Deepa Khanal

AI Engineer | Data Engineering for AI | ML Engineer

Nepal , Kathmandu • +977-9867244436 • deepakhanal25@gmail.com

@ linkedin.com/in/deepa-khanal-b660471b5 @ https://github.com/khanal-deepa

AI Engineer with hands-on experience in large-scale machine learning, deep learning, and embedding-based retrieval systems across NLP, document intelligence, and multimodal applications. Skilled in designing scalable data pipelines, vector search, AWS-based AI infrastructure, and evaluation-driven ML systems with strong focus on performance, interpretability, and production deployment.

EXPERIENCE

AI Research Engineer (Contract) — Document Intelligence / Retrieval Australia

Dec, 2024 - Present

- Designed and deployed an embedding-based document change analysis system for insurance policy revisions, modeling structural, semantic, and lexical changes for compliance review.
- Designed and implemented a large-scale data ingestion and indexing pipeline on AWS OpenSearch, processing 1.4M document chunks with efficient metadata schema and embedding generation; achieved p95 retrieval latency of 180 ms, enabling scalable evaluation and exploration workflows.
- Defined evaluation protocol on 1,200 expert-labeled pairs; achieved 94% accuracy (Structural 96%, Semantic 92%, Lexical 97%); produced SQL-based evaluation breakdowns and auditable evidence traces, reducing manual review time by 78%.

AI Developer — Applied NLP & Multimodal Systems Singapore

Sharelook Jun, 2024 - Nov, 2024

- Contributed to an LLM-driven educational assistant; reached 88% response relevance via human evaluation on 6000 test queries.
- Implemented semantic indexing and retrieval (embeddings + ranking); achieved Recall@10 = 0.81 and MRR = 0.67; reduced p95 latency by 47% (220 ms) through pipeline optimization and batching.
- Integrated text-to-image generation for educational content; achieved 85% alignment approval across 1200 evaluated samples.

AI/ML Intern — Data & Generative AI Lalitpur, Nepal

Prixa Technology Sep, 2023 - Dec, 2023

- Built preprocessing and analysis workflows in NumPy/Pandas, reducing data cleaning time by 40%.
- Supported a generative AI-based extraction system for Nepali-language documents; achieved 91% extraction accuracy and macro F1 of 0.89 on 500 annotated pages.
- Authored internal technical documentation (methodology, error cases, improvement roadmap) to enable reproducibility and handover.

TECHNICAL SKILLS

- **Programming**: Python
- **ML/AI:** Deep Learning, CNNs, Transfer Learning, Embeddings, Semantic Similarity, Information Retrieval, Generative AI
- **Frameworks**: PyTorch, TensorFlow
- **Data**: NumPy, Pandas, Matplotlib, SQL
- **IR/Systems:** Vector Databases, AWS, OpenSearch
- **Apps/Deployment**: Streamlit
- **Evaluation:** Accuracy, Macro F1, Precision@K, Recall@K, MRR, p95 Latency, Throughput

EDUCATION

Bsc.CSIT Kathmandu

Tribhuvan University Nov, 2019 - Nov, 2024

- Relevant coursework: Neural Networks, Data Mining, Artificial Intelligence, Data Structures & Algorithms, Database Management Systems, Discrete Math, Linear Algebra, Statistics.
- Undergraduate project: American Sign Language Recognition (CNN) — trained on 7,800 images across 24 gesture classes; achieved 92.6% test accuracy and macro F1 of 0.90; documented failure modes and mitigations.

PROJECTS

Image Classification using Transfer Learning (ResNet50)

- Built a ResNet50-based classifier on 5,200 images across 10 categories; achieved 93.4% validation accuracy and macro F1 of 0.91.
- Compared frozen vs fine-tuned strategies; selected configuration improving accuracy by 2.1% with +18% training time trade-off; performed error analysis by class to guide augmentation.

REFERENCES

Available On Request