CORLEONE YANG

GitHub | LinkedIn | Website | Google Scholar | yahe.yang@gwu.edu

EXPERIENCE

VEA — Los Angeles, United States
Full-Stack Engineer (React.js, node.js, RDS-pgvector, AWS, LangChain, Stripe API) — Sep 2024 – Present

- Collaborated in an **agile team** to fix 10–15 weekly bugs related to UI errors, API failures (core, AI, cloud, third-party, auth, Stripe), concurrency issues, incorrect data rendering, and caching inconsistencies. Used Postman and AWS CloudWatch for full-stack troubleshooting. Wrote unit tests using Jest to ensure code reliability, maintaining **65% SonarQube coverage**. Worked closely with the CI/CD team to ensure smooth periodic deployments from dev to sandbox to production.
- Developed a dynamic SOAP note sidebar, enabling simultaneous API fetching and form state updates to ensure seamless initial rendering and real-time synchronization. Designed an optimized data flow, structuring dependencies to ensure controlled updates, managing state propagation to prevent race conditions, and minimizing re-renders for a more responsive UI. Received positive user feedback after deployment to production.
- Extended the company's service layer by integrating **OpenAI APIs** and AWS Bedrock APIs as foundational components, enabling **Embedding Storage**, **Vector Search**, **Retrieval-Augmented Generation (RAG**), and **Rule-based Filtering**. Optimized the recommendation system to process both structured data and unstructured text, while maintaining a vector database to expand and evaluate new diagnostic records for continuous improvement.

Baynovation — San Jose, United States
Software Engineer Intern (Flask, FastAPI, AWS Glue, AWS SageMaker, Huggingface) — May 2024 – Aug 2024

- Developed an ETL pipeline (S3 → AWS Glue → Parquet) and integrated AWS SageMaker for feature engineering, model training, and inference. Trained a Transformer-based time-series forecasting model, improving prediction accuracy by 18%. Exposed a REST API for real-time predictions and leveraged backend APIs to access SageMaker inference endpoints for deployment.
- Fine-tuned a **LLaMA3-based** chatbot using company internal documents with **Lora** (Low-Rank adaptation) via the Hugging Face PEFT library. Collected and analyzed frequently asked user queries to evaluate the fine-tuned model's performance. Iteratively refined system prompts based on evaluation results and redeployed the chatbot via API for enhanced accuracy and user alignment.

Tencent — Shenzhen, China
Machine Learning Engineer Intern (Pytorch, TensorFlow, Gymnasium) — August 2022 - October 2022

- Built a reinforcement learning framework with OpenAI Gym and PyTorch on a Conda-based Ubuntu environment. Implemented the Deep Q-Network (DQN) algorithm, optimizing smart agent performance to achieve 86% task completion accuracy in simulated environments.
- Developed a ranking model inspired by **BERT4Rec**, leveraging self-attention mechanisms to model user click sequences. Optimized the ranking stage in a recommendation pipeline using TensorFlow, improving **CTR** by 7% in offline evaluations with historical interaction data.

XiuNeng Capital — Shenzhen, China
Quant Intern — May 2021 - August 2021

- Designed and implemented an **Event Study Algorithm** using Python, pandas, and matplotlib to analyze the impact of specific events on stock prices. Improved event impact prediction accuracy by **18%** through integrating data from Bloomberg Terminal and internal databases.
- Conducted factor selection and ranking using correlation analysis, PCA, and cross-sectional regression to identify predictive factors for portfolio optimization. Integrated statistically significant factors into a multi-factor model for asset allocation. Back tested the strategy using Python (Backtrader), achieving a **12% improvement** in the Sharpe Ratio compared to the benchmark.
- Developed a low-latency event-driven trading system by integrating real-time event streaming data from broker APIs. Leveraged **XGBoost** for event classification, enabling rapid decision-making in a high-frequency trading environment. Conducted simulated trading to evaluate strategy profitability, optimizing model parameters based on back-tested performance.

PROJECTS

Smart Inventory Management System | Link

- Engineered a full-stack inventory management system using **Next.js**, **Material UI**, and **Firebase**, integrating **OpenAI Vision** for image classification and **HuggingFace's Llama3.1** for inventory management suggestions, with **CI/CD** implementation via **GitHub Actions** and **Vercel** for streamlined deployment.

Document Chatbot

- Developed a full-stack document chatbot using **Express.js** and **React**, implementing **RAG** with **AWS Bedrock API** and **Pinecone** for efficient vector storage and retrieval. Integrated **LangChain** for enhanced language model interactions. Implemented **CI/CD** pipeline using **GitHub Actions** for automated testing and deployment to **AWS EC2**, ensuring seamless updates and scalability.

TikTok Shop Content Recommendation Engine

- Developed a content recommendation engine for TikTok Shop using **Golang**, **Python**, and **TypeScript**, with **Kafka** for real-time data streaming and **MySQL** for data storage. Implemented personalized product and content suggestions based on user keywords and tags, significantly improving user engagement and driving sales growth.

School Projects

- Implemented a polynomial derivative calculator, A* algorithm for 8-puzzle problem, lightweight database using AVL trees, and matrix operation algorithms, all utilizing **Java**.(Data Structure). Implemented a custom OS kernel with thread creation and signal handling, developed a multi-threaded frog river-crossing simulation, designed a GPU-accelerated virtual memory system using **CUDA** with page replacement algorithms, and created a GPU-based file management system featuring root directory and I/O operations. (OS)

CERTS

Cloud Certificates: Azure AI Engineer Associate | AWS Solution Architect Associate | AWS Data Engineer Associate

Kaggle Medals: Explainable AI for Patent Professionals | Chatbot Arena Human Preference Predictions

SKILLS

Programming Languages:	Python | Java | C | C++ | C# | CUDA | HTML/CSS | JavaScript/TypeScript | MATLAB | R | Go
Backend/Frontend:	Node.js | Express.js | Next.js | React | SpringBoot | Flask | Django | PostgreSQL | MongoDB | Redis
Tools:	AWS/Azure/GCP | Kubernetes | Docker | Git | Jenkins | Terraform | Postman | Huggingface | LangChain
Machine Learning/Data:	PyTorch | TensorFlow | NumPy | Pandas | Matplotlib | Scikit-learn | Databricks | Kafka

EDUCATION

The George Washington University — **Washington, DC**
Master of Science in Information Technology — *Aug 2023 - May 2025*

Chinese University of Hong Kong — **Shenzhen**
Bachelor of Science in Mathematics — *Sep 2019 - May 2023*