AI Engineer

Location: Bengaluru – JP Nagar (On-site)
Type: Full-time
Experience: 3–5 years

About SaralX

SaralX was founded in July 2023 with a simple belief: technology should include, not exclude. People still struggle to access digital services due to disabilities, SaralX is working to rewrite how the internet is experienced.

In less than three years, we have partnered with 150+ leading Indian brands — including MakeMyTrip, Yatra, Bandhan AMC, EaseMyTrip, redBus, Rapido, Zepto, Money Control, BSNL, CRISIL, Navi AMC, Karur Vysyay Bank, and Kuku FM — helping them make their digital platforms accessible to users of all abilities.

Our work is trusted and backed by institutions that matter. SaralX is incubated at NSRCEL - IIM Bangalore, and supported by respected global and national accelerators such as AssisTech Foundation, ATTVARAN, SACC, Wadhwani Foundation, Seed Stars, and YouthCo:Lab.

We are proud recipients of the I-Innovate Award, and have been empanelled with the Government of India since Feb 2025, enabling us to make government websites and mobile applications accessible at scale.

SaralX is a social impact company committed to building a more inclusive digital world by making technology accessible to people of all abilities. Our mission goes beyond business - we aim to create meaningful change by ensuring equal access to digital experiences for everyone.

Role Overview

We are building an in-app voice copilot for banks and BFSI companies — a personal assistant embedded inside banking apps and web portals that lets customers ask anything about their account and get things done by voice ("block my card", "what's my EMI date", "transfer ₹5,000") in English, Hindi, and Indic languages. We ship as an SDK that any bank can drop into its app, with a roadmap to our own fine-tuned speech models and on-prem deployment for banks.

What You'll Own

- The real-time voice pipeline end to end: build and own our speech-to-speech loop (VAD → STT → LLM agent → streaming TTS) on frameworks like Pipecat or LiveKit Agents, targeting sub-second voice-to-voice latency with barge-in/interruption handling.
- The agent brain: design the LLM layer — prompt architecture, function/tool calling against banking APIs (balance, transfers, card actions), confirmation flows for money movement, guardrails, and PII redaction.
- RAG & knowledge: build retrieval over bank product docs/FAQs (vector DB, embeddings, chunking, evaluation) so the assistant answers accurately, not confidently-wrong.
- Model layer (Phase 2, post-funding): fine-tune and distill open-weight models — Whisper/IndicConformer for banking-domain Hinglish ASR (LoRA/PEFT, Distil-Whisper-style distillation), Indic TTS fine-tuning, and self-hosted LLM serving on vLLM — moving us off third-party APIs onto our own models.
- Evaluation & data: set up transcription-accuracy (WER) and task-completion benchmarks, and build the pipeline that turns consented call audio into training data.
- Engineering quality: latency profiling, cost-per-minute optimization, and production reliability of the voice stack.

Must-Have

- 3–5 years in ML/AI engineering with at least 1–2 years hands-on with LLMs in production (prompting, function calling, RAG — not just notebook experiments).
- Strong Python plus solid software engineering fundamentals (APIs, async, WebSockets/WebRTC basics, Docker).
- Hands-on experience with speech models: fine-tuning or deploying Whisper (or similar ASR), and/or working with TTS models — you should know what WER, RTF, and streaming synthesis mean.
- Experience with the Hugging Face ecosystem (transformers, PEFT/LoRA fine-tuning, datasets).
- Comfort reading model cards and licenses, evaluating open-weight models, and making build-vs-API tradeoffs.
- Bias to ship: you've taken at least one ML-powered feature from idea to production users.

Good to Have

- Built a real-time voice agent before (Pipecat, LiveKit, Vocode, Vapi/Retell, or hand-rolled) - this is the single strongest signal for us.
- Indic language / Hinglish NLP or ASR experience (AI4Bharat models, IndicVoices, code-switching), or native fluency in Hindi plus another Indian language.
- Model distillation or quantization experience (Distil-Whisper recipe, GGUF/AWQ, TensorRT).
- Self-hosted LLM serving (vLLM, SGLang) and GPU infra experience; on-prem or VPC deployment exposure.
- Telephony/audio plumbing: Twilio/Exotel/Plivo media streams, codecs, echo/noise handling.
- BFSI or other regulated-domain exposure (understands why audit logs, consent, and data residency aren't optional).
- Open-source contributions or a public project we can look at.

Why Join Us

- You're not tuning someone else's pipeline, you're building ours from scratch.
- Work on genuinely hard, unsolved problems: production-grade Hinglish voice AI doesn't exist yet; you'll be building it.
- India's BFSI AI market is exploding (banks are declaring AI strategy in their annual reports); we're building the picks and shovels.
- Direct access to customers and the founder — your decisions ship to real banking users.

How to Apply

Interested candidates should send the following to career@saralx.com:

- Resume
- A short note on why you are interested in this role and SaralX's mission

Email Subject: Application for AI Engineer