

# YIXIAO(CLAIRE) LING

(+001) 267 206 2692

[yxling@alumni.upenn.edu](mailto:yxling@alumni.upenn.edu); <https://github.com/yixiao-ling?tab=repositories>; <https://www.linkedin.com/in/yixiao-ling/>

## EDUCATION

**University of Pennsylvania** | Penn Engineering

Aug. 2023 – May. 2025

**M.S.E.** in Data Science

**Relevant Courses:** Data Structure and Algorithms, Machine Learning, Distributed Systems, Deep Learning, Large Language Models, Computer Vision

**GPA:** 3.96/4.00

**University of Nottingham** | School of Mathematical Science

Sep. 2019-Jul. 2023

**B.S. in Mathematics and Applied Mathematics**

**Relevant Courses:** Statistical Models and Methods, Probability Models and Methods, Econometrics

**GPA:** 3.92/4.00 (Top 5%)

**Honors:** Dean's scholarship

**The London School of Economics and Political Science**

Jun. 2021 – Jul. 2021

Summer Course Program in Computational Methods for Financial Mathematics

**Grade:** A+

## TECHNICAL SKILLS

**ML & AI Engineering Skills:** Experience in end-to-end ML and implementing CI/CD pipelines. Experience in deploying models to cloud services (AWS, Azure). Experienced in working with ML platforms (Dataiku, SageMaker). Expertise in LLM technologies and AI infra. Strong experience with Generative AI models.

**Data Science Skills:** Proficiency in large-scale systems data analysis and statistical modelling.

**Programming Languages:** C++, Python, SQL, Java, Spark, MATLAB

**Platforms & Tools:** Google Cloud, Amazon Web Services, RDMS, MYSQL, R Studio, Tableau, Power BI

## WORK EXPERIENCE

**AI Engineer, SAP**

Jun. 2025 – Now

- Developed **full-stack distributed multi-modal AI system** (chatbot and **slides auto-generation**) for sales professionals, enabling rapid customer analysis with high precision and actionable insights.
- Engineered real-time data ingestion from multiple sources, including document processing and web scraping, incorporating **automated quality validation, S3 cloud storage**, and intelligent **metadata management** for seamless content extraction.
- Integrate multiple LLM providers (**OpenAI, Gemini, Claude**), **optimized algorithms** and built **caching** layer with Redis for session management, file uploads and API response caching.
- Developed **concurrency management** using **ThreadPool, asyncio, rate limiting**, Redis streams and **process tools** for CPU-intensive tasks, scaling system to handle 10,000+ concurrent requests with **queue management**.
- Established **monitoring and alerting** with custom metrics for API response times, error rates and resource utilization.

Built **Analytics dashboard** with real-time metrics, usage tracking, and performance monitoring for business teams.

**AI Engineer Intern, Alibaba Cloud**

Feb. 2024 – Sep. 2024

- Designed and built a **scalable RAG-based chatbot system** on distributed architecture using **cloud service**, deployed with **FastAPI** and maintained system scalability & performance.
- Developed a **multimodal module**, optimized **LLM reasoning** by **fine-tuning, prompting** techniques and developed agents to improve conversation quality.
- Developed ElasticSearch (ES) **vector database**, optimized database loader, semantic **search** and **ranking algorithms**, increasing the recall accuracy of 4.5% and recall speed of 15%.
- Used the **CI/CD pipeline** to automate deploying the optimized LLM model to production and created a diagnostics platform for real-time monitoring & analysis of LLM **distributed training**, ensuring effective **performance tracking**.

**Generative AI Intern, Wharton Analytics**

Sep. 2023 – Feb. 2024

- Developed a **Machine Learning** solution for Hearst Corporation to automatically align magazine content with marketplace taxonomy.
- Employed and optimized **PySpark** for **distributed data processing** on large volumes of HTML data. Leveraged Delta Lake to manage data versioning and maintain a reliable, auditable data pipeline.
- Used **AWS** for building, training, and deploying the NLP models. **Fine-tuned BERT** for **NLP** semantic analysis.
- Achieved a 78% tagging accuracy rate, leading to a 15% increase in user engagement and a 10% boost in sales.

**Machine Learning Engineer Intern, Experian**

Jan. 2023-Sep. 2023

- Developed data preprocessing pipeline on 13GB user interaction data to prepare for building risk control model.
- Conducted **feature engineering**, selecting 35 key features from 5000+ features by feature importance and variable correlation. Derived 89 new variables based on real business.
- Deployed **risk control model** to determine whether to give credits to clients by using machine learning models, such as **logistic regression, XGBoost** and **LightGBM**, achieving accuracy of 0.85 classifying good & bad clients.