David Dai

10 E Reed St, San Jose, California 95112
607-379-1652 yanfeidai0811@gmail.com https://www.linkedin.com/in/yanfeidai/

Education

Cornell University **Aug. 2023 – Dec. 2024**
Master of Computing and Information Science *Ithaca, New York*

Nanjing University **Sep. 2018 – Jun. 2023**
Bachelor of Engineering *Nanjing, Jiangsu*

Experience

PayPal Inc. **Feb. 2025 – Present**
Machine Learning Engineer *San Jose, CA*

- Contributed to candidate generation (recall) via a **two-tower embedding pipeline** and ANN retrieval to ensure high-quality candidate pools for the Shopping feed.
- Developed and productionized **fine-rank models** for Shopping feed with **advanced feature crosses** and Feature Store integration. Built **calibrated CTR/CVR predictors** addressing delayed-feedback bias and class-imbalance, supplying stable scores to ranking and auction systems for real-time ad allocation.
- Architected and deployed an end-to-end **Agentic fraud analysis system** where a **planning agent** orchestrates specialized agents via **MCP tool calls**. Optimized the pipeline with **sematic and retrieval layer**, and established **CI/CD** pipelines, reducing manual analysis **by ∼60%**.
- Developed a **mini-model system** for rapid **root-cause analysis** of novel fraud trends. Combined **Information Value** and **LightGBM feature importance ranking** to identify top risk drivers, providing actionable insights to downstream modeling teams and accelerating fraud response **from 3 days to 1 hours**.

ProtagoLabs/Netmind.AI **May. 2024 – Dec. 2024**
Machine Learning Engineer *Vienna, VA*

- Optimized LLMs for NetMind.AI's token-based model serving platform, enhancing performance and cost-efficiency across owned and rented infrastructure, supporting a daily active user base of over 10,000 on the NetMind Power platform.
- Implemented **quantization** using **TensorRT** across multi-GPU and multi-node environments with **DeepSpeed**, achieving a 6x throughput increase over baseline models with accuracy loss under 0.05 based on benchmark results.
- Focused on developing automated **fine-tuning** capabilities for the platform, leveraging PEFT techniques with **LoRA** and **SGLang** to enable end-users to customize models efficiently with a simple, click-to-tune option.
- Collaborated with the research team to explore **post-training** scale laws: used CPT for long reasoning outputs, SFT for multi-turn dialogue structuring, and RL with heuristic rewards to balance depth and efficiency, challenges with managing long outputs and reward hacking informed further refinements.

Boston Derm Advocate **Nov. 2023 – May. 2024**
Machine Learning Engineer Intern *Remote*

- Modeled personalized **skincare recommendations** as a supervised learning and ranking problem, predicting user-specific product preference scores to optimize downstream recommendation quality.
- Implemented and evaluated tree-based ensemble models (**XGBoost**) for preference prediction, performing systematic feature ablation and metric-driven model selection using **F1 and AUC**.
- Engineered structured features across user profiles, product attributes, and interaction signals, and leveraged **feature importance analysis** to refine representation design and reduce model complexity.
- Improved model generalization by mitigating **overfitting via regularization and early stopping**, and assessed scalability and inference stability on large datasets, resulting in significant uplift in user engagement.

Projects

LLM-Driven Categorization and Financial Information Extraction | *Microsoft* **Jan. 2024 – Jun. 2024**

∗ Simulated a delayed-feedback ad conversion dataset by modeling click-to-conversion latency with parametric distributions (e.g., exponential, Weibull) to reflect real-world conversion delays in online advertising.
∗ Implemented and compared multiple delayed-feedback strategies (naive labeling, fixed-window attribution, importance weighting, and survival analysis), analyzing bias–variance tradeoffs under different delay regimes.
∗ Built a modular experimentation framework with configurable delay distributions and evaluation protocols, and summarized findings on implications for CVR estimation and ads ranking systems.