4 The Effect of Reference Filtering We investigate the impact of reference filtering on research idea generation for lower-capacity models like GPT-4o Mini. As shown in Table 5, when fewer reference papers are provided, our reference filtering method improves alignment with human-authored ideas, as evidenced by higher semantic similarity and idea overlap scores compared to using the same number of randomly sampled references. This suggests that reference filtering helps lower-capacity models focus on the most relevant information, enhancing idea generation despite computational constraints.

Thus, when computational resources are limited, reference filtering can be an effective strategy for generating higher-quality ideas with a restricted number of references. However, when all references are available, GPT-4o Mini achieves the best semantic similarity and idea overlap, indicating that when computing resources is not a major concern, models benefit from broader contextual exposure.

4.3 Human Study: Effectiveness of Insight Score

To evaluate whether the Insight Score provided by LLMs aligns with human judgment, we conducted a rigorously designed blind human study. Eight immunology researchers ranked research ideas generated by both LLMs and humans without knowing the origin of each idea. To further eliminate bias, the human-generated ideas were extracted from recent immunology articles that are not used in LLM training data, and the human evaluators were kept unaware of the source article of the idea.

Four research ideas were presented to the human evaluators. Three research ideas were generated using the method discussed in Section 3.2, and the idea from an immunology article was the fourth. The human evaluators ranked the four ideas in order of novelty and feasibility, from most to least, using a four-level scale. They were not allowed to use external resources or assign ties. To minimize cognitive biases such as recency and primacy effects, the order of presentation was randomized for each human evaluator. The assessment was conducted through an online survey. More details on the human study along with an example of a immunology article and the LLM generated research ideas is provided in the Appendix.

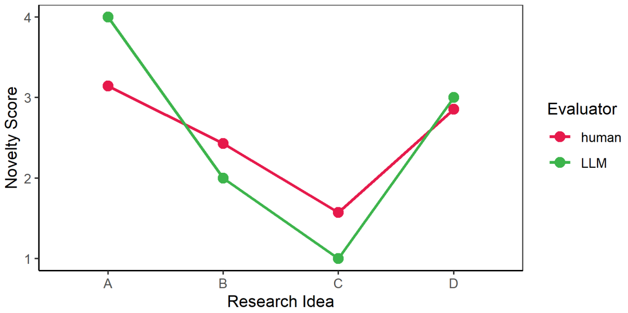


Figure 4: The difference in novelty rankings between the average human evaluation and the LLM evaluation. Research Idea A represents the main idea from the selected paper, with a novelty score of 4 indicating the highest level of novelty.

Figures 4 and 5 compare the average of the rankings of ideas assigned by human evaluators and the ranking of the LLM evaluator. In each figure, Idea A represents the target paper's idea, while Ideas B, C, and D correspond to the three generated ideas.

For novelty, as shown in Figure 4, the LLM evaluator's assessments closely align with human judgments, as indicated by the similar trends of the red line (the averaged human judgment) and the green line (the judgment of the LLM evaluator). However, for feasibility as shown in Figure 5, a notable discrepancy emerges, particularly in the ranking of the target paper's idea. While the ranking trends for the three generated ideas remain consistent, the target paper's idea shows a significant deviation between human and LLM assessments.

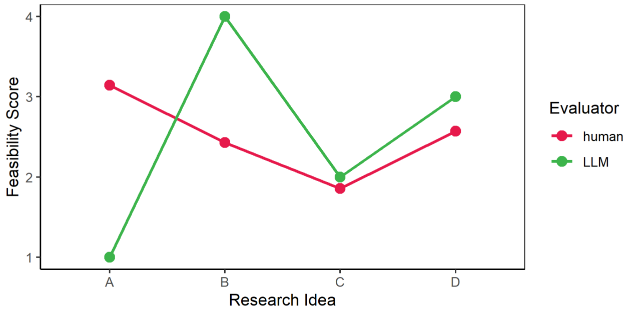


Figure 5: The difference in feasibility rankings between the average human evaluation and the LLM evaluation. Research Idea A represents the idea from the selected paper, with a feasibility score of 4 indicating the highest level of feasibility.

This divergence stems from the nature of feasibility as a quality indicator. Unlike novelty, which can be assessed independently, feasibility requires verification. Since the target paper's idea has been published and validated by the research community, it is intuitively more feasible than newly proposed ideas generated by LLMs. Additionally, the human evaluators in our study are trained researchers with domain expertise, whereas the LLM evaluator is a general-purpose model rather than one specifically fine-tuned for research evaluation. As a result, assessing feasibility is inherently more challenging for the LLM.

Despite this difference, the overall alignment between human judgments and the LLM evaluator within the scope of generated ideas demonstrates that our proposed Insight Score effectively approximates human assessments, making it a practical tool for evaluating future generated ideas.

4.4 Effect of the Number of Generated Research Ideas on Insight Score

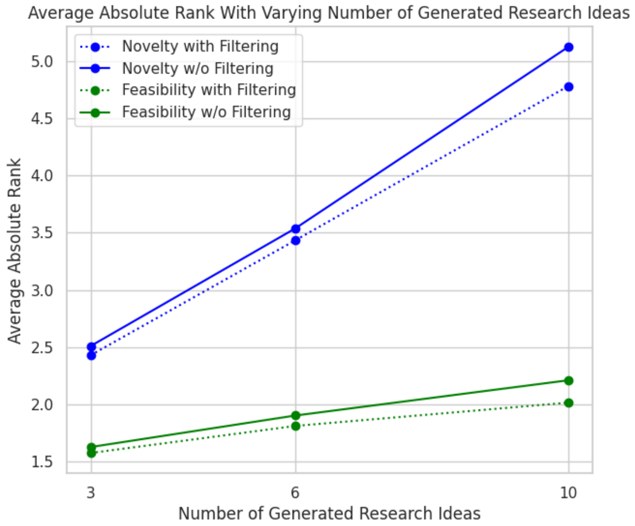


Figure 6: Effect that the number of generated research ideas has on the target paper's Average Absolute Rank.

We assess how the number of generated research ideas n affects the target paper's absolute rank. We use research ideas generated with GPT-4o Mini. Figure 6 shows that as n increases, the target paper is ranked farther in the list for both novelty and feasibility, regardless of reference filtering. Changes in the absolute rank of the target paper will affect the Insight Score.

To illustrate the effect varying n has on the Insight Score, consider we acquire the Insight Score for an LLM that generates three research ideas for one target paper. If the target paper ranks 3rd, then its Insight Score would be 0.667. Now, if we have the same LLM generate 10 research ideas and the target paper ranks 5th, as suggested by Figure 6, then its Insight Score would be 0.4. Despite using the same LLM, the variation in n causes different Insight Scores.

Another effect that n has on the Insight Score is the granularity effect, which arises from the discrete nature of the Insight Score. A larger n allows for a more granular measurement of an LLM's capability in generating research ideas, meaning that a single shift in ranking position results in a smaller change in the Insight Score compared to a smaller n. For example, consider Case $n = 6$ and Case $n = 10$. If the target paper ranks around the top 50%, that is, $\rho_i(q) = 4$ for $n = 6$ and $\rho_i(q) = 6$ for $n = 10$, both will have an Insight Score of 0.5. However, if the target paper's ranking position drops by one, then for $n = 6$, $\rho_i(q) = 5$, the Insight Score will increase to 0.667; whereas for $n = 10$, $\rho_i(q) = 7$, the Insight Score will be 0.6. This difference does not necessarily indicate that the LLM with an Insight Score of 0.667 is better than the one with 0.6. Instead, it reflects that the first model generated fewer research ideas, resulting in a less granular Insight Score.

Thus, using the same n for comparing different LLMs' Insight Scores is recommended to avoid unfair comparison.

5 Conclusion

In this work, we introduce *IdeaBench*, a benchmark system designed to systematically evaluate LLM-driven research idea generation, addressing the lack of standardized assessment in this domain. By constructing a structured dataset of influential papers and their referenced works, we provide a context-rich environment that mirrors human researchers' ideation processes. Our evaluation framework, which leverages LLMs to approximate human assessment, reveals that while LLMs excel at generating novel ideas, they often struggle with feasibility. *IdeaBench* serves as a valuable resource for benchmarking and comparing any technique using LLMs to generate research ideas, fostering further research into improving AI-assisted scientific discovery. We encourage the research community to build upon our evaluation framework, contributing new insights and methodologies to advance AI-assisted scientific discovery. We hope this work inspires future advancements in LLM ideation strategies, integrating external knowledge, and enhancing evaluation methodologies to better assess and guide AI-driven research ideation.

Acknowledgements

This work is supported in part by the US National Science Foundation under grants 2217071, 2213700, 2106913, 2008208, and NIH grant 1R01LM014012. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

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

[2] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738* (2024).

[3] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems* 36 (2024).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[6] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937* (2023).

[7] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6556–6576. doi:10.18653/v1/2024.naacl-long.365

[8] Google Scholar. 2024. Google Scholar Top Publications. https://scholar.google.com/citations?view_op=top_venues.

[9] Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, and Aidong Zhang. 2024. Embracing Foundation Models for Advancing Scientific Discovery. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 1746–1755.

[10] Sikun Guo, Guangzhi Xiong, and Aidong Zhang. 2025. Optimizing External and Internal Knowledge of Foundation Models for Scientific Discovery. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*. SIAM, 431–434.

[11] Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255* (2024).

[12] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023).

[13] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[14] Ruochen Li, Liqiang Jing, Chi Han, Jiawei Zhou, and Xinya Du. 2024. Learning to Generate Research Idea with Dynamic Control. *arXiv preprint arXiv:2412.14626* (2024).

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:964287

[16] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023).

[17] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). https://arxiv.org/abs/2303.08774

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[19] Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559* (2023).

[20] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[21] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227* (2025).

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

[23] Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *arXiv preprint arXiv:2004.14564* (2020).

[24] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[25] Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation. *arXiv preprint arXiv:2407.10817* (2024).

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

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

[28] Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2025. Toward Reliable Biomedical Hypothesis Generation: Evaluating Truthfulness and Hallucination in Large Language Models. *arXiv preprint arXiv:2505.14599* (2025).

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

[30] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems* 34 (2021), 27263–27277.

[31] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[32] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622* (2019).

[33] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis Generation with Large Language Models. *arXiv preprint arXiv:2404.04326* (2024).

A Appendix

A.1 Implementation Details

We describe the resources used to generate and evaluate research ideas. We used various API services to generate research ideas and to evaluate them. Additionally, we employed accelerated hardware to compute semantic similarity scores between generated research ideas and their corresponding target paper abstracts.

The OpenAI API service [1] was employed to generate research ideas that used the OpenAI suite of models. The service was also used for extracting research ideas from target paper abstracts and evaluating research ideas with the LLM similarity rating and the Insight Score. To generate research ideas with the Gemini family of LLMs, Google AI's API service [2] was used. To generate research ideas with the Llama 3.1 family of LLMs, DeepInfra's API service [3] was used.

To evaluate the semantic similarity between research ideas and their corresponding target paper abstracts, we computed BERTScores using one NVIDIA A6000 48GB GPU. This hardware allowed for the efficient computation of BERTScores.

A.2 Extracting Research Idea from a Target Paper

To ensure a fair comparison between the research ideas generated by LLMs and those in the target paper when ranking the ideas, we extract the core research idea from the target paper's abstract using GPT-4o with a specifically designed prompt. Abstracts often contain distracting information, such as detailed results, which

[1] More details about the OpenAI API service can be found here: https://platform.openai.com/docs/overview

[2] More details about the Google AI API service can be found here: https://ai.google.dev/gemini-api/docs/api-key

[3] More details about the DeepInfra's API service can be found here: https://deepinfra.com/

Prompt template used to extract the research idea from a given target paper's abstract.

Write a concise paragraph summarizing the following biomedical paper abstract as if you are proposing your own research idea or hypothesis. Focus on describing the main research idea and provide a high-level summary of the findings without detailed results or specific numerical data. Please begin the paragraph with "Hypothesis: " or "Given that ".

Abstract:
`{target_paper_abstract}`

Summary:

Figure 7: Prompt template used to extract a target paper's research idea.

may not directly reflect the central research idea and may bias the Insight Score when ranking research ideas. Therefore, we designed a prompt that focuses on summarizing the main research idea in a way that aligns with how our LLM generates ideas. Figure 7 shows the prompt template. This process enables a fair ranking of the target paper's idea alongside the LLM generated ideas.

A.3 Computing Idea Overlap Score

Prompt template used to measure the overlap of ideas between a generated research idea and its target paper.

You are an expert in understanding and analyzing scientific content. Your task is to evaluate the degree of overlap between the ideas presented in a hypothesis and the abstract of a scientific paper. Please read both the hypothesis and the abstract carefully. Then, rate the overlap on a scale of 1 to 10, where 1 indicates minimal or no overlap, and 10 indicates a perfect or nearly perfect overlap. Provide a brief explanation for your rating.

Hypothesis: `{generated_research_idea}`

Abstract: `{target_paper_abstract}`

Rating: On a scale of 1-10, rate the overlap between the ideas in the hypothesis and the abstract.

Explanation: In one sentence, provide a brief explanation for your rating, mentioning the key points of overlap and any significant differences you observed.

Figure 8: Prompt template to obtain LLM similarity rating.

Figure 8 shows the prompt used to obtain Idea Overlap scores between target paper abstracts and their corresponding research ideas, including both a similarity rating and an explanation of the overlap.

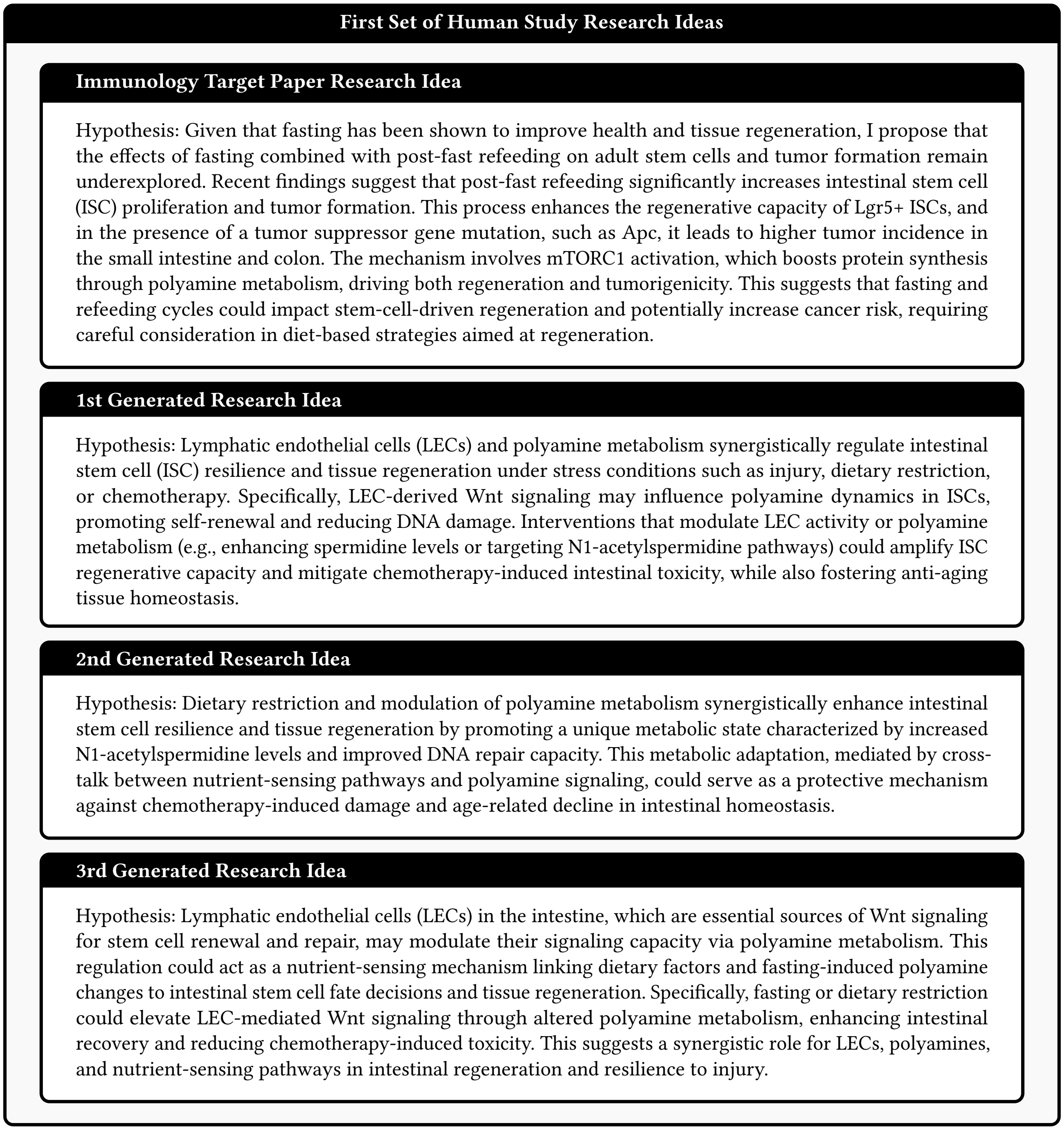
A.4 Additional Details for Human Study on Effectiveness of the Insight Score

To evaluate how well the Insight Scores provided by GPT-4o align with human judgment we design a rigorous blind human study where eight immunology researchers perform human evaluation on research ideas. We describe the main results of the human study in Section 4.3 of the main text. Here we provide additional details of the human study design and provide examples of the research ideas being evaluated.

Among the eight immunology researchers, four are doctoral candidates specializing in immunology, and four hold doctoral degrees in immunology related fields. For our study, two target papers were selected using a controlled process to mitigate bias. One researcher chose a target paper that met three key criteria: it was recent enough to be absent from the LLM's training data, it had not been previously read by the other seven researchers, and was relevant to the researchers' expertise in immunology. We report this target paper's research idea in Figure 9.

Research ideas were then extracted from the target paper, and GPT-4o (High Resource) generated three research ideas for each target paper following the process detailed in Section 3.2. We present the generated research ideas in Figure 9. Both the target paper–derived and LLM-generated ideas were presented in a randomized order via an online survey to human evaluators, who ranked them on novelty and feasibility. Importantly, evaluators were blinded to the origin of each idea, ensuring that their assessments were based solely on content rather than any preconceptions about the source. This blinding, coupled with randomization, minimizes potential evaluator bias by preventing any influence from the idea's provenance or presentation order.

The same set of research ideas was also ranked by GPT-4o using the method described in Section 3.3. We compare the human and GPT-4o rankings and present our findings in Section 4.3 of the main text.



First Set of Human Study Research Ideas

Immunology Target Paper Research Idea

Hypothesis: Given that fasting has been shown to improve health and tissue regeneration, I propose that the effects of fasting combined with post-fast refeeding on adult stem cells and tumor formation remain underexplored. Recent findings suggest that post-fast refeeding significantly increases intestinal stem cell (ISC) proliferation and tumor formation. This process enhances the regenerative capacity of Lgr5+ ISCs, and in the presence of a tumor suppressor gene mutation, such as Apc, it leads to higher tumor incidence in the small intestine and colon. The mechanism involves mTORC1 activation, which boosts protein synthesis through polyamine metabolism, driving both regeneration and tumorigenicity. This suggests that fasting and refeeding cycles could impact stem-cell-driven regeneration and potentially increase cancer risk, requiring careful consideration in diet-based strategies aimed at regeneration.

1st Generated Research Idea

Hypothesis: Lymphatic endothelial cells (LECs) and polyamine metabolism synergistically regulate intestinal stem cell (ISC) resilience and tissue regeneration under stress conditions such as injury, dietary restriction, or chemotherapy. Specifically, LEC-derived Wnt signaling may influence polyamine dynamics in ISCs, promoting self-renewal and reducing DNA damage. Interventions that modulate LEC activity or polyamine metabolism (e.g., enhancing spermidine levels or targeting N1-acetylspermidine pathways) could amplify ISC regenerative capacity and mitigate chemotherapy-induced intestinal toxicity, while also fostering anti-aging tissue homeostasis.

2nd Generated Research Idea

Hypothesis: Dietary restriction and modulation of polyamine metabolism synergistically enhance intestinal stem cell resilience and tissue regeneration by promoting a unique metabolic state characterized by increased N1-acetylspermidine levels and improved DNA repair capacity. This metabolic adaptation, mediated by cross-talk between nutrient-sensing pathways and polyamine signaling, could serve as a protective mechanism against chemotherapy-induced damage and age-related decline in intestinal homeostasis.

3rd Generated Research Idea

Hypothesis: Lymphatic endothelial cells (LECs) in the intestine, which are essential sources of Wnt signaling for stem cell renewal and repair, may modulate their signaling capacity via polyamine metabolism. This regulation could act as a nutrient-sensing mechanism linking dietary factors and fasting-induced polyamine changes to intestinal stem cell fate decisions and tissue regeneration. Specifically, fasting or dietary restriction could elevate LEC-mediated Wnt signaling through altered polyamine metabolism, enhancing intestinal recovery and reducing chemotherapy-induced toxicity. This suggests a synergistic role for LECs, polyamines, and nutrient-sensing pathways in intestinal regeneration and resilience to injury.

Figure 9: First Set of Human Study Research Ideas. This includes a research idea extracted from a recent immunology target paper along with three research ideas generated using GPT-4o and the target paper's references.