Xi (Robby) Qiu

San Jose, California
610-604-2243 robbyqiu@outlook.com linkedin.com/in/xi-qiu github.com/robnanarivo

Work Experience

Machine Learning Engineer – PayPal *San Jose, CA* | **Mar 2025 – Present**

- Designed and built a multi-agent AI system for Suspicious Actor Report (SAR) generation using Gemini, Next.js, and FastAPI. Improved agents' reasoning capability by 2.5x and raised report accuracy by 50% through prompt engineering, evaluation, and workflow enhancements.
- Built and optimized agentic workflows for entity extraction, data aggregation, and structured report generation. Developed a human-in-the-loop module that revises LLM-generated analyses based on investigator feedback.
- Evaluated and integrated vLLM and LMCache into production ML pipelines, reducing inference latency by 40%.
- Migrated 10+ production ML pipelines to modern, high-performance platforms using GenAI-assisted automation, significantly reducing manual migration effort.
- Ran knowledge-sharing sessions on frontier AI research and GenAI fundamentals, including paper reviews and LLM architecture deep dives.

Software Engineer – PayPal *San Jose, CA* | **Jul 2023 – Mar 2025**

- Spearheaded the development of 3 microservices for PayPal's rewards platform serving over 10 million users, achieving a 10x increase in rewarded transactions. Used Java, Spring Boot, Spring Batch, and Message Queue.
- Increased revenue by 30% through developing a feature to charge rewards commissions directly from merchants. Eliminated the cost of intermediary agencies for merchants.
- Led an initiative to improve the test framework. Reduced functional test execution time by 40%, increased test coverage for all components by 10%, and improved traceability of test case changes.

Education

University of Pennsylvania Aug 2021 – May 2023

- *Master of Computer and Information Technology* – GPA: 4.0

University of Oxford Oct 2015 – Jun 2019

- *BA in Philosophy, Politics and Economics* – First Class Honours

Projects

Transformer Interpretability Research | *PyTorch, TransformerLens, CircuitVis* | Link

- Recreated the GPT-2 model architecture from scratch based on the original paper. Trained on the `TinyStories` dataset and experimented with various sampling methods including Temperature, Top-k, and Top-p sampling.
- Detected induction heads with attention pattern visualizations and ablation in GPT2-small. Reverse-engineered QK-, OV-, and K-composition circuits using factorized matrices of query, key, and value weights of the transformer heads.

Alignment Evaluation for Frontier Models | *GPT, Claude, InspectAI* | Link

- Built a threat model on LLM's tendency for self-preservation and defined metrics to measure this property.
- Generated a multiple-choice question dataset of 300 questions using Claude and GPT. Scored and filtered questions with LLM judge to ensure that the dataset is high-quality, diverse, and unbiased.
- Ran evaluations on GPT-4o mini with the dataset using InspectAI evaluation framework. Analyzed model behaviors across different prompting strategies including zero-shot, chain-of-thought (CoT), and CoT with self-critique.

RLHF for GPT-2 | *PyTorch, Gymnasium, TransformerLens* | Link

- Built the foundational reinforcement learning skillset required for RLHF by implementing and training classic algorithms (Q-learning, SARSA, DQN, and PPO) from scratch using Gymnasium environments.
- Implemented end-to-end PPO and GRPO algorithms for training GPT-2 with RLHF, including value head modeling, rollout collection, advantage estimation, and KL-regularization.
- Fine-tuned GPT-2 with both full-parameter training and LoRA. Experimented with rule-based and sentiment-based reward functions to analyze how distinct reward signals steer model behavior.

RAG App for Climbing Trip Planning | *Python, LangChain, Vector DB, Scrapy* | Link

- Developed a retrieval augmented generation recommendation system for rock climbing trip planning. Used LangChain for prototyping and optimized performance with prompt engineering and fine-tuning.
- Created a web scraper with Scrapy and stored the data in a vector DB for retrieval.

Distributed Cloud Platform | *C++, Unix* | Link

- Created a distributed cloud platform that offers email and file storage services in a team of 4. The platform supports multiple frontend and backend servers, and features data replication, fault tolerance, and data recovery.