

Nazmul Huda Nabil

AI Engineer

github.com/NazmulHudaNabil | nabil648777@gmail.com | +880 1613813545 | Dhaka, Bangladesh

PROFESSIONAL SUMMARY

AI Engineer specializing in NLP, LLM applications, RAG, and backend AI systems. Hands-on experience building production-oriented applications with Python, FastAPI, PyTorch, Transformers, LangChain/LangGraph, Qdrant, PostgreSQL, and Redis. Experienced in developing and deploying AI solutions including resume screening platforms, production-grade LLM APIs, and predictive ML services, with a focus on scalable architecture, evaluation, security, and reliable AI integration.

TECHNICAL SKILLS

- AI & LLM:** LangChain, LangGraph, RAG, LLM Applications, Prompt Engineering, AI Agents, Guardrails, LLM Evaluation
- Machine Learning & Deep Learning:** PyTorch, TensorFlow, scikit-learn, Neural Networks, Hugging Face Transformers, XGBoost, CatBoost
- NLP:** BERT, Bangla-BERT, Text Embeddings, TF-IDF, Word2Vec, Semantic Search, Text Classification, Whisper
- Backend & APIs:** Python, FastAPI, REST APIs, Pydantic, SQLAlchemy, Docker
- RAG & Vector Databases:** Qdrant, ChromaDB, FAISS, PGVector, BM25, Hybrid Search, Semantic Chunking, Reranking
- Databases & Caching:** PostgreSQL, Supabase, Neon, Redis, Upstash Redis, MySQL
- MLOps & Deployment:** Docker, GitHub Actions, Render, Hugging Face Spaces, LangSmith
- Developer Tools:** Git, GitHub, VS Code, Jupyter, Google Colab

KEY PROJECTS

AI Resume Screening Platform | github.com/NazmulHudaNabil/AI-Resume-Screening-Platform

FastAPI · Streamlit · Neon Postgres · Qdrant Cloud · Upstash Redis · Groq Llama-3 · Google Gemini

- Built an end-to-end resume screening platform with a FastAPI backend and Streamlit frontend, using Groq Llama-3 for structured resume extraction and Google Gemini for embeddings.
- Designed a hybrid scoring engine combining semantic similarity search (Qdrant) with rule-based criteria to rank candidates, producing explainable AI justifications for each match.
- Architected the data layer on Neon Postgres for persistence and Upstash Redis for caching, and authored a full technical project report documenting system design and results.

Production-Ready LangGraph API | github.com/NazmulHudaNabil/production-langgraph-api

FastAPI · LangGraph · Mistral AI · LangSmith · Docker · Render

- Designed a production-grade LLM chat API around a LangGraph agent state machine with automatic primary-to-fallback model retry logic (Mistral Small → Mistral Medium), eliminating single-model failure points.
- Implemented a security pipeline that blocks prompt-injection attempts and masks PII (email, phone, SSN, credit card) on both input and output before/after LLM calls.
- Built TTL-based response caching and configurable rate limiting to cut redundant LLM calls and API costs, plus a metrics layer exposing latency, token usage, error rate, and cache hit rate via a /metrics endpoint.
- Integrated LangSmith tracing for full request-level observability and containerized the service with Docker and one-click Render.com deployment; covered the security and caching layers with an offline pytest suite requiring no live LLM calls.

PricePilot — Car Price Prediction API

Python · FastAPI · Machine Learning · MySQL · Render · Railway

- Built and deployed a FastAPI service that predicts car prices from a trained ML regression model, exposing a REST endpoint for real-time inference.
- Deployed the API to Render with Railway as an external MySQL host, handling production configuration and database connectivity across separate hosting providers.

EDUCATION

Daffodil Polytechnic Institute | Diploma in Computer Science | 2024 – Present

TRAINING & CERTIFICATIONS

Ostad AI Engineering Bootcamp | September 2025 – Ongoing

- Enrolled in an advanced AI Engineering program covering Python, Mathematics, Statistics, NLP, NLU, Computer Vision, LLMs, LangChain, Vector Databases, and Explainable AI.
- Developing 20 hands-on projects to deepen technical expertise in AI systems and agent development.