Sonal Singh

GitHub: github.com/sonallsinghh | LinkedIn: linkedin.com/in/sonallsinghh | sonalsinghh01@gmail.com | +91 8884903974

Summary

AI Engineer (Backend/GenAI) with 1+ year of experience building production RAG systems and real-time voice/chat agents. Develop low-latency LLM applications using FastAPI, Redis, and AWS, covering end-to-end document ingestion -> embeddings/indexing -> retrieval/reranking -> grounded generation with LangChain/LangGraph/LlamaIndex. Strong in event-driven and streaming architectures, LLM tool/function calling, observability-driven debugging, and full production ownership (reliability, evaluation, incident response).

Work Experience

Software Development Engineer — Webknot Technologies Jan 2025 - Present

AgentX

- Built an AI Agent Creation Platform for enterprise voice + chat agents; introduced reusable agent templates and automated provisioning (secrets, connectors, Twilio/webhook setup), reducing manual setup steps from **approx. 12** to **approx. 3** and enabling **40+** deployments.
- Implemented agentic tool/function calling with guardrails (validation, retries, fallbacks), integrating **RAG** via embeddings (text-embeddings-openai) plus API actions (CRM/ticketing/internal services); improved task completion rate by **approx. 18%** and reduced failed tool calls by **approx. 35%**.
- Engineered a real-time voice pipeline with Twilio WebSockets, Deepgram STT, and ElevenLabs TTS with streaming I/O, session state, and observability; achieved **p95 end-to-end voice turn latency approx. 1.4s** and handled **25+** concurrent calls with **99.5%** session success rate.

ScaleTrain

- Engineered a RAG-powered learning assistant using FastAPI, embeddings, and LangGraph to answer course-related queries, serving **approx. 1k–5k queries/day** (est.) with **p95 700–1200 ms**.
- Developed ingestion + chunking + metadata pipelines to convert course content into retrieval-ready context (**approx. 5k–20k chunks indexed**, est.) enabling knowledge tracing and personalized feedback.
- Created a low-latency event-driven backend with PostgreSQL, Redis, Kafka, and Server-Sent Events (SSE), reducing repeat-query latency by **approx. 30–50%** (est.)

Projects

Clear Visa

- Built a real-time mock visa interview platform with AI agents simulating visa officers for immersive voice/video practice.
- Developed FastAPI backend services with PostgreSQL schemas for secure user management, session tracking, and interview history.
- Integrated LiveKit and LiveKit Agents with LangChain and LLM orchestration to enable adaptive multi-turn interviews and real-time streaming interactions.

LexiLearn

- Designed LexiLearn using Gemini 2.0 Flash with diagnostic assessments, personalized learning goals, and adaptive learning tracks.
- Added long-context conversational memory retaining approx. 80% of multi-turn context to improve personalization quality and progress tracking accuracy.

Education

2021 - 2025 B.E Computer Science and Engineering at **RNSIT** (GPA: 9.2/10.0)

Skills

Languages	Python, JavaScript (Node.js)
Backend & AI	FastAPI, Agentic Workflows, Async Python, Hybrid Retrieval + Reranking, Prompt Injection Mitigation, LangChain, LangGraph, RAG, Embeddings, LLM Agents, WebSockets
Databases	PostgreSQL, MySQL, MongoDB, Redis
Developer Tools	Docker, REST APIs, Azure OpenAI, GCP, AWS, Git