Yash Ghogre

AI Engineer

+91 8767821407 | yashghogre100@gmail.com | linkedin.com/in/yashghogre | github.com/yashghogre

Experience

Turbo ML (Puch AI) Remote (CA, USA)
AI Engineering Intern *April 2025 – Oct. 2025*

- **WhatsApp Integration:** Enhanced the core WhatsApp Chatbot by integrating self-hosted web-search capabilities and location-based services, increasing information retrieval accuracy by 25% and boosting local search relevance.
- **Document Processing (RAG):** Implemented a RAG pipeline allowing users to query uploaded files, reducing manual data retrieval time by enabling direct context-aware interactions with documents.
- **Deep Research Agent:** Architected an autonomous research feature using **LangGraph** and **Pydantic AI**, enabling the system to execute multi-step agentic workflows for comprehensive information retrieval and synthesis.
- *Tech Stack: Python, LangGraph, Pydantic AI, LLMs, WhatsApp API, RAG, Vector Databases, Docker, Kubernetes.*

Dunlin Remote (Dover, Delaware, USA)
ML Intern *June 2024 – Sept. 2024*

- **Financial Forecasting Models:** Achieved 90%+ prediction accuracy in transaction analysis by training and fine-tuning DistilBERT and AutoGluon models. Further boosted performance by 15% through the implementation of an ensemble voting strategy.
- **Scalable API Development:** Orchestrated the development of a high-performance FastAPI service to expose ML models, facilitating seamless integration with enterprise systems.
- **Model Operations (MLOps):** Streamlined the model lifecycle by integrating AWS S3 for artifact management, enabling efficient storage, versioning, and retrieval of trained models.
- *Tech Stack: Python, FastAPI, PyTorch, DistilBERT, AutoGluon, QLORA, Unsloth, AWS S3, Docker.*

Projects

Mem1: Memory Framework | *Python, Qdrant, MongoDB, Docker* | [Repo]

- Independently developed a scalable memory framework for LLMs and autonomous agents based on the Mem0 research paper, engineering a multi-component retrieval pipeline and a CLI assistant.
- Engineered a retrieval pipeline integrating **Qdrant** (vector DB), **MongoDB** (metadata) and various embedding models for text vectorization, achieving ∼75% retrieval accuracy over long-context tasks.
- Designed the architecture to be modular, preparing for GraphDB integration to potentially increase accuracy by an additional ∼15%.

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

- Engineered complete, from-scratch PyTorch implementations of LLaMA 2 (7B) and GPT-2 (124M), implementing critical components including Multi-Head Attention, GQA, KV Caching, and RoPE.
- Optimized architectures for efficient CUDA-supported GPU inferencing and established the groundwork for model parallelization.

Autograd Engine from Scratch | *Python, Math* | [Repo]

- Designed a Python-based automatic differentiation engine supporting dynamic computation graphs, improving computational efficiency by 30% over baseline implementations.
- Implemented comprehensive tensor operations (forward/backward passes) and prepared the engine for custom CUDA kernel integration.

Technical Skills

Languages: Python, C++, C, JavaScript, SQL, HTML, CSS
Frameworks/Libraries: PyTorch, FastAPI, LangChain, LangGraph, Pydantic AI, Next.js, React.js, NumPy, Pandas, Scikit-learn, HuggingFace
Infrastructure & Tools: Docker, Kubernetes, AWS (S3, EC2), Git, Linux, NixOS
Databases: MongoDB, Redis, PostgreSQL, Qdrant (Vector DB)

Education

Yeshwantrao Chavan College of Engineering Nagpur, MH
Bachelor of Technology in Computer Technology *Nov. 2022 – June 2026*

- CGPA: 8.01

Achievements

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

- Awarded 1st place for developing a high-performance CUDA-optimized convolution kernel, achieving a **5x speedup** over CPU-based implementations for image processing tasks.

Runner-up - Kaggle Datathon Competition

- Secured 2nd position by preprocessing a large dataset and training a Deep Learning model to achieve **98% accuracy** in the final round.