

SAHIL CHAVAN

AI Systems Engineer

Navi Mumbai, India — +91 9834020445 — sahildchavan04@gmail.com

[LinkedIn](#) — [GitHub](#) — [Portfolio](#)

Summary

AI Systems Engineer with **3+ years of experience** architecting and deploying production-grade **Speech AI, Voice AI, Computer Vision, and LLM-powered systems**. Expertise in **Language Identification (LID), Wake-Word Detection, Prompt Engineering, Context Engineering, Memory Systems, Tool Calling, LLM Evaluation, and AI Agent Design**. Proven experience developing **Computer Vision-based AI agents**, fine-tuning **MMS, ECAPA-TDNN, and wav2vec 2.0** across **8 Indic languages**, and building **Gemini Live intelligence systems** for scalable real-time conversational experiences.

Technical Skills

AI / ML: Speech Recognition, Language Identification (LID), Wake-Word Detection, LLM Evaluation, Prompt Engineering, Computer Vision, Deep Learning, Model Fine-Tuning, Model Benchmarking

Speech & Audio: Facebook MMS, ECAPA-TDNN, wav2vec 2.0, Silero-VAD, Librosa

Frameworks: PyTorch, TensorFlow, OpenCV

Programming: Python

Tools & Cloud: AWS Sagemaker, Aethir, XOR, Langfuse, Firebase

Experience

AI Research & Development Engineer — Jio Platforms Limited (JPL)

Navi Mumbai, India

June 2025 – Present

- Fine-tuned **Language Identification (LID)** models using **Facebook MMS and ECAPA-TDNN** for **8 Indic languages**, improving robustness under noisy and real-world audio conditions.
- Designed and deployed a **wake-word detection system** using **wav2vec 2.0** for the production phrase “**Hello Jio**”, optimized for **low-latency, always-on inference**.
- Architected production-grade **prompt engineering systems** for **Gemini Live**, including **system prompts, developer prompts, web-search orchestration, tool-calling workflows, and response optimization strategies**, improving response quality, reliability, and latency.
- Architected an **AI intelligence layer** integrating **memory management, context engineering, tool calling, and LLM orchestration**, enabling production-grade voice AI agent workflows.
- Implemented **prompt versioning management** using **Langfuse and Firebase**, enabling controlled deployment, evaluation, and iteration of production prompts for LLM-based systems.
- Conducted **LLM evaluation and benchmarking** using **Perplexity, GPT-4, EXA AI, and LinkUp AI**, focusing on **latency, caching strategies, and response consistency**.

Artificial Intelligence Developer → Senior Artificial Intelligence Developer — iAgent-Protocol

Pune, India

May 2023 – May 2025

- Improved AI agent performance by **30%** through dataset optimization, algorithm tuning, and behavioral modeling.
- Reduced AI training time from **8 hours to 4.5 hours (43%)** by designing efficient metadata, labeling, and preprocessing pipelines.
- Developed **AI-driven FPS gaming agents** using **deep learning, behavioral cloning, and adaptive decision systems**, improving gameplay stability and realism.
- Applied **Computer Vision** using **OpenCV** for real-time perception, navigation, and decision-making for gaming AI bots.
- Collaborated within a **4-member AI team** to build scalable experimentation workflows, reducing gameplay bugs by **20%** and accelerating model iteration cycles.

Projects

- **LLM Intelligence System & Prompt Engineering:** Designed and maintained production-grade **Gemini Live** prompts, including **system, developer, tool-calling, web-search, and response orchestration prompts**. Built intelligence-layer capabilities around **memory management, context engineering, tool/function integration, and LLM routing**, while implementing **prompt versioning, evaluation, and deployment workflows** using **Langfuse and Firebase**.
- **Language Identification (LID) – Indic Languages:** Fine-tuned **MMS and ECAPA-TDNN** models for **8 Indic languages**, addressing accent variation, background noise, and code-mixed speech with evaluation pipelines for real-world deployment.
- **Wake-Word Detection – “Hello Jio”:** Built a **wav2vec-based wake-word system** optimized for **low false positives and low latency**.

Education

B.Tech in Computer Science & Engineering, Dr. D. Y. Patil International University, Pune, India

July 2019 – July 2023