4th International Conference on the Science of Science and Innovation, June 16-18, 2025
Copenhagen Business School, Copenhagen, Denmark

Academic Simulacra: Forcasting Research Ideas through Multi-Agent LLM Simulations

Yingrong Mao, Jiwoong Choi, Donghyun Kang, James Evans

University of Chicago

Keywords: Multi-Agent Simulation, Simulated Scholarship, Large Language Models

Background

This research revolves around the following question: *Can we simulate an agent that replicates the thinking process of a real-world scholar? Moreover, can we observe any patterns across gender, ethnicity, or affiliation in the publication process?* Building on Don Swanson's seminal work in Literature-Based Discovery (LBD) [3], recent efforts have increasingly incorporated Large Language Models (LLMs) to enhance knowledge discovery, assisting idea generation and optimizing novelty from literature [1, 4]. In this study, we apply a multi-agent framework powered by state-of-the-art LLMs to predict future research, using extensive bibliographic data from top-tier computer science conferences.

Dataset and Methods

We collected all papers published in top computer science conferences (ICML, ICLR, AAAI, CVPR, EMNLP, ECCV, ACL), including each paper's title, abstract, author names, and affiliations. To mitigate data leakage—given that we used the Llama 3.1 8B model with a knowledge cutoff in December 2023—we designated papers published before that date as our training set and aimed to predict papers published in 2024. Additionally, for downstream analysis, we used LLaMA 3.3 70B to infer authors' broader ethnic backgrounds (e.g., categorizing 'Chinese American' under 'Chinese').

Multi-Agent Simulation Workflow (see Figure 1): The multi-agent simulation environment was implemented using Autogen [5], integrated with a state-of-the-art LLM (e.g., Llama 3.1 8B) served via Ollama [2]. For each conference, 200 papers published in 2024 were randomly sampled. Agents were instantiated based on authorship positions: the first, second, and last authors (when a paper had more than three authors) or all authors (for papers with three or fewer authors). Each agent was initialized with the publication history (titles and abstracts) from the author's prior works in which they were listed as either the first or last author. Following the simulated scholars' discussion, a research-assistant agent summarized, extracted, and formatted the generated key ideas and paper titles.

To evaluate our approach, we compared the predicted (generated) paper titles and key ideas to their actual counterparts using cosine similarity computed in the SPECTER2 embedding space. SPECTER2 provides transformer-based document embeddings explicitly trained on citation graphs, effectively capturing semantic and topical relationships among scientific publications. Additionally, we conducted a regression analysis modeling the average cosine similarity (between the generated key ideas and actual ideas) as a function of the proportion of authors with Chinese ethnicity, the proportion of authors affiliated with institutions based in China, and their interaction, while controlling for conference fixed effects.

Key Findings

Across 1,400 sampled papers, the generated key ideas showed a mean cosine similarity of 0.860 (std = 0.027) to the actual ideas, ranging from 0.760 to 0.950. Half of the papers exceeded 0.862, suggesting that the multi-agent framework generally captured core thematic elements. The tight clustering around a high mean underscores the potential of multi-agent LLM simulations to forecast future research ideas effectively.

At lower proportions of Chinese authors, higher affiliation with institutions in China correlates with lower similarity between the generated and actual key ideas. (See Figure 2) However, as the proportion of Chinese authors increases, the effect of Chinese institutional affiliation decreases, leading to higher similarity scores. Under high ethnic homogeneity, the influence of institutional affiliation converges, indicating that author ethnicity plays a dominant role in shaping writing style or content, while institutional effects become negligible in homogenous author groups.

Given the variability among large language models, the optimal strategy would ideally involve running simulations across multiple models to verify consistent trends. However, due to limited GPU resources, conducting such extensive simulations would currently require months to complete. At present, each paper's simulation takes 1.5 minutes and involves roughly 15 conversational turns on average. Future work will aim to increase the number of papers analyzed and enhance interpretability by identifying the most influential segments of these discussions and evaluating their similarity to authentic scholarly interactions.

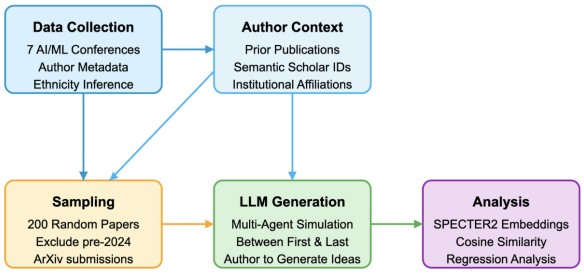

Figure 1: Multi-Agent Simulation Workflow

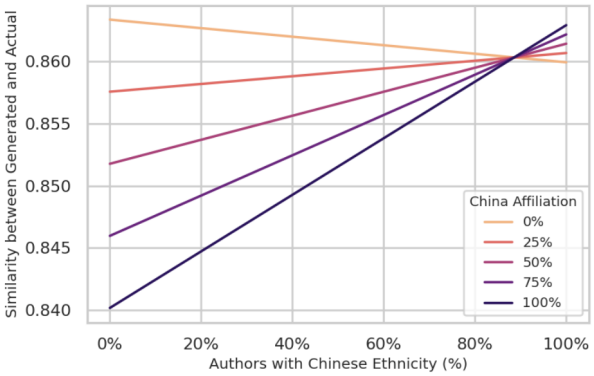

Figure 2: Interaction Between Author Ethnicity and Institutional Affiliation (Conference: ICML)

References

[1] J. Baek, S. K. Jauhar, S. Cucerzan, and S. J. Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[2] Ollama. Ollama. https://ollama.com, 2025.

[3] D. R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

[4] Q. Wang, D. Downey, H. Ji, and T. Hope. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259*, 2023.

[5] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 3(4), 2023.