

Hiteshkumar Gupta

+1 782-882-6861 | Saskatoon, SK | hiteshgupta6011@gmail.com | linkedin.com/in/hiteshkumar-gupta | github.com/hmgtech | guptahitesh.me

Software Engineer with 4 years' experience across backend, AI, and frontend systems, passionate about LLMs, developer tools, and CI/CD pipelines.

WORK EXPERIENCE

Siemens EDA

SOFTWARE ENGINEER – AI

Saskatoon, Canada

Sept 2024 – Current

- **Engineered** and optimized **hybrid RAG pipelines** using BM25 sparse retrieval, metadata filtering, Reciprocal Rank Fusion, and reranking with Milvus, **improving retrieval accuracy by 50%** and enabling context-aware responses.
- **Designed** LLM evaluation and observability pipelines to benchmark retrieval and response quality, monitor hallucinations, and guide iterative model improvements.
- **Fine-tuned** domain-specific LLMs using PyTorch and Hugging Face Transformers, adapting pretrained models to specialized datasets to **improve task-specific output quality and relevance**.
- **Designed and deployed** ReAct-based agentic workflows with tool orchestration and optimized prompting, **increasing task success rate by 70%** for multi-step LLM workflows.
- **Built and scaled** Ray-based distributed AI infrastructure on Kubernetes across CPU/GPU clusters, implementing autoscaling for **LLM training, inference, and agent workloads** to handle dynamic compute demand.
- **Maintained Jenkins CI/CD pipelines** integrated with Jira webhooks to automate deployments to Kubernetes-based test environments, using **Argo CD, Helm**, and blue-green deployment strategies to support continuous delivery with rollbacks.
- **Designed and implemented** a Valkey Pub/Sub coordination layer for horizontally scaled chatbot replicas, ensuring **reliable request/response** routing and consistent session handling across replicas.
- **Engineered sandboxed** MCP-based agents for File/Shell interaction within restricted environments, enabling autonomous development workflows while maintaining **controlled execution boundaries**.

Manentia AI

SOFTWARE ENGINEER

Delhi, India

Sept 2022 – July 2023

- **Architected and deployed** microservices on AWS using EC2, S3, and Lambda with automated CI/CD pipelines, **improving deployment reliability by 40%**.
- **Designed and implemented** high-throughput RESTful APIs using Django and Flask, optimizing backend services for scalable web-based data processing.
- **Engineered** Python/Pandas data-processing pipelines and optimized workflows, **reducing processing time by 30%**.
- **Developed and optimized** PyTorch-based classification models, improving inference efficiency while maintaining high model accuracy across domain-specific workloads.
- **Containerized** ML inference services using Docker and AWS Lambda and implemented automated testing and performance monitoring, **reducing deployment overhead by 50% and system downtime by 25%**.
- **Led peer code reviews and mentored junior engineers**, establishing maintainable and testable engineering practices across the AI team.

Mobiuso Technologies Pvt. Ltd.

SOFTWARE ENGINEER

Ahmedabad, India

June 2021 – Aug 2022

- **Developed** containerized microservices using Node.js, Express.js, and Docker, improving application scalability and service reliability.
- **Optimized** Sequelize ORM schemas and database queries, **improving query performance by 40%** and reducing API response latency.
- **Built** RESTful APIs supporting **20K+ monthly users** while maintaining less than **200ms average response latency** in production.
- **Optimized** React/Next.js applications using lazy loading, memoization, and code splitting, **reducing initial page load time by 15%**.
- **Streamlined** frontend data processing and date-handling workflows using Lodash and Moment.js, **reducing client-side processing overhead by 25%**.
- **Collaborated with cross-functional teams** to deliver production features under client SLAs, contributing to reliable and timely releases.

PROJECTS

AskHim (<i>FastAPI, React, ChromaDB, Tree-sitter, Llama 3.1B, CLI</i>)	GitHub Link
<ul style="list-style-type: none">• Built an LLM-powered code intelligence assistant for exploring and understanding large codebases, reducing developer code-search time by ~30% through semantic retrieval and contextual code generation.• Engineered an AST-aware code ingestion pipeline using Tree-sitter to parse and structurally chunk 50K+ code snippets, generating embeddings and indexing them in ChromaDB for efficient semantic code retrieval.• Developed streaming inference across both React Web UI and CLI interfaces, delivering responses in <2 seconds and enabling interactive code exploration with a locally hosted Llama 3.1B model.• Designed the retrieval and context pipeline to ground LLM responses in relevant codebase context, improving code understanding and reducing reliance on the model's parametric knowledge.	

QuickHire (<i>ReactJS, NodeJS, MongoDB, Stripe, CI/CD, Git</i>)	Website / GitHub Link
<ul style="list-style-type: none">• Built a full-stack service marketplace using ReactJS, NodeJS, and MongoDB, enabling users to discover, list, and purchase services through a scalable web application.• Designed and implemented RESTful APIs with OAuth 2.0 and JWT-based authentication, securing user registration, login, profile management, and protected application resources.• Engineered service discovery, search, wishlist, and feedback workflows, improving content discoverability and enabling users to evaluate and manage marketplace offerings.• Integrated Stripe-based payment workflows and automated CI/CD pipelines, streamlining production deployments and supporting reliable transaction processing.	

EDUCATION

Dalhousie University	Halifax, Canada
Master of Applied Computer Science	Sept 2023 – Dec 2024
Parul University	Vadodara, India
Bachelor of Technology in Computer Science and Engineering	June 2017 – May 2021

PUBLICATION

Title: Cleanliness Automation: YoloV3, IEEE	Paper Link
Publisher: IEEE at <i>2021 6th International Conference for Convergence in Technology</i>	April 2021
<ul style="list-style-type: none">• Designed and trained a custom 256-layer CNN-based object detection model using YOLOv3 and TensorFlow, achieving 98% detection accuracy with 0.2% loss on real-world waste datasets.• Engineered an automated data-labeling and preprocessing pipeline, reducing dataset preparation time by 75% and streamlining model training workflows.• Deployed the trained vision model through a Flask application for real-time waste detection and classification from live camera feeds.	

SKILLS

Languages: Python, C++, Java, JavaScript/TypeScript
LLM & AI: PyTorch, Transformers, LangChain, LangGraph, MCP, RAG, LLM Fine-Tuning, LLM Evaluation & Observability
AI Systems & Infrastructure: Ray, Kubernetes, Docker, FastAPI, Distributed Training & Inference, Model Serving, CI/CD
Retrieval & Data: PostgreSQL, Milvus, pgvector, ChromaDB, MongoDB, MySQL, DynamoDB
Cloud & Backend: AWS (EC2, Lambda, CloudFormation, CloudWatch), GCP, Node.js, Django, REST APIs, React.js