

Kunal Kashyap

Email: kunal.kashyap.8775@gmail.com **Phone:** +91-8743036307 **Portfolio:** kunal-1.vercel.app
GitHub: github.com/kunalkashyap-1 **LinkedIn:** linkedin.com/in/kunal-kashyap-a0798222b

Professional Summary

AI Systems Engineer with **2+ years of experience** building backend systems and LLM-powered applications. Experienced in Python, FastAPI, RAG, local LLM inference, real-time AI pipelines, and production API development, with additional full-stack experience across the MERN ecosystem.

Core Skills

- **AI/ML:** LLMs, RAG, Semantic Search, Prompt Engineering, Model Serving, Local LLM Inference
- **LLM Ecosystem:** Ollama, Hugging Face, Transformers, Whisper, Groq, ChromaDB
- **Backend:** Python, FastAPI, Flask, Django, Node.js, Express.js, REST APIs, WebSockets, SSE
- **Databases Infrastructure:** PostgreSQL, MySQL, MongoDB, Redis, Docker, Docker Compose, Linux
- **Frontend:** React.js, Next.js, SvelteKit, TypeScript, JavaScript
- **Engineering:** Git, GitHub Actions, API Design, Authentication/JWT, Caching, Async Programming, System Design

Work Experience

System Engineer

Tata Consultancy Services (TCS)

July 2025 – Present

- Delivered **AI/ML solutions** across Insurance and Financial Services, reducing manual processing effort by approximately **30%** across targeted automation workflows.
- Built asynchronous **Python/FastAPI** services using AsyncIO, Pydantic, and OracleDB, improving processing performance through asynchronous execution and caching.
- Developed **fraud detection and predictive analytics** solutions using XGBoost, Scikit-learn, and SHAP to support risk analysis and improve model interpretability.

STACKIOT TECHNOLOGIES Pvt. Ltd.

January 2024 – July 2025

- Developed and maintained production web applications and backend services using **Node.js, Express.js, React.js, and MySQL**.
- Designed and optimized **REST APIs**, refactored legacy modules, and improved backend performance, reliability, and maintainability.
- Reduced recurring application downtime by approximately **20%** through systematic debugging, backend optimization, and performance improvements.

Projects

Friday: Local AI Voice Assistant

- Built a **local-first AI voice assistant** with real-time speech recognition, LLM integration, text-to-speech, wake-word detection, and modular skill-based orchestration.
- Developed a modular pipeline integrating **faster-whisper, Silero VAD, Ollama, Kokoro TTS**, and extensible skills for web search, music control, reminders, and system automation.
- Implemented **hybrid local/cloud LLM support** with cross-platform audio and system integration for Windows and Linux.
- **Tech:** Python, Faster-Whisper, Silero VAD, Kokoro TTS, Ollama, Groq

Nova: AI Chat Assistant with RAG

- Developed a full-stack **AI chat platform** using **FastAPI, SvelteKit, PostgreSQL, Redis, and ChromaDB**, supporting multi-model conversations and real-time interactions.
- Implemented **RAG, hybrid search, semantic memory, web search, and context optimization** using ChromaDB, Redis caching, and external search capabilities.
- Built a modular **LLM service layer** supporting both Ollama and HuggingFace models, with JWT authentication, file attachments, model management, and token/cost tracking.
- **Tech:** Python, FastAPI, SvelteKit, Ollama, HuggingFace, ChromaDB, Redis,

Git-Go: A Simple Git Implementation

- Git-Go is a lightweight, custom-built implementation of few Git commands in Go, designed to replicate Git's core functionality for learning purposes.
- To gain a deeper understanding of how Git operates internally, improve problem-solving skills, and experiment with the Strategy Pattern for clean and maintainable command execution logic.
- **Tech:** Go, Git Internals, SHA-1, File I/O, Strategy Pattern

Vaartalap: Real-Time AI Translation and Transcription

- Led the development of a real-time AI system for translation and transcription in video conferencing using Whisper.
- Integrated multilingual language support, real-time processing, and dynamic language detection using Whisper AI.
- **Tech:** Python, Whisper, Flask, Next.js, TypeScript, Speech Processing

Education

B.Tech in Computer Science and Engineering

2021 - 2024

Dr. Akhilesh Das Gupta Institute Of Technology, GGSIPU

CGPA: 9.22 / 10

Achievements

-
- Published research paper: "VAARTALAP: Embedding Whisper AI-like Model into a Video-Conferencing System to Aid Real-Time Translation and Transcription" (2023)

Certifications

-
- **The AI Agent Engineer Course: Complete AI Agent Bootcamp** Jan 2026
 - Gained practical knowledge of AI agent development, including agent architectures, tool calling, workflows, and LLM-based automation.
 - Explored building intelligent agents capable of reasoning, interacting with tools, and executing multi-step tasks.
 - **The Complete Prompt Engineering for AI Bootcamp (2025)** Sep 2025
 - Developed practical skills in prompt engineering, covering structured prompting, task decomposition, and improving LLM response quality.
 - Applied prompting techniques for reliable and effective interaction with generative AI systems.
 - **Fundamentals of Back-end Engineering** Jan 2025
 - Gained in-depth knowledge of back-end engineering, including HTTP/HTTPS protocols, RESTful API design, database indexing, and distributed system architecture.
 - Developed expertise in performance optimization, encompassing load balancing, scaling, caching strategies (Redis), and security practices.
 - **The Complete 2023 Web Development Bootcamp** Nov 2023
 - Acquired expertise in full-stack web development using HTML, CSS, JavaScript, Node.js, and Express.js.
 - Built and deployed dynamic web applications incorporating MongoDB, RESTful APIs, and user authentication.