Karanbir Singh Pelia

Stony Brook, NY | +1 (631) 710-5508 | karanbirsinghpelia@gmail.com | karanbir-pelia | karanbir-pelia | karanbir-pelia

Education

Stony Brook University | MS, Computer Science (GPA: 3.62) — Stony Brook, NY | ***Aug 2024 – May 2026***
Coursework: Data Science, HCI, Computer Graphics, Advanced Algorithms, Distributed Systems, Data Management

Savitribai Phule Pune University | BE, Computer Engineering (GPA: 3.64) — Pune, MH | ***Sep 2020 – Jun 2024***
Coursework: Data Structures and Algorithms, Operating Systems, Computer Networks, Database Systems, Machine Learning

Technical Skills

Languages: Python, Java, JavaScript, C++, SQL
Frameworks & Libraries: Spring Boot, FastAPI, Flask, Kafka, gRPC, REST APIs, asyncio, React, Tailwind, NumPy, Pydantic, Micrometer
Databases & Data: PostgreSQL, MongoDB, MySQL, Redis, Parquet, Avro, Confluent Schema Registry, PyArrow
DevOps & Tooling: Docker, Kubernetes, Git, Jenkins, CI/CD, Prometheus, Grafana, Postman, Jira, Confluence, Agile, OOP Design Patterns
AI & Developer Tools: RAG, ChromaDB, sentence-transformers, Ollama, LLM Evaluation, Claude Code, GitHub Copilot, Codex, Google Antigravity

Experience

Infrrd Inc. | Software Engineer Intern – Product Engineering — San Jose, CA | ***May 2025 – Jan 2026***
Technologies: *Python, FastAPI, Flask, MongoDB, REST APIs, Grafana, Jenkins, Git*

- Built an internal FastAPI evaluation platform automating end-to-end accuracy benchmarking across 6+ document-processing pipelines and 100+ document types, standardizing QA workflows and reducing manual effort by ∼50%.
- Strengthened the No Touch Processing (NTP) LLM verification framework with directional page scanning and negative keyword filters, resolving near-duplicate misclassifications and expanding output qualifying for 2–3× premium billing.
- Engineered post-processing correction logic resolving page-interval split errors and blank-page voids present in 4–6% of LLM classification output, and improving Closing Disclosure tabular field extraction accuracy by 20+ percentage points.
- Implemented feature-flagged pipeline routing shipping 3+ production features — PDF processing, prompt-document list decoupling, and NTP toggling — with LLM output format validation and fallback handling for zero-downtime backward-compatible releases.
- Built a stateless async Flask replay endpoint eliminating full-pipeline re-execution for classification testing, reconstructing payloads from request IDs and tracking job status in MongoDB for rapid isolated iteration.

Projects

Entitle | Privacy-First Safety Net Benefits Navigator — ***Apr 2026 – Jun 2026***
Technologies: *Python, FastAPI, React, Tailwind CSS, Gemma 4, ChromaDB, sentence-transformers, SQLite, PyArrow, PyMuPDF*

- Built a full-stack RAG AI agent that transforms free-form text into structured profiles, retrieves top-8 policy chunks from ChromaDB (sentence-transformers), and matches users to 15 federal and 14 state programs, surfacing $2,500+/month in benefits.
- Engineered a batched eligibility inference pipeline with sequential Ollama execution and parallel async Gemini calls, combining a categorical rule-based pre-filter and structured JSON fallback per batch, reducing runtime from 13+ minutes to under 5 minutes.
- Integrated a multimodal document-reading pipeline compositing PDF pages via PyMuPDF before Gemma 4 Vision inference, enabling users to photograph denial letters and renewal notices for plain-language explanation across 9 languages.

MediGrid | Fault-Tolerant Medical Resource Availability Platform — ***Feb 2026 – Apr 2026***
Technologies: *Java, Spring Boot, Kafka, Avro, Schema Registry, gRPC, RAFT, PostgreSQL, Redis, Docker, Kubernetes, Prometheus*

- Implemented RAFT consensus from scratch — leader election and log replication over gRPC — benchmarked at ∼20K writes/sec under full replication, recovering to ∼31K writes/sec after a 2-node failure with quorum intact.
- Built an event-driven REST → Kafka → RAFT → PostgreSQL ingestion pipeline with Avro and Confluent Schema Registry for strict schema validation, routing malformed events to a DLQ to isolate bad data without stalling ingestion.
- Implemented Redis cache-aside (30s TTL) with automatic PostgreSQL fallback (median 5 ms, p99 14 ms on cache miss) and maintained an append-only time-series store powering OLAP analytics, shortage alerts, and Parquet exports.

Site to Slides | Webpage-to-Presentation Automation Service — ***Mar 2025 – Apr 2025***
Technologies: *Python, Firecrawl API, REST APIs, WebSockets, JWT, Postman*

- Reverse-engineered 8 undocumented endpoints via browser network inspection, implementing a 2-layer JWT auth flow with a 5-minute expiry buffer to convert any public URL into a shareable AI presentation via an undocumented third-party API.
- Built an LLM-powered extraction layer using Firecrawl to parse webpage content into 4–6 structured sections with per-section image mapping, fed into a retry-backed slide generator with up to 4 attempts per slide for reliable output.

Fatigue Sense | Real-Time Driver Fatigue Monitor — ***Oct 2024 – Nov 2024***
Technologies: *Python, TensorFlow/Keras, OpenCV, MediaPipe Pose/Hands, NumPy, Pygame, scikit-learn, Matplotlib*

- Built a real-time driver safety system unifying a Keras CNN with MediaPipe Pose and Hand tracking across 3 concurrent detection pipelines, flagging fatigue, posture, and hand-position risks with 95% classification accuracy.
- Engineered a multi-modal inference pipeline tracking 21 hand landmarks and 4 posture conditions per frame, with time-gated alert logic (2–3s sustained thresholds) to suppress false positives and deliver sub-second visual and auditory warnings.

Publications

Advances in Plant Disease Detection and Classification Systems — Scrivener Publishing LLC | ***Nov 2025***
Co-authored a chapter in **Optimizing AI Applications for Sustainable Agriculture**, based on my BE project **AgroDoc**, a CNN-powered plant disease detection web app achieving 97.38% accuracy across 38 plant conditions spanning 14 crops. The chapter surveys CNNs, transfer learning, and data augmentation for plant disease detection, and explores their role in IoT-enabled precision agriculture.