

Vinu Kevin Diesel

+1-520-500-4578 | vinukevin13@gmail.com | linkedin.com/in/vinu-kevin-diesel | github.com/Vinu-Kevin-Diesel | Richardson, TX

Experience

AI Engineer

Jun 2025 – Present

NEDL Labs

USA

- Architected a 4-stage data pipeline turning healthcare policy PDFs into executable SQL rules: LLM-based text parsing, logic-tree flattening into all possible decision paths, rule assembly by category (billing, clinical, reimbursement), and database ingestion.
- Built a 5-layer validation framework cross-checking LLM-extracted rules against source PDF text via fuzzy string matching, forward/reverse coverage analysis, and code-manifest reconciliation to detect missed or hallucinated policy logic.
- Developed a policy domain classifier with multi-tier scoring, 3 circuit breakers, and 3 confidence gates, auto-categorizing insurance policies across pharmacy, orthopedics, oncology, and 10+ medical specialties.
- Built a hybrid rule-based and LLM-powered quality validator cross-referencing document-layout output, extracted billing codes, parsed rules, and code groupings to flag coverage gaps and inconsistencies via weighted penalty scoring.

AI Intern

Jan 2025 – May 2025

NEDL Labs

USA

- Built a PDF-to-structured-data pipeline for IGT casino gaming reports using Detectron2 object detection, EasyOCR, and table-recognition models within a 4-container Docker microservice architecture.
- Implemented automatic column matching using 384-dim sentence embeddings stored in PostgreSQL (pgvector) with semantic search, mapping raw input fields to a standardized data dictionary across 39+ casino properties.

Big Data & Deep Learning Intern

Dec 2022 – Jan 2023

Hewlett Packard Enterprise

Singapore

- Collected a custom dataset via web scraping and used the MoveNet framework to extract 17 skeletal keypoints (shoulders, elbows, wrists, hips, knees), converting them into normalized coordinate vectors for model input.
- Trained a multi-layer CNN with dropout regularization and hyperparameter tuning (learning rate, batch size, epochs), achieving 98% classification accuracy across slouched, bent, and squatting postures.
- Deployed the trained model on Azure Databricks using Spark-based batch-inference pipelines with interactive dashboards for real-time visualization of posture classification results; presented findings in a seminar to faculty and industry professionals, demonstrating applications in workplace ergonomics and safety monitoring.

Technical Skills

Languages: Python, C, C++, SQL, YAML, Bash **Data & Databases:** PostgreSQL, pgvector
ML/AI: PyTorch, TensorFlow, scikit-learn, HuggingFace Transformers, SentenceTransformers, OpenAI SDK
LLM & NLP: Anthropic Claude API, NVIDIA NIM, Groq, LangGraph, RAG, Prompt Engineering
AI Security & Evaluation: LLM Red-Teaming, Adversarial Robustness, Prompt Injection Defense
Infrastructure & Cloud: Docker, Kubernetes (GKE), Helm, FastAPI, GitHub Actions, CI/CD, MLOps, GCP (Cloud Logging/Monitoring)

Education

Masters in Data Science, University of Arizona, Tucson

GPA: 3.88 | Aug 2023 – May 2025

Bachelors in Computer Science and Engineering, Vellore Institute of Technology, India

GPA: 8.61 | Jun 2019 – Aug 2023

Projects

FinePrint: AI Contract & Policy Analyzer | Python, FastAPI, Z3 SMT Solver, NVIDIA NIM, Gemini, Docker

- Built a domain-agnostic service flagging risky clauses in leases, policies, and contracts via two engines: an LLM red-flag reviewer using a **map-reduce clause pipeline** (exhaustive extraction, then batched per-clause analysis) that tripled clause coverage over single-pass prompting with citation-backed findings, and a **Z3 SMT solver** formally proving logical contradictions and unreachable clauses from LLM-extracted typed rules.
- Built a provider-agnostic LLM layer (**NVIDIA NIM**, Groq, Gemini, self-hosted vLLM via one OpenAI-compatible interface) with schema-validated output, retry, and model failover; hybrid PDF ingestion adds OCR fallback (Gemini/Tesseract) for scanned and garbled-font documents.

LLM Red-Teaming & Evaluation Platform | Python, asyncio, Claude Sonnet 4.6 (Anthropic), Groq, Cerebras, NVIDIA NIM

- Built a modular AI-security framework stress-testing LLMs across hallucination, instruction-following, reasoning consistency, and adversarial robustness via a weighted-rubric LLM judge and 6 mutation-based attack strategies.
- Architected an async, multi-provider orchestrator with a **Claude Sonnet 4.6** judge (prompt caching), per-provider rate limiting, and YAML-configurable suites; benchmarked Qwen3-32B (Groq), Llama 3.1-8B (Cerebras), and Mistral Medium 3.5 (NVIDIA NIM) over 500+ adversarial cases, surfacing a 21% hallucination gap on multi-step reasoning.

Autonomous Document Intelligence Agent | LangGraph, FastAPI, pgvector, React, Docker

- Built a multi-agent RAG system with a React frontend ingesting PDFs, DOCX, and images via layout detection, Tesseract OCR, and semantic chunking, storing 384-dim HNSW-indexed sentence-transformer embeddings in pgvector.
- Designed a LangGraph state machine routing queries to factual, comparative, and multi-hop agents, with a critic agent scoring grounding via cosine similarity and triggering re-retrieval below 0.78 confidence; deployed as a 3-container Docker stack.