Surya Kasturi

Email: kasturisurya@gmail.com
Website: https://suryakasturi.com
Location: Boston, MA

Summary

AI research engineer with 8 years of experience in language model research and engineering, with deep expertise in ML infrastructure. I enjoy collaborating across teams to identify high-impact problems and develop practical solutions, from quick wins to fundamental architectural changes.

Education

- M.S., Computer Science, University of Delaware, Newark, DE
- B.Tech., Electrical Engineering, JNTU Hyderabad, India

Research & ML Engineering Experience

Lendbuzz, Boston, MA, Senior Software Engineer, Machine Learning, 2021 – Present

- **Pre IPO unicorn** startup in AI risk modeling and auto finance
- **Generative AI Systems**: Designed and implemented a chatbot for car dealers using large language models. Developed dialog state tracking and slot extractor with in-context learning
- **Automations**: Co-owner of automated loan approval system processing thousands of applications daily, risk analysis of automated credit profiling.
- **Performance Optimization**: Reduced GRPC server latency by 50% through multiprocessing, efficient serialization, and database indexing.
- **Backend Infrastructure**: Containerized legacy services, authored 20k+ lines of API specifications, and created reproducible development environments adopted organization-wide, enabled distributed processing capability in ML data platform.

PingAn AI Institute, New York, NY, Machine Learning Engineer, 2018–2021

- **Founding Team**: Founding engineer of the New York office, led research initiatives, engineering development. Formerly OneConnect Research Institute.
- **Conversational AI Platform**: Led development of chatbot platform as alternative to Google DialogFlow and IBM Watson. Built the dialog state tracking, natural language understanding components and the UI.
- **Research & Development**: Conducted research on knowledge-grounded dialog systems (DSTC7) and schema-guided dialog state tracking (DSTC8). Best human evaluation results at DSTC7.
- **Machine Reading Comprehension**: Developed transformer-based question answering models, contributing to top SQuAD leaderboard results from PingAn team.
- **Domain Transfer Learning**: Worked on few-shot learning approaches for adapting task-oriented dialog systems across different domains.

Advanced Digital Sciences Center, UIUC, Research Assistant, Summer 2017

- **Neural Network Distillation**: Implemented knowledge distillation techniques for time series forecasting models using PyTorch. Conducted research on finding anomalies in HVAC systems sensor signals

Technical Skills

- **Deep Learning Frameworks**: PyTorch
- **Web Frameworks**: Flask, Django, React

- **Languages**: Python, JavaScript, SQL
- **Systems**: Docker

Publications & Research

Zheng, J., Kasturi, S., Lin, X. C. M., Salvi, O., & Wang, H. J. (2019). The oneconn-memnn system for knowledge-grounded conversation modeling. *33rd AAAI Conference on Artificial Intelligence.* (Best human evaluation results)

Awards and Recognition

- MIT Delta V Accelerator Fellowship, 2023
- Best in human evaluation at DSTC7, AAAI 2019
- Google Summer of Code, 2013

Activities & Service

- Currently learning compiler construction
- Open source contributions to Eleuther AI, 2025
- Reviewer: U.S. NSF 2025, ACL 2025, Dialog System Technology Conference 2019
- MIT MicroMasters (Microeconomics and Development Policy)
- Product Sense course (Maven)

Early Career & Internships

Core member of SciPy community, 2013 – 2016: Developed features for SciPy Central in Python/Django. Maintained production. Google Summer of Code.

PwC, Hyderabad, Software Engineer, 2015–2016: Cloud platform migration, proof-of-concept development using Google infrastructure.

Indian Institute of Science, Bangalore, Research Assistant, Winter 2013: Computer vision, face recognition research

Fraunhofer Institute for Communication Systems ESK, Munich, Research Intern, Summer 2013: Automotive AI research, traffic simulation.