Amit Timalsina

github.com/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 PDF parser for clinical protocols that returned layout conserved structured output. This improved F1 of chunk retrieval by 60% absolute and accuracy of entity extraction by 50% absolute.
- Processed 200k clinical protocols 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).
- Built RAG chat application over clinical protocol extracted data in Snowflake, achieving 95% recall@10. Internal stakeholders reported high satisfaction, product ready for customer deployment.
- Built OMR extraction using VLM, YOLO, and LLM in 1 week for a startup, deployed to production in 2 weeks.
- 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 dev tools for developer efficiency.

- Developed end-to-end pipleine for reproducibility of models and evaluation metrics by utilizing versioned data store (DVC, GCS, Gitlab). Reduced benchmarking experiment time by 50% and GPU resources waste which was caused by incorrect experiments and no bechmark datasets.
- Developed centralized evaluation suite ensuring 100% accurate metric comparison across experiements. Single-point bug fixes improved evaluation 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/few-shot prompting, and confidence scoring.
- 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 table header classifier with FastText and fuzzy 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 for weekly retraining and batch feature generation.
- Developed invoice information extraction system reducing processing time from 2 mins to 4 secs per document with 87% token-level accuracy. Built end-to-end pipeline using OpenCV, pyTesseract, spaCy, and Transformers.
- 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 blinticai.com | app.blinticai.com
No-Code AI Agent Platform for Customer Support January 2025 - Ongoing
Built end-to-end no-code platform enabling businesses to create AI agents for customer support without technical expertise.

- Developed full-stack platform with Next.js frontend and FastAPI backend, serving 10+ demo customers with custom agent orchestration and hybrid retrieval (dense embeddings + BM25 + Jina reranker) for comprehensive knowledge base search.
- Built custom agent orchestrator supporting multiple data sources (text, files, URLs, website scraping) and custom actions (API calls) for external integrations, enabling seamless workflow automation and third-party system connectivity.
- Implemented intelligent support escalation system with team tagging and custom chat embed UI customization, allowing businesses to maintain control over complex queries while providing personalized customer experiences.
- Architected scalable infrastructure on Vercel and Google Cloud Run with PostgreSQL database, Supabase auth, and comprehensive observability using Langfuse for production monitoring and performance tracking.
- Planned for open-source release to democratize AI agent development for customer support and enable community contributions to the platform.

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.

Cricket Fitness Analysis Github
FastAPI, Postgres, Supabase, Docker, WebSocket, Langfuse

- Log fitnes, coaching, match, and rest day using voice input with AI follow up questions. Input is transcribed using OpenAI whisper and then converted into structured ouptut using OpenAI structured output.
- Developed progress tracking dashboard with charts for weekly performance visualization, enabling players to monitor training, coaching, matches, and rest days systematically.

Skills

- **Language and Technology:** Python, Bash, Sqlite, Postgres, Docker, AWS, GCP
- **Framework & Library:** FastAPI, Flask, Pydantic, PyTorch, TensorFlow (Keras), Transformers, HuggingFace Hub, FastText, XGBoost, OpenCV, Scikit-learn, Instructor, OpenAI, Langchain, Llama-index, CrewAI, Autogen, Qdrant VDB, Pinecone VDB, Amazon SageMaker, MLFlow, DVC, NumPy, Pandas, NLTK, spaCy
- **Machine Learning:** Prompting, Agentic Systems, RAG, Information Extraction, Document AI, Transfer Learning, PEFT, Fine-tuning, MLOps, Applied Statistics and Probability, Machine Learning and Deep Learning algorithms, NLP and NLU, Recommendation Systems

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