Prashant Kafle

Kathmandu, Nepal (Remote) | +977-9741897738 | prashantkafle7738@gmail.com | linkedin.com/in/prashantkafle33 | github.com/prashant7738 | prashantkafle33.com.np

Summary

AI Engineer and Computer Engineering student who ships production RAG pipelines, multi-agent LLM systems, and full-stack applications for freelance clients and in faculty-supervised team projects. Validates AI-generated code and LLM outputs before they reach production. Wrote a research paper on evaluating AI agents with LLM-based judges. Focused on reliable, cost-efficient LLM systems. Available for remote full-time, contract, or internship work.

Education

Thapathali Engineering College, IOE – Tribhuvan University Kathmandu, Nepal
Bachelor of Engineering in Computer Engineering *Expected Jan. 2028*

Experience

Freelance Software Engineer (AI & Full-Stack, Contract) Mar. 2025 – Present
Self-Employed *Remote*

- Delivered production AI features for client applications – RAG document search and LLM chat pipelines on Groq/OpenAI with fallback strategies – to reduce single-provider dependency and keep client costs predictable
- Owned full-stack delivery end-to-end for remote clients: React front ends, Django REST Framework APIs, JWT/RBAC auth, and client-facing documentation for non-technical stakeholders
- Reviewed and validated AI-generated code and LLM outputs before client delivery, catching hallucinated logic and edge-case failures ahead of production

Backend & LLM Pipeline Owner – AnalyzeAI (Minor Project) Jan. 2026 – Apr. 2026
Thapathali Engineering College, IOE – Tribhuvan University (Team of 4, Faculty-Supervised) *Kathmandu, Nepal*

- Shipped a six-stage LangGraph data analyst with a 4-person team, splitting backend, agent-pipeline, and frontend ownership
- Owned the FastAPI backend and LLM validation gates; validated AI-generated code through git-based review before merging

Projects

Contextra AI | *FastAPI, RAG, pgvector, PostgreSQL, OpenAI Embeddings, Astro, React* | Live 2026

- Architected a production RAG system for PDF Q&A; hybrid OCR pipeline (PyMuPDF + EasyOCR) feeds OpenAI text-embedding-3-small embeddings into pgvector for grounded, citation-backed LLM answers
- Added Pareto 80/20 summary and smart flashcard generation features; JWT multi-workspace auth; FastAPI backend with SQLAlchemy + Alembic

AnalyzeAI – Multiagent Analyst *(Team of 4)* | *Python, FastAPI, React, LangGraph, Groq/Gemini, Pandas* | Live 2026

- Built and shipped a full-stack, six-stage LangGraph data analyst that turns messy CSV, Excel, JSON, JSONL, and Parquet files into cleaned data, quality scores, statistical findings, recommendations, and decision-ready HTML/PDF reports
- Combined deterministic Pandas/SciPy analysis with bounded Groq/Gemini assistance; added signal-driven chart planning, grounded narrative validation, ECharts/Matplotlib dual rendering, FastAPI jobs with SSE progress, dataset chat, and authenticated history
- Engineered failure-aware delivery with LLM fallbacks, validation gates, upload limits, cooperative cancellation, isolated run artifacts, and a focused 205-test backend regression suite

The Automaton Boardroom | *Python, LangGraph, LangSmith, E2B, Streamlit, Groq, Gemini* | GitHub 2026

- Built an autonomous multi-agent software micro-agency that turns app ideas into specs and code, tests it in isolated E2B sandboxes, feeds failures back for up to 5 fix iterations, and stops for human review
- Added prompt-injection guards, build-decision questions, LangSmith tracing, and Groq/Gemini fallback for resilience and cost

Live Bidding NP | *React, Django Channels, SQLAlchemy Core, PostgreSQL, WebSockets, Redis* 2025 – 2026

- Real-time auction platform with concurrent WebSocket bid handling via Django Channels + Redis; SQLAlchemy Core for raw-SQL performance; JWT in HttpOnly cookies against XSS

Publications

Evaluating AI Agents with LLM-Based Judges 2026
Prashant Kafle, Sunil Paudel *Undergraduate research paper*

- Compared ROUGE-L, Bradley-Terry, Elo-style ranking, and judge-human agreement for evaluating agent responses on GAIA-style tasks: paper, code

Technical Skills

Languages: Python, JavaScript (ES6+), TypeScript, SQL, HTML/CSS, C/C++
AI & LLM: RAG, LangGraph, LangChain, Multi-Agent Systems, LLM Evaluation, LangSmith, E2B, Prompt Engineering, HuggingFace, Groq API, OpenAI API, pgvector, ChromaDB
Backend & APIs: FastAPI, Django, Django REST Framework, Django Channels, WebSockets, JWT Auth
Data & Databases: Pandas, NumPy, Matplotlib, Scikit-learn, PostgreSQL, SQLAlchemy Core, Redis
Frontend & Tools: React, Astro, Git, Docker, Linux