Every colored text is a link.

Ayrton San Joaquin

Scientist & ML Engineer, Trustworthy AI

ayrton@aya.yale.edu | Singapore | ajsanjoaquin | ajsanjoaquin | Values

Education

Yale-NUS College *Singapore*

BSc. (Honors) in Data Science, Minor in Philosophy. **(Scholar, With High Distinction)** *August 2018 - May 2023*

Semester Abroad at the University of Copenhagen, Denmark

Experience

French National Centre for Scientific Research (CNRS)@CREATE *Singapore*

AI Scientist, DesCartes program (https://descartes.cnrsatcreate.cnrs.fr/) *September 2023 - Present*

- Leading a study on efficient fine-tuning of open-source Large Language Models (LLMs) for instruction-following.
- Fine-tuned dozens of SOTA LLMs (Llama-{3,2}, Mistral, Mixtral, Phi-2, Gemma, TinyLlama) on Ultrachat dataset (200k) and Alpaca (52.2k)
- Implemented automated training scripts in distributed settings ranging from 1-2 node clusters provided by the Singapore National Supercomputer.

Machine Learning Safety Scholars Program, Center for AI Safety *Palo Alto, United States*

Scholar *June 2022 - August 2022*

- Performed prompt-injection attacks against LLMs (GPT-3, LaMDA, T5) via API access.
- Implemented various strategies in **robustness** (PGD, adversarial training), **anomaly detection** (AUROC, ViM), **calibration** (RSME, Brier scores), and **trojan attacks** (data poisoning).

Data Privacy and Trustworthy Machine Learning Lab, NUS *Singapore*

Research Engineer *May 2021 - March 2022*

- Collaborated with **Google DeepMind** on privacy and adversarial machine learning research for my bachelor's thesis in a team across 4 time zones. **Published in a top security conference (ACM CCS) as the youngest and only undergraduate co-author.**

Arterys (Freelance) *San Francisco, United States*

Deep Learning Engineer *March 2020 - June 2020*

- Created a COVID-19 Pneumonia classifier **4 days after pandemic declaration in collaboration with A.I. Singapore** based on a ResNet-38.
- Collaborated with Arterys to deploy the model in their platform for use by American hospitals and researchers. Model engineer in a team of 4 across 3 time zones.

AIEngineering Projects

DesCartes Program Semantic Search Engine *LLMs*

• Deployed a system internally to perform document retrieval via vector-based semantic search. Accepts and returns multilingual queries in English and French. The system is composed of a Llama-3 (8B) model for chat and bge-small model for embeddings. Orchestrated via LlamaIndex.

Meta Project Aria Workshop 2023 *LLMs, Contextual AI*

• **Invited by Meta Reality Labs** to design a use-case for Project Aria. Created an LLM assistant that uses social media data and real-time visual context from smart glasses. Video demo.

Document Summarization *LLMs*

• Fine-tuned GPT-J (6B) on 100k arxiv pre-prints to summarize research papers used by my lab.

Eplaining Neural Networks with Meaningful Perturbations *Explainable AI*

• For explaining an image classifier's prediction, I implemented the algorithm described in *Explanations of Black Boxes by Meaningful Perturbation (Fong, et. al., 2018)*.

Equitable Valuation of Data Using Shapley Values *Explainable AI*

• I did a Pytorch implementation of computing Shapley values via Truncated Monte Carlo sampling from *What is your data worth? Equitable Valuation of Data (Ghorbani and Zou, 2019)*.

COVID-19 Medical Triage Model *Computer Vision*

• I developed a model meant to help triage patients (prioritize certain patients for testing, quarantine, and medical attention) that require diagnosis for COVID-19. Trained on 26k x-ray images.

Open-Source AI *DevOps*

• Added new features for major machine learning projects including **Pytorch, HuggingFace Transformers, and YOLOv4** (object detection model).

Skills

Machine Learning: Pytorch, Tensorflow, LlamaIndex, JAX, HuggingFace, Langchain, NLTK, Spacy

Data: Pandas, PySpark, Querying (SQL, MongoDB), Vector Database (Qdrant, Pinecone)

MLOps: Linux, Databricks, GCP, AzureML, Snowflake, Docker, Flask, Continuous Integration, Kubernetes