MANINDER SINGH

Data Scientist | Production ML & LLM Systems

Bengaluru, India | manindersingh120996@gmail.com | +91-9041416238
GitHub | Medium | Portfolio

PROFESSIONAL SUMMARY

Data Scientist specializing in **production-grade ML systems** with deep expertise in **Large Language Models, Transformers, and inference optimization**. Currently consulting at Uber via Indium Software, architecting scalable ML pipelines that process millions of documents globally. Proven track record of translating research into production systems with measurable business impact, including reducing operational costs through automation and achieving human-level accuracy on complex tasks. Strong foundation in **first-principles thinking**, system design, and open-source contributions with 350k+ downloads.

PROFESSIONAL EXPERIENCE

Indium Software (Client: Uber) *Feb 2025 – Present*
Data Scientist (External Consultant) Bengaluru, India

- Architected and deployed **Transformer-based auto-transcription pipeline** processing driver documents (licenses, registrations, insurance) across global markets, handling millions of documents with sub-second latency requirements
- Built end-to-end ML workflows including **model training, ONNX optimization, and production deployment** for real-time inference systems serving Uber's global earner network
- Fine-tuned vision and language models achieving **human-level+ accuracy** on structured field extraction, reducing annotation errors by 35% and enabling automated quality validation
- Reduced manual review effort by **40%** and operational costs through intelligent automation of document understanding workflows, directly impacting platform onboarding velocity

LTIMindtree *Jul 2022 – Jan 2025*
Senior Software Engineer Bengaluru, India

- Developed **multimodal LLM-based video analysis system** for detecting development tools and summarizing coding workflows, achieving 83% accuracy and reducing manual video review time by 60%
- Built production **speech-to-SQL system** enabling natural language database queries with 85% precision, improving non-technical user productivity by 50%
- Led development of internal **RAG-based GenAI chatbot** integrating vector search and retrieval pipelines with LLMs, serving 500+ internal users with 90% query satisfaction rate
- Improved Intelligent Ticket Dispatcher accuracy from **20% → 78%** through architecture redesign, feature engineering, and training pipeline optimization, reducing misdirected tickets by 70%
- Delivered transfer-learning classification models with **up to 95% accuracy** using limited labeled data through strategic fine-tuning and data augmentation techniques

EDUCATION

Indian Institute of Technology (IIT) Palakkad *2020 – 2022*
M.Tech in Data Science

- Post-Graduate Affairs Secretary - Led PG policy design and academic initiatives

TECHNICAL SKILLS

Machine Learning & AI

Transformers, Large Language Models (LLMs), Natural Language Processing, Vision Transformers, Multimodal Models, Retrieval-Augmented Generation (RAG), Parameter-Efficient Fine-Tuning (LoRA, QLoRA)

Programming & Tools

Python (Advanced), SQL, Git, Linux, Model Deployment Pipelines

ML Systems & MLOps

PyTorch, Hugging Face Transformers, ONNX Runtime, Model Serving & Inference Optimization, MLFlow, Distributed Training (DDP), Docker, FastAPI, Cloud Platforms (GCP, Azure ML Studio)

Core Expertise

Production ML Systems, Model Optimization, System Design, Research-to-Production Translation, First-Principles Problem Solving

OPEN SOURCE CONTRIBUTIONS & IMPACT

Programming Language Identification Model – Open-sourced on Hugging Face with **3,50,000+ downloads**, adopted by Protect AI for LLM Guard production system. Fine-tuned CodeBERTa on Rosetta Code dataset achieving state-of-the-art accuracy for code language detection.

SELECTED PROJECTS

- **Vision Transformer (ViT) from Scratch** – Complete PyTorch implementation from first principles with detailed mathematical foundations
- **Transformer (Encoder-Decoder) from Scratch** – Full implementation for machine translation with attention visualization
- **Distributed Training with PyTorch DDP** – Multi-GPU training systems with gradient synchronization and optimization
- **LLM Fine-Tuning with LoRA** – Practical implementation notebooks with theory breakdowns and best practices
- **Programming Language Detection Model** – Fine-tuned CodeBERTa achieving 95%+ accuracy across 25+ languages

TECHNICAL PUBLICATIONS & WRITING

- Detailed Explanation of Self-Attention Mechanisms
- Fine-Tuning LLMs with LoRA: Theory & Practice
- Optimizing LLM Inference for Production
- Mathematics of Convolution & Deconvolution
- PCA: Intuition, Mathematics, and Practice
- Categorical Correlations (Chi-Square & Cramer's V)
- Video Frame Deduplication Techniques

CERTIFICATIONS

- Generative AI with Large Language Models – DeepLearning.AI & AWS
- Natural Language Processing Specialization – DeepLearning.AI
- AI for Medicine Specialization – DeepLearning.AI
- Prompt Engineering for ChatGPT – Vanderbilt University