Tathagat

AI/ML Engineer | LLMs • Agentic AI • RAG • FastAPI • Python

$25/hr · Open To Offers

Delhi, India (IST)

SKILLS

Python, LLMs & Agentic AI, RAG & Vector Databases, FastAPI, PyTorch, Docker, PostgreSQL, Git

ABOUT

I'm an MCA graduate passionate about building AI systems that solve real-world problems. I enjoy experimenting with LLMs, retrieval systems, and AI agents, and I've built projects including a GPT-style language model, a RAG-based DevOps AI Agent, and a proposition-based RAG indexing approach. I'm always exploring new ideas, reading research papers, and turning them into working systems.

EXPERIENCE

Data Analyst Intern · PUCHO Online

May 2024 – Jun 2024

internship

- Cleaned and analyzed datasets with over 50,000 records using Python and Excel.
- Built automated Excel dashboards to track business KPIs, reducing weekly reporting time by 20%.
- Worked with the team to understand reporting requirements, prepare data, and improve reports based on feedback.
- Performed data cleaning, validation, and basic analysis to support business decisions.

EDUCATION

R.D. Engineering College, Ghaziabad · Master of Computer Applications (MCA), Computer Applications

Jan 2024 – Dec 2026

Ghaziabad

D.K. College, Dumraon · Bachelor of Computer Applications (BCA), Computer Applications

Jan 2019 – Dec 2023

Dumraon

PROJECTS

MiniLLM

github.com/tathagat-git/MiniLLM

Python · PyTorch · FastAPI · PostgreSQL · Docker

Built a GPT-style language model from scratch with 57.5M parameters, trained on 150M tokens. Implemented pre-training, QLoRA fine-tuning, DPO alignment, and deployed it as a FastAPI service with streaming inference.

DevOps AI Agent

github.com/tathagat-git/DevOps-AI-Agent

Python · FastAPI · FAISS · ChromaDB · Ollama · Docker · QLoRA

Built a RAG-based AI assistant that helps troubleshoot DevOps issues using retrieval, tool calling, and LLM reasoning. Reduced response latency from 28s to 10s through retrieval optimization and memory improvements.

Proposition RAG

github.com/tathagat-git/proposition-rag

Python · FAISS · Sentence Transformers · Ollama · LangChain

Designed a proposition-based RAG indexing approach inspired by Dense X Retrieval. Instead of embedding raw chunks, the system indexes atomic facts and retrieves full context through metadata, improving retrieval hit rate from 80% to 90% while reducing latency.