Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

Abstract

Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation. Finally, we acknowledge that human judgements of novelty can be difficult, even by experts, and propose an end-to-end study design which recruits researchers to execute these ideas into full projects, enabling us to study whether these novelty and feasibility judgements result in meaningful differences in research outcome. [1]

1 Introduction

The rapid improvement of LLMs, especially in capabilities like knowledge and reasoning, has enabled many new applications in scientific tasks, such as solving challenging mathematical problems (Trinh et al., 2024), assisting scientists in writing proofs (Collins et al., 2024), retrieving related works (Ajith et al., 2024, Press et al., 2024), generating code to solve analytical or computational tasks (Huang et al., 2024, Tian et al., 2024), and discovering patterns in large text corpora (Lam et al., 2024, Zhong et al., 2023). While these are useful applications that can potentially increase the productivity of researchers, it remains an open question whether LLMs can take on the more creative and challenging parts of the research process.

We focus on this problem of measuring the *research ideation* capabilities of LLMs and ask: are current LLMs capable of generating novel ideas that are comparable to expert humans? Although ideation is only one part of the research process, this is a key question to answer, as it is the very first step to the scientific research process and serves as a litmus test for the possibility of autonomous research agents that create their own ideas. Evaluating expert-level capabilities of LLM systems is challenging (Bakhtin

[1] Interested researchers can sign up for this end-to-end study at: https://tinyurl.com/execution-study. We release our agent implementation and all human review scores at: https://github.com/NoviScl/AI-Researcher.

[*] The last two authors advised this project equally.

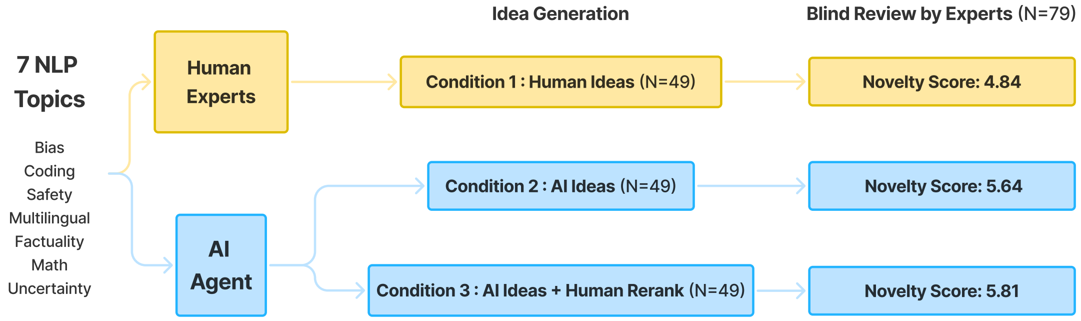

Figure 1: Overview of our study: we recruit 79 expert researchers to perform blind review of 49 ideas from each of the three conditions: expert-written ideas, AI-generated ideas, and AI-generated ideas reranked by a human expert. We standardize the format and style of ideas from all conditions before the blind review. We find AI ideas are judged as significantly more novel than human ideas ($p < 0.05$).

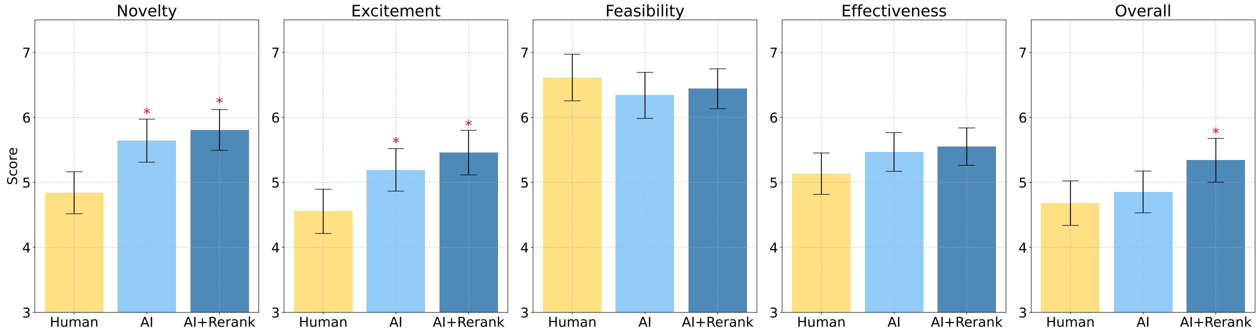

Figure 2: Comparison of the three experiment conditions across all review metrics. Red asterisks indicate that the condition is statistically better than the `Human` baseline with two-tailed Welch's t-tests and Bonferroni correction. All scores are on a 1 to 10 scale. More detailed results are in Section 5.

et al., 2022, Collins et al., 2024), and research ideation takes this to an extreme. Qualified expert researchers are difficult to recruit at scale, evaluation criteria can be highly subjective, and it is difficult for even the best experts to judge the quality of an idea (Beygelzimer et al., 2021, Simsek et al., 2024).

We address these challenges directly, recognizing that for important, high-stakes tasks like research ideation, there is no substitute for a large-scale expert evaluation. We design a carefully controlled comparison of human and LLM ideas that overcomes sample size and baseline problems present in earlier small-scale evaluation studies. Our study recruited a large pool of over 100 highly qualified NLP researchers to produce human baseline ideas and perform blind reviews of human and LLM ideas. To reduce the possibility that confounding variables affect our outcome measures, we enforce strict controls that standardize the styles of human and LLM ideas and match their topic distribution.

We compare our human expert baseline with a simple and effective LLM agent that incorporates retrieval augmentation and adopts recent ideas in inference-time scaling, such as overgenerating and reranking LM outputs. These measures allow us to make statistically rigorous comparisons between human experts and state-of-the-art LLMs (Figure 1).

Our evaluation-centric approach complements many recent methods-centric works that attempt to instantiate research agents (Baek et al., 2024, Li et al., 2024, Lu et al., 2024, Wang et al., 2024, Yang et al., 2024). The majority of these works rely on fast and lower-cost evaluation surrogates – either by decreasing the number of expert reviewers (Baek et al., 2024, Li et al., 2024, Wang et al., 2024, Yang et al., 2024), constraining the length and detailedness of the ideas (Wang et al., 2024, Yang et al., 2024), or relying on LLM-as-a-judge (Lu et al., 2024). They do not perform the large-scale human comparison studies that are needed to answer the motivating question of our work. Our work takes the opposite approach, performing a year-long and high-cost evaluation that provides human expert baselines and a standardized evaluation protocol to serve as a foundation for future follow-up studies and methods work.

Through nearly 300 reviews across all our conditions, we find that AI-generated ideas are judged as more novel than human expert ideas ($p < 0.05$), which holds robustly under multiple hypothesis correction and across different statistical tests. We find some signs that these gains are correlated with excitement and overall score, and may come at the slight expense of feasibility, but our study size did not have sufficient power to conclusively identify these effects (Figure 2).

Qualitative analysis of free-text responses in our review corroborates these findings on novelty and feasibility. Apart from evaluating the ideas, we also analyze the LLM agent, showing limitations and open problems – despite excitement about inference-time scaling of LLMs, we find that they lack idea diversity when we scale up idea generation, and they cannot currently serve as reliable evaluators.

2 Problem Setup

The central experiment of our work is a comparison of human- and LLM-generated ideas. While this goal is simple, there is no existing consensus on how to formulate the task of research ideation and evaluation, and we begin by defining the key aspects of our experiment design.

We think of research idea evaluation as consisting of three separate components: 1). the idea itself, generated in response to our instructions, 2). the writeup which communicates the idea, and 3). the evaluation of the writeup by experts. We outline our experiment design in each of these three parts with particular focus on potential confounders, such as the area of research, the format of a research idea, and the evaluation process.

Ideation Scope and Instructions Research ideas can take many different forms. They can be simple tricks to improve model performance, or they may be large-scale research programs that form the basis of a Ph.D. thesis. Any experiment on ideation must carefully balance the realisticness and interestingness of a research idea with the practical realities of eliciting ideas from a large population. In our case, these tradeoffs are even more pronounced, as we have designed our ideation experiments so that the resulting ideas can be executed by experts in a follow-up set of experiments.

These constraints have led us to study prompting-based NLP research as a testbed for our study. Prompting research has been popular in recent years of NLP and AI research (e.g., Chen et al., 2023, Diao et al., 2024, Madaan et al., 2023, Qin et al., 2024, Schulhoff et al., 2024, Si et al., 2023, Wang et al., 2023, Wei et al., 2022, Yao et al., 2023, Yasunaga et al., 2024, Zhou et al., 2023, inter alia). This class of projects strikes a reasonable trade-off among our constraints. The most impactful prompting projects like chain-of-thought have had a major influence on LLM performance (Wei et al., 2022), and prompting projects are executable with minimal computing hardware.

We further structure our ideation process to avoid selection-bias-based confounders in ideation. If we simply ask LLMs and humans to produce ideas on 'prompting topics', we may find that LLMs and humans differ in the types of research ideas they produce (for example, LLMs may naturally suggest more projects on safer topics, which might be judged as less exciting by humans). This would lead us to simply measure misalignment in research topic preference between LLMs and humans, which is not the goal of our study. To address this possibility, we define a set of seven specific research topics extracted from the Call For Papers page of recent NLP conferences such as COLM. [2] Specifically, our topics include: Bias, Coding, Safety, Multilinguality, Factuality, Math, and Uncertainty (see Appendix A for a complete description of these topics).

Each human and LLM participant of the ideation experiment receives the same set of natural language instructions including the same topic description, idea template, and demonstration example to ensure a fair comparison. For human participants, we additionally allow them to select a preferred topic from the list, and for each selected topic, we generate a corresponding LLM idea. This exactly matches the idea topic distribution between the LLM and human participants, while ensuring that human experts are able to select topics according to their expertise.

Idea Writeup An idea can only be evaluated if it is written up to be communicated, but this writing process introduces many additional potential confounders. Human researchers may write in ways that subtly signal quality research, such as including more examples and implementation details. The format of the writeup functions as a way to scaffold what contents should be included and the level of detailedness. Ideally, we want both human and LLM participants to provide all the necessary implementation details for their generated ideas.

We take inspiration from guidelines used in grant submissions and introduce a template to specify the structure and detailedness of idea proposals. Specifically, we construct a template that includes fields for the title, problem statement, motivation, proposed method, step-by-step experiment plan, test case examples, and the fallback plan. Both the LLM agent and the human idea writers are instructed to follow this template and our provided demonstration examples to produce a project proposal as the output (see Appendix B for the full template and Appendix C for the demo example).

Even with these templates, there may be subtle writing style cues that affect the outcome measure. For example, humans may tend to write in a more engaging and informal tone. To reduce this possibility further, we developed a style normalization module that uses an LLM to convert all ideas into the same writing and formatting style without changing the original content. Our small-scale human study shows that such a normalization approach leads to a 50% accuracy for expert human judges who are asked to distinguish AI ideas from human ideas. Finally, the use of an LLM style anonymizer has the possibility of substantively changing the content of the ideas. To rule this out, the first author of this paper manually verified each human idea proposal to ensure all contents of the original ideas were preserved. We present the full prompt used in Appendix D.

Review and Evaluation Reviewing research ideas is notoriously subjective, so we want to design a review form that defines all review criteria clearly to standardize and anchor the evaluations as much as possible. At the same time, we want our review criteria and measured variables to capture all the desiderata of high-quality research ideas.

We follow best practices from AI conference reviewing (e.g., ICLR and ACL) when designing the review form, where we define four breakdown metrics including novelty, excitement, feasibility, and expected effectiveness, apart from the overall score. For each metric, we ask for a numerical score on a 1-10 scale along with a free-text rationale. We provide clear definitions and grounding for each numerical scale to calibrate all reviewers' standards (see Appendix E for the full review form).

Our blind review evaluation will compare ideas from three different conditions:

1. `Human Ideas`: Idea proposals written by our recruited expert researchers.

[2] https://colmweb.org/cfp.html

2. `AI Ideas`: Idea proposals generated by our LLM agent. We directly take the top-ranked ideas from the agent's output.
3. `AI Ideas + Human Rerank`: Idea proposals generated by our LLM agent. The first author of this paper manually selected the top-ranked ideas out of all the LLM agent's generations rather than relying on the LLM ranker in order to better estimate the upper-bound quality of AI ideas.

In the next two sections, we instantiate how our LLM agent generates ideas and how our expert participants generate and review the ideas.

3 Idea Generation Agent

We build a simple but effective LLM ideation agent to compare with the human expert baseline. Rather than focusing on innovating the agent itself, we adhere to a minimalist design principle, aiming to understand the current capabilities of LLMs in idea generation. Our research ideation agent has three essential components: paper retrieval, idea generation, and idea ranking, which we will describe in detail below.

3.1 Paper Retrieval for RAG

To ground idea generation, the agent needs to retrieve papers related to the given research topic, so that it will be aware of related works when generating new ideas. To do so, we leverage retrieval-augmented generation (RAG), which has demonstrated effectiveness on many knowledge-intensive tasks (Lewis et al., 2020, Shi et al., 2024). Concretely, given a research topic (e.g., "novel prompting methods that can improve factuality and reduce hallucination of large language models"), we prompt an LLM to generate a sequence of function calls to the Semantic Scholar API. We use `claude-3-5-sonnet-20240620` as the backbone model for our agent but the pipeline should generalize to other LLMs as well. The paper retrieval action space includes: {`KeywordQuery(keywords), PaperQuery(paperId), GetReferences(paperId)`}. Each action generation is grounded on the previous actions and executed results. We keep the top $k = 20$ papers from each executed function call and stop the action generation when a max of $N = 120$ papers have been retrieved. We then use the LLM to score and rerank all retrieved papers based on three criteria: 1) the paper should be directly relevant to the specified topic; 2) the paper should be an empirical paper involving computational experiments;[3] 3) the paper is interesting and can inspire new projects. The LLM is prompted to score each retrieved paper on a scale of 1 to 10 based on these criteria and we use the top-ranked papers for the next step of idea generation.

3.2 Idea Generation

Our key insight for idea generation is to generate as many candidate ideas as possible. Our intuition is that only a small fraction of all generated ideas might be high-quality, and we should be willing to expend inference-time compute to generate more candidates so that we can later use a reranker to discover the "diamond in the rough". This aligns with existing results showing that scaling inference compute with repeated sampling can boost LLM performance on various coding and reasoning tasks (Brown et al., 2024, Li et al., 2022). Specifically, we prompt the LLM to generate 4000 seed ideas on each research topic. The idea generation prompt includes the demonstration examples and the retrieved papers. We craft $k = 6$ demonstration examples by manually summarizing exemplar

[3] Note that we exclude position papers, survey papers, and analysis papers throughout this study since their evaluation tends to be very subjective.

papers (Dhuliawala et al., 2023, Madaan et al., 2023, Weller et al., 2023, Weston and Sukhbaatar, 2023, Yasunaga et al., 2024, Zheng et al., 2024) into our desired idea format. For retrieval augmentation, we randomly select $k = 10$ papers from the top-ranked retrieved papers and concatenate their titles and abstracts to prepend to the idea generation prompt. We also append the titles of all previously generated ideas to the prompt to explicitly ask the LLM to avoid repetitions.

To remove duplicated ideas from this large pool of candidate ideas, we first perform a round of deduplication by encoding all seed ideas with `all-MiniLM-L6-v2` from Sentence-Transformers (Reimers and Gurevych, 2020) and then computing pairwise cosine similarities. We set a similarity threshold of 0.8 for the idea deduplication based on manual inspection. [4] This leaves about 5% non-duplicated ideas out of all the generated seed ideas. We expand more on this duplication issue later in Section 7.1.

3.3 Idea Ranking

The next step is for our ideation agent to rank all the remaining ideas so that we can find the best ones among them. To build such an automatic idea ranker, we use public review data as a proxy. Specifically, we scraped 1200 ICLR 2024 submissions related to LLMs (with keyword filtering) along with their review scores and acceptance decisions. We explored multiple ways of predicting the scores and decisions of these submissions and found that LLMs are poorly calibrated when asked directly to predict the final scores or decisions, but can achieve non-trivial accuracy when asked to judge which paper is better in pairwise comparisons.

We converted the ICLR submissions into our standard project proposal format and randomly paired up accepted and rejected papers and asked LLMs to predict which one is accepted. On this task, `Claude-3.5-Sonnet` achieves an accuracy of 71.4% with zero-shot prompting. For comparison, `GPT-4o` achieves 61.1% and `Claude-3-Opus` achieves 63.5%, and we do not observe significant gains from additional prompting techniques like few-shot or chain-of-thought prompting. We therefore choose the `Claude-3.5-Sonnet` zero-shot ranker.

N	Top-10	Bottom-10	Gap
1	6.28	5.72	0.56
2	6.14	5.24	0.90
3	5.83	4.86	0.97
4	5.94	4.99	0.95
5	6.42	4.69	1.73
6	6.11	4.81	1.30

Table 1: Average ICLR review scores of top- and bottom-10 papers ranked by our LLM ranker, with different rounds (N) of pairwise comparisons.

In order to obtain reliable scores for all project proposals based on pairwise comparisons, we adopt a Swiss system tournament where all project proposals are paired with those whose accumulated scores are similar, and if the proposals are judged to be better, they gain an additional point. We repeat this for N rounds so the total score of each project proposal will be within the [0, N] range. As a sanity check, we use the `Claude-3.5-Sonnet` ranker to rank the 1.2K ICLR LLM-related submissions and compare the average review scores of the top 10 ranked papers and the bottom 10 ranked papers in Table 1. We see a clear separation between the top and bottom ranked papers, indicating the effectiveness of the LLM ranker. We choose $N = 5$ for all our experiments since it gives the best ranking result on this validation set. The top-ranked project proposals from the agent will be directly used for the `AI Ideas` condition of the human study.

Since our AI ranker is still far from perfect, we also introduce another experiment condition where the first author of this paper manually reranked the generated project proposals instead of relying on the LLM ranker, and we call this the `AI Ideas + Human Rerank` condition. As we show in

[4] We provide randomly sampled idea pairs and their similarities in Appendix H. We also provide additional implementation details about the ideation agent in Appendix F.

Table 12, 17 out of the 49 ideas in the `AI Ideas + Human Rerank` condition overlap with the `AI Ideas` condition, while the other 32 are different, indicating the discrepancy between the LLM ranker and the human expert reranking.

4 Expert Idea Writing and Reviewing

In this section, we shift focus to the human branch of idea generation comparison. We present the details of our human study, including information about the recruited experts, the human idea generation task, and the subsequent review process.

4.1 Expert Recruitment

We recruit our expert participants (including for idea writing and reviewing) by sending sign-up forms to several channels, including: 1) the OpenNLP Slack channel with 1426 NLP researchers from 71 institutions (with consent from the channel manager); 2) Twitter (X); 3) Slack channels of various NLP groups by direct communication with the group members; and 4) official chat app of the NAACL 2024 conference. We also conducted in-person recruitment by giving out name cards and wearing T-shirts [5] with sign-up links at the NAACL 2024 conference as well as various other local NLP social events. Our study has been approved by the Stanford IRB (ID 74246).

We performed screening on all the US participants [6] based on their provided Google Scholar profiles. We set a minimum requirement of having published at least one paper at a major AI venue. [7] We reached out to all participants who satisfied this requirement with the consent form and followed up with the annotation documents for those who consented to participate.

In the end, we recruited $N = 49$ experts for writing ideas, and $N = 79$ experts for reviewing ideas. Note that 24 out of the 79 reviewers also participated in the idea writing, and we made sure no reviewer would review their own idea. This results in $N = 104$ total participants across the two tasks. Each idea writer is asked to write one idea within 10 days and we compensate $300 for each, with a $1000 bonus for the top 5 ideas as scored by the expert reviewers. Each idea reviewer is assigned 2 to 7 ideas to review and we collected $N = 298$ unique reviews in total. They are given one week to finish the reviews and we compensated $25 for each review written by the idea reviewers.

4.2 Expert Qualifications

Our pool of participants is highly qualified and diverse. The 49 idea writers come from 26 different institutions (Table 15) and the majority of them are current PhD students (Figure 3 left). The 79 reviewers come from 32 institutions (Table 16) and are mostly PhD students and Postdocs (Figure 3 right). We use their Google Scholar profiles to extract several proxy metrics, including the number of papers, citations, h-index, and i10-index at the time of their submission. Table 2 shows that our idea writers have an average of 12 papers and 477 citations, while every reviewer has published at least two papers and has an average citation of 635 and h-index of 7. Moreover, based on their survey

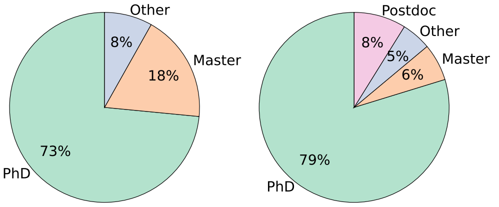

Figure 3: Positions of our idea writer (left) and reviewer (right) participants.

[5] https://x.com/ChengleiSi/status/1804273510656749649

[6] We have to recruit participants located in the US due to logistical reasons.

[7] E.g., *ACL, NeurIPS, ICLR, ICML, AAAI.

	Idea Writing Participants (N=49)					Idea Reviewing Participants (N=79)				
Metric	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
papers	12	10	2	52	9	15	13	2	52	10
citations	477	125	2	4553	861	635	327	0	7276	989
h-index	5	4	1	21	4	7	7	0	21	4
i10-index	5	4	0	32	6	7	5	0	32	6

Table 2: Research profile metrics of the idea writing and reviewing participants. Data are extracted from Google Scholar at the time of idea or review submission.

Metric	Mean	Median	Min	Max	SD
`Human` Ideas					
Familiarity (1-5)	3.7	4.0	1.0	5.0	1.0
Difficulty (1-5)	3.0	3.0	1.0	5.0	0.7
Time (Hours)	5.5	5.0	2.0	15.0	2.7
Length (Words)	901.7	876.0	444.0	1704.0	253.5
`AI` Ideas					
Length (Words)	1186.3	1158.0	706.0	1745.0	233.7
`AI + Human Rerank` Ideas					
Length (Words)	1174.0	1166.0	706.0	1708.0	211.0

Table 3: Statistics of the 49 ideas from each condition.

responses, 72 out of the 79 reviewers have previously reviewed for major AI conferences or journals. These statistics indicate that our participants are highly qualified and have substantial research experience. [8]

4.3 Idea Writing

We report statistics of our idea writers' ideas to measure their quality. As shown in Table 3, idea writers indicate a moderately high familiarity with their selected topic (3.7 on a 1 to 5 scale), and indicate the task as moderately difficult (3 on a 1 to 5 scale). They spent an average of 5.5 hours on the task and their ideas are 902 words long on average. These indicate that participants are putting substantial effort into this task. [9] We also show the distribution of their selected topics in Table 4.

Topic	Count
Bias	4
Coding	9
Safety	5
Multilingual	10
Factuality	11
Math	4
Uncertainty	6
Total	49

Table 4: Idea topic distribution.

4.4 Idea Reviewing

Review Assignment We let all reviewer participants select their top two preferred topics as well as their preferred reviewing load (from 2 to 7). We then randomly assign them to ideas within their selected topics and all ideas are anonymized. In the assignment, we balance the number of ideas from each condition for each reviewer and ensure that each reviewer gets at least one human idea and one AI idea. Every idea is reviewed by 2 to 4 different reviewers. We also avoid assigning ideas written by authors from the same institution to avoid any potential contamination. Table 5 shows that each reviewer wrote an average of 3.8 reviews from 2 or 3 conditions, across 1 to 3 topics.

[8] Detailed breakdown of participant positions is in Appendix K.

[9] See Appendix J for more details on the quality control of human ideas.

Metric	Mean	Median	Min	Max	SD
Ours					
Familiarity (1-5)	3.7	3.0	1.0	5.0	0.9
Confidence (1-5)	3.7	4.0	1.0	5.0	0.7
Time (Minutes)	31.7	30.0	5.0	120.0	16.8
Length (Word)	231.9	208.0	41.0	771.0	112.1
ICLR 2024					
Confidence (1-5)	3.7	4.0	1.0	5.0	0.8
Length (Word)	421.5	360.0	14.0	2426.0	236.4
Length (Word; Strengths & Weaknesses)	247.4	207.0	2.0	2010.0	176.4

Table 6: Statistics of our collected reviews, with ICLR 2024 reviews as a baseline (for the 1.2K submissions that mentioned the keyword "language models").

Metric	Mean	Min	Max	SD
# Reviews	3.8	2.0	7.0	1.3
# Conditions	2.5	2.0	3.0	0.5
# Topics	1.5	1.0	3.0	0.6

Table 5: Statistics of the review assignment.

Review Quality Check Apart from ensuring reviewer qualifications, we also compute statistics to measure the quality of the reviews in Table 6. On average, the reviewers indicated a familiarity of 3.7 (out of 5) in their selected topic and a confidence of 3.7 (out of 5) in their reviews. This is comparable with the 1.2K ICLR 2024 submissions related to language models, where the reviewers also have an average confidence of 3.7 out of 5. Moreover, reviewers spent an average of 32 minutes on each review, with each review being about 232 words long.

Since our review forms are different from the ICLR review forms, we compare them with the ICLR reviews where we remove the summary and question sections and only count the lengths of the strengths and weaknesses sections. This way, the ICLR reviews have an average length of 247, similar to our collected reviews. As an additional measure of review quality, out of the 298 unique reviews that we have collected, 80 of them provided links to existing papers in their rationales to justify why the proposed method is not novel. These results further validate the high quality of our review data.

5 Main Result: AI Ideas Are Rated More Novel Than Expert Ideas

In this section, we present our main finding on whether LLMs can generate better research ideas than experts. Consistently across three different statistical tests accounting for the possible confounders, we find that AI ideas have higher novelty scores than human ideas while being comparable on all other metrics.

5.1 Test 1: Treating Each Review as an Independent Datapoint

In Test 1, we treat each review as an independent datapoint and aggregate all reviews from the same condition. We treat the `Human Ideas` as the baseline condition and compare it with `AI Ideas` and `AI Ideas + Human Rerank` using two-tailed Welch's t-tests with Bonferroni correction. We show the barplot in Figure 2 and the detailed numerical results in Table 7. Both `AI Ideas` ($\mu = 5.64 \pm \sigma = 1.76$) and `AI Ideas + Human Rerank` ($\mu = 5.81 \pm \sigma = 1.66$) are significantly better than `Human Ideas` ($\mu = 4.84 \pm \sigma = 1.79$) on the novelty score ($p < 0.01$). In this particular test, the AI ideas in both conditions are also significantly better than human ideas on the excitement score ($p < 0.05$), and the `AI Ideas + Human Rerank` condition is also significantly better than `Human Ideas` in terms of

Condition	Size	Mean	Median	SD	SE	Min	Max	p-value
Novelty Score								
Human Ideas	119	4.84	5	1.79	0.16	1	8	–
AI Ideas	109	5.64	6	1.76	0.17	1	10	**0.00****
AI Ideas + Human Rerank	109	5.81	6	1.66	0.16	2	10	**0.00*****
Excitement Score								
Human Ideas	119	4.55	5	1.89	0.17	1	8	–
AI Ideas	109	5.19	6	1.73	0.17	1	9	**0.04***
AI Ideas + Human Rerank	109	5.46	6	1.82	0.17	1	9	**0.00****
Feasibility Score								
Human Ideas	119	6.61	7	1.99	0.18	1	10	–
AI Ideas	109	6.34	6	1.88	0.18	2	10	1.00
AI Ideas + Human Rerank	109	6.44	6	1.63	0.16	1	10	1.00
Expected Effectiveness Score								
Human Ideas	119	5.13	5	1.76	0.16	1	8	–
AI Ideas	109	5.47	6	1.58	0.15	1	10	0.67
AI Ideas + Human Rerank	109	5.55	6	1.52	0.15	1	9	0.29
Overall Score								
Human Ideas	119	4.68	5	1.90	0.17	1	9	–
AI Ideas	109	4.85	5	1.70	0.16	1	9	1.00
AI Ideas + Human Rerank	109	5.34	6	1.79	0.17	1	9	**0.04***

Table 7: Scores across all conditions by treating each review as an independent datapoint (Test 1). Size is the number of reviews for each condition and the p-values are computed with two-tailed Welch's t-tests with Bonferroni correction. We **bold** results that are statistically significant ($^{*}p<0.05$;$^{**}p<0.01$;$^{***}p<0.001$). AI ideas are judged as significantly better than human ideas in terms of novelty and excitement while being comparable on all other metrics.

the overall score ($p<0.05$). We do not observe significant differences between AI-generated ideas and human-written ideas on the other metrics.

5.2 Test 2: Treating Each Idea as an Independent Datapoint

Since we collect multiple reviews for each idea, one could argue that we should not treat each review as an independent datapoint. To account for this potential confounder, we perform Test 2 where we average the scores of each idea and treat each idea as one datapoint. This way, the sample size for every condition will be $N = 49$, namely the number of ideas. We treat the Human Ideas as the baseline condition and compare it with AI Ideas and AI Ideas + Human Rerank using two-tailed Welch's t-tests with Bonferroni correction. As shown in Table 8, we still see significant results ($p < 0.05$) where both AI Ideas ($\mu = 5.62 \pm \sigma = 1.39$) and AI Ideas + Human Rerank ($\mu = 5.78 \pm \sigma = 1.07$) have higher novelty scores than Human Ideas ($\mu = 4.86 \pm \sigma = 1.26$).

5.3 Test 3: Treating Each Reviewer as an Independent Datapoint

Another possible confounder is that different reviewers might have different biases, for example, some reviewers may be more lenient than others. To account for such reviewer biases, we perform Test

Condition	Size	Mean	Median	SD	SE	Min	Max	p-value
Novelty Score								
`Human Ideas`	49	4.86	5.00	1.26	0.18	1.50	7.00	–
`AI Ideas`	49	5.62	5.50	1.39	0.20	1.50	8.33	**0.03***
`AI Ideas + Human Rerank`	49	5.78	6.00	1.07	0.15	3.00	8.33	**0.00****
Excitement Score								
`Human Ideas`	49	4.56	4.33	1.16	0.17	2.00	7.00	–
`AI Ideas`	49	5.18	5.50	1.33	0.19	2.50	7.33	0.08
`AI Ideas + Human Rerank`	49	5.45	5.50	1.36	0.19	1.00	7.33	**0.00****
Feasibility Score								
`Human Ideas`	49	6.53	7.00	1.50	0.21	3.00	9.00	–
`AI Ideas`	49	6.30	6.00	1.27	0.18	2.50	8.50	1.00
`AI Ideas + Human Rerank`	49	6.41	6.50	1.06	0.15	4.00	9.00	1.00
Expected Effectiveness Score								
`Human Ideas`	49	5.10	5.33	1.14	0.16	3.00	7.00	–
`AI Ideas`	49	5.48	5.50	1.23	0.18	2.00	7.50	0.58
`AI Ideas + Human Rerank`	49	5.57	5.50	0.99	0.14	3.00	7.50	0.17
Overall Score								
`Human Ideas`	49	4.69	4.67	1.16	0.17	2.00	6.67	–
`AI Ideas`	49	4.83	5.00	1.34	0.19	1.50	7.50	1.00
`AI Ideas + Human Rerank`	49	5.32	5.50	1.24	0.18	2.00	7.50	0.06

Table 8: Scores across all conditions by averaging the scores for each idea and treating each idea as one data point (Test 2). Size is the number of ideas for each condition, and the p-values are computed with two-tailed Welch's t-tests with Bonferroni correction. We **bold** results that are statistically significant ($^*p<0.05$;$^{**}p<0.01$). AI ideas are judged as significantly better than human ideas in terms of novelty while being comparable on all other metrics.

3 where we treat each reviewer as one datapoint and compute their average score on each condition. Then for each reviewer, we get their mean score difference between the `AI Ideas` condition and the `Human Ideas` condition, as well as the difference between the `AI Ideas + Human Rerank` condition and the `Human Ideas` condition. This way, we only analyze the differences among the different conditions. That is, if the differences are significantly higher than zero under the one-sample t-test, that indicates reviewers are giving higher scores to one condition compared to the other. The results are shown in Table 9, and we see significant results ($p<0.05$) that AI ideas in both the `AI Ideas` and `AI Ideas + Human Rerank` conditions are rated more novel than `Human Ideas`. Therefore, we conclude that AI ideas generated by our ideation agent are judged as more novel than human expert generated ideas, consistently across all three different statistical tests. [10]

6 In-Depth Analysis of the Human Study

While the above main results highlight the promise of LLMs in generating novel research ideas, there are some additional nuances. In this section, we move beyond the statistical comparisons and dive

[10] We also include results of fitting linear mixed-effects models in Appendix N, which reinforces our conclusions. Additionally, we plot the breakdown of all metrics by topic in Appendix O.

	N	Mean Diff	p-value
Novelty Score			
`AI Ideas` vs `Human Ideas`	70	0.94	**0.00****
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.86	**0.00****
Excitement Score			
`AI Ideas` vs `Human Ideas`	70	0.73	**0.01***
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.87	**0.00****
Feasibility Score			
`AI Ideas` vs `Human Ideas`	70	-0.29	0.36
`AI Ideas + Human Rerank` vs `Human Ideas`	65	-0.08	0.74
Effectiveness Score			
`AI Ideas` vs `Human Ideas`	70	0.42	0.16
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.39	0.16
Overall Score			
`AI Ideas` vs `Human Ideas`	70	0.24	0.36
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.66	**0.01***

Table 9: Mean score differences between AI ideas and human ideas by treating each reviewer as a data point (Test 3). All p-values are computed with one-sample t-tests with Bonferroni correction. We **bold** results that are statistically significant ($^*p<0.05$;$^{**}p<0.01$).

into other aspects of our collected data. Specifically, we focus on the quality of human ideas, reviewer preferences, and the extent of reviewer agreement.

6.1 Human Experts May Not Be Giving Their Best Ideas

We first investigate whether human experts are submitting their best ideas to us. We did a post-study survey to understand how idea-writing participants came up with their ideas. Out of the 49 participants, 37 of them came up with the idea on the spot, while the other 12 already had the idea before the study. Furthermore, we asked the survey question: *"How does this idea compare to your past research ideas (ideas that you actually worked on)? Please answer with a percentile. E.g., this idea is one of my top 10% ideas."* Our participants indicated that on average their submitted ideas are about the top 43% of all their past ideas. This implies that our collected ideas are likely the median-level ideas from these expert researchers, which is reasonable given that most of them came up with the idea within the 10-day time constraint of the task.

6.2 Reviewers Tend to Focus More on Novelty and Excitement

To gain a deeper understanding of the dynamics between the different metrics in the review process, we explore whether reviewers focus on specific aspects when evaluating the ideas. We compute the pairwise correlation between different metrics in Table 10. The overall score mostly correlates with the novelty score ($r=0.725$) and excitement score ($r=0.854$), while having almost no correlation ($r<0.1$) with the feasibility score. This implies that reviewers might be paying more attention to the novelty and excitement aspects of the ideas when they are reviewing.

	Overall	Novelty	Excitement	Feasibility	Effectiveness
Overall	–	0.725	0.854	0.097	0.642
Novelty	0.725	–	0.719	-0.073	0.357
Excitement	0.854	0.719	–	-0.031	0.565
Feasibility	0.097	-0.073	-0.031	–	0.251
Effectiveness	0.642	0.357	0.565	0.251	–

Table 10: Pairwise correlation between different metrics (symmetric matrix).

6.3 Reviewing Ideas is Inherently Subjective

Finally, we acknowledge that reviewing is inherently subjective, and reviewing based on ideas rather than executed papers might be even more subjective. We investigate this using inter-reviewer agreement. Specifically, we randomly split reviewers of each paper into half, use one half to rank the top and bottom 25% of all ideas, and then measure agreement with the held-out set of reviewers. [11] As shown in the first block of Table 11, reviewers have a relatively low agreement (56.1%) despite the fact that we have provided detailed explanations for each metric in our review form. As a baseline comparison, the NeurIPS 2021 reviewer consistency experiment found 66.0% accuracy using this reviewer agreement metric in the balanced setting (Beygelzimer et al., 2021, Lu et al., 2024). We also computed the reviewer agreement using the same metric on the 1.2K ICLR 2024 submissions related to language models, which has a balanced accuracy of 71.9%. While our reviewer agreement is higher than random (50%), it is generally lower than conference reviewing, most likely due to the higher subjectivity involved when evaluating ideas without seeing the actual experiment results.

7 Limitations of LLMs

With our findings from the human study in mind, we now turn to LLM performance to provide insights that could inform future methods for improving idea generation systems. Our ideation agent is motivated by two potential strengths of LLMs: their ability to scale by generating a vast number of ideas - far more than any human could - and the possibility of filtering these ideas to extract the best ones from the large pool. In theory, this approach could lead to high-quality ideas by leveraging inference scaling. However, we present empirical evidence that this naive assumption about scaling idea generation has significant limitations.

7.1 LLMs Lack Diversity in Idea Generation

We adopted an over-generate and rank paradigm in idea generation. This raises the question: is there an upper limit to how many new ideas LLMs can generate? To answer this question, we take a closer look at 4000 generated seed ideas for each topic.

We encode all raw ideas with `all-MiniLM-L6-v2` from Sentence-Transformers. For each idea, we compute its cosine similarity with all previously generated ideas on the same topic. We consider an idea as a duplicate if it has a similarity of above 0.8 with any of the previously generated ideas. In Figure 4, we show that as the agent keeps generating new batches of ideas, the percentage of non-duplicates in newly generated batches keeps decreasing, and the accumulated non-duplicate ideas eventually plateau. In fact, out of the 4000 generated seed ideas, there are only 200 non-duplicate

[11] This metric follows the balanced accuracy metric as used in Lu et al. (2024) and avoids the limitations of other agreement metrics like Krippendorff's alpha, which require overlapping reviews and would result in a sparse matrix due to the non-overlapping nature of our reviewer assignments. We do the random splitting 20 times and report the average to reduce variances.