Prashant Raj Bista

Biratnagar, Nepal | prashant.bista.18@gmail.com | +977 9842503707 | prashantrajbista
prashantrajbista | prashantrajbista | prashantrajbista

Summary

Work spanning research, client demos, and production — exploring the approach, showing early versions to clients, and helping ship and maintain what came out of it.

Professional Experience

Junior AI Engineer *Jun 2025 – Jun 2026*
IdeaBoxAI · Outsourced through Azminds Services · Remote

- Built LangGraph agent pipelines orchestrating multi-step LLM reasoning with tool calls, retrieval, business-rule guards, retries, and fallbacks, so failures degrade instead of propagating
- Designed structured-extraction pipelines over multilingual packaging-label PDFs (OCR + LLM) with automated QA/QC gates on model output, evaluated against sets built from reviewed production samples; drove false positives to zero on held-out data
- Instrumented every LLM call with Langfuse tracing (latency, token cost, output quality) and used production traces as the feedback loop driving prompt and pipeline iteration
- Shipped FastAPI REST microservices for a multi-tenant platform (usage metering, Redis caching, token streaming) and a ClickHouse time-series service for demand and shipment-delay forecasting

Associate Software Engineer *Dec 2024 – Jun 2026*
Azminds Services · Full-time · Kathmandu, Nepal

- Worked on *Agent Next Door*, a geospatial matching service for Singapore property: PostGIS nearest-neighbour distance queries over agents' rented/sold transaction records rank the best-matched agents for a given location
- Built a RAG assistant whose router dispatches each query to the right specialist backend (SQL databases, document stores, live web search), grounding every answer in retrieved evidence
- Engineered large-scale ETL/ELT and scraping pipelines on Databricks and Apache Spark, plus the backend services and third-party/partner API integrations around them

Personal Projects

LLM Misuse Monitor | *Python, FastAPI, Ollama, Llama Guard 3, Streamlit, Docker* *Code*

- Self-hostable monitoring for a served LLM: captures every prompt/completion off the request critical path and scans it for jailbreaks, prompt injection, PII leakage, and system-prompt leaks
- Pluggable detector ensemble plus a logprob-anomaly tier, feeding explainable risk scoring that auto-opens triage cases above threshold; 50 tests, Docker Compose deploy

DeepSC-S: Semantic Communication for Speech (reproduction) | *PyTorch, Weights & Biases* *Documentation* | *Weights*

- Reproduced DeepSC-S (arXiv:2012.05369) from an incomplete paper spec, resolving every underspecified choice by hypothesis testing against the reference implementation
- Experiment grid across AWGN, Rayleigh, and Rician channels at SNR 0–20 dB (tracked in W&B): 7.5–10.5 dB SDR on AWGN, under 1 dB matched-vs-mismatched gap; open-source checkpoints and staged write-up published

Publications

Mechanical Automation with Vision: A Design for Rubik's Cube Solver *May 2024*
Proceedings of the 15th IOE Graduate Conference, Vol. 15 (ISSN 2350-8914) Paper

Education & Research Training

Tribhuvan University, IOE, Thapathali Campus *Apr 2021 – Apr 2025*
B.E. – Electronics, Information and Communication Engineering
AI Fellowship 2024 | *Fusemachines Nepal* | Certificate *Apr 2024 – Oct 2024*
Research Trainee | *Nepal Research and Collaboration Center (NRCC)* *May 2025 – Aug 2025*
Authored a five-year review paper on SLAM and drone autonomy *Review Paper*

Technical Skills

LLM Systems: LangGraph, LangChain, RAG, prompt design, evals, Langfuse observability
Backend & Data: Python, FastAPI, Django, REST APIs, PostgreSQL, ClickHouse, Redis, SQL, ETL/ELT pipelines (Spark, Databricks), PyTorch
Geospatial & Frontend: PostGIS; TypeScript, JavaScript, Vue.js
Infrastructure: AWS (Lambda, S3, RDS, EC2, IAM, ELB, Route 53, Amplify), Docker, CI/CD, Git, Linux
Languages: Nepali (native), English, Hindi