Amit Timalsina

linkedin.com/in/amittimalsina | selftaughtamit.medium.com | amittimalsina21@gmail.com

Experience

Consultant Remote
Senior Software Consultant (GenAI, ML, Backend) September 2024 - Ongoing
Consulting on LLM integration for extraction, RAG, Agentic use-cases.

- Built custom PDF parsing solution for clinical trial documents that improved entity extraction accuracy by 50%+ by solving cross-page entity detection and reducing LLM tokens by 70%. Enabled enterprise client to build analytics product for stakeholders.
- Led end-to-end data extraction systems for clinical trials and document processing startups, improving accuracy vs. previous LayoutLM models. Built OMR extraction solution in 1 week for startup client, deployed to production in 2 weeks.
- Processed 200k clinical trial documents enabling new analytics services (e.g., "How often is age inclusion criteria used in trials?"). Handled rate limits, edge cases, and tight deadlines across multiple LLM providers (Anthropic, Llama, Mistral, Qwen).
- Built RAG chat application over clinical trial data in Snowflake achieving 95% recall@10. Internal stakeholders reported high satisfaction, product ready for customer deployment.
- Implemented Langfuse for LLM observability, reducing debugging time by 80% and enabling customer-level spend tracking. Led cross-functional collaboration with SRE team for in-house deployment.

Kniru Bangalore, India
Founding Engineer, SDE-2 January 2024 - September 2024
First full-time hire at pre-seed fintech startup. End-to-End (AI, Backend, Frontend) productionalization of core features.

- Built multi-agentic RAG chat system serving 80 MAU with 10 daily queries, achieving 70%+ user satisfaction and 50% retention. System handled spending analysis, tax advising, investment, and retirement planning through specialized agents accessing PostgreSQL, vector database, and Web Search APIs.
- Architected notification engine serving 300 users with 95% delivery success rate and 20% engagement. Built event-based architecture with 100+ notification templates, delivering 3 notifications per user daily.
- Designed chat-based onboarding experience improving completion rate from 70% to 90%. Created "WOW moment" by providing personalized financial insights during onboarding.
- Implemented OWASP guardrails against LLM vulnerabilities, Langsmith tracing, and cost-efficient evaluation suite. Mentored junior engineers in AI and backend development.

Docsumo Bangalore, India
Team Lead April 2023 - January 2024
Lead team of two engineers to build and maintain internal engineering tools for data assets, and model evaluation.

- Built end-to-end ML pipeline with versioned data store (DVC, GCS, Gitlab) and MLFlow logging, reducing benchmarking experiment time by 50% and preventing costly GPU experiments through centralized configuration management.
- Developed centralized evaluation suite ensuring 100% guaranteed metric comparison across teams. Single-point bug fixes improved system reliability and reduced maintenance overhead.
- Created data assets dashboard with metadata across prod, staging, and testing environments. It also had weekly model performance tracking of all customer accounts, used by teams to evaluate improvements and degradation.
- Led architectural decisions for tracing, evaluation frameworks, and uniform benchmarking datasets across multiple document types.

Machine Learning Engineer May 2022 - January 2024

- Built LLM-powered key-value extractor reducing annotation time by 70% and enabling faster support for new document types. Implemented layout parsing, zero-shot prompting, and confidence scoring for USBS, 1040 forms, Acords, Invoices, and Utility bills.
- Fine-tuned Visual Document Understanding models for NER tasks achieving 91%+ accuracy across document types. Deployed 10+ models to production serving enterprise customers with 500k+ ARR.
- Led multilingual KV extraction research for Spanish bank statements and Chinese invoices, achieving 7% accuracy improvement from baseline. Improved customer retention through enhanced international document support.
- Built rule-based table header classifier with FastText ensemble achieving 96% accuracy. Created inference API for production deployment.

Nepal Can Move Remote
Junior Machine Learning Engineer July 2021 - March 2022

- Built and deployed delivery failure prediction system processing 100k-150k monthly deliveries, reducing failure rate from 25% to 15% and saving Re 6 lakhs (5k USD) in first month. Optimized for 87% recall with limited features, implemented automated CI/CD pipeline with weekly retraining and batch feature generation for production stability.
- Developed invoice information extraction system reducing processing time from 2 minutes to 4 seconds per document with 87% token-level accuracy. Built end-to-end pipeline using OpenCV, pyTesseract, spaCy, and Transformers for critical logistics document processing.
- Researched recommendation engine prototype achieving 75% accuracy on top@10 retrieval using TensorFlow Recommenders, benchmarked on 100k MovieLens ratings for potential e-commerce integration.
- Built medical diagnosis sequence generation model using BERT architecture for 25 diseases, achieving 0.83 F1 score as POC for healthcare business venture.

Products

Blintic AI Remote
AI-Powered Document Processing Platform 2023 - Ongoing
To be written

- To be written

ORGO Earth Remote
Sustainable Agriculture AI Platform 2022 - Ongoing
To be written

- To be written

Projects

Document Classification Github

- Common OCR interface for closed-source providers like Google Vision and open source providers like Pytesseract.
- Training, Evaluation, and Inference support for Fasttext, Small Language Models (BERT, etc), Vision Language Models (like LayoutLM), and LLMs.

Predicting and Forecasting Happiness Github

- Predictor model that detects mood of a person based on questionnaire and data tracked by fitness tracker.
- Forecasting model to forecast individual actions several timestamps ahead with minimum past data (Time Series Forecasting).
- This is a research project which resulted in a research paper submitted to Young Scientists Submit Nepal 2020.

Skills

Language and Technology: Python, Bash, Docker, AWS, GCP, PostgreSQL, Snowflake
Framework & Library: FastAPI, Flask, PyTorch, TensorFlow (Keras), Transformers, HuggingFace Hub, FastText, XGBoost, OpenCV, Scikit-learn, Instructor, OpenAI, Langchain, Llama-index, CrewAI, Autogen, Qdrant VDB, Pinecone VDB, Atlas Vector Search, Amazon SageMaker, MLFlow, DVC, NumPy, Pandas, NLTK, spaCy
Machine Learning: Prompting, Agentic Systems, RAG, Information Extraction, Document AI, Transfer Learning, Fine-tuning, MLOps, Applied Statistics and Probability, Machine Learning and Deep Learning algorithms, NLP and NLU, Recommendation Systems, Generative AI, PEFT
Project Management: Agile methodologies like Scrum and Lean, End-to-End project management, Communication

Awards and Achievements

Young Scientist of Nepal Young Scientists Summit (RECAST)
Won the first position in Computer Science department of 5 th Young Scientist Summit (YSS). 2020

Glocal 20under20 of 2020 Glocal Teen Hero
Selected as Glocal 20Under20 for 2020 Batch as a "The Early Change Makers". 2020

Education

Sainik Awasiya Mahavidyalaya Bhaktapur, Nepal
Diploma Computer Engineering July 2019 - Oct 2021

Coursera Remote

Deep Learning Specialization Aug 2020 - Nov 2020

Coursera Remote
DeepLearning.ai TensorFlow Developer Specialization Aug 2020 - Nov 2020

Coursera Remote
Google Project Management: Professional Certificates Oct 2022 - Nov 2023