Xinhao Jiang

bennyjxh@gmail.com | 510-207-4905 | linkedin.com/in/benny-jiang | bennyjiang.com | SF Bay Area

EXPERIENCES

Software Engineer, Monetization Generative AI | Meta, Menlo Park, CA — Apr 2024 - present

- **Overview:** Leveraging large language models (LLMs) to optimize ad creation, delivery, and performance.
- **Reinforcement Learning:** Applied Proximal Policy Optimization (PPO) to train Llama3 for generating engaging ad texts, resulting in 1% lift in ad Click-Through Rates and 1.5% incremental revenue.
- **LLM Engineering:** Architected semantic extraction and summarization models, scaling serving infrastructure to process 100M+ ad copies and websites weekly.
- **GenAI Product Development:** Summarized lengthy ad texts to improve user digestibility on FB Reels and IG Stories, boosting conversion rates by over 1% and driving a 1.5% increase in in-segment revenue.
- **LLM Serving Optimization:** Implemented persistent KV caching and speculative decoding for LLM inference on GPU clusters, reducing P50 latency by 60% and enabling real-time ad content generation.

Machine Learning Engineer, Recommender System | Meta, Menlo Park, CA — Jan 2023 - Apr 2024

- Optimized Facebook's notification ranking system, achieving 80k+ increase in Daily Active People.
- **ML**: Enhanced early-stage ML ranking model architecture and features, resulting in ~10% peak capacity savings.
- Developed the retrieval and rendering process for group activity notifications, boosting user engagement by ~5%.

Machine Learning Engineer Intern, Recommender System | Meta, New York, NY — May 2022 - Aug 2022

- **Ranking Optimization:** Improved viewer watch time by over 2% through the design and implementation of value model logic promoting fresh, timely media in the ranking system.
- **Data Engineering:** Assembled data pipelines calculating engagement time for ~20 million IG Reels posts daily.
- **ML Modeling:** Developed multitask multi-label deep learning models predicting media time-sensitivity.
- **Signal Serving:** Deployed trained models as public features and onboarded reliable real-time signals.
- Launched language mismatch filter in indexing system, boosting viewer engagement metrics by over 3%.

Software Engineer Intern, Financial Technologies | Tencent, Shenzhen, China — May 2021 - Aug 2021

- Collaborated with a 15-member team to enhance the transaction and user management of digital currency platforms.
- Developed an RPC server in C++ enabling customers to forge secure payment channels with merchants.
- Upgraded an RPC framework supporting multi-process reverse proxy and server load balancing.

Software Developer Intern | Hulu, Santa Monica, CA — Jun 2020 - Aug 2020

- Constructed an integration platform monitoring the status of seven micro-services, reducing time for engineers across five teams for server inspection and debugging.
- Led a project team of three in developing a Chrome extension using JavaScript, allowing users to collect and share video clips in TV shows.

EDUCATION

Carnegie Mellon University — Pittsburgh, PA

M.S. in Computational Data Science, Analytics Concentration — May 2021 - Dec 2022

- **QPA: 4.01/4.00**
- Relevant Coursework: Search Engine, Deep Learning, Advanced Natural Language Processing, Machine Learning, Cloud Computing, Distributed Systems, Large-Scale Multimedia Analysis, Interactive Data Science.

University of California, Berkeley — Berkeley, CA

B.A. in Computer Science & Data Science, emphasis in Applied Mathematics & Modeling — Aug 2017 - May 2021

- **GPA: 3.93/4.00** (High Distinction in General Scholarship)
- Relevant Coursework: Deep Reinforcement Learning, Database Systems, Operating Systems, Parallel Programming.

SKILLS

Tools & Technologies: Spark(ETL, SparkML), Cloud Computing(AWS, Azure, GCP, Kubernetes), Docker, MySQL.

Programming Languages: Python(PyTorch, Scikit-learn, NumPy, SciPy, Pandas), PHP, Java, C++.

PUBLICATIONS & AWARDS

Publication

- **Jiang, X.** ***"Diagnosing root causes of intermittent slow queries in cloud databases"***, Proc. VLDB Endow. 13, 8 (April 2020), 1176–1189.

Awards

- **Top-5 Finalist,** Inaugural Alexa Prize SimBot Challenge, April 2023;
  - Modular design with Fine-tuned LLM and VLM to control simulated robot. (Link)
- **3rd Place Winner**, East Coast Regional Datathon Presented by Citadel in partnership with Correlation One, April 2022.