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.

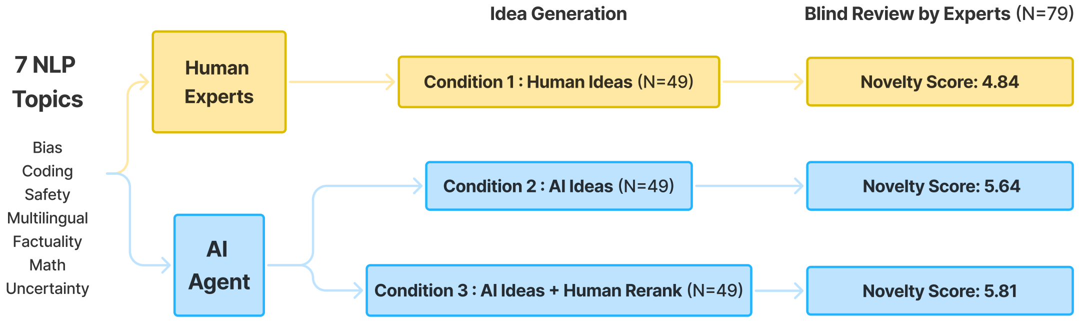

Figure 1: Overview of our study: we recruit 79 expert researchers to perform blind review of 49 ideas from each of the three conditions: expert-written ideas, AI-generated ideas, and AI-generated ideas reranked by a human expert. We standardize the format and style of ideas from all conditions before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

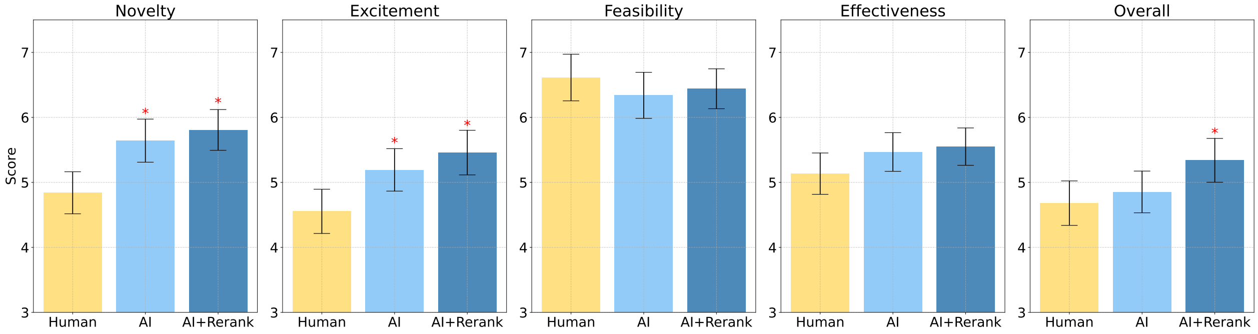

Figure 2: Comparison of the three experiment conditions across all review metrics. Red asterisks indicate that the condition is statistically better than the `Human` baseline with two-tailed Welch's t-tests and Bonferroni correction. All scores are on a 1 to 10 scale. More detailed results are in Section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