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

THE UNIVERSITY OF CHICAGO

Jiwoong Choi [1] Yingrong Mao [1] Donghyun Kang [1] James Evans [1]

[1]Knowledge Lab, University of Chicago

KNOWLEDGE LAB

Methods

We introduce a multi-agent simulation framework for forecasting research ideas using "scholar agents" powered by large language models (LLMs). We instantiate approximately 10,000 + scholar agents based on their publication histories prior to 2024 and simulate discussions to collectively generate key research ideas for every papers published in seven major computer science conferences in 2024.

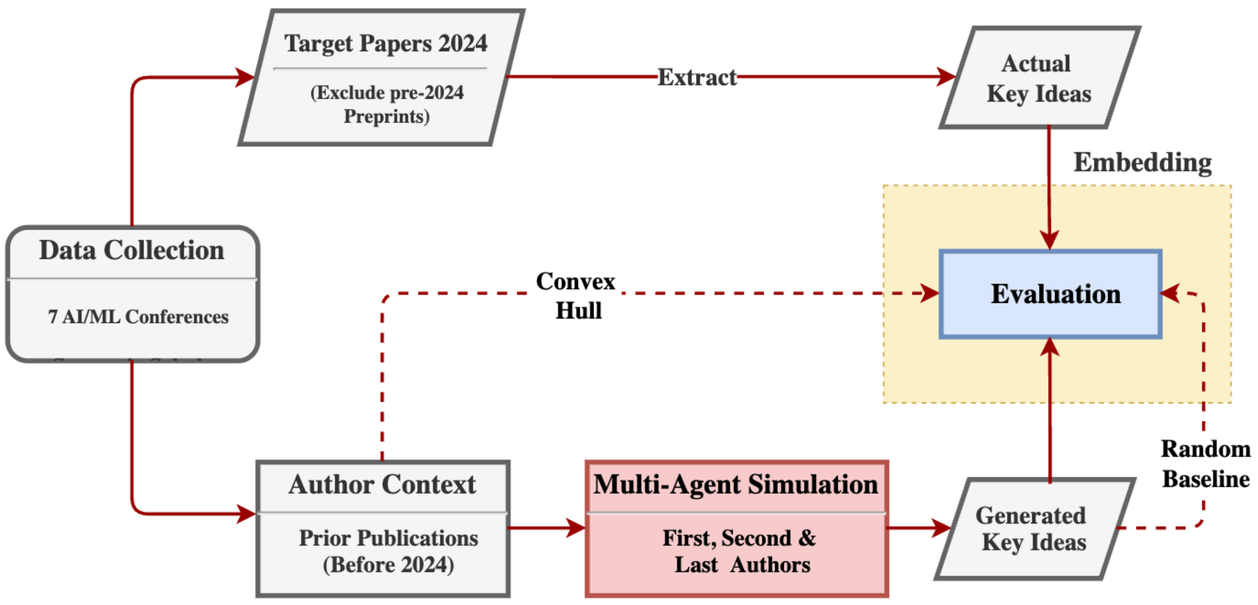
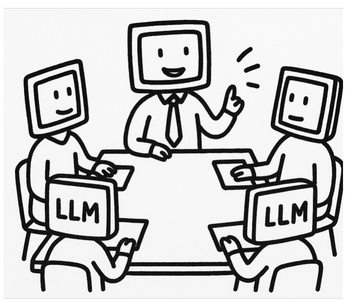


Figure 1. Simulation Flowchart

- Instantiated scholar agents plus a research-assistant agent inside the AutoGen framework; first (& second) authors propose ideas & last authors critique and refine "key ideas" for each target paper.
- Embedded generated and actual key ideas with all-mpnet-base-v2 and computed cosine similarity, benchmarking against the random baseline.
- Compared the distance & direction of each generated result (relative to the convex hull of prior publications) with those of the target paper.



Today, let's come up with some research ideas about ICML Conference based on your previous experience and knowledge.

Future Research Forecast

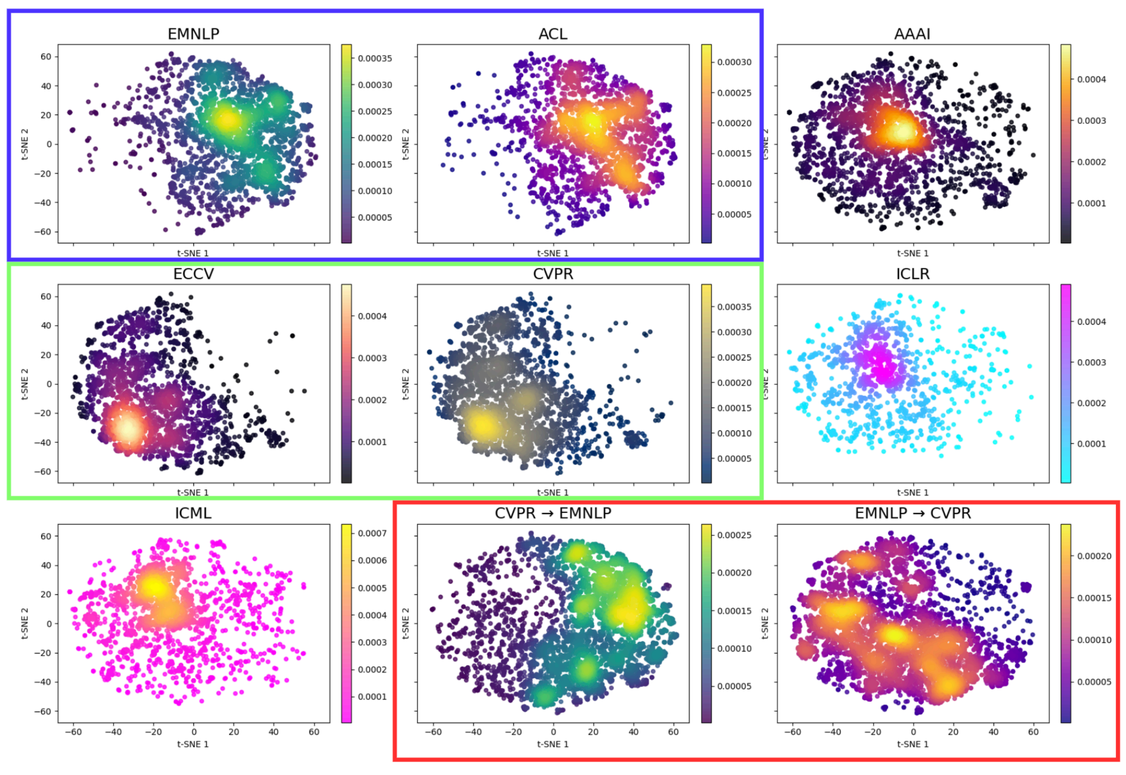


Figure 2. t-SNE of Generated Key-Idea Embeddings

t-SNE on generated idea embeddings shows an upper-right NLP cluster and a lower-left CV cluster. In a counterfactual swap (let CVPR authors to generate EMNLP papers and vice versa), a bridging region emerges between the CV and NLP areas.

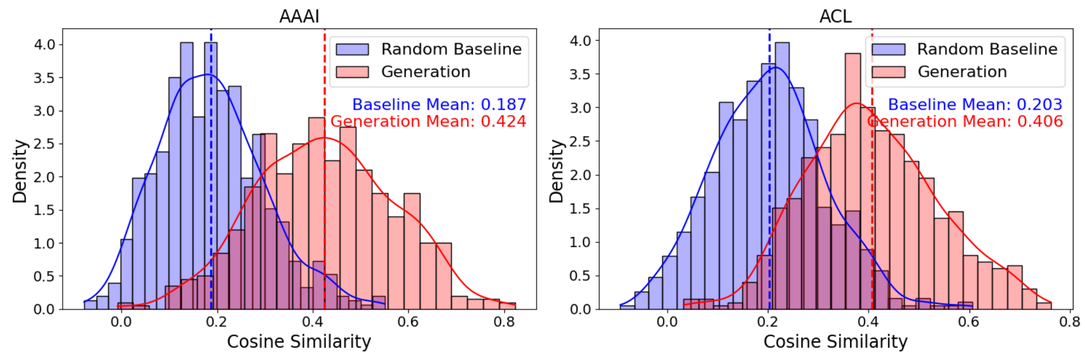


Figure 3. Cosine Similarity to Target Papers vs. Random Baseline

Venue	Random Baseline	Model Generation
ACL	0.2033	0.4061
CVPR	0.1913	0.4317
EMNLP	0.1923	0.3897
ECCV	0.1877	0.4265
ICLR	0.1984	0.4110
ICML	0.1819	0.4065

Table 1. Mean cosine similarity between each target paper's embedding and (1) randomly selected generated key idea (baseline) versus (2) key idea generated specifically for that paper

Generated results outperform the random baseline in predicting the target papers.

Can LLMs conduct exploratory research?

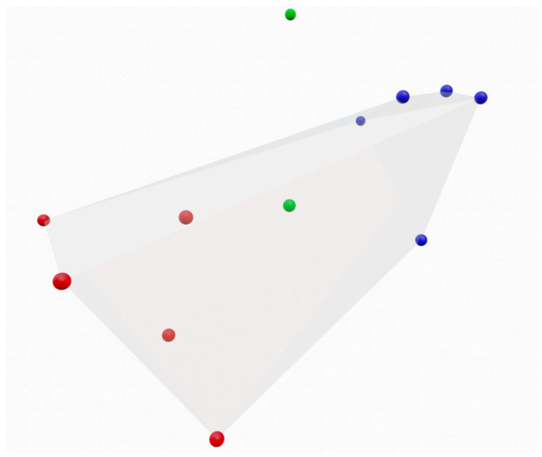


Figure 4. Conceptual Representation of Convex hull of the two authors' past works (red & blue), with target papers (green)

- Defining "exploratory"— We label a target paper **exploratory** when its embedding lies in the **top-10%** of Euclidean distances beyond the convex hull of the authors' past works.
- **What counts as a correct prediction.** A generated abstract is deemed to *match* its target when (i) it lands at the *similar* distance from the convex hull and (ii) its **residual vector** [a] forms a cosine ≥ 0.65 with the target residual, meaning both leap in nearly the same geometric direction.
- **Exploratory vs. Non-exploratory Papers.** For exploratory papers, the model is not creative—**55%** of its outputs stay near the convex hull (mere interpolation), another **38%** leap out but off-angle, and only **7%** reproduce the human leap.
  For non-novel papers, mere interpolation falls to **14%**; the model leaps beyond the hull in **74%** of cases (off-angle) and achieves a correct match in **11%**.
- **Thus LLMs readily predict beyond past work when the target itself is non-exploratory, yet rarely align with the precise creative direction required for genuine exploration.**

[a] $r = y - \hat{y}$, where $\hat{y}$ is the projection of y onto the hull; r encodes the direction of the leap beyond interpolation.

giochoi@uchicago.edu