

HET PATEL

linkedin.com/in/het1074 | github.com/patelhet04 | hetpatel.dev | Boston, MA | hetpatel0499@gmail.com

SUMMARY

Software Engineer with 4+ years building production systems and AI solutions. Specialized in GenAI/LLMOps, full-stack development, and DevOps. Track record of implementing **cost-effective engineering solutions** achieving \$30K+ savings and 10x scalability through optimized GPU infrastructure and system design.

TECHNICAL SKILLS

Generative AI & ML: Agentic AI, vLLM, MCP (Model Context Protocol), LangChain, LangGraph, RAG Architecture, Pytorch, Transformers, Ollama, LlamaIndex, Hugging Face, Prompt Engineering, CUDA, Load Testing (k6)

LLMOps & Cloud: AWS (Bedrock, SageMaker, Lambda, EC2, S3, ECR, Route 53, CloudWatch), Docker, CI/CD (Jenkins, GitHub Actions), Nginx, PM2, Terraform, Docker, GitHub, Git

Programming & Web: Python, JavaScript, TypeScript, React, Next.js, Node.js, Express, Flask, FastAPI, Shell Scripting

Databases: Vector DB (FAISS, ChromaDB), DynamoDB, MongoDB, PostgreSQL, MySQL, Redis, GraphQL, RESTful APIs

PROFESSIONAL EXPERIENCE

DMSB AI Strategic Hub | AI Engineer | Boston, MA

Jan 2025 – Present

- Built an end-to-end GenAI grading solution with custom LangGraph agents and anti-gaming guardrails to automate rubric creation and bulk essay evaluation, **reducing faculty grading time by 95% (8+ hours to 30 mins)**.
- Engineered RAG pipeline (LlamaIndex + Qdrant HNSW) with semantic chunking and metadata filtering to index 30+ PDFs per assignment, **validated through A/B testing** against human grading to **achieve 90% accuracy**.
- Established on-premises infrastructure (NGINX + GitHub Actions + MinIO) with SSO authentication (JWT + 2FA) serving 1000+ users at 99.8% uptime, **reducing cloud costs by \$30K annually** while maintaining FERPA compliance.
- Led biweekly UAT sessions with faculty and RAs through a **3-month rapid prototyping cycle**, incorporating feedback to **deliver MVP 50% faster** than initial timeline (6 months to 3 months).
- Designed an AI presentation coach (Next.js + Flask API + F5-TTS) providing real-time audio analysis via containerized Whisper microservices and 20 Celery workers, **scaling from 10 to 100+ concurrent users**.
- Optimized GPU inference using vLLM docker container on NVIDIA Blackwell infrastructure after benchmarking against Ollama, achieving **5x faster throughput (8.5s to 1.7s)** for 100+ concurrent requests.
- Deployed application observability using Loki/Promtail/Grafana stack, **reducing MTTR from 2+ hours to under 30 mins**.

Silver WebBuzz Private Limited | Software Engineer II | Ahmedabad, Gujarat, India

Sept 2020 – Aug 2023

- Collaborated with 5 engineers and product team** on Node.js microservices migration using Docker containerization, Jenkins CI/CD, and Jest/Supertest automated testing, **reducing deployment from 4 hours to 15 minutes**.
- Developed batch processing pipeline using Python & GPT-4 to extract and normalize data from 200+ unstructured PDFs, automating manual data entry and reducing pharmacy **onboarding from 3 weeks to 2 days**.
- Spearheaded Node.js media processing pipeline for marketing website using AWS (Lambda, SQS, EC2) with FFmpeg video compression and S3 lifecycle automation, **achieving 40% storage cost reduction**.

EDUCATION

Master of Science - Information Systems, Northeastern University

Sep 2023 – Present

Bachelor of Engineering - Computer Engineering, Gujarat Technological University

Aug 2016 – Jun 2020

PROJECT WORK

Brevity - Smart News Summarizer | [github/Brevity](https://github.com/Brevity)

Apr 2025 – Jul 2025

- Constructed full-stack news summarizer with smart chatbot integration using Next.js, FastAPI, and DynamoDB, deployed serverlessly via AWS Lambda with Terraform IaC, **processing 500+ news articles weekly**.
- Utilized DistilBART for text summarization and **AWS Bedrock foundation models** for embeddings and LLM tasks.
- Integrated self-RAG system (LlamaIndex + ChromaDB) with query routing and web search, **achieving 85% hallucination reduction** through self-testing validation and Redis session management.