Nabin Rai

Lalitpur, Nepal +977 9848852690 kulyangrai10@gmail.com nabin-rai

Professional Summary

Senior AI Engineer with **5 years** of experience architecting enterprise AI systems, specializing in **Agentic Workflows** and bare-metal LLM orchestration. Proven expertise in building custom **ReAct control loops**, enforcing strict structured outputs (Hermes, Pydantic), and engineering automated self-correction mechanisms. Strong foundation in multi-modal computer vision, low-latency edge AI, and deploying production machine learning infrastructure (AWS, Docker).

Core Technical Skills

Agentic AI & LLMs: Custom ReAct Loops, Hermes 3, Pydantic, Tool Calling, OpenAI API, Task Decomposition, RAG.
Observability & Evals: LangFuse, RAGAS, Component-Level Evals, LLM-as-a-Judge, Semantic Similarity, TheFuzz.
LLM Infra & Optimization: Local Inference (Ollama, vLLM), LoRA, DAPT, FastAPI, AWS (SageMaker).
ML, Vision & Edge: PyTorch, YOLO, ResNet, EfficientNet, PaddleOCR, W&B, Edge AI (ONNX, TFLite, TF.js).

Selected Agentic AI Projects

Custom Autonomous Execution Engine & Self-Correction Loop

- Engineered a custom **ReAct control loop** using local inference (**Ollama/Hermes 3**) and **Pydantic**, bypassing framework overhead to strictly manage state arrays and enforce structured tool schemas.
- Implemented dynamic "Code as Action" execution wrapped in try/except blocks; successfully fed Python stack traces back into the LLM context window to trigger automated reflection and self-correction.

Agentic Workflow Observability (LangFuse POC)

- Integrated **LangFuse** tracing to monitor token usage, manage context window truncation, and conduct detailed error analysis on failing tools across multi-turn agent interactions.

Professional Experience

Senior Freelance AI & Computer Vision Engineer **Mar 2025 – Present**
Self-Employed *Remote*

- **LLM Evaluation Frameworks (RAGAS):** Engineered a FastAPI microservice using RAGAS/LangChain to benchmark RAG pipelines (accuracy, relevance, groundedness). Built a multi-metric CV extraction evaluation framework (semantic similarity, TheFuzz) achieving a **91% F1-score** across 16+ structured attributes.
- **Real-Time Edge Vision (ValiditAI):** Fine-tuned ResNet-34 (Liveness) and EfficientNet-B0 (Face Swap) via Weights & Biases sweeps. Optimized forehead detection models for edge environments (ONNX, TFLite, TF.js), achieving 0.62 mAP and **<20ms inference latency**.
- **Sports Analytics Workflows (KickAI):** Architected sequential computer vision pipelines for Top 5 European leagues, integrating YOLO ball detection, ResNet-50 quadrant classification (via marked frames), and ROI-based OCR for real-time score/time extraction.

Team Lead **Mar 2024 – Mar 2025**
TAI Inc. Pvt Ltd *On-Site*

- Led 5 AI engineers delivering enterprise automation solutions, scaling AWS/Docker deployment infrastructure.
- Fine-tuned 7B/16B parameter LLMs (StarCoder, StarChat) using DAPT/LoRA, **cutting legacy COBOL documentation time by 70%**.
- Architected real-time webhook receivers on AWS/Docker to handle conversational triggers with **<5s response times**.

Mid-Level ML Engineer **Aug 2022 – Mar 2024**
TAI Inc. Pvt Ltd *On-Site*

- Developed multi-task MobileViT models achieving industry-competitive ±5-year MAE (age) and 92% gender accuracy.
- Pioneered SMPL Gait implementation for person re-ID without 3D modeling, achieving **95% surveillance precision**.

Associate ML Engineer **Dec 2021 – Aug 2022**
TAI Inc. Pvt Ltd *On-Site*

- Built OCR pipelines using PaddleOCR and OpenVINO, boosting text extraction accuracy by **12% (CER/WER)**.
- Implemented an active learning pipeline, streamlining dataset curation and significantly reducing manual annotation.
- Enhanced preprocessing workflows (dewarping, deskewing, noise reduction) with OpenCV to improve OCR input quality.

Education

Tribhuvan University **2016 – 2021**
Bachelor of Electronics & Communication Engineering *Lalitpur, Nepal*

- **Relevant Coursework:** Engineering Mathematics, C++, Probability and Statistics.

Languages

English: Professional • **Nepali:** Native