

Pukar Chalise

AI/ML Engineer · Voice AI · Agents · RAG

AI Engineer specialising in AI agents, large language models, voice AI and retrieval-augmented generation. I design agentic workflows with LangGraph, LangChain, tool calling and vector databases, and ship them as scalable production systems with FastAPI, Docker and AWS. I'm passionate about automating software engineering with intelligent agents.

EXPERIENCE

AI Engineer · MeetGabbi

Feb 2026 - Jun 2026 · Cincinnati, USA (remote)

- Built a HIPAA-compliant AI voice receptionist for a US orthopedic hospital on an agentic architecture: scheduling, patient intake, triage routing and follow-ups.
- Delivered a production voice AI system for ARC Psychiatry, with clinical-grade prompt engineering and conversation design that accounts for compliance.
- Engineered the core voice engine on the GPT Realtime API: VAD gating against false triggers, multi-speaker awareness for turn-taking, and domain-specific prompt architecture.
- Designed the Twilio telephony architecture: inbound/outbound routing, PSTN bridging and real-time audio streaming to the Realtime backend.
- Deployed real-time voice AI for US law firms (client intake, legal Q&A, call analytics), and built LangGraph agent workflows with tool calling and API integrations.
- Supported the onboarding team: standardised client workflows, mentored members and automated onboarding.

AI Developer / ML Engineer · Code Himalaya

Nov 2024 - Feb 2026 · Kuponhole

- Designed and built AI systems including Video KYC, intelligent document parsing and conversational chatbots, using computer vision, OCR, NLP and LLMs.
- Built AI agents and RAG pipelines with LangChain, LangGraph, ChromaDB and Qdrant; implemented tool calling and agent orchestration with external APIs.
- Developed recommendation and predictive analytics models using vector embeddings and similarity search.
- Deployed AI services with Docker and AWS (EC2, S3) and GitHub Actions CI/CD, following MLOps practices; reduced hallucinations through retrieval, prompt and embedding optimisation.

Junior Backend Engineer · Digital Horizons

Mar 2024 - Nov 2024 · Kathmandu

- Developed RESTful FastAPI services for microservices; designed optimised MongoDB databases on AWS (EC2, S3).
- Automated deployment with CI/CD pipelines, reducing manual deployment time by 80%.

SELECTED PROJECTS

Smart Traffic Signal Control, Kathmandu (Final Year Project)

7-junction SUMO corridor; TGNN (GRU → GAT) + MaskablePPO within a safety envelope. Adaptive control: -48.1% waiting time, +8.4% completed trips, -14.8% CO₂ vs fixed-time (9 seeds).

SUMO · PyTorch Geometric · Stable-Baselines3 · Gymnasium

Video KYC

Biometric verification with ArcFace + RetinaFace (90%+ accuracy); Silent-Face anti-spoofing with geometric liveness at 97.8% TPR; FastAPI on AWS EC2 with S3.

PyTorch · OpenCV · DeepFace · MediaPipe · Tesseract

PersonaPlex Full-Duplex Conversational AI

Deployed NVIDIA's 7B speech-to-speech model across Lightning AI, AWS, Kaggle and Colab; managed a 16.7 GB model load, CUDA limits, CPU offloading and streaming inference.

PyTorch · CUDA · HuggingFace Hub

Real-Time Voice AI Agent

Kyutai STT/TTS with an LLM dialogue engine; multi-turn context, interruption handling and silence detection.

Kyutai · FastAPI · LangGraph · OpenAI / Gemini

AI Code Assistant

GitHub agent with RAG over codebases; tool-calling workflows for repo search, summaries, docs and PR analysis; webhooks for real-time updates.

LangGraph · LangChain · GPT · Mistral · ChromaDB · Docker

AI-Powered Resume Parser

PDF, DOCX and image parsing at 95%+ accuracy with Gemini Vision OCR and LLM extraction to JSON.

Gemini · LangChain · FastAPI · MongoDB

SKILLS

Agents & LLMs: LangGraph, LangChain, tool calling, RAG, prompt engineering, fine-tuning, GPT-4, Gemini, Groq, Ollama

Voice AI: GPT Realtime API, VAD gating, multi-speaker awareness, ASR, faster-whisper, Twilio (telephony, PSTN, streaming)

ML & vision: PyTorch, TensorFlow, Scikit-learn, HuggingFace Transformers, CNN/RNN/LSTM/Transformer, OpenCV, OCR, NER, semantic search

Engineering: Python, FastAPI, REST APIs, Docker, Git, GitHub Actions, AWS (Cloud Practitioner), CUDA, Lightning AI, MongoDB, SQL, ChromaDB, Qdrant, Linux

EDUCATION

BSc IT (Hons) · TechSpire College

Sep 2023 – present · Kathmandu

Research Analyst in the college club: data-driven research, trend analysis and presenting insights.

Languages: Nepali, English, Hindi

References available on request