

KISHOR HANGE

GenAI / LLM & Machine Learning Engineer · 3+ Years · Bengaluru

+91 70282 85855 | kishorhange46@gmail.com | linkedin.com/in/kishorhange | github.com/kishorhange111 | kishorhange.vercel.app

SUMMARY

GenAI and Machine Learning Engineer with 3+ years building and shipping production ML and LLM systems at E2open: RAG pipelines evaluated with Ragas and DeepEval, AI agents (LangChain, LangGraph, CrewAI) with MCP, LLM fine-tuning, inference optimization and MLOps. Cut inference latency by 38% and improved classifier accuracy by 9.5%. Outside work, trains and benchmarks models end to end on cloud GPUs (SFT, DPO, QLoRA, GRPO, distillation), published on GitHub.

WORK EXPERIENCE

E2open | Software Engineer

Jul 2023 – Present

Promoted to Software Engineer in Sep 2025 (Associate Software Engineer, Jul 2023 – Sep 2025)

Generative AI & LLMs

- Built an **LLM-powered chatbot** that automatically resolves customer support tickets, reducing manual support workload.
- Built **RAG pipelines** with LangChain, FAISS and Chroma, evaluated with **Ragas and DeepEval** for faithfulness and relevance.
- Fine-tuned large language models** with PEFT (LoRA, QLoRA) for domain-specific and **multilingual** use cases.
- Designed **agentic LLM workflows** with LangChain, CrewAI and n8n, and implemented **Model Context Protocol (MCP)** for secure LLM access to enterprise databases; tracked experiments in MLflow and Comet ML.
- Built **semantic document matching** on Vertex AI (Universal Sentence Encoder, Sentence Transformers).

ML Systems, Optimization & MLOps

- Cut model inference latency by 38%** with quantization and pruning (ONNX Runtime, Hugging Face Optimum), saving one full sprint of delivery time.
- Fine-tuned BERT** for business-document classification, and replaced a legacy classifier with a Random Forest on **custom Word2Vec embeddings, improving accuracy by 9.5%**.
- Built **real-time model monitoring** and a feedback loop with FastAPI, Airflow, **Prometheus** and Grafana.
- Built **Spark, Hadoop and Kafka** data pipelines with **Snowflake** analytics; automated onboarding and operations workflows.

D-Smart Innovative Technology | Web Development Intern

Nov 2021 – Feb 2022

- Built reusable UI components with React.js, Bootstrap, jQuery and Sass; cut page load time by ~5% by removing unused CSS.

PROJECTS

github.com/kishorhange111

mini-chatgpt | Base LLM → SFT → DPO → benchmarks · PyTorch, Transformers, Im-evaluation-harness

- SFT of **Qwen2.5-0.5B** on 160k conversations (127M tokens) with assistant-only loss masking, then **DPO written from scratch** on 12k preference pairs.
- Matched Qwen's official Instruct model** on IFEval (36.7 vs 36.3) and **beat it on ARC-Easy (+5 pts)**; held-out preference accuracy 52% → 68%.

generative-ai-engineering | 8 LLM projects · QLoRA, TRL, FAISS, Whisper, Qwen2-VL

- QLoRA text-to-SQL** (Qwen2.5-7B): execution accuracy **60.0% → 66.2%**; **distilled** into a 0.5B model (43.4% → 52.2%), no human labels.
- Function-calling agent**: exact tool calls **16.7% → 94.8%**; **GRPO** on GSM8K (format 55% → 98%); RAG retrieval eval (nDCG@10 0.403 vs BM25 0.217); Whisper for Marathi (CER 53.6% → 14.7%).

deep-learning-portfolio | 14 CV, NLP & generative projects · PyTorch, Keras, ONNX Runtime

- BERT distillation + ONNX + int8: 4.8x faster, 85% smaller** (−0.6 pts); GPT-2 LoRA (−69% GPU memory); ViT, U-Net, Music Transformer and molecule generation from scratch.

cnn-edge-optimization-tflite-onnx-tvm | Final-year project · TFLite, ONNX, Apache TVM

- Compiled a Keras CNN for edge devices through TFLite, ONNX and Apache TVM pipelines; models **~12x smaller**.

TECHNICAL SKILLS

GenAI & LLMs: LangChain, LangGraph, CrewAI, MCP, OpenAI API, AWS Bedrock, RAG, FAISS, Chroma, Ragas, DeepEval

LLM Training: Hugging Face Transformers, PEFT (LoRA, QLoRA), TRL, SFT, DPO, GRPO, Distillation, Im-eval-harness

ML / Deep Learning: PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, BERT, Vision Transformers, Whisper, NLP

Model Optimization: Quantization (int8, 4-bit NF4), Pruning, ONNX Runtime, Hugging Face Optimum, TFLite, Apache TVM

MLOps & Deployment: FastAPI, Docker, Kubernetes, GitHub Actions, MLflow, Comet ML, Airflow, Prometheus, Grafana

Data & Cloud: Python, SQL, Java, C++, PySpark, Hadoop, Kafka, Snowflake, GCP (Vertex AI), AWS (SageMaker, S3)

EDUCATION & CERTIFICATIONS

B.Tech, Computer Science & Engineering | DKTE's Textile & Engineering Institute, Ichalkaranji | CGPA 8.31

2019 – 2023

Certifications: DeepLearning.AI (Coursera): Machine Learning, NLP, CNNs and Time Series in TensorFlow · AWS Academy Cloud Foundations