

SHUBHAM PAGARE

AI Platform & Performance Engineer

+91 7028119806 | shubhamp2017@gmail.com | linkedin.com/in/shubhampagare/ | https://github.com/sudoShubham | Pune, Maharashtra, India

SUMMARY

Senior Performance & AI Engineer with 5+ years of experience designing scalable AI systems, LLM observability platforms, and enterprise performance solutions across Azure, AWS, and GCP. Proven track record of improving system performance, reducing cloud costs, and delivering AI-powered solutions for Fortune 500 clients.

TECHNICAL SKILLS

- **Generative AI & Agentic Systems:** LangChain, LangGraph, Langfuse, AI Agent Development, LLM Observability, E2E Trace Analysis, Prompt Engineering
- **Performance & Monitoring:** JMeter, Gatling, Azure Load Testing, LoadRunner, Datadog, Dynatrace, Grafana, CloudWatch, Azure App Insights
- **Cloud & DevOps:** Azure AKS, AWS ECS, GCP, Jenkins, Azure ADO Pipelines, Git, GitHub, Docker, Microservices
- **Programming:** Python, JavaScript, React, Django, C++ | **Analysis:** Thread/Heap Dump, AWR, Memory Leak, LLM Token Profiling

WORK EXPERIENCE

Waynautic Technologies | Senior Performance & AI Engineer | Dec 2024 – Present

Client: PwC US — AI-Based Product on Azure Cloud

- Designed and deployed multi-agent AI workflows using LangChain, LangGraph, Azure OpenAI, and enterprise data sources to automate QA and business processes.
- Developed a custom **E2E trace tracking tool** with **Langfuse** — providing deep LLM pipeline visibility, identifying latency hotspots, token inefficiencies, and failure points.
- Spearheaded AI performance engineering covering LLM analysis, load testing strategy, and Azure infrastructure optimization, resulting in measurable **cost reduction**.
- Identified critical **architectural bottlenecks**; participated in architectural reviews and design decisions with direct communication with **PwC US stakeholders**.
- Built **AI-powered utilities** for business users and QA teams, streamlining workflows and accelerating testing cycles.

Zensoft Services | Performance Engineer | Sept 2023 – Nov 2024

- Achieved **120% improvement** in response time and ~0.1% error rate resolving multi-layer WebSocket performance issues.
- Built a custom JMeter + Selenium framework replacing LoadRunner TrueClient, delivering **100% cost savings** on tool licensing.
- Automated test execution and analysis on Azure ADO using Python, reducing insight delivery time by **50%**.

NTT Data Services | Performance Engineer | Aug 2021 – Sept 2023

- Led development of an in-house open-source monitoring & auto-analysis tool, reducing a 1-hour task to **~6–7 minutes**.
- Diagnosed **memory leaks**, network packet drops, and microservices issues across AWS and GCP, significantly improving production stability.
- Developed precision JMeter load models for AWS systems; implemented database replication for improved reliability.
- **Star Award** (Foresight & Hard Work) | **KK Award** (Outstanding Performance of the Year) — NTT Data Services

Softnautics LLP | Engineering Intern | Jan 2021 – June 2021

- Migrated codebase from Python to C++, achieving a **10x improvement** in execution speed.

EDUCATION & CERTIFICATIONS

MIT ADT University — B.Tech CSE (AI/ML), CGPA: 8.32/10 (Distinction) Aug 2017 – July 2021 | Pune

Site Reliability Engineering — Google / Coursera (2024)

Data Science Orientation — IBM / Coursera (2019)

KEY PROJECTS

LLM Observability Platform

- Built E2E trace analysis platform using Langfuse
- Tracked latency, token consumption, failures, and costs
- Enabled root-cause analysis across AI pipelines

AI Agent Platform

- Developed LangGraph-based AI agents for QA and business automation
- Integrated enterprise tools and workflows
- Reduced manual effort and improved operational efficiency