Shehryar Sohail

Senior AI Engineer | LLM Systems and Voice AI

Lahore, Pakistan | +92 316 411 2721 | sherryyatwork@gmail.com
linkedin.com/in/shehryar718 | github.com/itsmesherry | itsmesherry.com

Summary

AI engineer who ships production LLM and voice AI systems. Built a multi-agent voice platform for an insurance company that handles 10% of live inbound calls at sub-800 ms average latency, with under 10% human handoff. Former co-founder and CTO of an AI services firm building LLM, speech, and computer vision systems for clients. Author of three open-source Python packages with 26,000+ combined PyPI downloads. Strong in RAG, agentic workflows, structured extraction, real-time speech pipelines, and LLM fine-tuning.

Experience

Paklogics Oct 2024 – Oct 2026
Senior Python AI/ML Engineer, AI Team Lead *Lahore, Pakistan*

- Architected and shipped a multi-agent voice AI system for an insurance company: 1 orchestrator routing to 4 specialist agents (health, car, travel, bike) on ElevenLabs, automating 10% of live inbound call traffic and capturing qualified leads across insurance lines.
- Held sub-800 ms average response latency on complex multi-step agentic workflows using asynchronous tool calls, with human handoff under 10% for the lines of business covered.
- Built a SIP bridge (PJSIP/PJSUA2) connecting the client's on-premise 3CX phone system to the OpenAI Realtime API, then migrated the stack to ElevenLabs Conversational AI.
- Tested and evaluated agents with automated voice QA (Cekura), ElevenLabs evaluation criteria, and LLM-as-judge scoring; monitored production calls through Axiom logs.
- Owned deployment end to end before handing operations to a dedicated DevOps team; built content search and curation automation for an ed-tech platform.
- Led a 3-engineer AI team, trained 5+ junior engineers and interns, and served as primary technical contact for clients, writing proofs of concept and feasibility reports with business development.

VECTOR Inc. Jan 2023 – Aug 2024
Co-founder and CTO *Islamabad, Pakistan*

- Co-founded an AI services firm delivering LLM, speech, and computer vision work, from applied research builds to production systems for business clients; ran engineering, hiring, client delivery, and B2B partnerships; won the DevFest 2023 pitch competition presenting the company's product.
- Shipped SkyLocker, a WhatsApp assistant that stores, searches, and returns users' files using multimodal RAG over text and images (Milvus, ChromaDB, Hugging Face embeddings, AWS S3, OpenAI, Ollama).
- Fine-tuned LLMs including Falcon-7B, RedPajama, and CodeGen-2B (LoRA) and GPT-3.5 / GPT-4o (OpenAI fine-tuning); built natural-language-to-SQL systems.
- Led R&D on Converso, a real-time voice bot (Faster-Whisper, ElevenLabs and Coqui TTS, Silero VAD for interruptions, FastAPI over WebSockets); contributed to the open-source OpenVoiceChat library.
- Built automatic closed-caption generation and a Python video-processing backend powering editing tools for a media client; implemented lane detection and trajectory prediction on nuScenes.

AIMs Lab, FAST-NUCES Jan 2022 – Jan 2023
AI/ML Research Trainee *Islamabad, Pakistan*

- Fine-tuned BERT, XLM-RoBERTa, GPT-2, ResNet, and ViT models for text and image classification; built tweet sentiment models and contributed POS tagging and NER work to an Urdu NLP library.

Projects and Open Source

ocr-llm | Python, OpenAI and Gemini vision APIs, pypi.org/project/ocr-llm

- Built a PDF-to-Markdown and information-extraction library using vision LLMs after Mistral OCR and CPU-based OCR proved unreliable on a client's scanned credit card statements; validated extraction by recomputing sub-table totals and reconciling transactions against the net amount. 3,000+ PyPI downloads.

WebLM | Python, FastAPI, Pydantic, React, weblm.dev

- Built and launched a web-to-Markdown API with an open-source Python SDK that cleans pages for LLM ingestion and extracts typed structured output into Pydantic models; shipped API-key auth, Paddle subscription billing, and a microservice setup with caching and analytics.

async-limiter | Python, asyncio, pypi.org/project/async-limiter

- Created an asyncio rate limiter combining time-window limits (X calls per N minutes) with concurrency caps, fixing leaks in existing async limiters while batching OpenAI and Gemini API calls. 15,000+ PyPI downloads.

titans-unofficial | PyTorch, github.com/itsmesherry/titans-unofficial

- Implemented Google Research's Titans paper (Learning to Memorize at Test Time) in PyTorch, covering all three variants (Memory as Context, Gate, and Layer) with CUDA support, tests, and CI. 8,000+ PyPI downloads.

claude-audit | TypeScript, Node.js, github.com/itsmesherry/claude-audit

- Published an AI-powered codebase auditing CLI on npm (`npx claude-audit`); 1,200+ npm downloads.

Skills

Languages: Python, TypeScript, JavaScript, SQL, C++
LLMs and agents: OpenAI, Claude, Gemini, Llama; LangChain, LlamaIndex, CrewAI; RAG, tool and function calling, structured output, prompt engineering, multi-agent orchestration
Voice AI: ElevenLabs Conversational AI, OpenAI Realtime API, Faster-Whisper, Silero VAD, SIP (PJSIP/PJSUA2), 3CX
ML and fine-tuning: PyTorch, Hugging Face Transformers, LoRA/PEFT, OpenAI fine-tuning, scikit-learn
Evaluation and observability: LangSmith tracing, Axiom logging, LLM-as-judge, Cekura voice QA
Data: PostgreSQL, MongoDB, SQLite, Milvus, ChromaDB, Pinecone
Backend and cloud: FastAPI, Flask, Next.js, Docker, AWS (S3, EC2, Lambda), GCP, Nginx, CI/CD

Education

National University of Computer and Emerging Sciences (FAST-NUCES) Sep 2020 – Jun 2024
BS Artificial Intelligence *Islamabad, Pakistan*

- CGPA 3.5/4.0; Dean's Honor List (4 semesters); Silver Medal for second-highest semester GPA in the university.