Bibidh Subedi
bibidhdub@gmail.com | github.com/bibidhSubedi0

Skills

Languages: Python, SQL, C++
AI & LLM: OpenAI API, Hugging Face Transformers, LangChain, LangGraph, RAG, Fine-tuning, Prompt Engineering
ML & Data: PyTorch, scikit-learn, Pandas, NumPy
Deployment & Tools: FastAPI, Docker, Git, Linux

Projects

RedBoxDb – High-Performance Vector Database **pip install redboxdb**
bibidhSubedi0/RedBoxDb

- Developed a vector database engine in C++17 with a Python SDK, achieving an ingestion throughput of ~2M vectors/sec (128 dim floating point) and ~50,000 vectors/sec over a custom TCP server.
- Implemented a custom binary TCP protocol for client-server communication to eliminate HTTP/JSON overhead, resulting in 1ms p50 search latency.
- Engineered an append-only log persistence system and a Python wrapper that manages the C++ backend lifecycle via subprocesses and socket communication.

NepalLawFt – Domain-Specific Fine-tuned LLM **https://huggingface.co/spaces/Bibidh/NepalLawFT-Demo**
bibidhSubedi0/NepalLawFT

- Fine-tuned Llama-3.2-3B-Instruct on a Nepali legal Q&A dataset using QLoRA (4-bit NF4 + PEFT), reducing trainable parameters to ~24M (~1% of base) while achieving +51% LLM-as-Judge and +47.5% ROUGE-L improvement over the base model.
- Built an automated evaluation pipeline using ROUGE-L, character-level BLEU, multilingual semantic similarity, and LLM-as-Judge scoring via Groq API, with all results tracked and exported for analysis.
- Merged and deployed the fine-tuned model to HuggingFace Hub and served it via a Gradio chat interface on HF Spaces, handling the full inference pipeline from adapter merging to live demo.

CivicLens Nepal – RAG-Powered Legal & Governance Assistance **https://bibidh-civiclens-nepal.hf.space**
bibidhSubedi0/CivicLensNepal

- Built a retrieval-augmented assistant that indexes Nepal's constitution, laws, budgets, and governance documents, enabling factual Q&A in both Nepali and English with precise source citations.
- Engineered a multilingual pipeline to process legacy Preeti-encoded and scanned PDFs into 67,000+ searchable chunks using multilingual-e5 embeddings and ChromaDB.
- Integrated Llama 3.1 8B via Groq for low-latency reasoning, delivering cited answers with relevance scores through a FastAPI backend and lightweight web UI.

Autonomous PR Reviewer **https://github.com/apps/autonomous-pr-reviewer-bot**
bibidhSubedi0/Autonomous-PR-Reviewer

- Built a multi-language GitHub App that triggers on PR webhooks, shallow-clones the target commit, and dispatches flake8, cppcheck, and ESLint based on detected file extensions, orchestrated as a stateful LangGraph agent with conditional routing that skips AI review on clone failure and passes lint errors into Gemini's context.
- Preprocessed diffs by stripping context lines (60–70% size reduction) with partial-diff annotations, enforced structured JSON output from Gemini 1.5 Flash, and posted capped inline comments via the GitHub Reviews API using per-installation JWT tokens, supporting Python, C++, and JavaScript codebases out of the box.

Education

IOE Pulchowk Campus: BE in Computer Engineering Expected 2027
GPA: 3.68 / 4.0