Siyang Wu Email: siyangwu@uchicago.edu

Kaggle GitHub Phone: 814-862-8792

EDUCATIONAL BACKGROUND

- Sep/2024 - Dec/2025. **University of Chicago** Chicago, IL, USA
  M.S. in Data Science, Overall GPA: 3.92/4.00
- Jan/2021 - Jun/2024. **University of Michigan, Ann Arbor** Ann Arbor, MI, USA
  B.S. in Data Science with High Distinction, Overall GPA: 3.98/4.00

Related Courses: Human-Centered Machine Learning (A); Bayesian Data Analysis (A); Machine Learning (A); Applied Regression (A+); Statistics & AI (A+); Database Management Systems (A+).
Honors: **University Honors** (3 terms), **James B. Angell Scholar** (2 terms)

PUBLICATIONS & PREPRINTS

- **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: 2509.15518.
- **Siyang Wu***, H. Bao*, S. Li*, Ari Holtzman, James A. Evans. Mapping Overlaps in Benchmarks through Perplexity in the Wild. . arXiv preprint: 2509.23488.
- **Siyang Wu**, H. Bao, N. Kunievsky, and James A. Evans. Automatically Advancing LLM Expertise in Technology Judgment. *Submitting to ACL 2026.* arXiv preprint: 2505.12452.
- H. Bao, **Siyang Wu**, J. Choi, Y. Mao, and James A. Evans. Language Models Surface the Unwritten Code of Science and Society. *Submitting to ACL 2026.* arXiv preprint: 2505.18942.

RESEARCH EXPERIENCES

- Is hesitation a reliable signal of credibility in LLMs? Jun/2025 - Current
  *Project Lead. Supervised by **Prof. Bryon Aragam**, Booth, University of Chicago*
  - **Introduced a post-hoc method to assess LLM credibility in black-box settings** via behavioral robustness, quantifying the output distribution gap between semantically equivalent perturbed and unperturbed prompts as a reference-free estimation for credibility.
  - Our method addresses common limitations in existing methods, such as overconfidence and repetition bias, while extending credibility assessment to long-form generation where these methods fail.
  - Demonstrated that behavioral robustness under premise perturbations yields strong alignment between estimated credibility and correctness/factuality, offering a principled foundation for credibility assessment in closed LLMs.
- Can LLMs uncover the tacit rules behind human preferences? Nov/2024 - May/2025
  *Researcher. Supervised by **Prof. James Evans**, Knowledge Lab, University of Chicago*
  - Resulting publication: arxiv 2505.18942. Under review ACL 2026. **Developed a hypothesis-generation framework utilizing LLMs to surface "unwritten codes" and heuristics in societal decision-making**, demonstrated within scientific peer review.
  - Developed an iterative consistent-hypothesis mining framework that compares paired papers to diagnose why one outscores the other. This process successfully mapped the shift from 'priors' (initial assumptions, such as theoretical rigor) to 'posteriors' (data-driven heuristics, like storytelling and contextualization), which better explain actual review outcomes.
  - Demonstrated that while human reviewers explicitly reward normative traits like theoretical rigor, their actual scoring patterns rely heavily on implicit factors like contextualization and storytelling, which they rarely articulate.
- Do LLMs know what they know in judging scientific patents? Sep/2024 - May/2025
  *Project Lead. Supervised by **Prof. James Evans**, Knowledge Lab, University of Chicago*
  - Resulting publication: arxiv 2505.12452. Under review ACL 2026. **Developed a proactive agentic framework for scientific understanding that decomposes LLM errors into missing vs. unused knowledge**, showing that models often possess relevant internal understanding, but fail to apply.

- Introduced a novel patent differentiation benchmark of 1.3 million computer-science patents, testing LLMs' ability to detect nuanced conceptual distinctions between semantically similar inventions masked by strategically complex writing.
- Demonstrated that smaller models' questions are better query guides for larger models' reasoning through scalable introspection.

Do benchmarks genuinely test the capabilities they claim to assess? June/2025 - Oct/2025

- *Project Lead. Supervised by* ***Prof. Ari Holtzman****, Conceptualization Lab;* ***Prof. James Evans****, Knowledge Lab, University of Chicago*
  - Resulting publication: arxiv 2509.23488. Under review ICLR 2026. Top 2% rating. **Introduced a statistical framework that reveals essential functions of the benchmarks being tested**, what we term as "benchmark signatures", i.e., sets of salient in-the-wild tokens whose LLMs' perplexity patterns predict benchmark performances.
  - Engineered a large-scale meta-evaluation framework using `vLLM` to extract perplexity distributions at a billion token scale, benchmarking 89 datasets across 32 models.
  - Demonstrated that signature-level analysis reveals more genuine benchmark overlaps and mitigates biases from benchmark families and formats, offering a principled way to scratch the interconnected LLM capability space.
  - Revealed that LLM benchmarks have strong mismatch issues between the design and execution, especially in logic, instruction following, and language.

Do LLMs create informal usages of words as humans? Jan/2025 - Apr/2025

- *Project Lead. Supervised by* ***Prof. Zhewei Sun****, Speech&Language Group, Toyota Technological Institute at Chicago*
  - Resulting publication: EMNLP 2025 Findings. **Introduced a systematic comparison between human and machine-generated term creations.** Designed a unified evaluative framework to assess linguistic alignment between human-attested terms and LLM-generated terms across characteristics, creativity, and informativeness.
  - Developed quantitative metrics for morphological and semantic creativity. Proposed new measures to operationalize linguistic creativity and evaluate how LLMs coin and reuse terms compared to human speakers.
  - Developed a framework for iterative collection of machine-generated word usages and fine-tuned eighteen 8B-parameter models via `LoRA` to investigate knowledge transferability of informal term creation.

EXTRACURRICULAR EXPERIENCES

- UBC Ovarian Cancer Subtype Classification and Outlier Detection Dec/2023 - Jan/2024
  - Optimized a preprocessing pipeline to handle medical scans (∼30 GB per image), performing efficient segmentation and stain normalization under memory constraints.
  - Inferenced with ensemble learning by `ResNeSt-200e`, `Efficientnet-V2`, `SEResNeXT-101d` under multi-instance-learning framework and achieved **top 5%** on leaderboard.
- Kaggle Competition - Large Language Model Science Exam Aug/2023 - Oct/2023
  - Built a large-scale scientific RAG system achieving 0.905 MAP@3 (**top 3%**) on the Kaggle private test leaderboard.
  - Optimized the efficiency of retrieval methods using `GTE-base` embeddings and `FAISS` indexing.
  - Synthesized 320K multiple-choice questions from external science texts using the `LangChain GPT-3.5 pipeline`, and trained a `DeBERTa-V3-Large model` via full-parameter fine-tuning.

INTERNSHIP

- Research Intern at China Life Asset Management Company Jun/2024 – Sep/2024
  - **Developed and evaluated multimodal agentic systems** by integrating open-source Text-to-Speech and Automatic-Speech-Recognition models (`FishSpeech`, `FunASR`) into an automated reasoning–action loop.
  - Engineered an end-to-end orchestration pipeline with the foundation model, `deepseek-v3`, parsing 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.

OTHER INFORMATION

- Computer Skills: C++, Python, MySQL, R, MATLAB, JavaScript, LaTeX, HTML/CSS, Docker
- Languages: Chinese (native), English (proficient)