Pankaj Yadav

(+91) 9140809846 pankajydv3451@gmail.com LinkedIn GitHub

EXPERIENCE

Machine Learning Engineer — KlearNow.AI **June 2025 – Present**

- Designed and deployed an **AGNO**-based multi-agent **Document AI system** using **Claude 4.5 Sonnet**, coordinating **6 microservices** for document classification, OCR, and structured data extraction, processing **3,000+ documents/day** with **99% reliability**.
- Modernized document extraction workflows by migrating from **LayoutLM** and **BERT** models to an **LLM-powered architecture**, enabling seamless adaptation to **1000+ document templates** and removing the need for template-specific retraining.
- Built a **Validation Agent** enforcing **14+ schema rules** on **LLM outputs**, mitigating hallucinations, reducing downstream error rates by **30%**, and cutting manual correction effort by **85%**.
- Built event-driven pipelines using **Apache Kafka** and **gRPC** across **6 microservices** for end-to-end processing; implemented **AWS SQS**-based architecture reducing queue backlog by **60%** at peak load.
- Designed prompt-based extraction workflows generating structured **JSON outputs** from OCR text, achieving **92%+** field-level accuracy across invoice and customs formats.

Machine Learning Engineer Intern — KlearNow.AI **March 2025 – May 2025**

- Fine-tuned **Qwen** and **LLaMA** for structured document extraction using **LoRA/QLoRA** on **4-bit quantized** models, reducing GPU memory by **60%** and improving field-level accuracy by **18%** over zero-shot baselines across **5+ document categories**.
- Designed one-shot and few-shot prompting strategies for document extraction, reducing prompt token usage by **30%** while maintaining extraction accuracy across **5+ document categories**.
- Developed image preprocessing pipelines including orientation correction, denoising, and OCR optimization, reducing document processing latency by **35%**.

Software Engineering Intern — Beans.ai **June 2023 – August 2023**

- Created indoor maps for **10+ facilities** using ArcGIS and geospatial datasets, improving delivery routing accuracy.
- Developed a CNN-based malaria detection app with **Grad-CAM** visualization, achieving **94%** accuracy on **27K+** blood smear images.

TECHNICAL SKILLS

- **Programming:** Python, Java, C
- **Machine Learning & NLP:** PyTorch, Scikit-learn, XGBoost, SVM, Transformer Models (BERT, LayoutLM, GeoLayoutLM), Hugging Face Transformers, LLMs (LLaMA, Qwen), NER, OCR, Document AI
- **LLM & Generative AI:** RAG, Prompt Engineering, LLM API Integration (Claude Sonnet), Agentic AI Systems (AGNO), Structured Output Extraction, Prompt Optimization, LangChain, Ragas
- **Backend & Distributed Systems:** FastAPI, Apache Kafka, gRPC, AWS SQS, REST APIs, Microservices
- **Cloud & DevOps:** AWS, Docker, Kubernetes, CI/CD, SageMaker, Git, Streamlit

PROJECTS

RAG-based Document QA System

- Designed and developed a production-ready **RAG** system leveraging **FAISS** vector search and **Claude/LLaMA** LLMs for document question answering.
- Optimized retrieval performance through advanced chunking and embedding strategies, achieving a **30% improvement** in retrieval relevance; integrated **Ragas** for automated evaluation and benchmarking.
- Containerized and deployed services with **FastAPI** and **Docker**, enabling low-latency, real-time document Q&A through REST APIs.

Invoice & Non-Invoice Information Extraction

- Fine-tuned **GeoLayoutLM** for key-value extraction from invoice and customs documents; trained on **3,100 samples** with **135 classes**, achieving **93% labeling accuracy**.

Document Classification System

- Built a multi-model pipeline classifying **10 document classes** via textual and visual features, ensembling **BERT**, **LayoutLM**, and **Detectron2** on **150K+ images** with distributed multi-GPU training, achieving **98% accuracy**.

NER-Based Key-Value Extraction

- Built a **BERT-based NER pipeline** for OCR documents using **BIO tagging**; extended positional embeddings from **512 to 1024 tokens** on **30K+ samples**, achieving **91% NER accuracy**.

Malaria Disease Detection System GitHub

- CNN-based system identifying infected blood smear cells from **27K+ images** at **94% accuracy**; implemented **Grad-CAM** visualization and adversarial robustness testing (**FGSM**, one-pixel).

ACHIEVEMENTS

- Solved **500+ DSA problems** across **LeetCode** and **GeeksforGeeks**.
- Ranked in the **Top 20%** among **50,000+ participants** in **LeetCode Weekly Contests**.
- Authored technical articles on **BERT** and **Kimi-VL**, covering architecture, fine-tuning strategies, quantization, and cost-efficient inference on **NVIDIA A10/L4 GPUs**.

EDUCATION

B.Tech, Computer Science Engineering, SRM University, Chennai **2022 – 2026**

POSITIONS OF RESPONSIBILITY

President, CS Club — led **3+ technical events**, managed **50+ member** team **May 2022 – May 2025**

Event Coordinator, Annual Tech Fest — coordinated **8+ sessions**, **200+ student** participation **Feb 2023 – May 2023**