Shreehari Vaasistha L

shreehari9481@gmail.com | github.com/ShreehariVaasishta | shreehari.hashnode.dev |
linkedin.com/in/shreehari-vaasistha-l/

Overview

AI & Systems Engineer with 4+ years of experience architecting scalable ML infrastructure and high-performance web applications and AI/ML platforms for Fortune 500 clients. Proven success in leading engineering initiatives, optimizing system performance, and deploying end-to-end solutions across cloud infrastructure (AWS, GCP). Passionate about building impactful user-facing systems and contributing to open-source projects.

Experience

Limbik | AI Engineer, Full-Stack & MLOps — September 2024 - Present

- Architected a provider-agnostic **LLM Inference Gateway** from the ground up. Implemented a **Factory design pattern** to unify integrations across **OpenAI, Gemini, and Perplexity**, enabling seamless model switching.
- Engineered a high-throughput Predictive Simulation Engine to forecast content resonance and amplification metrics. Leveraged **asyncio** to orchestrate **concurrent** model execution, maximizing I/O throughput and significantly reducing latency for end-users.
- Diagnosed and fixed a severe memory fragmentation issue in the central production service. Replaced the default allocator with **Mimalloc**, eliminating **97%** of memory overhead and increasing request throughput by **5x** on identical hardware.
- Built a self-serve **Metaflow** environment on Kubernetes with **Karpenter** auto-scaling, enabling the team to independently benchmark open-source models (Qwen, Llama 3) against proprietary APIs and run various other experiments, effectively reducing cloud idle costs.

AI Planet (formerly DPhi) | Software Engineer — October 2021 - August 2024

- Led engineering for a product ecosystem impacting 300,000+ AI community users.
- Architected and implemented efficient LLM deployment pipelines that optimized deployment workflow.
- Improved model inference API performance by **60%**, reducing latency and improving user experience.
- Cut cloud service costs by **12%** by identifying and scaling down underused resources.
- Owned the development of key products including: app.aimarketplace.co, aiplanethub/openagi.

AI Planet (formerly DPhi) | Software Engineer - Intern — January 2021 - September 2021

- Directly reported to the CTO and engineered scalable REST APIs and database schemas for the community LMS platform, ensuring data integrity and efficient information retrieval.
- Contributed to the development of key products including: aiplanet.com/courses, skillspace.ai

Open Source Contributions

OpenAI Python SDK | Python, API, SDK

- Contributed a bug fix to the official **OpenAI Python client**, supporting raw responses for *parse()*
- Contribution: https://github.com/openai/openai-python/

Metaflow | Python, Kubernetes, MLOps

- Added support for custom Title/Description for Flows running through Argo Workflow.
- Fixed accessibility bugs in Markdown render components.
- Collaborated with maintainers through code reviews and discussions on GitHub Issues.
- Contribution: https://github.com/Netflix/metaflow/

KServe | Python, Kubernetes, MLOps

- Integrated support for loading Kubernetes configs using a dictionary in the Python SDK.
- Contribution: https://github.com/kserve/kserve/pull/2924

DjangoPackages | Python, Django, Django Rest Framework

- Enhanced package versatility by implementing dual licensing capabilities and advancing licensing functionalities.
- Contribution: https://github.com/djangopackages/djangopackages/

Technical Skills

Languages: Python (3.12+), Golang, JavaScript, SQL
AI & GenAI: LLMs (OpenAI, Gemini, LLAMA, Qwen), Synthetic Data Generation, Prompt Engineering, RAG
MLOps & Cloud: Kubernetes, Metaflow, Karpenter, Docker, AWS, GCP (Vertex AI), ArgoCD, GitHub Actions
Backend & Systems: FastAPI, Django, Asyncio, Pydantic, SQLAlchemy, Factory Pattern, Microservices, Memory Optimization (Mimalloc)
Tools & Databases: uv (Astral), Ruff, PostgreSQL, Redis, MongoDB, Linux (Fedora/Arch)

Education

University of Mysore — Mysore, India
Master of Computer Applications (MCA); GPA: 9.2/10 — *Jan. 2023 – Dec. 2025*

Relevant Coursework: Machine Learning, Artificial Intelligence, Cloud Computing, Advanced Data Structures, Advanced Software Engineering

Bangalore University — Bangalore, India
Bachelor's Degree; GPA: 8.18/10 — *Aug. 2018 – Sep. 2021*