# Jean Michel Amath Sarr, PhD

**Research Software Engineer**

jeanmichelamathsarr@gmail.com | https://jmamath.github.io/ | Google Scholar | Linkedin

## Professional Summary

**Research Software Engineer (PhD)** specializing in scalable synthetic data infrastructure and MLOps for frontier LLMs. Core contributor to **Gemini's multilingual capabilities**, architecting the end-to-end synthetic data pipeline that scaled instruction-following across **25 languages**. Expert in decoupling experimental logic from infrastructure to accelerate research velocity (O(1) dataset scaling). Proven ability to bridge the gap between rigorous research methodology and production-grade distributed systems.

## Experience

### Google Research Africa

**Research Software Engineer**

**Foundation Models for Plant Phenotyping (Oct 2024 - Present)**
Building evaluation and experimentation infrastructure for Vision-Language Models and multimodal Foundation Models (PaliGemma, Gemini, Gemma) across computer vision tasks.

Config-Driven Experimentation Framework

- Introduced and championed the concept of config-driven experimentation, architecting a framework that became the standard for the Artemis team.
- Enforced version-controlled hyperparameter tracking in the codebase, eliminating ad hoc workflows and ensuring 100% reproducibility.
- Designed the framework's modularity to allow immediate extension through custom launchers, effectively decoupling experimental logic from infrastructure execution and enhancing experimentation efficiency.
- The framework was adopted by the whole team and is now the standard to run experiments.

Universal Data Pipeline and Abstraction

- Architected a generalized TensorFlow Datasets (TFDS) data pipeline that decoupled ingestion from data loading, enabling O(1) scaling for new datasets.
- Standardized the team's data infrastructure using a unified COCOA schema, expanding data capacity from 4 to 10 datasets a week before our last Q2 workshop, unblocking the team to deliver experimental results on 2.5x more data.
- Supported immediate integration of 50+ new datasets in Q3 with zero additional code changes.
- The pipeline was further extended to support fine-tuning and now serves as primary infrastructure for 60+ datasets across inference and fine-tuning for for object detection and instance segmentation

Improving Developer Velocity through Code Consolidation

- Consolidated fragmented research codes into a single task-centric binary, reducing technical depth and eliminating code duplication.
- Enabled a surge in research velocity, with more than 100 experiments launched in Q3, by refactoring the codebase.
- Delivered a unified interface that provides a stable foundation for rapid prototyping of new models without significant maintenance overhead.
- Drove a significant increase in productivity reducing the time to add a new model by 50% from a week on average to 2-3 days.

**Gemini Multilinguality (Sept 2023 - March 2025)**
Owned and shipped the synthetic data generation pipeline for Gemini's multilingual instruction-following , directly contributing to production release and enabling scale across 25 languages.

- Architected a scalable pipeline generating high-quality instruction-response pairs across 25 languages, improving win rates by an average of 0.04 on internal LMSYS evaluation against production baselines.
- Owned the complete MLOps lifecycle: data generation methodology, prompt engineering for quality, model ablations to validate improvements, inference optimization for cost efficiency, and rigorous large-scale evaluation.
- Executed 50+ fine-tuning experiments with systematic hypothesis testing, discovering and implementing interventions that significantly improved data quality through advanced prompting techniques and leveraging more powerful generation models.
- Adapted pipeline for next-generation models, ensuring continued relevance as model architectures evolved and maintaining production readiness.

Impact: Data directly shipped in Gemini, demonstrating production-scale synthetic data effectiveness for multilingual capabilities.

**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.

- Generated multilingual synthetic data tailored for specific use cases: essay writing, poetry, creative storytelling across multiple languages.
- Created a Multilingual Creativity test set based on the Bard with translocalization (translation + localization).
- Fine-tuned PaLM-2 on generated datasets, achieving measurable quality improvements in Japanese, Hindi, and Korean via automated side-by-side evaluation using Slim-Flow.
- Pioneered early synthetic data approaches that informed subsequent Gemini multilinguality work and established best practices for multilingual data generation.

**XTREME-UP Benchmark (Sept 2022 - March 2023)**

- Contributed to XTREME-UP, a user-centric multilingual and multimodal benchmark for under-represented languages, leading the autocomplete task development.
- Created multilingual datasets from Universal Dependencies spanning 23 languages and fine-tuned mT5/ByT5 baselines using T5X and SeqIO.
- 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.
- Co-authored benchmark paper and analysis presented at industry conference.

### Institute of Research for Development (IRD), Dakar, Senegal

**Machine learning research engineer (Dec 2017 - Dec 2018)**

- Benchmarked machine learning algorithms (Random Forests, CNNs, fully connected networks) for bottom sea estimation in West African waters using multispectral acoustic data.
- Implemented automated hyperparameter tuning using Bayesian optimization techniques (GPyOpt library).
  Modeled fishing effort and climate impacts on Senegalese Octopus vulgaris stock using advanced statistical models, quantifying changes in critical population parameters to provide actionable fishery management advice.

## Education

### PhD, Computer Science — Sorbonne University, Paris, France & Cheikh Anta Diop University, Dakar, Senegal (2019–2023)

- *Thesis*: "Study of Data Augmentation for the Robustness of Deep Neural Networks."
- *Focus*: Leveraging synthetic data to improve robustness under distribution shift and predict deployment performance on unlabeled domains.

### Master of Research, Applied Mathematics — Cheikh Anta Diop University, Dakar, Senegal (2015–2017)

### Bachelor, Pure Mathematics — Paul Sabatier Universty, France, Toulouse (2009–2012)

## Technical Skills

| Area | Skills |
|---|---|
| LLMs & Synthetic Data | Synthetic data generation for supervised fine-tuning and preference learning, fine-tuning and evaluation of LLMs, prompt engineering, model ablation, inference optimization |
| MLOps Infrastructure | Full-stack MLOps (data pipelines, experiment tracking, evaluation frameworks), distributed training, production deployment |
| Programming & Frameworks | Python (expert-level proficiency), TensorFlow, JAX/Flax, PyTorch, Keras, T5X, SeqIO, Pandas, NumPy, Scikit-learn |
| Research Methods | Statistical modeling, deep learning architectures, research design, academic publishing |

## Selected Publications & Writing

**Technical Reports (6000+ combined citations):**

- **Gemini Team, Google** (2025) — "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities."
  *Contributed to multilingual instruction-following capabilities through synthetic data generation pipeline*
- **Gemini Team, Google** (2023) — "Gemini: A Family of Highly Capable Multimodal Models"
  *Contributed to multilingual data generation and evaluation infrastructure*

**Research Writing:**

- **"Synthetic Alignment Research: Key Insights for AI Leaders"** (2025) — Four-part research series synthesizing insights from 20+ papers on RLHF limitations and synthetic alignment methods. Published at https://jmamath.github.io/.

**Conference Publications:**

- **XTREME-UP Benchmark** (2023) — Co-authored multilingual and multimodal benchmark paper for under-represented languages.

## Awards & Honors

- **UNESCO Top 100** outstanding projects using Artificial Intelligence for Sustainable Development Goals (2021) - *for Project Djehuty*.
- **Google PhD Fellow** (2020)
- **Programme Doctoral International Modélisation des Systèmes Complexes** (2019)

## Interests

I am an avid runner, I ran a semi-marathon this year in around 2:08, and ran a 5k under 22 minutes. I also like to dance (afrobeat, salsa, kizomba, bachata), and I read a lot (psychology, biology, investment, health).