Sai Krishna Ponnam

ponnam2@wisc.edu | (608) 395-1868 | Madison, WI | LinkedIn | GitHub | Website

EDUCATION

University of Wisconsin-Madison **Sep 2024 – May 2026**
Master's in Computer Science GPA – 4.0 / 4.0
Courses: Advanced NLP, Machine Learning, Image Processing, Big Data Systems, High Performance Computing.

National Institute of Technology (NIT), Calicut, India **July 2015 – May 2019**
Bachelor of Technology in Computer Science GPA – 8.43 / 10.0
Courses: Data Structures and Algorithms, Discrete Structures, Operating Systems, DBMS, Computer Networks, Distributed Computing, Computer Intelligence, Pattern Recognition, Graph Theory.

SKILLS

- Programming languages: C, C++, Java, Kotlin, Python, Scala, JavaScript, SQL, Bash.
- Databases: Redis, Cassandra, MongoDB, Hive.
- Search Technologies: Lucene, Elasticsearch, Solr.
- Deep Learning: TensorFlow, PyTorch, Keras, HuggingFace, CUDA.
- Other Technologies: Spark, Kafka, OpenMP, Kubernetes, Docker, Git.
- Soft Skills: Team Player, Agile, Proactive, Adaptable.

PROFESSIONAL EXPERIENCE

Target Corporation, India **July 2019 – July 2024**

Senior Software Engineer Nov 2021 – July 2024
Guided Search (Facets, Autocomplete, Related Suggestions) - Search

- Architected and built a lexical-based NER system in **Python** to identify attributes within user queries, reducing manual tagging by 50%.
- Fine-tuned Llama2-70B for NER task using **PyTorch** and **QLoRA**, boosting attribute coverage by 12%.
- Migrated autocomplete from **Trie** to **Lucene FST**, reducing response time by 10ms and cutting CPU usage by 30%.
- Automated user interaction metrics pipeline for facets using **Scala Spark**, enabling data-driven roadmap planning and improving feature targeting, which reduced analysis time by 80% and accelerating planning cycles.
- Designed and developed an **A/B test** analysis framework with **Spark** and **Python** to streamline impact assessment and accelerate product decisions, reducing turnaround time by 40% and supporting 10+ experiments per year.
- Reimplemented related queries pipeline in **Scala Spark**, achieving 8x faster data generation (from 8 hours to 1 hour).

Software Engineer Jun 2019 – Nov 2021
Autocomplete - Search

- Implemented a low-latency context model using LSTM with **PyTorch** and **Python** to personalize autocomplete suggestions, increasing attributable demand by 2% (~$30M/year) and CTR by 10%.
- Built a real-time Kafka pipeline in **Kotlin**, with monitoring via **Grafana,** processing 150M+ user events per day, powering timely and accurate guest context generation for improved system responsiveness.
- Automated deployment of a multi-node **Redis** cluster on **Kubernetes**, cutting deployment time from 1 hour to 10 minutes.
- Ran performance profiling and optimization for search services fusing JMeter, reducing monthly operational costs by 15%.

ACADEMIC PROJECTS

Mini Torch framework (GitHub)

- Developed in C++ with CUDA employing shared memory and reduction techniques for computational optimization.
- Implemented core neural network operations such as matrix multiplication, activation functions, and backpropagation.

Tokenizer Transfer for Multilingual Factual Knowledge Retrieval Task (GitHub)

- Implemented a Vocabulary-Free Multilingual Neural Tokenizer.
- Utilized FVT and FOCUS to adapt a large tokenizer into a language-specific tokenizer.

Benchmarking Distributed Training Frameworks for Large-Scale Models: FSDP vs. DeepSpeed (GitHub)

- Conducted an empirical evaluation of Vision Transformer training on the CIFAR dataset using FSDP and DeepSpeed.
- Demonstrated that FSDP achieved 13% faster performance on smaller models in single-node environments, while DeepSpeed improved memory efficiency by up to 30% on multi-node configurations.

LEADERSHIP & ACHIEVEMENTS

- Led a cross-functional learning group since 2021, organizing bi-weekly research paper discussions to foster knowledge sharing on machine learning and search systems.
- Secured 3rd place in the Target Hackathon (Oct 2023) for developing an LLM-powered NER system for search queries.
- Awarded for end-to-end implementation and optimization of the context feature in autocomplete (2020).
- Recognized with the “Be One Team” award in 2019 for exceptional collaboration and teamwork.
- Volunteered as a Python programming mentor in the Uplift program, helping individuals from non-technical backgrounds.