Siddarth Madala

smadala2@illinois.edu | (412) 324-2020 | GitHub: gh/sidmadala | LinkedIn: in/sidmadala

EDUCATION

University of Illinois Urbana-Champaign | *PhD Computer Science* — Urbana-Champaign, IL
Advisor: Julia Hockenmaier, PhD — Aug 2022 - May 2027 (Expected)
Concentration: Natural Language Processing, Information Retrieval, Embodied AI

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 learning and reinforcement learning to improve grounded language generation.

Graduate Teaching Assistant — Aug 2022 - Present
- TA for CS 441 (Applied Machine Learning) and CS 447 (Natural Language Processing), online courses with 300+ students per semester.
- Led office hours, graded and 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.

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.

Mi-Corporation — Durham, NC
Software Engineering Intern — May 2020 - Aug 2020
- Developed and deployed .NET‑based agricultural verification platforms in over half of USA's State Departments of Agriculture, streamlining compliance inspections for small and midsize farms and enabling offline data capture with asynchronous uploads in low‑connectivity zones.
- Implemented a server‑side Word‑templating engine to dynamically generate and populate inspection forms with embedded imagery and tables for 500+ clients, replacing static PDFs and empowering customizable, data‑rich reporting.

RESEARCH

- Huang, J. et al. (2025). Tackling Distractor Documents in Multi-Hop QA with Reinforcement and Curriculum Learning. In submission at AACL 2026. Preprint available at arXiv:2503.12759.
- Huang, J. et al. (2025). Contextual Relevance and Adaptive Sampling for LLM-Based Document Reranking. In submission at EACL 2026.
- Chaudhuri, A., et al, (2021). "Efficient Fault-Criticality Analysis for AI Accelerators using a Neural Twin," IEEE International Test Conference (ITC), pp. 73-82, doi: 10.1109/ITC50571.2021.00015.

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