


Jean Michel Amath Sarr, PhD

Research Software Engineer

Accra / Ghana

 jeanmichelsarr@google.com
jeanmichelamathsarr@gmail.com

 (+233) 5046 30427

Experience

Google

Accra, Ghana

April 2024 -
present

Research Software Engineer

- Multilingual Instruction Backtranslation (MIB) is an agentic workflow extracting multilingual pieces of text from the web and generating plausible instructions out of it. These tuples (instructions, text) are then used for instruction tuning of Gemini. As part of MIB I had the following contributions:
 - Added a language identifier to the pipeline, allowing to keep only data in the target language.
 - Implemented the pipeline to support old and new versions of Gemini.
 - Leveraged the latest instruction tuned Gemini (gemit) to generate responses and paraphrase content from the Web.
 - Resulted in a 16% increase in side by side evaluation against Gemini multilingual.
 - MIB is a candidate pipeline to support scaling Gemini multilinguality to 100 languages

Sept. 2022
- March
2024

Research Resident

- Multilingual instruction tuning aims to fill the performance gap of LLM in English and other languages. As part of this effort, I extended [Self Instruct](#), a recent instruction tuning methodology to the multilingual context. 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 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](#).

Dakar, Senegal

Research and Development Institute

Dec 2017 -
Dec 2018

Machine learning research engineer

- 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

Dec 2015 -
Dec 2016

Data analyst intern

- 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 ICWA 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

2019-2023

PhD, Computer Science

- 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

2015-2017

Master of Research, Applied Mathematics

Paul Sabatier University

Toulouse, France

2009-2012

Bachelor, Fundamental Mathematics

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)