MAHESH CHAVAN

Bengaluru, India | +91 7676604917 | dscsmaheshchavan@gmail.com | linkedin.com/in/maheshchavan- | github.com/navahchseham

Summary

AI Engineer, 3 years building production **generative AI** systems (RAG, LLM agents, fine-tuned open-weight models) in **PyTorch** and **LangGraph**, deployed on-prem with full **MLOps** ownership.

Experience

Samsung Electro-Mechanics, Bengaluru, Karnataka *Aug 2023 – Present*

AI Engineer

- Built and maintained **RAG** pipelines on **LangChain**, **Elasticsearch** (dense vector + BM25 hybrid search), and OpenAI embeddings to enable engineers to query semiconductor datasheets and process documents; achieved **85% answer accuracy** on a **50K+ document corpus** evaluated against a curated test set.
- Designed the document ingestion pipeline with **metadata-aware chunking** (splitting by section headers and page boundaries before recursive character splitting), improving retrieval precision by roughly **20%** over fixed-size chunking on held-out benchmarks; indexed nearly **500K chunks** in Elasticsearch with HNSW-based kNN search.
- Developed **LLM-based agents** in **LangGraph** with **ReAct** prompting to automate engineering workflows such as production log analysis and ticket triage, integrating tools for SQL querying, REST API calls, and file parsing, reducing average issue resolution time from **40 min to 22 min**.
- Implemented **human-in-the-loop guardrails** in agent workflows through LangGraph conditional edges, requiring engineer approval before destructive actions (Jira ticket creation, email dispatch), with full execution traces logged to **PostgreSQL**.
- Fine-tuned **Mistral-7B** with **QLoRA (4-bit)** on domain-specific semiconductor QA data (8K samples) in **PyTorch** via **HuggingFace PEFT** and **bitsandbytes**; the model matched **GPT-3.5 level performance** on domain evaluation sets while running entirely on-prem on a single **A100 GPU**.
- Set up **LLM evaluation workflows** around **RAGAS** (faithfulness, context precision, answer relevancy) integrated into **Jenkins CI** pipelines, enabling the team to catch prompt regressions before deployment across 3 active projects.
- Deployed **PyTorch** model inference services as **Docker** containers on the on-premise **Kubernetes** cluster with **Helm** charts and Nginx ingress, handling around 200 daily requests with HPA-based auto-scaling; migrated one batch workload to **vLLM**, improving throughput by **2.5x**.
- Owned the **MLOps lifecycle** for 2 machine learning projects (**PyTorch**, **scikit-learn**): versioned datasets and models with self-hosted **MLflow**, automated training and deployment pipelines via **Jenkins**, and managed rollouts through Helm on the on-prem K8s cluster with staged promotion (staging to production).

Projects

FRONTIER: ML DEPLOYMENT PLATFORM (Desktop Application) *2023 – Present*

C#, WPF (.NET), Python, PyTorch, FastAPI, Redis, gRPC, Docker

- Co-built an end-to-end **ML training and deployment platform** with a **WPF (.NET)** desktop client and a **FastAPI** backend, compressing time-to-production from **3 days to under 12 hours** across 5 teams.
- Engineered a **Redis**-backed distributed job queue with priority scheduling and auto-retry logic, managing up to **100 concurrent PyTorch training jobs** with zero task loss under peak load.
- Containerized all backend services behind **Docker Compose**, enabling one-command local environment setup and cutting new developer onboarding time from **2 days to under 2 hours**.

DOCQA: INTERNAL DOCUMENT INTELLIGENCE SYSTEM *2025*

Python, PyTorch, LangChain, Elasticsearch, sentence-transformers, FastAPI, PostgreSQL, Kubernetes

- Shipped a document Q&A system on **LangChain RetrievalQA** with **Elasticsearch** as the hybrid retrieval backend (dense vector + BM25), serving 50 daily users across 3 engineering teams.
- Rolled it out as a **FastAPI** service on the on-prem **Kubernetes** cluster behind Nginx ingress, with **PostgreSQL** for user sessions and audit logs.

TASKPILOT: AGENTIC WORKFLOW AUTOMATION SERVICE *2026*

Python, LangGraph, GPT-4o-mini, PostgreSQL, Samsung Knox APIs, Kubernetes

- Architected a **Knox-integrated workflow automation agent** in **LangGraph**; designed a **3-node agent graph** (Planner -> Executor -> Reviewer) with conditional retry and human-in-the-loop approval on destructive actions.
- Integrated 4 tools (Jira, Confluence, Knox mail, Knox messaging) with execution logs in PostgreSQL; deployed on **internal K8s** and adopted by **12 engineers**, saving **4 hours/week** of sprint admin.

Technical Skills

Languages: Python, C#, Go, JavaScript
Machine Learning & Deep Learning: PyTorch, TensorFlow, Keras, scikit-learn, NumPy, Pandas, ONNX Runtime
LLM Frameworks: LangChain, LangGraph, HuggingFace Transformers, PEFT, TRL, Accelerate, bitsandbytes
LLMs & Techniques: Generative AI, GPT-4o, Mistral, LLaMA-3, RAG, ReAct, QLoRA, vLLM, NLP
MLOps & Infra: Docker, Kubernetes, Helm, Jenkins, Git, MLflow, Nginx, Linux, CUDA
Web & APIs: FastAPI, gRPC, WPF (.NET), REST, WebSockets
Databases & Retrieval: PostgreSQL, Redis, Elasticsearch (dense + BM25), FAISS, pgvector
Practices: Prompt Engineering, RAGAS Evaluation, Agentic AI Design, CI/CD, On-Prem Deployment

Achievements

- Awarded **Best Employee for H1-2024** at Samsung Electro-Mechanics Bangalore out of **200+ engineers** for delivering the internal LLM-based document search platform.
- Secured **1st Place at CentuRITon Hackathon (National Level)** – M S Ramaiah Institute of Technology, competing against 100+ teams, endorsed by Devfolio & MLH (Nov 2022).

Education

Dayananda Sagar College of Engineering *Aug 2019 – Jun 2023*

Bachelor of Engineering in Computer Science and Engineering, GPA: 8.12, Bengaluru, Karnataka