

Siddarth Madala

smadala2@illinois.edu | (412) 324-2020 | GitHub: [gh/sidmadala](https://github.com/sidmadala) | LinkedIn: [in/sidmadala](https://in.linkedin.com/in/sidmadala) | Homepage: sidmadala.github.io

EDUCATION

University of Illinois Urbana-Champaign | *PhD Computer Science*

Advisor: Julia Hockenmaier, PhD

Concentration: Natural Language Processing, Information Retrieval, Embodied AI

Urbana-Champaign, IL

Aug 2022 - May 2028 (Expected)

Duke University | *B.S. Statistics and Computer Science, Minor: Mathematics*

Aug 2018 - May 2022

EXPERIENCE

University of Illinois Urbana-Champaign

Urbana-Champaign, IL

Graduate Research Assistant

Aug 2022 - Present

- Researching information retrieval and embodied systems under Professor Julia Hockenmaier. Focused on spatio-temporal reasoning for multimodal large language models, using curriculum and reinforcement learning to improve temporal and spatial grounding.

Graduate Teaching Assistant

Aug 2022 - Present

- TA for CS 441 (Applied Machine Learning) and CS 447 (Natural Language Processing), online courses with 500+ students per semester.
- Led office hours, graded/proctored assessments, and guided student projects; translated abstract concepts into clear, actionable explanations and gave constructive feedback on design and debugging to improve student performance and comprehension.

Pinterest

New York, NY

Homefeed Ranking/Personalization Research Intern

May 2026 - Aug 2026

- Integrated PinCLIP, Pinterest's foundational vision-language model (VLM), into Homefeed's L2 late-stage ranking pipeline using RQ-VAE-derived semantic IDs, driving cross-team infrastructure and PinCorpus backfill efforts to serve 600+ million MAUs.
- Designed lightweight VLM embedding aggregation techniques that doubled training/inference throughput on large-scale distributed GPU clusters and lifted online repins by 1% at RecSys scale, sustaining 15M queries/sec with minimal latency tradeoffs in online experiments.
- Propagated experimental findings into Pinterest's flagship multi-modal Conditional Foundation Model (mCFM), extending impact across all major surfaces including Homefeed, Notifications, Board More Interests, and Search.

Epsilon

Chicago, IL

Decision Sciences/Personalization Research Intern

May 2025 - Aug 2025

- Architected a vector-based similarity search and LLM-driven filtering/reranking pipeline to synthesize labels for tens of millions of URLs across 200+ category tags, enabling scalable surrogate model distillation on Databricks and Apache Spark clusters.
- Engineered a hash embedding-based URL and metadata token embedding system to compress 750 GB of daily ingested URL browsing data by 70% and accelerate embedding throughput by 93x, powering both real-time and batch advertisement personalization workflows.
- Developed and benchmarked surrogate multi-label classifiers (logistic regression, XGBoost, neural networks) on LLM-generated synthetic labels, driving a 50% uplift in AUROC compared to legacy, hand-tuned mapping systems.
- Scaled distributed synthetic labeling, hash-embedding and multi-label classification infrastructure to support bi-weekly batch jobs on Databricks and local clusters, labeling and indexing trillions of URLs to feed as features in high-throughput, real-time ad-bidding systems.

Elemental Cognition

New York, NY

Natural Language Processing Research Intern

May 2024 - Aug 2024

- Extended Semantic Role Labeling and Named Entity Recognition from sentence to document level by unifying 10+ corpora (PropBank, CoNLL, OntoNotes, etc.) and fine-tuning GPT-3.5, slashing manual annotation costs by 92% while maintaining high-quality labels.
- Deployed a real-time token classification pipeline for document-level SRL/NER, processing hundreds of user-uploaded documents per second to amplify LLM retrieval/reasoning and cut document query latency by 83%.

IBM Research

Yorktown Heights, NY

Hybrid Cloud Research Intern

May 2022 - Aug 2022

- Engineered a Kubernetes-based Ray cluster pipeline on IBM Cloud to parallelize Optical Proximity Correction workloads across 20,000+ cores, boosting job throughput by 47% and cutting lithography correction times by 34%.
- Designed a containerized Electronic Design Automation (EDA) workflow platform to unify multiple research projects into a single build for testing and deployment, potentially reducing licensing costs for small chip designers by over 60%.

Quantum Computing/Machine Learning Research Intern

May 2021 - Aug 2021

- Pioneered a novel Qiskit-based quantum twin generator to model digital circuits in the quantum domain, delivering logic verification speeds 200% faster than traditional boolean EDA workflows.
- Developed a graph neural network system leveraging macro and connectivity data from intermediate EDA logs to predict micro-processor die size, cutting design iteration cycles from weeks to days.

RESEARCH

- Huang, J. et al. "Contextual Relevance and Adaptive Sampling for LLM-Based Document Reranking." *ACL 2026*. [2026.acl-long.94](#)
- Huang, J. et al. "Tackling Distractor Documents in Multi-Hop QA with Reinforcement and Curriculum Learning." *EACL Findings 2026*. [2026.findings-eacl.294](#)
- Chaudhuri, A. et al. "Efficient Fault-Criticality Analysis for AI Accelerators using a Neural Twin." *IEEE ITC 2021*. [doi:10.1109/ITC50571.2021.00015](#)

Skills: Python, Java, C, C++, Rust, R, Bash, PostgreSQL, PyTorch, Git, Databricks, Spark, Hadoop, AWS, Docker/Kubernetes, Figma