ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition

Yujie Liu[1*], **Zonglin Yang**[2,1*], **Tong Xie**[3], **Jinjie Ni**[4], **Ben Gao**[5,1],
Yuqiang Li[1], **Shixiang Tang**[1], **Wanli Ouyang**[1], **Erik Cambria**[2†], **Dongzhan Zhou**[1†]
[1] Shanghai Artificial Intelligence Laboratory [2] Nanyang Technological University
[3] University of New South Wales [4] National University of Singapore [5] Wuhan University
liuyujie.cs@gmail.com, {zonglin.yang,cambria}@ntu.edu.sg, zhoudongzhan@pjlab.org.cn

Abstract

Large language models (LLMs) have demonstrated potential in assisting scientific research, yet their ability in discovering high-quality research hypotheses remains unexamined due to the lack of a dedicated benchmark. To address this gap, we introduce the first large-scale benchmark for evaluating LLMs with a sufficient set of sub-tasks of scientific discovery, which are inspiration retrieval, hypothesis composition, and hypothesis ranking. We develop an automated framework that extracts critical components—research questions, background surveys, inspirations, and hypotheses—from scientific papers across 12 disciplines, with expert validation confirming its accuracy. To prevent data contamination, we focus exclusively on papers published in 2024, ensuring minimal overlap with LLM pretraining data. Our evaluation reveals that LLMs perform well in retrieving inspirations, an out-of-distribution task, suggesting their ability to surface novel knowledge associations. This positions LLMs as "research hypothesis mines", capable of facilitating automated scientific discovery by generating innovative hypotheses at scale with minimal human intervention.

1 Introduction

Large language models (LLMs) have shown their potential to assist scientist's research as a copilot (Luo et al., 2025). One of the most challenging copilot tasks is to help scientists discover new valid research hypotheses, where a typical setting is to only provide a research question and a small background survey as input. Understanding how LLMs perform on this task is crucial for selecting the appropriate models and evaluating how different training strategies influence their effectiveness in scientific discovery.

However, although there are efforts benchmarking LLM's performance to the general task, such as Chatbot Arena (Chiang et al., 2024) and MixEval (Ni et al., 2024), it still lacks understanding of the scientific discovery ability of each LLM.

One of the main reasons for this vacancy is the lack of understanding of the scientific discovery process, i.e., how each research hypothesis is formulated. Recently, Yang et al. (2024c) decomposed this research hypothesis formulation process into a sufficient set of sub-tasks, which are (1) retrieving inspirations based on the research question, (2) properly mixing the research background information with the retrieved inspirations to compose research hypotheses, and (3) ranking the composed hypotheses to provide one with the highest confidence. This decomposition is viable because of a fundamental assumption that a majority of hypotheses can originate from a research background and several inspirations.

[*]Both authors contributed equally to this work. [†]Corresponding author.

Preprint. Under review.

Discipline	Cell	Chem	ETS	MS	Phys	EGS	EVS	BL	BS	Law	Math	AT	Overall
Paper Number	152	113	114	116	132	117	116	115	115	97	113	86	1386

Table 1: Disciplines and paper number distribution. Chem=Chemistry, ETS=Earth Science, MS=Material Science, Phys=Physics, EGS=Energy Science, EVS=Environmental Science, BL=Biology, BS=Business, AT=Astronomy.

This fundamental assumption is supported by cognitive science findings that *creative ideas often result from the cohesive association of two (or more) seemingly unrelated pieces of knowledge* (Koestler, 1964; Benedek et al., 2012; Lee & Chung, 2024). The cognitive science findings are discipline-independent and widely applicable. For example, the proposal of backpropagation is a research hypothesis. In this case, the research background is about multi-layer logistic regression, and the inspiration is the chain rule in calculus.

This research aims at filling the research gap, by providing a benchmark specifically designed to evaluate LLM's performance in terms of the three decomposed tasks of scientific discovery. This benchmark covers 12 disciplines, selecting papers published on top venues such as Nature, Science, or journals of a similar level. The statistics of the benchmark is shown in Table 1.

To construct the benchmark, we download 1386 papers and develop an automated LLM-based agentic framework to analyze each paper into research question, background survey, inspirations, and main hypothesis. We invited five experts in Physics, Chemistry, Astronomy, and Material Science disciplines to check whether the decomposition is accurate. Among the randomly sampled 62 papers checked by the experts, the decomposition accuracy was 91.9% considering only major issues and 82.3% when including both major and minor issues. It shows that the automated framework can accurately extract these components from a paper.

To prevent the data contamination problem, we only select those papers published in 2024. The advantage of our LLM agentic framework for the extraction is that as the LLM's pretraining data cutoff date moves forward, the framework can automatically extract more recent papers to avoid overlapping.

Based on the benchmark, we systematically compare popular LLMs across the three decomposed tasks. We find that current LLMs perform well in retrieving inspirations across disciplines, despite the inclusion of carefully crafted challenging negative inspiration examples. For example, when we ask GPT-4o to select the top 4% of inspiration candidates, the probability that a ground truth inspiration will be included in the top 4% is 45.7%. It is surprising because the inspiration retrieval task is actually an out-of-distribution (OOD) task since inspiration is supposed to be *not known* to be related to the research question but in fact *can assist* it. Otherwise, the resulting hypothesis won't be novel. In addition, we find that LLMs also have a good performance on the hypothesis composition and hypothesis ranking task. This suggests that LLMs could be leveraged as research hypothesis mines, where higher-performing LLMs on the three fundamental tasks of scientific discovery act as richer mines, and more inference compute corresponds to more miners.

Overall, the contributions of this paper are:

- We introduce the first large-scale benchmark for evaluating LLMs' capabilities in scientific discovery with a sufficient set of sub-tasks: inspiration retrieval, hypothesis composition, and hypothesis ranking.
- We develop an automated agentic framework to extract essential components—research questions, background surveys, inspirations, and hypotheses—from scientific papers, enabling the scalable and contamination-resistant construction of benchmarks.
- We conduct a comprehensive analysis of LLM performance using our benchmark, presenting the first large-scale study on out-of-distribution (OOD) inspiration retrieval. Our findings demonstrate that LLMs can effectively retrieve inspirations beyond established associations, positioning LLMs as "research hypothesis mines" capable of generating novel scientific insights at scale with minimal human involvement.

2 Related Work

2.1 LLMs for Scientific Discovery

CoLM (Yang et al., 2024a) investigates generative inductive reasoning, which is to propose (commonsense) hypothesis from observations. SciMON (Wang et al., 2024) introduces literature-based discovery, showing how retrieved concepts aid hypothesis composition. MOOSE (Yang et al., 2024b) and MOOSE-Chem (Yang et al., 2024c) find that most hypotheses in social science and chemistry can be seen as emerging from a research background and several inspirations. In addition, many researchers have the belief that "An idea is nothing more nor less than a new combination of old elements" (Young, 1975; Kumar et al., 2024). Yang et al. (2024c) further decompose the hypothesis formulation process into a sufficient set of sub-tasks: inspiration retrieval, hypothesis composition, and hypothesis ranking.

2.2 Benchmarking LLMs

Most existing benchmarks assess the general intelligence of LLMs (Chiang et al., 2024; Ni et al., 2024). IdeaBench (Guo et al., 2024) is designed for biomedical idea generation and does not evaluate LLMs based on the full set of sub-tasks in scientific discovery. Their focus is limited to generating hypotheses from background knowledge rather than retrieving and integrating inspirations. Additionally, their reference extraction relies on rule-based methods, which are less accurate than the LLM-based agentic framework. Moreover, their benchmark is restricted to the biomedical domain, whereas ours spans 12 disciplines. DiscoveryBench (Majumder et al., 2024) and ScienceAgentBench (Chen et al., 2024) identify and pick specific discovery-relevant tasks (e.g., write a specific code) from 20 papers and 44 papers correspondingly. They do not analyze the fundamental decomposition of the scientific discovery task itself.

3 Benchmark Construction

3.1 Preliminary

Yang et al. (2024c) propose a fundamental assumption that a majority of chemistry and material science hypotheses can originate from a research background and several inspirations. Here research background refers to a research question and/or a background survey; inspiration is a piece of knowledge, and it can be also represented by a research paper that discusses this piece of knowledge. They propose this assumption based on extensive discussions with domain experts, and the cognitive science findings that *creative ideas often result from the cohesive association of two (or more) seemingly unrelated pieces of knowledge* (Koestler, 1964; Benedek et al., 2012; Lee & Chung, 2024). Denoting background knowledge as b, inspiration knowledge as i, and hypothesis as h, this assumption can be represented in Equation 1:

$$h = f(b, i_1, \ldots, i_k) \tag{1}$$

Based on this assumption, they decompose the research hypothesis formulation process into a sufficient set of sub-tasks, which are (1) retrieving inspirations based on the research background, (2) properly mixing the research background information with the retrieved inspirations to compose research hypotheses, and (3) ranking the composed hypotheses to provide one with the highest confidence. This process can be represented in Equation 2 and Equation 3. Here I represents the full literature corpus to retrieve inspiration.

$$P(h \mid b) \approx \prod_{j=1}^{k} P(i_j \mid b, h_{j-1}, I) \cdot P(h_j \mid b, i_j, h_{j-1}) \tag{2}$$

$$H = \{h_1, h_2, \ldots, h_n \mid \hat{R}(h_i) > \hat{R}(h_{i+1}) \text{ for all } i\} \tag{3}$$

The cognitive science findings are not limited to any single discipline. For example, Yang et al. (2024b) shows that this assumption can also be leveraged in social science disciplines to generate

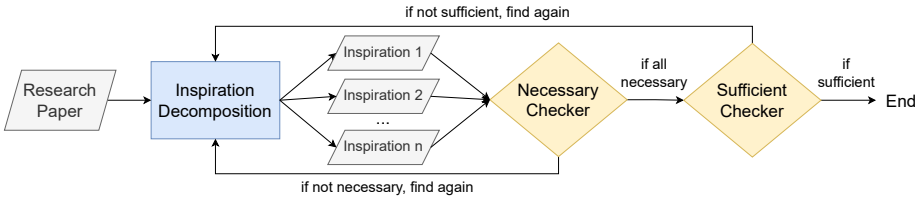

Figure 1: Overview of the inspiration retrieval framework.

high-quality research hypothesis. After extensive discussions with researchers in other disciplines such as Physics, Biology, Earth Science, Astronomy, and Math, we find that this assumption is largely and widely true across disciplines.

Based on this observation, we construct this benchmark collecting papers across 12 disciplines in top-ranked venues, develop an agentic framework to automatically extract each paper's research background, inspirations, and hypothesis (§ 3.2), discuss how negative inspiration papers are selected to evaluate LLMs' performance on inspiration retrieval (§ 3.3), and present expert evaluation on the extracted information to illustrate the quality of the benchmark (§ 3.4).

Different from directly assigning a score to each hypothesis and ranking the hypotheses based on their scores (Equation 3), this benchmark adopts a pairwise evaluation for ranking (Equation 4), since pairwise evaluation is widely reported as more robust and reliable (Si et al., 2024). $R(h_i, h_{i+1}) = h_i$ represents h_i is selected as a better hypothesis.

$$H = \{h_1, h_2, \ldots, h_n \mid R(h_i, h_{i+1}) = h_i \text{ for all } i\} \tag{4}$$

3.2 LLM-Based Agentic Framework

We develop a LLM-based agentic framework to automatedly extract the research question, background survey, inspirations, and hypothesis.

The extraction of research question, background survey, and hypothesis is relatively straightforward for LLMs. Specifically, we carefully design prompts and adopt iterative self-refine (Madaan et al., 2023) to extract them. The background survey is summarized based on the information in the introduction section and the related work section.

The extraction of inspirations is not so straightforward compared to the other components. Figure 1 shows the inspiration extraction framework. We have simplified its design as much as possible, retaining only the essential components to ensure efficiency and accuracy.

The inspirations are mostly described in the introduction section, usually starting with "Motivated by", and can be also described in the related work and methodology sections. As shown in Figure 1, given the full passage of a research paper, the "inspiration decomposition" module first iteratively extracts several potential inspirations. Here each potential inspiration is represented by the title and abstract of a referenced research paper. Therefore after the "inspiration decomposition" module suggests the title of a referred paper as inspiration, we use Semantic Scholar and Crossref to find the abstract of the referred paper to compose an inspiration. Then the "necessary checker" module examines whether each extracted inspiration is needed to formulate the hypothesis and not redundant, and the "sufficient checker" is to check whether all necessary inspirations have been extracted to be enough to be possible to formulate the hypothesis. Here by "enough", we mean the information in the research question, background survey, and the inspirations can cover the information scope of the hypothesis.

To prevent data contamination, we apply the agentic framework only to papers published in 2024 or later, thereby minimizing overlap with the pretraining data of LLMs. The cutoff dates for each model's pretraining data are summarized in Table 8.

3.3 Negative Inspiration Selection

Although ground-truth inspirations can be extracted, we need negative inspirations to calculate the performance of LLMs on inspiration retrieval. Here our goal is to provide an in-depth analysis on the inspiration retrieval ability by carefully composing a negative inspiration paper set for each paper in the benchmark.

Specifically, for each paper in the benchmark, we collect three levels of negative inspiration papers, based on their distance to the benchmark paper. The first-level is papers that are cited and referred to by the benchmark paper, or papers that have high semantic similar titles to the benchmark paper. For each benchmark paper, we collect 100 citation-adjacent papers with Crossref API, and 50 semantic adjacent papers with Semantic Scholar API. The second-level are papers that are in the same discipline with the benchmark paper, and the third-level are papers that belong to completely different disciplines (randomly selected). We randomly collect 2000 papers for each discipline by Web of Science, which can be used for both the second-level and the third-level. During experiment, for each benchmark paper, we randomly select 25 negative inspiration papers from each of the distance level to compose the negative inspiration set.

The purpose of the three-level design is two-fold. Firstly, real inspiration can come from each of the levels. By the splitting, LLM's preference in terms of distance can be analyzed. Secondly, the incorporation of closely related papers makes the negative inspiration papers non-trivial: we find that LLMs tend to select papers that are close to the benchmark paper. If the negative papers are only from irrelevant disciplines, then the retrieval success rate will be very high and less meaningful.

3.4 Expert Evaluation

We invited five PhD students from diverse disciplines—Physics (1), Chemistry (2), Materials Science (1), and Astronomy (1)—to evaluate the accuracy of our decomposition framework. Specifically, we randomly sampled 62 papers from the benchmark dataset, each accompanied by its extracted research question, background survey, inspirations, and hypothesis, and presented them in the form of a questionnaire. An example of the questionnaire is provided in Appendix A.2.

For each paper, the experts first read the full text of the original research paper and then assessed the accuracy of the extracted components. Each inspiration was evaluated for its necessity, while the entire set of inspirations was assessed for its sufficiency in supporting the research hypothesis.

Overall, the evaluation identified five cases with issues in inspiration identification (three) or hypothesis extraction (two), and six minor issues in research question extraction. The decomposition accuracy was 91.9% considering only major issues and 82.3% when including all issues.

4 Experiments

4.1 Inspiration Retrieval

For each benchmark paper, with the extracted research question, groundtruth inspirations, and the negative inspirations, we can calculate the accuracy of an LLM to retrieve the groundtruth inspiration with the research question.

During the experiment, for each benchmark paper, we prepare an inspiration candidate set consisting of 75 papers, including 2–3 groundtruth inspiration papers and about 25 negative inspiration papers from each distance level. Each paper is represented by its title and abstract. The retrieval is performed in several rounds, where in each round, the inspiration candidate set is randomly split into several groups, where each group contains 15 papers. Then LLM is instructed to select the top 3 of each group that it thinks can best serve as inspirations for the background question. In a new round, the selected papers are combined into a new inspiration candidate set and split into groups again.

Following this iterative selection process, only 20% (15 out of 75) of the papers are retained after the first round, and only 4% (3 out of 75) remain after the second round. Thus, the LLM is tasked with selecting 3 papers from an initial pool of 75. We use Hit Ratio as the evaluation metric, it is calculated as the number of groundtruth inspiration papers selected by the LLM divided by the total number of inspiration candidates.

(a) The accuracy (%) of LLMs in retrieving the groundtruth inspiration while only **20%** of inspiration candidates are selected.

Model	Cell	Chem	ETS	MS	Phys	EGS	EVS	BL	BS	Law	Math	A	Overall
Llama-3.2-1B	34.65	34.80	32.57	30.26	30.25	34.75	35.43	33.21	41.09	29.74	36.22	30.10	33.68
Llama-3.1-8B	74.08	78.00	79.69	74.54	76.75	84.56	75.20	75.81	80.00	65.95	75.59	68.37	75.92
Qwen Turbo	74.37	77.20	80.08	72.69	75.80	88.03	78.35	74.01	82.18	67.24	74.80	66.84	76.17
GPT-4o Mini	76.06	83.20	82.76	77.49	81.53	89.96	79.92	70.76	84.00	70.69	74.80	71.94	78.74
Gemini 2.0 FT	74.65	79.60	80.84	73.43	78.34	90.35	76.77	75.09	85.09	**80.17**	76.38	**77.55**	78.89
Gemini 2.0 Flash	75.77	76.40	85.82	75.28	79.94	91.89	75.98	75.09	86.91	78.02	76.77	71.94	79.24
Qwen Plus	79.15	82.00	82.76	75.28	80.57	91.12	81.10	76.53	84.73	75.00	79.53	73.98	80.27
DeepSeek-V3	80.00	83.60	85.44	76.01	79.94	91.51	79.53	76.90	86.91	75.86	77.56	73.98	80.74
Claude 3.5 Haiku	**80.56**	85.20	85.06	77.86	79.94	90.35	83.07	75.81	87.27	70.69	77.56	75.51	80.89
Llama-3.1-70B	78.31	84.00	84.67	80.07	80.25	89.58	81.10	**79.42**	86.91	75.43	77.95	75.51	81.18
Claude 3.5 Sonnet	78.31	78.40	85.06	76.75	81.53	91.51	**85.04**	77.62	**88.00**	77.59	79.53	**77.55**	81.43
GPT-4o	80.00	**87.20**	**89.27**	**80.81**	**84.39**	**93.05**	81.89	77.98	87.64	79.74	**83.07**	75.00	**83.43**

(b) The accuracy (%) of LLMs in retrieving the groundtruth inspiration while only **4%** of inspiration candidates are selected.

Model	Cell	Chem	ETS	MS	Phys	EGS	EVS	BL	BS	Law	Math	A	Overall
Llama-3.2-1B	10.70	11.60	12.26	9.59	11.15	8.49	14.57	13.00	17.09	12.50	11.42	10.71	11.91
Llama-3.1-8B	32.39	38.00	40.61	31.37	32.80	59.85	36.61	28.52	55.64	28.88	36.22	34.69	37.87
Gemini 2.0 FT	31.27	41.20	40.61	30.63	32.48	71.04	39.37	33.57	59.64	37.07	34.65	33.16	40.18
GPT-4o Mini	30.42	43.60	41.00	34.69	33.44	66.80	40.16	28.88	64.73	32.76	37.80	35.71	40.59
Qwen Turbo	35.49	42.40	42.15	33.95	35.03	66.80	43.31	33.21	61.45	29.74	36.61	34.69	41.21
Gemini 2.0 Flash	31.55	38.80	44.06	34.32	34.39	74.52	37.40	32.49	64.00	37.50	37.80	32.65	41.46
Claude 3.5 Sonnet	36.34	41.20	42.91	30.63	36.31	67.57	40.55	34.30	63.64	34.91	37.40	33.67	41.62
Qwen Plus	36.06	47.20	45.21	33.58	34.39	72.97	43.31	35.38	64.36	34.91	39.37	36.22	43.43
Claude 3.5 Haiku	41.13	**48.40**	45.98	34.69	33.44	69.88	44.09	34.30	64.00	37.93	38.19	**41.33**	44.28
DeepSeek-V3	38.87	46.00	44.06	36.90	**36.62**	**75.29**	41.73	**40.07**	65.45	36.64	38.58	37.76	44.78
Llama-3.1-70B	**41.41**	44.00	**47.51**	36.90	34.39	70.66	**45.28**	37.18	65.45	**39.22**	38.19	39.29	44.87
GPT-4o	39.44	46.40	47.13	**38.38**	35.35	**75.29**	44.88	38.63	**65.82**	**39.22**	**40.16**	38.78	**45.65**

Table 2: Performance of LLMs in hypothesis retrieve task. Gemini 2.0 FT=Gemini 2.0 Flash Thinking; Chem=Chemistry, ETS=Earth Science, MS=Material Science, Phys=Physics, EGS=Energy Science, EVS=Environmental Science, BL=Biology, BS=Business, A=Astronomy.

Model	Distance Level 1		Distance Level 2		Distance Level 3	
	(top 20%)	(top 4%)	(top 20%)	(top 4%)	(top 20%)	(top 4%)
Llama-3.2-1B	23.57%	6.33%	15.52%	2.93%	14.46%	2.85%
Qwen Turbo	52.72%	12.05%	9.45%	1.11%	4.46%	0.34%
Claude 3.5 Sonnet	53.96%	10.15%	10.16%	0.70%	2.40%	0.13%
Llama-3.1-8B	53.69%	11.17%	10.65%	0.77%	2.94%	0.14%
Gemini 2.0 Flash Thinking	54.49%	10.59%	10.34%	0.58%	2.24%	0.11%
GPT-4o	54.90%	10.02%	9.84%	0.47%	2.09%	0.09%
Llama-3.1-70B	55.32%	10.04%	9.82%	0.55%	2.16%	0.09%
DeepSeek-V3	55.74%	10.22%	9.80%	0.43%	1.79%	0.07%
GPT-4o Mini	55.90%	10.67%	9.54%	0.47%	2.12%	0.09%
Claude 3.5 Haiku	55.70%	10.19%	9.51%	0.49%	2.00%	0.07%
Gemini 2.0 Flash	55.91%	10.63%	9.63%	0.42%	2.03%	0.09%
Qwen Plus	56.11%	10.57%	9.52%	0.50%	2.16%	0.16%

Table 3: Analysis of negative inspiration retrieval in the inspiration retrieval task. Each value represents the average percentage of negative inspirations retrieved across three distance levels, under two settings where only 20% and 4% of the candidate inspirations are selected, respectively.

The Hit Ratio results are presented in Table 2. The values in "Overall" column represent averages across 12 disciplines. Overall, LLMs demonstrate surprisingly high retrieval accuracy.

In the first selection round, where 20% of the papers are retained, most LLMs successfully identify around 80% of the groundtruth inspiration papers. Even in the final round, where only 3 papers are selected from the initial set of 75, most LLMs maintain an accuracy exceeding 40%, with GPT-4o remaining the best model at 45.65%. These findings show that LLMs can identify papers that were not known as relevant but have the potential to contribute to solving the background question. We

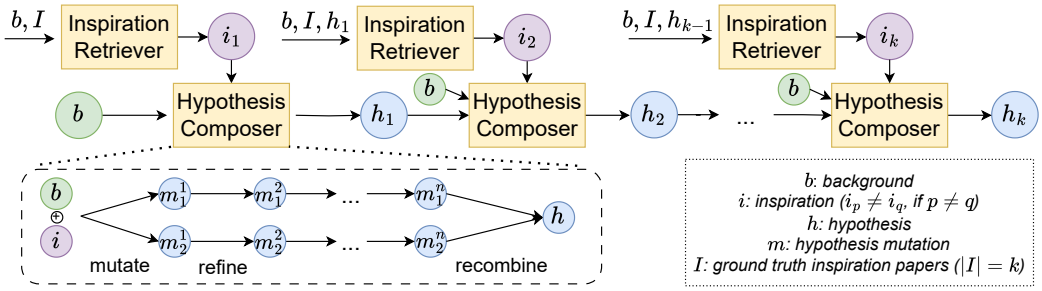

Figure 2: Overview of the hypothesis composition process.

think this high retrieval accuracy stems from LLMs' pretraining process, during which they may have already captured latent knowledge associations that are not yet recognized by scientists.

Table 2 also indicate the scaling law of LLMs for inspiration retrieval: the inspiration retrieval ability grows up very fast before and during 8B parameters, while stuck in a bottleneck at around 70B parameters. No matter how the LLMs are trained with different strategies, they seem to be bottlenecked at the same performance.

Furthermore, we analyze the Hit Ratio of negative inspirations across the three distance levels. The results are presented in Table 3. It indicates that regardless of the percentage of the papers selected, the closer an inspiration is to the benchmark paper, the higher the probability it is selected as an inspiration. We attribute it to two reasons. Firstly, statistically closer papers objectively have a better chance of being an inspiration; Secondly, if certain papers often appear together in the training data, the LLM may see them as more possibly contributing to each other.

Model	Cell	Chem	ETS	MS	Phys	EGS	EVS	BL	BS	Law	Math	A	Overall
Claude 3.5 Haiku	40.42	40.87	38.71	46.75	45.00	45.34	48.00	46.15	35.14	37.85	43.59	34.29	42.56
Llama-3.1-8B	44.58	47.83	42.78	46.04	45.05	44.30	46.47	47.37	44.21	47.58	48.21	45.14	45.68
Gemini 2.0 FT	45.67	39.79	48.48	47.22	48.77	49.24	48.57	48.02	41.47	47.03	42.81	40.00	46.30
Gemini 2.0 Flash	46.25	45.63	48.64	51.63	47.97	51.47	49.41	48.77	47.03	55.91	**56.24**	49.71	50.15
Llama-3.1-70B	46.67	49.86	50.83	51.53	50.60	50.61	52.10	54.36	49.47	53.94	51.11	49.14	50.92
GPT-4o Mini	46.67	49.42	50.91	52.63	53.82	53.33	54.86	54.36	46.92	56.97	52.48	53.14	52.47
Qwen Turbo	52.92	51.45	49.55	51.06	52.64	50.97	52.57	56.92	**53.16**	55.76	55.38	53.14	52.71
GPT-4o	55.00	53.04	54.09	53.95	53.82	52.97	53.14	55.38	46.15	53.99	54.53	52.57	53.37
DeepSeek-V3	52.78	52.27	53.18	54.25	54.91	53.91	53.71	56.32	50.27	55.15	52.14	53.71	53.79
Qwen Plus	**60.00**	**53.72**	**57.27**	**56.63**	**58.14**	**56.63**	**60.57**	**58.97**	51.05	**62.19**	55.90	**56.57**	**57.46**

Table 4: Performance of LLMs in hypothesis composition task. Each number represents the normalized performance of LLMs in composing hypothesis. Gemini 2.0 FT=Gemini 2.0 Flash Thinking; Chem=Chemistry, ETS=Earth Science, MS=Material Science, Phys=Physics, EGS=Energy Science, EVS=Environmental Science, BL=Biology, BS=Business, A=Astronomy.

4.2 Hypothesis Composition

With the retrieved inspirations, the next step is to associate them with the research background to compose research hypothesis. Figure 2 shows the framework we use for the hypothesis generation process. This framework strictly follows Equation 2, and is designed to be as simple as possible, avoiding unnecessary components. We use the evolutionary unit (Yang et al., 2024c) to associate the research background (b) and inspiration (i), shown in the bottom-left rectangle in Figure 2. Specifically, "mutate" means creating different ways to combine b and i together and "recombine" tries to keep the merits of different combination ways to compose a final hypothesis. The prompts for mutate, refine, and recombine are provided in Appendix A.5. In this step, we only measure the LLM's ability on hypothesis composing. For the LLM-generated hypotheses to be comparable with the groundtruth hypothesis for evaluation, here I represents the groundtruth inspiration papers (usually 2 to 3 papers), while "inspiration retriever" module each time retrieve only one inspiration, and will not retrieve the same inspiration again.

The detailed scoring criteria is shown in Appendix A.3, where we use a 6-point Likert scale (from 0 to 5) to measure whether the generated hypothesis has covered the key points in the groundtruth

Model	Cell	Chem	ETS	MS	Phys	EGS	EVS	BL	BS	Law	Math	A	Overall
Llama-3.1-70B	36.94	35.57	30.57	37.71	43.35	47.18	36.02	43.11	41.63	46.09	30.73	25.40	38.06
GPT-4o Mini	42.25	39.94	34.39	42.98	39.78	43.78	40.63	43.72	45.03	42.24	32.67	31.50	40.13
Gemini 2.0 Flash	43.73	44.38	35.95	51.86	54.63	55.16	40.98	44.00	46.88	48.31	38.24	35.75	45.11
Qwen Turbo	46.42	45.11	42.88	48.74	45.61	46.40	45.26	49.20	50.92	49.27	37.15	37.62	45.48
Gemini 2.0 FT	43.52	44.96	36.88	52.81	54.08	54.95	42.27	44.53	46.15	48.09	37.80	38.40	45.49
Qwen Plus	46.00	46.00	41.72	49.35	50.64	49.11	44.80	46.93	43.36	45.43	40.16	41.97	45.56
Claude 3.5 Haiku	48.15	46.88	45.55	52.45	54.10	52.48	48.83	48.06	51.23	52.93	44.49	40.27	48.86
Llama-3.1-8B	55.48	54.20	55.90	56.60	54.35	55.48	55.91	56.71	54.69	55.55	55.60	55.49	55.65
GPT-4o	60.75	60.99	53.24	61.69	61.34	61.20	60.52	64.11	64.67	61.14	52.60	51.80	59.60
DeepSeek-V3	**80.88**	**82.03**	78.85	**83.63**	80.82	81.47	**83.98**	81.77	83.48	80.69	**76.78**	75.88	80.99
Claude 3.5 Sonnet	80.23	80.83	**80.93**	83.20	**84.33**	**84.72**	82.63	**82.48**	**84.87**	**81.81**	76.20	**76.51**	**81.59**

Table 5: Performance of LLMs in hypothesis ranking task. Each number represents the accuracy (%) of LLMs in ranking ground-truth hypothesis among negative hypothesis. Chem=Chemistry, ETS=Earth Science, MS=Material Science, Phys=Physics, EGS=Energy Science, EVS=Environmental Science, BL=Biology, BS=Business, A=Astronomy.

hypothesis. To compute the generation accuracy, we normalize the average score by dividing it by the maximum possible score (5). The final results are summarized in Table 4. Table 4 shows that (a) all LLMs preserve a certain kind of ability to associate the research background and inspirations to compose hypothesis; (b) the hypothesis composition task remains challenging, as none of the models achieve consistently high performance.

4.3 Hypothesis Ranking

In this section, we evaluate the ability of LLMs to rank hypotheses pairwisely. Specifically, based on the hypothesis generation method in § 4.2, we use the top-ranked negative inspirations and background question to construct a set of negative hypotheses. From this set, we randomly sample 5 negative hypotheses. Additionally, with subsets of groundtruth inspirations and the research question, we use the hypothesis composition framework to generate another set of negative hypotheses. In this set, we randomly sample 10 as negative hypotheses for ranking. As a result, for each benchmark paper, we compose a set of 16 hypotheses, including one groundtruth one and 15 negative ones for ranking. To evaluate ranking performance, we use the groundtruth hypothesis to pairwisely compare with each of the 15 negative ones. The prompt for pairwise ranking is provided in the Appendix A.4.

Accuracy is used as the evaluation metric, which is calculated as the proportion of correct pairwise rankings out of 15 comparisons. During the pairwise evaluation, we find that many LLMs have strong position bias: they largely prefer the first hypothesis than the second. To avoid this bias, for each hypothesis pair, we compare them twice with reverse positions, and the results are averaged.

Table 5 presents the ranking accuracy of each LLM. This ranking results indicate a different scaling law with the scaling law we find in the inspiration retrieval task: more parameters and better pretraining strategies can significantly improve over the hypothesis ranking task, while might lead to less improvements in the inspiration retrieval task.

Model	✗✗	✓✗	✓✓
GPT-4o Mini	33.83	64.83	1.33
Qwen Plus	25.00	69.33	5.67
Llama-3.1-8B	2.50	**91.67**	5.83
Llama-3.1-70B	**52.67**	39.17	8.17
Gemini 2.0 Flash	35.50	51.67	12.83
Claude 3.5 Haiku	28.17	58.17	13.67
Gemini 2.0 FT	36.50	49.67	13.83
Qwen Turbo	39.33	45.67	15.00
GPT-4o	11.50	61.50	27.00
DeepSeek-V3	1.74	21.83	76.44
Claude 3.5 Sonnet	3.17	19.17	**77.67**

Table 6: Analysis of *position bias* in hypothesis ranking task. Each hypothesis pair is compared twice, with three possible outcomes: both wrongly ranked (✗✗); one right one wrong (✓✗); both rightly ranked (✓✓). Numbers are averaged percentages (%).

Table 6 analyzes the position bias problem in the hypothesis ranking task. Specifically, each hypothesis pair is compared twice, with three possible outcomes, and the table shows the averaged percentage of each outcome. It shows that many LLMs are hugely influenced by position bias (e.g., Llama-3.1-8B has 91.67% of the time reaching self-contradictory results), and some are less influenced (e.g., Claude 3.5 Sonnet only 19.17%). The large proportion of self-contradictory

results might be one of the main reasons that many LLMs in Table 5 reach a ranking accuracy around 50%.

5 Analysis

5.1 LLMs as Research Hypothesis Mines

The results show that (1) LLMs can already capture many unknown association of knowledge so to retrieve inspirations relatively accurately; (2) Given groundtruth inspirations, many LLMs can compose a hypothesis that capture at least a subset of main innovations in the groundtruth one; (3) with improved scale and better training strategies, LLMs' performance on hypothesis ranking can grow very quickly, and we have not seen the limit.

Also considering that the three tasks of inspiration retrieval, hypothesis composing, and hypothesis ranking are a sufficient set of sub-tasks of scientific discovery, it might indicate that given only a research background, LLMs can already discover hypothesis autonomously: it can screen lots of inspiration candidates to choose the good ones autonomously, associating the research background with the good inspirations autonomously, and autonomously rank those composed hypotheses to provide the scientists with the best ones.

In short, the only input scientists need to provide such a copilot is the research background and enough papers to serve as inspiration. In this view, LLMs can be regarded as research hypothesis mines: models with stronger performance on the three fundamental tasks of scientific discovery represent richer mines, while more inference compute corresponds to deploying more miners.

5.2 The Bottleneck Towards Automated Discovery

Across the three sub-tasks, the inspiration retrieval task appears to be the most challenging. Although performance improves rapidly even with relatively small models (e.g., 8B parameters), it quickly plateaus. Scaling up model size or enhancing pretraining strategies yields only marginal gains in retrieval performance.

We attribute this to the nature of the task: inspiration retrieval fundamentally requires deep domain understanding, which is primarily acquired during the pretraining phase as the model ingests millions of papers. In other words, success in this task may rely more on the "intuition" developed through large-scale pretraining rather than the enhanced reasoning abilities typically refined during post-training. Understanding the fundamental mechanisms behind how LLMs retrieve inspirations may help address a key bottleneck in advancing toward fully automated scientific discovery.

6 Conclusion

We introduced the first large-scale benchmark for evaluating LLMs in scientific discovery in terms of a sufficient set of sub-tasks, covering inspiration retrieval, hypothesis composition, and hypothesis ranking. Our benchmark, ResearchBench, spans 12 scientific disciplines and utilizes an automated, LLM-based agentic framework, significantly contributing to scalable and contamination-resistant dataset construction.

Our evaluation shows that LLMs achieve promising results in inspiration retrieval, effectively surfacing novel, out-of-distribution knowledge associations. They also demonstrate moderate capabilities in hypothesis composition and ranking tasks; however, performance in these two tasks indicates considerable room for improvement. Notably, we identify inspiration retrieval as a key bottleneck, where accuracy quickly plateaus with increasing model scale, underscoring the need for deeper domain-specific understanding primarily acquired during pretraining rather than fine-tuning.

These findings point toward a promising pathway, positioning LLMs as potential "research hypothesis mines". By systematically addressing the bottlenecks identified, LLMs hold great promise for becoming powerful tools capable of autonomously discovering high-quality scientific hypotheses, ultimately facilitating a paradigm shift towards fully automated scientific exploration.

References

Mathias Benedek, Tanja Könen, and Aljoscha C Neubauer. Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):273, 2012.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *CoRR*, abs/2410.05080, 2024. doi: 10.48550/ARXIV.2410.05080. URL https://doi.org/10.48550/arXiv.2410.05080.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3MW8GKNyzI.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. *CoRR*, abs/2411.02429, 2024. doi: 10.48550/ARXIV.2411.02429. URL https://doi.org/10.48550/arXiv.2411.02429.

Arthur Koestler. The act of creation. *London: Hutchinson*, 1964.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. Can large language models unlock novel scientific research ideas? *CoRR*, abs/2409.06185, 2024. doi: 10.48550/ARXIV.2409.06185. URL https://doi.org/10.48550/arXiv.2409.06185.

Byung Cheol Lee and Jaeyeon Chung. An empirical investigation of the impact of chatgpt on creativity. *Nature Human Behaviour*, pp. 1–9, 2024.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. *CoRR*, abs/2407.01725, 2024. doi: 10.48550/ARXIV.2407.01725. URL https://doi.org/10.48550/arXiv.2407.01725.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 279–299. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.18. URL https://doi.org/10.18653/v1/2024.acl-long.18.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209–225, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.13/.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13545–13565, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.804. URL https://aclanthology.org/2024.findings-acl.804/.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *ICLR 2025*, 2024c.

James Webb Young. A technique for producing ideas, 1975.

A Appendix

A.1 Prompt for Retrieving Inspirations

You are helping with the scientific hypotheses generation process. Given a research question, the background and some of the existing methods for this research question, and several top-tier publications (including their title and abstract), try to identify which publication can potentially serve as an inspiration for the background research question so that combining the research question and the inspiration in some way, a novel, valid, and significant research hypothesis can be formed. The inspiration does not need to be similar to the research question. In fact, probably only those inspirations that are distinct with the background research question, combined with the background research question, can lead to a impactful research hypothesis. The reason is that if the inspiration and the background research question are semantically similar enough, they are probably the same, and the inspiration might not provide any additional information to the system, which might lead to a result very similar to a situation that no inspiratrions are found. An example is the backpropagation of neural networks. In backpropagation, the research question is how to use data to automatically improve the parameters of a multi-layer logistic regression, the inspiration is the chain rule in mathematics, and the research hypothesis is the backpropagation itself. In their paper, the authors have conducted experiments to verify their hypothesis. Now try to select inspirations based on background research question. The background research question is: ", "The introduction of the previous methods is:", "The potential inspiration candidates are: ", "Now you have seen the background research question, and many potential inspiration candidates. Please try to identify which three literature candidates are the most possible to serve as the inspiration to the background research question? Please name the title of the literature candidate, and also try to give your reasons.

A.2 Guideline Format for Expert Checking

Titile:

Background question decomposed by automated framework:

Whether the background question correct?

<Reply fill in here. Required a detailed analysis>

ground-truth hypothesis decomposed by automated framework:

Whether the ground-truth hypothesis accurately reflect the main proposal of the paper?

<Reply fill in here. Required a detailed analysis>

Inspiration paper 1 title: Relation between the inspiration 1 paper and the decomposed paper:

Whether the collected inspiration paper 1 compose of a set of necessary conditions to reach to the ground-truth hypothesis?

<Reply fill in here. Required a detailed analysis>

Inspiration paper 2 title: Relation between the inspiration 2 paper and the decomposed paper:

Whether the collected inspiration paper 2 compose of a set of necessary conditions to reach to the ground-truth hypothesis?

<Reply fill in here. Required a detailed analysis>

Inspiration paper 3 title: Relation between the inspiration 3 paper and the decomposed paper:

Whether the collected inspiration paper 3 compose of a set of necessary conditions to reach to the ground-truth hypothesis?

<Reply fill in here. Required a detailed analysis>

Whether the collected inspirations paper compose of a set of sufficient conditions to reach to the coarse-grained hypothesis?

<Reply fill in here. Required a detailed analysis>

A.3 Prompt for Evaluating Generated Hypothesis

You are helping to evaluate the quality of a proposed research hypothesis by a phd student. The groundtruth hypothesis will also be provided to compare. Here we mainly focus on whether the proposed hypothesis has covered the key points of the ground-truth hypothesis. You will also be given a summary of the key points in the ground-truth hypothesis for reference. The evaluation criteria is called 'Matched score', which is in a 6-point Likert scale (from 5 to 0). Particularly, 5 points mean that the proposed hypothesis (1) covers three key points (or covers all the key points) in the ground-truth hypothesis, where every key point is leveraged nearly identically as in the ground-truth hypothesis, and (2) does not contain any extra key point(s) that is redundant, unnecessary, unhelpful, or harmful; 4 points mean that the proposed hypothesis (1) covers three key points (or covers all the key points) in the ground-truth hypothesis, where every key point is leveraged nearly identically as in the ground-truth hypothesis, and (2) but also contain extra key point(s) that is redundant, unnecessary, unhelpful, or harmful; 3 points mean that the proposed hypothesis (1) covers two key points in the ground-truth hypothesis, where every key point is leveraged nearly identically as in the ground-truth hypothesis, (2) but does not cover all key points in the ground-truth hypothesis, and (3) might or might not contain extra key points; 2 points mean that the proposed hypothesis (1) covers one key point in the ground-truth hypothesis, and leverage it nearly identically as in the ground-truth hypothesis, (2) but does not cover all key points in the ground-truth hypothesis, and (3) might or might not contain extra key points; 1 point means that the proposed hypothesis (1) covers at least one key point in the ground-truth hypothesis, but all the covered key point(s) are used differently as in the ground-truth hypothesis, and (2) might or might not contain extra key points; 0 point means that the proposed hypothesis does not cover any key point in the ground-truth hypothesis at all. Usually total the number of key points a ground-truth hypothesis contain is less than or equal to three. Please note that the total number of key points in the ground-truth hypothesis might be less than three, so that multiple points can be given. E.g., there's only one key point in the ground-truth hypothesis, and the proposed hypothesis covers the one key point nearly identically, it's possible to give 2 points, 4 points, and 5 points. In this case, we should choose score from 4 points and 5 points, depending on the existence and quality of extra key points. 'Leveraging a key point nearly identically as in the ground-truth hypothesis means that in the proposed hypothesis, the same (or very related) concept (key point) is used in a very similar way with a very similar goal compared to the ground-truth hypothesis. When judging whether an extra key point has apparent flaws, you should use your own knowledge and understanding of that discipline to judge, rather than only relying on the count number of pieces of extra key point to judge. Importantly, we should focus on whether the fundamental key points match, rather than being influenced by how complex, sophisticated, or advanced the proposed methods appear. A hypothesis that introduces high-level techniques or intricate methodologies does not necessarily mean it is a disadvantage with the ground-truth hypothesis. The core concern is whether the essential key points are correctly captured and utilized.Please evaluate the proposed hypothesis based on the ground-truth hypothesis. The proposed hypothesis is: ", "The ground-truth hypothesis is: ", "The key points in the ground-truth hypothesis are: "

A.4 Prompt for Pairwise Ranking

You are assisting scientists with their research. Given a research question and two research hypothesis candidates proposed by large language models, your task is to predict which hypothesis is a better research hypothesis. By 'better', we mean the hypothesis is more valid and effective for the research question. Please note the following:

(1) Neither hypothesis has been tested experimentally. However, some large language model generated hypothesis might contain expected performance of the hypothesis. Well, just do not believe any of the descriptions of the expected performance or the effect of the hypothesis. Instead, only focus on the technical contents and predict which hypothesis you think will be more effective for the research question if tested in real experiments.

(2) You should remember that, here, we only focus on whether the general direction or major components of the hypothesis are more effective. Providing additional details or making the content more comprehensive is neither an advantage nor a disadvantage. More detailed and multifaceted strategies, additional complexity, and potential challenges are neither advantages nor disadvantages. What truly matters is the fundamental, intrinsic core idea. The research question is: <the background

Score	Criteria
5 Points	(1) Covers three key points (or all key points) in the ground-truth hypothesis, with each key point leveraged nearly identically to the ground-truth hypothesis. (2) Does not contain any extra key point that is redundant, unnecessary, unhelpful, or harmful.
4 Points	(1) Covers three key points (or all key points) in the ground-truth hypothesis, with each key point leveraged nearly identically to the ground-truth hypothesis. (2) However, it also contains extra key point(s) that are redundant, unnecessary, unhelpful, or harmful.
3 Points	(1) Covers two key points in the ground-truth hypothesis, with each key point leveraged nearly identically to the ground-truth hypothesis. (2) Does not cover all key points in the ground-truth hypothesis. (3) May or may not contain extra key points.
2 Points	(1) Covers one key point in the ground-truth hypothesis and leverages it nearly identically to the ground-truth hypothesis. (2) Does not cover all key points in the ground-truth hypothesis. (3) May or may not contain extra key points.
1 Point	(1) Covers at least one key point in the ground-truth hypothesis, but all the covered key points are used differently from the ground-truth hypothesis. (2) May or may not contain extra key points.
0 Points	The proposed hypothesis does not cover any key point in the ground-truth hypothesis.

Table 7: Scoring criteria for hypothesis evaluation.

question of this paper> Research hypothesis candidate 1 is: <the ground-truth hypothesis> Research hypothesis candidate 2 is: <the negative hypothesis>

Now, please predict which hypothesis you think will be more effective for the research question if tested in real experiments.

A.5 Prompts for Mutate, Refine, and Recombine

Prompt for mutation: You are helping with the scientific hypotheses generation process. We in general split the period of research hypothesis proposal into three steps. Firstly it's about the research background, including finding a good and specific background research question, and an introduction of the previous methods under the same topic; Secondly its about finding inspirations (mostly from literatures), which combined with the background research question, can lead to a impactful research hypothesis; Finally it's hypothesis generation based on the background research question and found inspirations. Take backpropagation as an example, the research question is how to use data to automatically improve the parameters of a multi-layer logistic regression with data, the inspiration is the chain rule in mathematics, and the research hypothesis is the backpropagation itself. nNow we have identified a good research question, an introduction of previous methods, and a core inspiration in a literature for this research question. The experts know that a proper mixture of these components will definitely lead to a valid, novel, and meaningful research hypothesis. In fact, they already have tried to mix them to compose some research hypotheses (that are supposed to be distinct from each other). Please try to explore a new meaningful way to combine the inspiration with the research background to generate a new research hypothesis that is distinct with all the previous hypotheses in terms of their main method. The new research hypothesis should ideally be novel, valid, ideally significant, and be enough specific in its methodology. Please note that here we are trying to explore a new meaningful way to leverage the inspiration along with the previous methods (inside or outside the introduction) to better answer the background research question, therefore the new research hypothesis should try to leverage or contain the key information or the key reasoning process in the inspiration, trying to better address the background research question. It means the new research hypothesis to be generated should at least not be completely irrelevant to the inspiration or background research question. In addition, by generating distinct hypothesis, please do not achieve it by simply introducing new concept(s) into the previous hypothesis to make the difference, but please focus on the difference on the methodology of integrating or leveraging the inspiration to give a better answer to the research question (in terms of the difference on the methodology, concepts can be introduced or deleted).

Prompt for refine: You are helping with the scientific hypotheses generation process. We in general split the period of research hypothesis proposal into four steps. Firstly it's about finding a good and specific background research question, and an introduction of the previous methods under the same topic; Secondly its about finding inspirations (mostly from literatures), which combined with the background research question, can lead to a impactful research hypothesis; Thirdly it's about finding extra knowledge that work along with the inspiration can lead to a more complete hypothesis. Finally it's hypothesis generation based on the background research question, the found inspirations, and the extra knowledge. Now we have identified a good research question, a core inspiration in a literature for this research question, and extra knowledge. With them, we have already generated a preliminary research hypothesis. We have also obtain feedbacks on the hypothesis from domain experts in terms of novalty, validity, significance, and clarity. With these feedbacks, please try your best to refine the hypothesis. Please note that during refinement, do not improve a hypothesis's significance by adding expectation of the performance gain of the method or adding description of its potential impact, but you should work on improving the method itself (e.g., by adding or changing details of the methodology).

Prompt for recombine: You are helping with the scientific hypotheses generation process. We in general split the period of research hypothesis proposal into three steps. Firstly it's about the research background, including finding a good and specific background research question, and an introduction of the previous methods under the same topic; Secondly its about finding inspirations (mostly from literatures), which combined with the background research question, can lead to a impactful research hypothesis; Finally it's hypothesis generation based on the background research question and found inspirations. Now we have identified a good research question, an introduction of previous methods, and a core inspiration in a literature for this research question. In addition, several experts have already come out of several different hypotheses on how to leverage the inspiration to generate a novel, valid, and significant research hypothesis for the background research question. Please find the bright parts in these hypotheses, leverage the bright parts from them, modify and combine the good parts of them to generate a better research hypothesis in terms of clarity, novelty, validness, and significance (ideally than any of the given hypotheses). It is not necessary to include methods from every given hypothesis, especially when it is not a good hypothesis. But in general you should try your best to benefit from every given hypothesis. In fact, a researcher has already tried to propose hypothesis based on these information, and we have obtained the feedback to his hypothesis, from another respectful researcher. Please try to leverage the feedback to improve the hypothesis, you can leverage all these provided information as your reference.

A.6 LLM Knowledge Cutoff Date

Model	Cutoff Date	Release Date
GPT-4o	Oct 2023	May 2024
GPT-4o Mini	Oct 2023	Jul 2024
Llama-3.1-8B	Dec 2023	Jul 2024
Llama-3.1-70B	Dec 2023	Jul 2024
Gemini 2.0 Flash	Jun 2024	Dec 2024
Gemini 2.0 FT	Jun 2024	Dec 2024
Claude 3.5 Sonnet	Apr 2024	Jun 2024
Claude 3.5 Haiku	Jul 2024	Oct 2024
Qwen Plus	\	Nov 2024
Qwen Turbo	\	Nov 2024
DeepSeek-V3	\	Dec 2024

Table 8: LLM's pretraining data cutoff date. Symbol '\' means the official cutoff date is not specified.

A.7 Limitation

While our benchmark and framework offer a scalable and contamination-resistant means of evaluating LLMs in scientific hypothesis generation, several limitations remain. First, the decomposition accuracy, though high, is not perfect, and minor errors in component extraction may affect downstream evaluations. Second, our expert validation is limited to select disciplines; extending the benchmark's

coverage and validation to more domains would further strengthen its generality. Lastly, although we use 2024 publications to reduce pretraining overlap, we cannot fully guarantee the absence of indirect data leakage via related works. We leave these challenges to future work, including more refined validation pipelines and expansion into underrepresented scientific fields.

A.8 Experiment Compute Resources

The construction of our benchmark and subsequent evaluation of LLMs were conducted using publicly accessible APIs, including `Gemini 2.0 Flash` and `GPT-4o-mini`. All experiments were run using pay-per-use endpoints without private fine-tuning or proprietary infrastructure. The end-to-end cost of benchmark construction and large-scale LLM evaluation across the three sub-tasks totaled approximately $4,000 USD. This highlights the feasibility of replicating or extending our study within a reasonable compute and budget envelope, making our framework accessible to a broad range of academic and industrial researchers.

A.9 Experiment Statistical Significance

While we do not conduct formal statistical significance testing for individual comparisons, our results are aggregated over a large and diverse benchmark comprising 1,386 scientific papers across 12 disciplines. This large-scale evaluation mitigates the impact of outlier cases and provides a stable estimate of model performance. The consistency of trends observed across tasks and domains suggests that the reported differences are meaningful and generalizable. Future work may incorporate bootstrap resampling or hypothesis testing to provide stronger statistical guarantees for fine-grained comparisons.

A.10 Broader Impact

This work introduces a scalable and transparent benchmark for evaluating LLMs in scientific hypothesis generation, supporting more systematic and interpretable assessment of their research capabilities. By covering 12 disciplines and relying on publicly available data, our framework promotes reproducibility and broad accessibility.

The agentic extraction pipeline also enables future extensions to newer literature, making the benchmark adaptable over time. While LLMs should not replace expert judgment, our structured decomposition encourages human-in-the-loop use, where models serve as tools for amplifying creativity rather than replacing it.

We hope this work fosters responsible development of LLM-based tools that support scientific innovation across domains.