# SHUSHANT KAMATAR

✉ shushantkamatar195@gmail.com  📞 +91-8088348874  in LinkedIn  �github Github  ⦿ Coding Profile

## Experience

**Forin Agency**                                                                                  Nov 2025 – Present
*Machine Learning Engineer*

- Architected a high-concurrency real-time **Interview Intelligence System** using **FastAPI** and **WebSockets** to enable full-duplex communication, achieving **< 800ms latency** by parallelizing audio transcription **(Deepgram)** and LLM inference to support seamless , lag-free live interactions.
- Engineered a context-aware **Retrieval-Augmented Generation (RAG)** with **Supabase (pgvector)** power a live recommendation engine to dynamically retrieve candidate resume data and historical "winning questions" during live calls, enhancing interviewer decision-making speed.
- Developed a secure, production-grade Chrome Extension utilizing the specialized **'Offscreen' API** to capture high-fidelity system audio streams while effectively bypassing standard browser sandbox restrictions, ensuring reliable audio data ingestion for the inference engine.

**Lenovo**                                                                                  May 2024 – Sep 2025
*Software Engineer*

- Engineered a robust real-time **NLP Inference Pipeline** using **Python** and **Flask** to continuously stream and process sentiment analysis on high-volume product feedback, effectively transforming raw unstructured text into live , actionable strategic intelligence for product stakeholders.
- Optimized critical legacy ETL pipelines by re-architecting iterative Python loops into efficient **vectorized Pandas operations**, reducing processing time by **99%** (slashing execution windows from multiple hours to mere seconds) and enabling **real-time analytics** for the leadership team.
- Orchestrated the complex technical deployment of data-driven features across **Western Europe (15 markets)**, meticulously managing diverse localization requirements and ensuring **system reliability** during high-traffic regional rollouts and peak user activity periods.

## Skills

- **Languages:** Python, JavaScript, SQL, HTML/CSS
- **Frameworks:** FastAPI, Flask, PyTorch, Deepgram, LangChain
- **Tools:** Docker, Kubernetes, Github Actions, MLflow, DVC, Prometheus
- **Machine Learning & AI:** Sciketlearn, NLP, RAG, Vector Embeddings
- **System Design & Backend:** AsyncIO, Redis, WebSockets, Supabase, FAISS

## Education

**RV College of Engineering**                                                                   Jun 2019 – Aug 2023
B.E in Aerospace Engineering;  *CGPA: 7.07*                                                      *Bangalore Urban*

## Projects

**Enterprise Natural Language-to-SQL Engine** ⧉                                                      Dec 2025

- Architected a semantic schema retrieval system using **FAISS** to dynamically inject relevant database context into the LLM context window, reducing token usage by 90% and enabling robust support for complex schemas **(100+ tables)** by eliminating hallucination risks.
- Migrated inference stack from GPT-4 to **Llama 3.3** on **Groq LPUs**, reducing query latency to **<200ms** and operational costs by **75%** while maintaining SQL syntax accuracy through the implementation of advanced prompt distillation techniques and rigorous regression testing.
- Engineered a deterministic **Self-Correction Loop** that autonomously parses database execution errors (e.g., 'no such column' or syntax faults) to iteratively repair and re-execute malformed SQL queries, successfully recovering **40%** of failures without human intervention.

**Tech Stack:** SQLite, Python, Redis, Docker, FAISS , Streamlit

**High-Throughput Inference Gateway** ⧉                                                             Nov 2025

- Developed a resilient, unified API gateway using **FastAPI** and **Async Queues** to efficiently route concurrent requests across multiple heterogenous LLM providers, ensuring high availability through the implementation of automated failover strategies and intelligent load balancing .
- Implemented a hybrid semantic caching layer combining **Redis and Vector Embeddings** to intercept and serve recurring natural language queries , improving **P99** latency for recurring queries by **600ms** and reducing external API costs by **30%** by serving responses directly from the cache.

**Tech Stack:** FastAPI, Python, Redis, Docker

## Certificate

- Nasscom Certified Data Scientist ⧉