Yash Ghogre

AI Engineer

+91 8767821407 | yashghogre100@gmail.com | linkedin.com/in/yashghogre | github.com/yashghogre | ysh.is-a.dev

Experience

Turbo ML (Puch AI) Remote (CA, USA)

AI Engineering Intern (Core LLM & Agent Systems) *April 2025 – Oct. 2025*

- **Deep Research Agent:** Architected a multi-agent system using **LangGraph** that autonomously plans, searches, and synthesizes information, reducing research time for users by ∼**60%** compared to standard search.
- **Self-Hosted Search Infrastructure:** Designed and deployed a **self-hosted search system** for the WhatsApp chatbot, enabling low-latency retrieval over internal and external sources without reliance on third-party search APIs.
- **Retrieval-Augmented Generation (RAG):** Implemented a production **RAG pipeline** for document ingestion, chunking, embedding, and retrieval, improving factual accuracy by **40%** as measured via offline evaluations and reducing hallucination rates in real-time queries.
- **Geolocation System:** Engineered a location-aware recommendation engine that parses unstructured user intent to trigger geospatial queries, boosting local search relevance by **30%**.
- **Production Deployment:** Deployed stateful **agentic workflows to Kubernetes**, enabling horizontal scaling for **concurrent WhatsApp user sessions** while maintaining deterministic execution and failure recovery.

Dunlin Remote (Delaware, USA)

ML Intern (Production Model Serving & MLOps) *June 2024 – Sept. 2024*

- **Financial Forecasting Models:** Engineered an ensemble voting system combining **DistilBERT** and **AutoGluon**, achieving a **20% improvement** in transaction classification accuracy over baseline logic.
- **High-Performance Serving:** Reduced **P95 inference latency** by implementing async **FastAPI** endpoints with request batching and optimized model utilization under concurrent load.
- **MLOps Infrastructure:** Implemented **AWS S3** based model versioning and artifact management, streamlining the retraining pipeline and ensuring 100% reproducibility.

Projects

Rivet: Autonomous AI Software Engineer | *LangGraph, Docker, Pydantic v2, OpenAPI, GitHub Actions* | [Repo]

- Architected an **agentic workflow** that parses raw **OpenAPI specifications** and autonomously generates **strictly-typed Pydantic v2 SDKs**, significantly reducing manual API integration effort for developers.
- Implemented a **self-healing execution sandbox** using **Docker**, where the agent spins up ephemeral containers to run **pytest** on generated code, analyzes stderr logs, and iteratively refactors syntax and type errors without human intervention.
- Developed a **dependency graph walker (context slicing)** that analyzes API schema references to prune unused components, producing lightweight, endpoint-specific micro-SDKs.
- Built an end-to-end **CI/CD pipeline** using **GitHub Actions** to automatically build, test, version, and publish SDK releases to **PyPI**, enabling reproducible and automated delivery.

Mem1: Long-Term Memory Framework for LLMs | *Python, Qdrant, GraphDB, MongoDB, RAG* | [Repo]

- Engineered a scalable **long-term memory framework** inspired by the **Mem0** research paper, enabling autonomous agents to retain persistent, structured knowledge beyond fixed context-window limits.
- Architected a **hybrid retrieval system (GraphRAG)** combining **Qdrant** for semantic vector search, **MongoDB** for structured metadata storage, and a **graph database** for explicit entity and relationship modeling.
- Implemented **Reciprocal Rank Fusion (RRF)** to merge vector-based and graph-based retrieval signals, improving long-context retrieval quality over vector-only baselines.
- Built an **LLM-as-a-judge evaluation harness** to assess retrieval accuracy, factual consistency, and long-context recall, achieving ∼**75% retrieval accuracy** on multi-turn, long-context benchmarks.
- Designed a developer-friendly **API and CLI** that abstracts entity extraction and memory writes, allowing developers to inject stateful memory into stateless LLM applications with minimal integration effort.

Core LLM Architecture Implementation | *PyTorch, CUDA, Transformers* | [LLaMA 2 Repo] , [GPT 2 Repo]

- Implemented state-of-the-art LLM architectures (LLaMA 2-7B, GPT-2) from scratch in **PyTorch**, engineering core components like **Rotary Positional Embeddings (RoPE)**, **Grouped Query Attention (GQA)**, and **KV Caching**.
- Optimized inference performance by writing efficient tensor operations and managing memory allocation for GPU execution, mirroring production-grade transformer implementations.
- Validated implementation correctness by loading official pre-trained weights and achieving parity in **perplexity** and output generation against HuggingFace reference models.

Technical Skills

Agentic AI & LLMs: LangChain, LangGraph, Pydantic AI, RAG (GraphRAG), Tool Calling, Prompt Engineering, LLM Evaluation, OpenAI API, Anthropic API, DSPy
Machine Learning: PyTorch, Transformers, HuggingFace, Scikit-learn, NumPy, Pandas, CUDA
Backend & Systems: Python, C++, Docker, Kubernetes, FastAPI, NixOS, Linux, Git, CI/CD (GitHub Actions)
Databases & Search: Qdrant (Vector), Neo4j/GraphDB, MongoDB, Redis, PostgreSQL
Clouds & MLOps: AWS (EC2, S3), Model Serving, Experiment Tracking, Latency Optimization

Education

Yeshwantrao Chavan College of Engineering Nagpur, India
Bachelor of Technology in Computer Technology *Nov. 2022 – June 2026 (Expected)*

- CGPA: 8.01
- **Relevant Coursework:** Distributed Systems, Deep Learning, NLP, Database Systems

Achievements

Winner - GPU-Accelerated Computing Codeathon | *KPR Institute*

- Achieved **5× speedup** over CPU baselines by writing custom CUDA kernels for convolution operations, optimizing shared memory usage and thread block configurations.

Runner-up - Kaggle Datathon Competition

- Secured top rank (98% accuracy) by engineering a robust data preprocessing pipeline and finetuning deep learning models on a high-dimensional dataset.