# YIXIAO LING

yxling@alumni.upenn.edu • Linkedin • Github • Personal Website

San Francisco Bay Area (open to relocation at own expense) • (267) 206-2692

## EDUCATION

**University of Pennsylvania |** Penn Engineering *Aug. 2023 – May. 2025*
**M.S.E.** in Data Science
**Relevant Courses:** Operating Systems, Software Systems, Big Data Analytics, Machine Learning, Principle of Deep Learning, Large Language Models, Computer Vision
**GPA:** 3.90/4.00

**University of Nottingham |** School of Mathematical Science *Sep. 2019-Jul. 2023*
**B.S. in** Mathematics and Applied Mathematics
**Relevant Courses:** Data Structure and Algorithms, Optimization, 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 Software Engineer,** Wejob *Present*
**AI Engineer intern,** SAP *Jun.2025 – Oct.2025*

- Architected full-stack Business AI system with **multi-modal chatbot** and **automated executive slide generation**, serving 500+ sales professionals and reducing presentation prep time by 75% and cutting cost by over 90%.
- **Worked with business users** to **gather business requirements** and **translate** them into technical specifications for a Business AI system. Led **stakeholder communication** throughout the software implementation lifecycle
- Engineered real-time data ingestion from multiple sources, incorporating **S3 cloud storage**, multiple LLM providers, and built **caching** layer with **Redis**.
- Developed **concurrency management**, for CPU-intensive tasks, scaling system to handle 1,000+ concurrent requests with **queue management**, reducing API costs by 40% and improved response times from 8s to 1.2s.
- 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%.
- Deployed containerized LLM models on **Kubernetes** with auto scaling and used **CI/CD** pipeline to automate deploying the optimized LLM model to production and created a diagnostics platform for real-time monitoring.

**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*

- Built end-to-end data and ML pipeline: processed 13GB of user data, engineered 124 key features, implemented PostgreSQL queries and Django REST APIs supporting 200K daily transactions at 99% uptime, and deployed credit risk models (logistic regression, XGBoost, LightGBM) achieving 85% accuracy in classifying clients.