ASHISH KSHIRSAGAR

Pune, India ◇ 7276422736 ◇ kshirsagarashish149@gmail.com ◇ in: ask149 ◇ gh: Ask149

SUMMARY

AI/ML Engineer with 5+ years at Microsoft, Amazon, and Barclays, specializing in LLMs, multi-agent systems, NLP, and production ML infrastructure. Reduced LLM inference cost by 9× via prompt orchestration and routing at Microsoft. Built distributed services processing 10B+ annual requests at Amazon. Proficient in Python, PyTorch, LangChain/LangGraph, conversational AI, and cloud platforms (Azure, AWS, GCP).

WORK EXPERIENCE

Microsoft Corporation July 2024 - Present
Software Engineer *Seattle, WA*

· Reduced LLM log-triage cost by 9× via routing and prompt templates, preserving signal.
· Mitigated 15+ Windows OS update production issues by leading incident response and WinDbg-driven post-mortem debugging.
· Reduced recovery-update MTTR by 30% by building a C++ CI validation pipeline in Azure DevOps.

Amazon Development Center India Pvt Ltd October 2021 - August 2022
Software Development Engineer II *Bangalore, India*

· Enabled UK payment-method launch for 80K customers by owning delivery and ops readiness.
· Scaled distributed payment components to 10B+ annual requests by architecting Java services backed by DynamoDB and automating telemetry/on-call triage.

Barclays Global Service Pvt Ltd July 2019 - September 2021
Software Developer *Pune, India*

· Reduced QA effort by 33% by automating CI/CD and test pipelines for Java Spring services.
· Automated 20% of fraud workflows by delivering APIs and integrating Python services with React/Node.js and MySQL.

INTERNSHIP EXPERIENCE

Tesla Inc January 2024 - May 2024
Software Engineering Intern *San Francisco Bay Area, California*

· Reduced quoting manual effort by 60% and increased throughput by 30% by building a containerized .NET/C# service for Megapack (Docker, Kubernetes, Kafka; PostgreSQL, MongoDB).

Google LLC May 2023 - August 2023
Software Engineering Intern *Sunnyvale, California*

· Improved p99 latency by 5% for a product serving 1B+ users by developing journey APIs (Java, Protobuf/gRPC) on GCP.

EDUCATION

Arizona State University, Tempe, Arizona, USA August 2022 - May 2024
Master of Science in Computer Science (GPA: 4.0/4.0)

Pune Institute of Computer Technology, Pune, India July 2015 - May 2019
Bachelors of Engineering in Computer Engineering

TECHNICAL SKILLS

Programming Languages	Python, Java, C++, C#, Scala, JavaScript
ML & AI Frameworks	PyTorch, HuggingFace Transformers, LangChain, Scikit-learn, NLTK, SpaCy
Frameworks & Libraries	FastAPI, Flask, .NET, Java Spring, React, Node.js
Datastores	DynamoDB, PostgreSQL, MongoDB, MySQL, Neo4j (Graph)
Tools & Platforms	Docker, Kubernetes, Git, Kafka, MLflow, Azure, AWS, GCP, Azure DevOps

PROJECTS

klarEDA
· Automated common EDA steps by building an open-source Python library with one-call preprocessing and visualization.

High-Throughput Graph Processing Pipeline
· Improved graph-processing throughput by 50% by designing a real-time pipeline (C++, Docker, Kafka, Neo4j).

CO-CURRICULAR ACTIVITIES

Open Source Mentorship
· Mentored OSS contributors (OpenSSF @ GHC'23; CERN-HSF **GSoC'21**) on scalable ML tooling.