Jialiang Xu

liamjxu@gmail.com | liamjxu.github.io | Google Scholar | linked.in/xjl

RESEARCH SUMMARY

Research Engineer focused on improving the capabilities and reliability of LLMs through post-training and evaluation. Designed and deployed data curation pipelines and benchmarks at Google Cloud AI to address agentic reasoning and tool use capability gaps in deployed LLMs. Contributed to Stanford HELM suite and led development of HELM Capabilities, a benchmark for reasoning and instruction-following. Published at ACL, EMNLP, and NAACL, with **ACL Outstanding Paper Award** and **330+ citations**.

EDUCATION

Stanford University, M.S. Computer Science, AI Track, GPA: 3.95/4.00 **2025**

University of Illinois Urbana-Champaign, B.S. Computer Engineering, GPA: 3.98/4.00 **2022**

EXPERIENCE

Google Cloud AI, *Software Engineer (AI/ML)* **Jun 2026 – Present**

- Develop and evaluate agentic capabilities of Gemini, with emphasis on knowledge work and tool use. Design targeted evaluations and ablation studies to identify model capability gaps and failure modes in task-oriented agent settings.
- Develop post-training data and methodologies, including supervised fine-tuning and reinforcement-learning-oriented data, in collaboration with Google DeepMind teams.

Amazon Alexa, *Applied Scientist* **Aug 2025 – May 2026**

- Designed and deployed a semi-automated SFT and DPO data generation pipeline to improve alignment and personalization of production LLM systems, achieving **+16% improvement** in targeted personalization quality metrics.
- Created an agentic tool-use benchmark for post-training and systematic model evaluation, covering delightful behaviors, defect detection, regression testing, and domain-specific use cases through sampling and mimicking real-traffic user memories, interaction histories, and contextual signals.

Stanford University, *Research Assistant* **Dec 2023 – Jun 2025**

- Developed evaluation infrastructure for HELM, a widely used evaluation framework for foundation models. Enabled reproducible benchmarking of **10+ frontier language models and benchmarks**, with features including chain-of-thought prompting and LLM-as-a-judge evaluation.
- Led development of HELM Capabilities, a benchmark evaluating 22+ frontier language models across reasoning, general knowledge, instruction following, dialogue, and mathematical problem solving.

Amazon Alexa, *Applied Scientist Intern* **Jun 2024 – Sep 2024**

- Led research on LLM-based persona modeling for user response prediction, framing persona inference as a structured latent representation learning problem.
- Developed a persona extraction and retrieval-augmented generation system using large language models and causal inference to model user-conditioned responses. Achieved **+12% accuracy improvement over prior state-of-the-art**.

Amazon Alexa, *Applied Scientist Intern* **Jun 2023 – Sep 2023**

- Led a research project on large language models for recommendation systems.
- Built a framework combining a GNN-based candidate short-lister with a prompt-augmented, domain-adapted, and instruction-tuned LLaMA-2 13B re-ranker using PyTorch and HuggingFace.
- Improved Recall@k and NDCG@k by approximately **10%** over internal product baselines including LightGCN and BGCF.

Bosch Research, *Research Intern* **Mar 2023 – Jun 2023**

- Adapted open-source language models to domain-specific enterprise chat applications as privacy-preserving alternatives to API-based closed-source systems.
- Fine-tuned Flan-T5 models using domain adaptation and prompt tuning to inject domain knowledge while preserving general language capabilities; the resulting models outperformed vanilla baselines and popular API-based chat systems on corporate-owned data.

Microsoft Research Asia, *Research Intern* **Jul 2021 – Jul 2022**

- Led research on numerical reasoning and semi-structured data understanding, resulting in one preprint and two papers at ACL and EMNLP.
- Collaborated with product teams to transfer research into Bing, Azure Synapse Notebook, and Excel, including tools for table-field identification, automated pivot-table generation, and spreadsheet visualization and analysis.

PUBLICATIONS

** denotes equal contribution.*

Peer-Reviewed Conference Publications

[P8] **Profiling LLM's Copyright Infringement Risks under Adversarial Persuasive Prompting**
Jikai Long, Ming Liu, Xiusi Chen, **Jialiang Xu**, Shenglan Li, Zhaozhuo Xu, Denghui Zhang.
Findings of EMNLP, 2025

[P7] **Do LLMs Know to Respect Copyright Notice?**
Jialiang Xu, Shenglan Li, Zhaozhuo Xu, Denghui Zhang.
EMNLP, 2024

[P6] **SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions**
Shicheng Liu, Sina Semnani, Harold Triedman, **Jialiang Xu**, Isaac Dan Zhao, Monica Lam.
Findings of EMNLP, 2024

[P5] **Word Embeddings Are Steers for Language Models**
Chi Han, **Jialiang Xu**, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, Heng Ji.
ACL, 2024, Outstanding Paper Award

[P4] **SPAGHETTI: Open-Domain Question Answering from Heterogeneous Data Sources with Retrieval and Semantic Parsing**
Heidi C. Zhang, Sina J. Semnani, Farhad Ghassemi, **Jialiang Xu**, Shicheng Liu, Monica S. Lam.
Findings of ACL, 2024

[P3] **SUQL: Conversational Search over Structured and Unstructured Data with Large Language Models**
Shicheng Liu, **Jialiang Xu**, Wesley Tjangnaka, Sina J. Semnani, Chen Jie Yu, Monica S. Lam.
Findings of NAACL, 2024

[P2] **Inferring Tabular Analysis Metadata by Infusing Distribution and Knowledge Information**
Xinyi He, Mengyu Zhou, **Jialiang Xu**, Xiao Lv, Tianle Li, Yijia Shao, Shi Han, Zejian Yuan, Dongmei Zhang.
Findings of ACL, 2023

[P1] **Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems**
Jialiang Xu, Mengyu Zhou, Xinyi He, Shi Han, Dongmei Zhang.
EMNLP, 2022

Manuscripts and Preprints

[M3] **Reverse Image Retrieval Cues Parametric Memory in Multimodal LLMs**
Jialiang Xu*, Michael Moor*, Jure Leskovec.
arXiv, 2024

[M2] **InfoPattern: Unveiling Information Propagation Patterns in Social Media**
Chi Han*, **Jialiang Xu***, Manling Li*, Hanning Zhang*, Tarek Abdelzaher, Heng Ji.
arXiv, 2023

[M1] **LUNA: Language Understanding with Number Augmentations on Transformers via Number Plugins and Pre-training**
Hongwei Han*, **Jialiang Xu***, Mengyu Zhou, Yijia Shao, Shi Han, Dongmei Zhang.
arXiv, 2022

Research Artifacts and Benchmarks

[R2] **HELM Capabilities: Evaluating LMs Capability by Capability**
Jialiang Xu, Yifan Mai, Percy Liang.
Stanford Center for Research on Foundation Models, 2025

[R1] **HELM (Holistic Evaluation of Language Models)**
Contributor to an open-source framework for transparent, reproducible evaluation of foundation models.

PROFESSIONAL SERVICE

- **NeurIPS:** Reviewer (2024, 2025, 2026).
- **ICLR:** Reviewer (2025, 2026).
- **ICML:** Reviewer (2025).
- **AAAI:** Reviewer (2026).
- **ACL:** Reviewer (2023).
- **EMNLP:** Reviewer (2022, 2023); Conference Volunteer (2022).
- **NAACL:** Reviewer (2024).
- **ACL Rolling Review (ARR):** Reviewer (2024–2026), including February, April, June, August, October, and December 2024; February and July 2025; and January 2026.

HONORS AND AWARDS

- ACL 2024 Outstanding Paper Award **2024**
- Brown Institute Magic Grant **2024**
- Microsoft Stars of Tomorrow Award **2022**
- Horace and Kate Wu Scholarship **2022**
- Daniel W. and Carol A. Dobberpuhl Scholarship **2022**
- ECE Visionary Scholarship **2022**
- Yunni and Maxine Pao Memorial Scholarship **2021**
- ECE Alumni Association Scholarship **2021**
- UIUC Original Undergraduate Research Award **2021**
- Omron Scholarship **2020**
- Edmund J. James Scholarship **2018**

MEDIA COVERAGE

- **SPINACH paper** featured in *Wikipedia Signpost*'s "Recent Research" section, November 18, 2024.
- **Brown Institute Magic Grant** featured on the Stanford School of Engineering website as part of the 2024–2025 Magic Grant recipient cohort.