MUXIN LIU

muxinliu@seas.upenn.edu
https://maxinemuxin.github.io/
U.S. Permanent Resident (no sponsorship required)

EDUCATION

University of Pennsylvania *Philadelphia, PA*
Ph.D. in Computer and Information Science *Expected 2030*
Advisor: Chris Callison-Burch
NSF Graduate Research Fellowship Program (GRFP) Fellow

Harvey Mudd College *Claremont, CA*
B.S. in Computer Science and Mathematics *May 2025*
Second Major: Cognitive Science (at Pomona College)
GPA: 3.96/4.0, Dean's List All Semesters
Finalist for the CRA Outstanding Undergraduate Researcher Award *Sep 2024*

PUBLICATIONS

What Does Production AI SQL Really Look Like?
Muxin Liu, Matthew Russo, Varich Boonsanong, Jiayao Wang, Pawel Liskowski, Weicheng Zhao, Hossein Ahmadi, Mona Attariyan, Anupam Datta, Boxin Jiang, Aditya Parameswaran. *Under Review*

Distill, Debate, Discover: Grounding Multi-Agent Research Ideation in Graph-Distilled Insights.
Qianran Ma, **Muxin Liu**, Grace Kim, Xiao Liang, Chris Callison-Burch. *Under Review*

When Verification Fails: How Compositionally Infeasible Claims Escape Rejection.
Muxin Liu, Delip Rao, Grace Kim, Chris Callison-Burch. *COLM 2026; ACL 2026 SURGE Workshop*

Internal Mechanisms for Tracking Belief State in Language Models.
Muxin Liu, Jacob Andreas, Belinda Li. *ToM4AI Workshop at AAAI 2025*

Stream of Search (SoS): Learning to Search in Language.
Kanishk Gandhi, Denise Lee, Gabriel Grand, **Muxin Liu**, Winson Cheng, Archit Sharma, Noah D. Goodman. *COLM 2024 (Oral)*

LILO: Learning Interpretable Libraries by Compressing and Documenting Code.
Gabriel Grand, Lionel Wong, Matthew Bowers, Theo X. Olausson, **Muxin Liu**, Joshua B. Tenenbaum, Jacob Andreas. *ICLR 2024*

RESEARCH EXPERIENCE

Snowflake *Menlo Park, CA*
Research Intern *Oct 2026 – Dec 2026*

- Developing agentic optimization methods for LLM pipelines on complex structured tasks such as document processing, combining agentic rewriting with evolutionary search (GEPA), adaptively selecting the level at which rewrite rules are applied, and allowing the rules themselves to self-evolve.
- Characterized how AI SQL is used in real-world production workloads (under review).

Allen Institute for AI (AI2) *Seattle, WA (Remote)*
Part-time Student Researcher, Open-Ecosystem Multimodal Team *Jul 2026 – Oct 2026; returning Jan 2027*
Advisor: Christopher Clark

- Designing synthetic data generation pipelines for visual reasoning tasks that demand intensive reasoning, search, and grounding.

- Applying reinforcement learning (GRPO) to strengthen vision-language models' grounded reasoning, enabling complex search on puzzle-solving and navigation tasks; work in progress, with potential contribution to a future Molmo release.

University of Pennsylvania *Philadelphia, PA*
PI: Chris Callison-Burch *Aug 2025 – Present*

DARPA SciFy — Scientific Claim Verification & Automated Research

- **Scientific claim verification.** Analyzed how existing benchmarks fail to reveal models' shortcut reasoning; designed a data generation method that produces claims grounded in tables and charts requiring compositional reasoning, and characterized model failures through both quantitative and qualitative comparison (COLM 2026; ACL 2026 SURGE Workshop).
- **Auto research ideation.** Grounded agent expertise in research taste by distilling insights from the structure of the literature citation graph, and used multi-agent debate to generate ideas that are both feasible and novel (under review).
- **Auto research system.** Built an end-to-end, multi-stage system that verifies both factual and speculative AI research claims, from experiment design and execution to evidence synthesis and report generation, with engineering on retrieval, skill acquisition, and rubric-based LLM-judge evaluation.

Agentic Reasoning *(ongoing)*

- Improving multi-agent orchestration efficiency across deep-research, general agentic, and coding tasks by encouraging parallel subagent execution and better context management, and applying on-policy self-distillation (OPSD) to learn from hindsight.

Hallucination Detection & Mitigation *(ongoing)*

- Leveraging concepts from human cognitive biases to classify and detect hallucinations. Using this taxonomy to make reward signals denser, better defined, and more consistent across LLM judges, and applying RL methods such as GRPO and DPO to mitigate hallucination.

Massachusetts Institute of Technology, CSAIL *Cambridge, MA*
PI: Jacob Andreas

Internal Mechanisms for Tracking Belief State in Language Models *Jun 2024 – Jun 2025*

- Constructed a novel Theory of Mind benchmark via a Markov Decision Process formulation, framing the task as a world model that tracks state transitions of both the ground-truth world state and the agent's belief state across narrative contexts.
- Trained linear probes and devised activation patching metrics to localize and characterize mechanistic circuits underlying belief state updates in transformer models (ToM4AI Workshop at AAAI 2025).

Stream of Search (SoS): Learning to Search in Language *Jan 2024 – Aug 2024*

- Designed a scalable data generation pipeline producing synthetic search trajectories for linear equations via BFS with a neural heuristic, integrating a custom parser for high-throughput synthesis.
- Fine-tuned language models on trajectories spanning varied difficulty levels and step granularities; conducted fine-grained evaluations of search and backtracking behavior and assessed cross-domain generalization on the countdown task (COLM 2024 Oral).

LILO: Learning Interpretable Libraries *Jun 2023 – Nov 2023*

- Studied a neurosymbolic approach to program synthesis in which a symbolic compression algorithm extracts reusable functions from solved programs and an LLM annotates them to drive iterative synthesis.
- Compared the symbolic method as a library learner against an LLM few-shot library learner, and ran extensive ablations including few-shot example retrieval and synthetic vs. natural-language task framing (ICLR 2024).