Yuxin (Eric) Miao

LLM Engineer / Machine Learning Engineer

17317151652 | Shanghai, China | yuxin_miao@berkeley.edu

Education

University of California - Berkeley Dec. 2023

M.DevEng in Development Engineering (AI & Data Analytics), GPA: 3.96/4.0

ShanghaiTech University July 2021

Bachelor's in Computer Science, GPA: 3.63/4.0; Minor in Finance, GPA: 3.70/4.0

Skills

Languages: Python(7yr+), SQL(5yr+), Linux Shell, R, C++, Assembly

Applications: RAG, LLM deployment & finetune, Machine Learning, Deep Learning, MLOps, Data Analytics & Visualization

Tools: Azure, Kubernetes, Git, Docker, Ollama, OpenAI, Huggingface, PyTorch, Tensorflow, Pandas, Numpy, Scikit-learn, Visio, SPSS, etc.

Experience

Senior LLM Engineer May 2024 - present

Accenture China Shanghai, China

LLM

- **Text2SQL:** Developed agentic workflows application for a leading pharmaceutical client, enabling intelligent database interaction. Led backend algorithm development, prompt engineering, and data analysis. Integrated private knowledge bases to resolve domain business queries. Achieved more than 90% accuracy.
- **AI-scientist**: Deployed and customized an autonomous research assistant system for biomedical research, supporting end-to-end workflows from data preprocessing to algorithm design and adaptive model evolution based on user-defined research goals. Enabled expert-level data analysis in a fraction of the time, achieving comparable performance to human data scientists using only 5% of the effort.
- **NL2Slide:** Built a presentation generation system that converts natural language descriptions and medical articles into professional presentation slides. Supported various layouts, tables and charts extracted from text or scientific papers, and paper-derived images. Engineered slot-filling mechanisms and direct XML manipulation techniques.
- **AI-diagnostics:** Customized RAG framework to generate preliminary hospital exam result diagnostics based on medical paper. Outperform Baidu's similar products on response proficiency and quality.

ML

- **CV Disease Detection:** Designed and implemented deep learning and transfer learning algorithms (ResNet, DANN, ViT) and data pipelines for multiple serum disease multi-level classifications from medical images. Achieved significant performance improvement over existing solutions, effectively enhancing diagnostic accuracy.
- **Customer Data Modeling Pipeline:** Collaborated with a leading luxury group to harness cross-maison customer data, leveraging machine learning models to accurately evaluate customer value across brands with absent purchase behavior. Designed and implemented a comprehensive data pipeline and automation framework on Dataiku, streamlining data integration and model deployment for scalable, high-impact insights.

Machine Learning Engineer Intern Jun. 2023 - Aug. 2023

Walmart Sourcing Technology Shanghai, China

- Designed and implemented a LLM Chatbot for technical support, incorporating RAG architecture and leveraging state-of-the-art pre-trained models such as BERT and GPT, utilizing 4 Tesla-T4 GPUs, employing LoRA and ZeRO3 techniques, achieving an impressive model accuracy of 70%.
- Executed deployment on Azure with Kubernetes, complete with robust monitoring capabilities and automated retraining based on user feedback for continuous model improvement and optimization.

Junior Data Scientist July 2021 - Mar. 2022

Buy-Quickly Inc. Shanghai, China

- Directed a team of three in developing unsupervised customer segmentation models utilizing k-means clustering, neural networks, and strategic feature engineering
- The results introduced an innovative advertising aspect for online merchants, enabling precise identification of potential customers. Over 5000 online luxury retailers embraced our tagging toolkits, enhancing audience targeting efficacy across diverse categories, showing a 10% increase in ROI in A/B test.
- Empowered the team through the utilization of crawlers and NLP-based consumer review analyzers, enhancing operational efficiency and reducing workload. Streamlined workflow processes, significantly reducing the weekly time commitment from 10 hours to just 0.5 hours without compromising productivity