Maniraj Sai

+31684011721 manirajadapa@gmail.com https://www.linkedin.com/in/manirajsai/ github.com/thecr7guy2

Professional Summary

AI Engineer with 3+ years of experience developing and deploying **production-ready AI applications**. Built end-to-end solutions spanning **fine-tuning, evaluation, and agentic workflows**, with a strong focus on **scalability, reliability, and real-world impact**.

Experience

aXite Security Tools **July 2024 - Present**

AI Engineer *Schiphol, The Netherlands*

- Led end-to-end development of **AX-Office.ai**, a **fully on-premise AI platform** delivering **5 production LLM applications** (private assistant, Agentic RAG system, ASR transcription, document OCR, and AI coding agent) across the organization, eliminating all external LLM API dependencies on sensitive security data.
- Engineered **LLM inference infrastructure** to support **12+ concurrent users** with long-context workloads, optimizing KV cache utilization, evaluating quantization strategies (FP8, AWQ, FP4), and building a custom retrieval pipeline with embedding search and reranking.
- Established an internal **LLM evaluation framework** benchmarking **10+ open-weight models** across reasoning, instruction following, code quality, and summarization, enabling data-driven model selection and production stack improvements.

Labelfuse **April 2024 – July 2024**

LLM Engineer *Eindhoven, The Netherlands*

- Architected a Text-to-SQL system using locally hosted LLMs to enable non-technical users to query private databases via natural language, with schema-aware prompting, query validation, and retrieval over table metadata.

Projects

Autonomous AI Trading Bot | *Python, Anthropic API, GitHub Actions, GitHub Pages* **2026**

- Built an autonomous trading system that scrapes **SEC filings and Congressional disclosures**, scores insider conviction using a custom weighted model (stake increase, role seniority, recency decay), and places real trades on a demo account **twice weekly with zero human intervention.**
- Integrated **Claude Opus as the portfolio manager**, feeding enriched candidate data (fundamentals, RSI, price history, news) to generate **up to 5 picks per run** with allocation percentages, confidence scores, and written rationale referencing specific insider signals.

GPT-2 Small (124M): Pretraining & Instruction Tuning | *PyTorch, RunPod, HuggingFace* **2025**

- Pretrained GPT-2 Small from scratch on a **curated 18B token dataset** assembled from 7 sources (FineWeb-EDU, arXiv, RefinedWeb, Gutenberg) across **8 x A100 GPUs**, building a custom **memory-mapped dataloader** to handle large-scale data streaming beyond RAM capacity.
- Instruction-finetuned the pretrained model on **17M tokens across multiple SFT datasets**, outperforming the original GPT-2 Small on **TruthfulQA and MMLU** benchmarks and deployed an interactive demo on HuggingFace Spaces for public evaluation.

Education

University of Groningen **September 2021 – September 2023**

Masters in Artificial Intelligence, GPA: 8.0/10 *Groningen, The Netherlands*

- Thesis: **Estimating Uncertainty in GANs for Super-resolution** - incorporated uncertainty estimation techniques into GAN-based super resolution, with applications in medical imaging and computer vision.

Technical Skills

Languages: Python, SQL, Bash
LLM Techniques: Pretraining from scratch, SFT/Instruction Tuning, LoRA/QLoRA, Evaluation, Agentic RAG, Prompt Engineering, Tool Calling
LLM Frameworks: HuggingFace Transformers, PEFT, TRL, Accelerate, vLLM, PyTorch
Classical AI/ML: Scikit-learn, XGBoost, OpenCV
MLOps: MLflow, Weights & Biases, Airflow, Prometheus, Grafana, Docker, Terraform