

S M RUSSEL

Full-Stack AI Engineer & Systems Architect

Website: smrussel.dev | Email: smrusselofficial@gmail.com | GitHub: github.com/smrussel | LinkedIn: linkedin.com/in/smrussel

EXECUTIVE SUMMARY

Specialist Full-Stack AI Engineer and Systems Architect with deep expertise in autonomous multi-agent systems, sub-700ms real-time voice AI agents, enterprise retrieval-augmented generation (RAG) pipelines, and containerized cloud SaaS solutions. Proven track record of designing high-concurrency microservices, stateful graph orchestrations, and production AI architectures combining Python, Next.js, FastAPI, and Docker.

CORE TECHNICAL COMPETENCIES

- **AI & Multi-Agent Systems:** LangGraph, CrewAI, AutoGen, OpenAI Realtime, LiveKit WebRTC Voice Agents, Semantic Router, Tool Calling, DSPy
- **RAG & Knowledge Retrieval:** Hybrid Search (Dense + BM25), Vector DBs (Qdrant, Pinecone, ChromaDB), Cohere Rerank, Document Chunking &
- **Backend & Infrastructure:** Python, FastAPI, Node.js, Express, PostgreSQL, Redis, Docker, Container Orchestration, Linux, AWS, GCP, CI/CD
- **Frontend & Web Engineering:** TypeScript, Next.js, React, Astro, Tailwind CSS, WebSockets, State Management, Responsive & High-Performance
- **Systems & Architecture:** Event-Driven Microservices, Async Queues (Celery/BullMQ), API Gateway Design, Rate Limiting, High-Availability Systems

FEATURED SYSTEMS & ARCHITECTURE HIGHLIGHTS

Autonomous Multi-Agent Task Orchestration Engine

Lead Architect | 2024 - Present

- Architected cyclic, stateful multi-agent workflows using LangGraph and Python, managing human-in-the-loop validation and distributed checkpointing.
- Engineered dynamic tool-routing and sub-agent task delegation patterns, reducing workflow latency by 42% across multi-step execution graphs.

Real-Time Voice AI Assistant Pipeline (<700ms Turnaround)

Systems & AI Engineer | 2024

- Built an end-to-end streaming conversational voice agent integrating WebSockets, WebRTC (LiveKit), VAD (Silero), and ultra-fast streaming TTS/STT.
- Achieved sub-700ms voice-to-ear latency with adaptive audio frame chunking, interruption handling, and streaming token ingestion.

Enterprise Multi-Tenant RAG Knowledge Base Platform

Full-Stack AI Architect | 2023 - 2024

- Constructed hybrid semantic retrieval pipeline using Qdrant vector database and cross-encoder rerankers, ensuring deterministic source citations.
- Implemented tenant-isolated document partitioning, background embedding workers via Celery/Redis, and real-time streaming Next.js dashboard.

Containerized Cloud SaaS & Microservices Infrastructure

Full-Stack Engineer | 2023

- Standardized multi-stage Docker build containers with automated CI/CD deployment pipelines, zero-downtime rolling updates, and health monitoring.
- Developed type-safe REST & GraphQL APIs using FastAPI, Pydantic, and PostgreSQL with connection pooling and Redis cache invalidation.

EDUCATION & PROFESSIONAL FOUNDATION

Bachelor of Science in Computer Science & Engineering

Continuous practical research in Distributed Systems, Generative AI Agent Architectures, and Cloud Computing.