

Siyang Wu

Email: siyangwu@uchicago.edu
[Kaggle](#)

Phone: 814-862-8792
[GitHub](#)

EDUCATIONAL BACKGROUND

Sep/2024 - Jun/2026 University of Chicago	Chicago, IL, USA
• Master Science in Data Science, Overall GPA: 3.92/4.00	
Jan/2021 - Jun/2024 University of Michigan Ann Arbor	Ann Arbor, MI, USA
• Degree: B.S. in Data Science with High Distinction, Overall GPA: 3.98/4.00	

PUBLICATION

- **Siyang Wu**, H. Bao, N. Kunievsky, and James A. Evans. *Automatically Advancing LLM Expertise in Technology Judgment*. arXiv preprint: <https://arxiv.org/abs/2505.12452>.
- **Siyang Wu***, H. Bao*, S. Li*, Ari Holtzman, James A. Evans. *Mapping Overlaps in Benchmarks through Perplexity in the Wild*. arXiv preprint: <https://arxiv.org/abs/2509.23488>.
- **Siyang Wu**, Zhewei. Sun. *How Do Language Models Generate Slang: A Systematic Comparison between Human and Machine-Generated Slang Usages*. EMNLP 2025 Findings, to appear. arXiv preprint: <https://arxiv.org/abs/2509.15518>.
- H. Bao, **Siyang. Wu**, J. Choi, Y. Mao, and James A. Evans. *Language Models Surface the Unwritten Code of Science and Society*. arXiv preprint: <https://arxiv.org/abs/2505.18942>.

RESEARCH EXPERIENCE

Is hesitation a reliable signal of uncertainty in LLMs? Project Lead, Supervised by <u>Prof Bryon Aragam</u> , Booth, University of Chicago	Jun/2025 - Current
• Introduced a black-box, post-hoc method to assess LLM confidence via behavioral robustness, quantifying the output gap between semantically equivalent perturbed and unperturbed prompts as a reference-free estimation for confidence. It mitigates overconfidence and repetition bias and extends naturally to long-form, open-ended generation, overcoming key limitations of prior methods.	
• Demonstrated that behavioral robustness under premise perturbations yields strong alignment between estimated credibility and correctness/factualty, offering a principled foundation for credibility assessment in closed LLMs.	
Can LLMs uncover the tacit rules behind expert evaluation? RA, Supervised by <u>Prof James Evans</u> , Knowledge Lab, University of Chicago	Nov/2024 - May/2025
• Proposed a conceptual framework using LLMs as diagnostic tools to reveal tacit norms and implicit biases—the “unwritten code”—that govern human evaluation in science and society.	
• Designed an iterative self-hypothesis generation algorithm where LLMs compare peer-reviewed papers to extract and amplify hidden evaluative heuristics underlying scientific judgment, distinguishing normative priors (rigor) from posterior heuristics (storytelling, contextualization).	
• Demonstrated that human reviewers and LLMs share explicit priors but diverge in implicit posteriors—revealing how unspoken storytelling and positioning cues drive peer-review outcomes, extending to domains such as hiring and admissions.	
Do LLMs know what they know in scientific patents? Project Lead, Supervised by <u>Prof James Evans</u> , Knowledge Lab, University of Chicago	Sep/2024 - Jan/2025
• Introduced a novel patent differentiation benchmark of 1.3 million computer-science patents, testing LLMs’ ability to detect subtle conceptual distinctions between semantically similar inventions.	
• Developed a self-questioning framework that decomposes LLM errors into missing vs. unused knowledge, showing that models often possess—but fail to apply—relevant internal understanding.	
• Demonstrated that structured self-generated questions and scientific retrieval significantly improve model reasoning and conceptual accuracy, and that smaller models’ questions are better queries guide for larger models’ reasoning through scalable introspection.	
Do benchmarks genuinely test the capabilities they claim to assess? Project Lead, Supervised by <u>Prof Ari Holtzman</u> ; Prof James Evans, Conceptualization Lab; Knowledge Lab, University of Chicago	June/2025 - Oct/2025
• Developed a large-scale meta-evaluation framework connecting <i>benchmark semantics, performance, and perplexity signatures</i> to reveal true overlaps across 88 LLM benchmarks.	
• Introduced the concept of benchmark signatures, sets of salient in-the-wild tokens whose perplexity patterns predict benchmark performances, and reveal essential functions of the benchmarks test for.	

- Demonstrated that *signature-level analysis* robustly distinguishes benchmark functions and mitigates biases from benchmark families and question formats, offering new insights into the LLM capability space and benchmark validity.

Do LLMs create informal usages of words as humans?

Jan/2025 - Apr/2025

Project Lead, supervised by Prof Zhewei Sun, Research Professor, Speech&Language Group, TTIC

- Introduced the first systematic comparison between human and machine-generated slang usages. Designed a unified evaluative framework to assess linguistic alignment between human-attested slang and LLM-generated slang across three dimensions—characteristics, creativity, and informativeness.
- Built a paradigm of collecting machine-generated slang usages, which enables reproducible analyses of LLMs' internal knowledge of informal and creative language, filling a major gap in computational sociolinguistics.
- Developed quantitative metrics for morphological and semantic creativity. Proposed new measures to operationalize linguistic creativity and evaluate how LLMs coin and reuse slang terms compared to human speakers.

3D Handpose Parametrization and Reconstruction

May/2023 - Sep/2023

Project Lead, supervised by Prof Matt Reed, The University of Michigan Transportation Institute

- Implemented an end-to-end pipeline for parametrizing 3D hand scans and reconstructing 3D hand models based on dual quaternion parameters/Joint location representation.
- Developed an 8-way tree structure for voxel sampling, improving time complexity from $O(n^3)$ to $O(\log n)$.
- Synthesizing hand pose training dataset for supervised learning and data augmented hand pose set to 4 thousand to improve the robustness and generalization of the model.
- Designed and built a novel, light-weight model (7k parameters) with Keras/TensorFlow for a dual-quaternion matrix regression task.

EXTRACURRICULAR EXPERIENCE

UBC Ovarian Cancer Subtype Classification and Outlier Detection (69th/1346)

Dec/2023 - Jan/2024

- Classified five types of ovarian cancer from microscopy scans of biopsy samples.
- Segmented and preprocessed images, including resizing and staining, to prepare data for further analysis.
- Inferred with ensemble learning by ResNeSt-200e, Efficientnet-V2, SEResNeXT-101d under multi-instance-learning framework.

Kaggle Competition Large Language Model Science Exam (93rd/ 2663)

Aug/2023 - Oct/2023

- Built a large-scale scientific RAG system achieving 0.905 MAP@3 on the Kaggle private test leaderboard.
- Generated embeddings with GTE-base and indexed via FAISS for efficient retrieval.
- Expanded dataset to 320K MCQs using external science texts and fine-tuned DeBERTa-V3-Large (304M) through a LangChain-based GPT-3.5 pipeline.

INTERNSHIP

Research Intern at Financial Technology Dept of China Life Insurance Group

Jun/2024 - Sep/2024

- Developed and evaluated multimodal agentic systems by integrating open-source TTS and ASR models (FishSpeech, FunASR) into an automated reasoning-action loop.
- Engineered an end-to-end orchestration pipeline with foundation model, deepseek-v3, parses incoming text into structured JSON tasks, enabling autonomous scheduling and workflow execution through the WeCom API.
- Designed deployment and performance-monitoring protocols with synthesizing data benchmarking, and documentation to ensure reliability and reproducibility across multimodal components.

Data Analyst at Inspur Group

Jun/2019 - Jul/2019

- Conducted a survey investigating labor hours of workers and collected a clean dataset.
- Processed and analyzed collections of datasets with linear regression, inferring expected labor hours at each position.
- Generated an optimization scheme and gave a presentation to the workshop chief.

OTHER INFORMATION

- **Computer Skills:** C++, Python, MySQL, R, MATLAB, JavaScript, Latex, HTML (familiar); Docker (Beginner)
- **Languages:** Chinese (native), English (proficient)