Jean Michel Amath Sarr, PhD

Research Software Engineer

Accra / Ghana

https://jmamath.github.io/
jeanmichelamathsarr@gmail.com
Google Scholar
Linkedin
(+233) 5046 30427

Experience

Google Accra, Ghana

Research Software Engineer

Foundation Models for plant phenotyping (March 2025 - Present)
Artemis helps plant breeders find climate resistant crops by accelerating plant phenotyping using Foundation models. The goal is to use computer vision to accelerate the discovery of desirable traits in thousands of crops at various stage of growth.

Designed and implemented a reproducible evaluation infrastructure from the ground up to standardize and accelerate model assessment.

- Created a config driven infrastructure to track input variable change in our experimental workflow. Allowing to
- Authored a suite of libraries to automate the data ingestion pipeline. This work transformed raw COCO-formatted datasets into an efficient TensorFlow Datasets (TFDS) format, resolving loading errors on GPUs/TPUs and enabling the team to double the number of datasets supported before presenting our results in an industry workshop.
- Championed and enforced coding best practices, introducing a standardized lib/lib_test/binary structure, and implementing configuration files to manage model-specific variables.
- Streamlined team execution by organizing development tasks from a list to a tree with relevant dependencies
- Migrated a disorganized code base including many subdirectories for various experiments into a single abstraction to run inference at scale.

Gemini multilinguality (Sept 2023-March 2025)
Led the end-to-end development of a data generation pipeline to create instruction-response pairs for fine-tuning large language models. The resulting data boosted model performance by an average of 0.03 across 25 languages on a standard multilingual evaluation set, directly contributing to Gemini.

- Owned the project from design to production, establishing a systematic experimentation flywheel to measure and validate progress.
- Executed over 50 fine-tuning experiments, analyzing results to validate hypotheses and ensure performance gains were stable on next-generation models.
- Discovered and implemented a robust intervention that significantly improved data quality by leveraging more powerful models and advanced prompting techniques during generation.
- Adapted the pipeline to work with the latest models, ensuring its continued relevance for cutting-edge model development.

Research Resident

Multilingual Self-Instruction (March 2023 - Sept 2023)
Extended Self Instruct, a recent instruction tuning methodology to create multilingual instruction/response pairs to finetune LLM. The goal being to fill the performance gap of LLM in English and other languages. Multilingual Self Instruction uses LLMs to generate multilingual data tailored to improve performance on specific use cases like essays writing, poems, short stories, In the following I detail my contributions:

- Created a Multilingual Creativity test set based on the Bard. Creativity test set with translocalization (translation + localization).
- Generated data using PALM-2.
- Finetune PALM-2 on my dataset.
- Used silm-flow for automatic side by side evaluation of the trained models.
- Final dataset improved automatic sid by side quality in Japanese, Hindi and Korean.

XTREME-UP (Sept 2022 - March 2023)

- XTREME-UP is a user centric multilingual and multimodal benchmark for under-represented languages. As part of the team I was in charge of the autocomplete task. In the following I detail my contributions:
  - Created a dataset including 23 languages from Universal Dependencies.
  - Finetuned mT5 and ByT5 baseline using T5X and SeqIO.
  - Added top-k decoding to public library SeqIO in order to compute top-3 accuracy for mT5 and ByT5.
  - Contributed to the writing and analysis of the paper.

Research and Development Institute

Dakar, Senegal

Machine learning research engineer (Dec 2017 - Dec 2018)

- Benchmarked machine learning algorithms: Random Forests, Convolutional Neural Networks, Fully connected neural networks for bottom sea estimation in West African waters using multispectral acoustic data.
- Tuned Hyperparameters (learning rate and numbers of neurons in a hidden layer) automatically with Bayesian Optimisation techniques using the library GyOpt.
- Wrote paper: Complex data labeling with deep learning methods: Lessons from fisheries acoustics.

Oceanographic Research Center of Dakar-Thiaroye

Dakar, Senegal

Data analyst intern (Dec 2015 - Dec 2016)

- Implemented Generalized Linear Models (GLM) and Generalized Additive Models (GAM) in the field of fisheries to estimate the effect of climatic indices (MEI, AMO, SST, CUI) on abundance indices (recruitment, biomass, fertile biomass, etc) of the Octopus vulgaris.
- Result presented in : Kamarel BA, Jean Michel Amath SARR, et al. Fishing effects on Senegalese Octopus stock in the context of climate variability. International conference ICAWA 2016 : extended book of abstract : the AWA project : ecosystem approach to the management of fisheries and the marine environment in West African waters (p 65).

Education

Sorbonne University, Cheikh Anta Diop University — Paris, France

PhD, Computer Science (2019 - 2023)

- Investigated the role of data augmentation to improve robustness of neural networks under distribution shift. The thesis is available here.

Cheikh Anta Diop University — Dakar, Sénégal

Master of Research, Applied Mathematics (2015 - 2017)

Paul Sabatier University — Toulouse, France

Bachelor, Fundamental Mathematics (2009 - 2012)

Skills & Interests

- Python, tensorflow, keras, pytorch, jax, flax,
- Fine-tuning and evaluation of Large Language Models, statistical methods, research methods
- Interests: I like to dance (afrobeat, salsa, kizomba, bachata), I read a lot (psychologie, biologie, investment, health).

Awards & Honors

- My project Djehuty was selected in the UNESCO Top 100 outstanding projects using Artificial Intelligence for Sustainable Development Goals (2021)
- I was selected by the Google PhD Fellow (2020)
- I was selected by the Programme Doctoral International Modélisation des Systèmes Complexes (2019)