

Ayrton San Joaquin

Every colored text is a link.

🧪 [SCIENTIST, TRUSTWORTHY AI](#) | ✍️ [WRITER](#)

✉️ ayrton@aya.yale.edu | 📍 Singapore | 🔄 [ajsanjoaquin](#) | 🌐 [ajsanjoaquin](#) | 📌 [Values](#)

Education

Yale-NUS College

BSC. (HONORS) IN DATA SCIENCE, MINOR IN PHILOSOPHY. (**SCHOLAR, WITH HIGH DISTINCTION**)

Semester Abroad at the University of Copenhagen, Denmark

Singapore

August 2018 - May 2023

Experience

AI Standards Lab

RESEARCH ANALYST

• Contributing to the Codes of Practice of the EU AI Act.

Singapore

August 2024 - Present

French National Centre for Scientific Research (CNRS)@CREATE

AI SCIENTIST, [DESCARTES PROGRAM](#)

• Led a Franco-Singaporean team studying efficient fine-tuning of Large Language Models (LLMs). Published at EMNLP Findings 2024.
• Created and curated the Tagalog dataset of <https://seaeval.github.io/>, a multilingual benchmark for Southeast Asian Languages.

Singapore

September 2023 - August 2024

Machine Learning Safety Scholars Program, Center for AI Safety

SCHOLAR

• Studied model failures (CV and NLP), and led research on analyzing LLMs using few-shot learning.
• Implemented various strategies in **robustness** (PGD, adversarial training), **anomaly detection** (AUROC, ViM), **calibration** (RSME, Brier scores), and **trojan attacks** (data poisoning).

Palo Alto, United States

June 2022 - August 2022

Data Privacy and Trustworthy Machine Learning Lab, NUS

UNDERGRADUATE RESEARCHER

• 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.**

Singapore

May 2021 - March 2022

Arterys (Freelance)

DEEP LEARNING ENGINEER

• Created a COVID-19 Pneumonia classifier **4 days after pandemic declaration in collaboration with A.I. Singapore.**
• 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.

San Francisco, United States

March 2020 - June 2020

Open-Source Projects & Public Service

The AI Summit Singapore 2024

• Invited as a featured speaker to discuss about data privacy for Generative AI at Asia Tech x Singapore, co-hosted by the Singapore Government.

Information Security

Public AI Advocacy

• Co-authored the foundational document of a global network of AI researchers, policymakers, and practitioners to center AI research and development in the public interest and as public infrastructure.

Trustworthy AI

Project Aria Timeline Builder Workshop 2023

• Invited by Meta Reality Labs to design a use-case for [Project Aria](#). Participated as the only bachelor's graduate among 20 graduate and postdoctoral researchers. [Video demo.](#)

LLMs, Contextual AI

Open-Source AI

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

DevOps

Publications

November 2024 **San Joaquin, A., et. al.**, In2Core: Leveraging Influence Functions for Coreset Selection in Instruction Finetuning of Large Language Models. *EMNLP Findings 2024*

[link](#)

November 2022 Tramer, F., ..., **San Joaquin, A.**, et.al., Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. *ACM Conference on Computer and Communications Security (CCS) 2022*

[link](#)

March 2020 **San Joaquin, A.**, Using Deep Learning to Detect Pneumonia caused by COVID-19. *Towards Data Science (Editor's Choice)*

[link](#)

Skills

Machine Learning: Pytorch, Sickit-Learn, JAX, HuggingFace, Langchain, NLTK, Spacy

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

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