

Pukar Kafle

📍 Kathmandu, Nepal ✉️ kaflepukar47@gmail.com 🌐 pukarkafle.com.np 🔗 [Pukar Kafle](#) 🖱️ [puka-jpg](#)

About me

I build AI applications with Python and LLMs. I've shipped production features for SaaS products, mostly RAG systems, document processing, and agentic workflows. I care about writing code that actually works and solving real problems, not chasing AI hype.

Currently building a resume generator because I'm tired of seeing AI-generated resumes that all sound the same.

Work experience

Smart Contents Nepal, AI Application Developer

Kathmandu, Nepal

July 2025 – present

- Built end-to-end business plan generation system using LangGraph with integrated web search for domain-specific market research with complete ownership of product.
- Implemented progress tracking for long-running document generation processes with real-time updates to improve user experience.
- Refactored API architecture from multiple endpoints to unified route structure for Laravel backend integration.
- Fixed RAG chatbot's document parser which couldn't handle complex Japanese PDFs with images and unselectable text. Switched from traditional document parsing methods to LlamaParser.
- Developed FastAPI middleware and core features for multi-tenant chatbot application serving enterprise clients.
- Currently redesigning document generator from linear workflow to human-in-the-loop agentic system with interactive, section-by-section generation.

Flowstation AI, AI Engineering Intern

Remote | Canada

Nov 2024 – June 2025

- Built foundational understanding of neural networks by implementing character-level RNN, autograd engine, and transformer architecture from scratch
- Trained language models on custom datasets and explored model optimization techniques
- Implemented YOLO-based object detection for real-time image analysis

Personal projects

Resume Builder SaaS

Oct 2025 – present

Building AI-powered resume generation platform that creates tailored, ATS-optimized resumes from user profiles and job descriptions

- Developed backend API with FastAPI for resume parsing, PostgreSQL for user data storage, and custom features for job description analysis
- Designing AI enhancement system that generates professional resume content without generic or obvious AI patterns
- Building agentic workflow to match candidate qualifications with job requirements and produce customized resumes

GPT-1 Transformer from Scratch

Jan 2025

Implemented GPT-1 architecture in PyTorch to understand transformer mechanics and attention mechanisms from first principles

- Built complete implementation including multi-head attention, positional encodings, and training pipeline
- Gained practical understanding of backpropagation and mathematical foundations underlying modern LLMs

Skills _____

Languages & Frameworks: Python, FastAPI, Pydantic, SQLAlchemy, PyTorch
AI/ML Technologies: LangGraph, OpenAI API, Claude API, RAG, Transformers, Hugging Face
Data & Storage: PostgreSQL, pgvector, ChromaDB, Alembic, LlamaParser
DevOps: Docker, Git, GitHub CI/CD, Gunicorn, Redis, UV

Education _____

BTech Kalinga Institute of Industrial Technology , Computer Science and Engineering	Bhubaneswar, India
<ul style="list-style-type: none">GPA: 8.6/10.00Fulbright Scholarship recipient for undergraduate studies	Oct 2020 – May 2024