GANTUMUR BATTUMUR

Seattle, WA | 509-251-6832 | ganabattumur@gmail.com |LinkedIn | Github

SUMMARY

AI/ML Engineer specializing in production LLM systems, personalization pipelines, and scalable inference services. Experienced in designing end-to-end AI applications, including data ingestion, model integration, evaluation, and cloud deployment. Strong hands-on background in building reliable backend services and deploying AI-driven applications.

EXPERIENCE

Software Engineer Intern

F5 Networks - Liberty Lake, WA June 2024 - Oct 2024

- Automated Power Supply Unit validation and diagnostics workflows, reducing manual QA testing by approximately 70 percent and preventing recurring calibration failures.
- Developed Python-based automation for hardware calibration, fault detection, and structured data logging to support reproducible validation and debugging.
- Built real-time health-check pipelines and internal dashboards to improve system observability, reliability, and reporting for engineering teams.

Technologies: Python, internal automation frameworks, data logging pipelines, monitoring dashboards

Software Engineer

Seattle Mongolian Youth Center - Seattle, WA Jan 2025 - Jan 2026

- Built and maintained a production web application that scaled to over two thousand active users, contributing across frontend, backend, and data layers.
- Designed data models and retrieval logic to reduce query latency and improve responsiveness.
- Implemented backend APIs and database optimizations to support interaction tracking and content delivery.

Technologies: React, TypeScript, FastAPI, PostgreSQL, Python, Docker, AWS

PROJECTS

AI Vocabulary Learning Tool

- Built a Generative AI application using LLMs and retrieval-augmented generation (RAG) for stories and images for personalized vocabulary learning.
- Designed a full-stack AI system with a Next.js frontend, FastAPI backend, and PostgreSQL for the database.
- Implemented batched, cached, and asynchronous LLM inference pipelines, reducing generation latency by approximately 40 percent and lowering operational costs.

Technologies: Python, FastAPI, Next.js, PostgreSQL, Gemini API, RAG, Docker, REST APIs, asynchronous inference

Memory Training Web App

- Designed and implemented a personalization and recommendation pipeline using user interaction data and RAG-based context retrieval to adapt training content over time.
- Built Python-based backend services with FastAPI to support real-time inference, interaction tracking, and authenticated API access.
- Orchestrated LLM prompt pipelines and retrieval workflows using LangChain to improve relevance.

Technologies: Python, FastAPI, LangChain, PostgreSQL, React, Next.js, RAG, CI/CD, Docker, Render, Vercel

EDUCATION

Whitworth University - Spokane, WA | Sept 2021 - May 2025

Bachelor of Science in Computer Science, Minor in Mathematics