Jean Michel Amath Sarr, PhD

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

Professional Summary

Impact driven 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. Proven ability to bridge the gap between rigorous research methodology and production-grade distributed systems.

Experience

Google Research Africa

Research Software Engineer

Multimodal tool use (Oct 2024 - Present)
Building evaluation and fine-tuning infrastructure for multimodal tool use in Vision-Language Models (PaliGemma, Gemini, Gemma), enabling models to reliably follow visual instructions across computer vision tasks in plant phenotyping.

10x Research Velocity through Consolidation and Decoupling

- Achieved 10x experiment velocity (from ~10 experiments in Q2 to ~100 experiments in Q4) through two complementary architectural interventions:
  - Consolidated fragmented research codebases into a single task-centric binary, eliminating code duplication and reducing model integration time by 50% (from 1 week to 2-3 days).
  - Decoupled experimental logic from inference infrastructure via a configuration-driven framework, enabling a teammate to build an experimental launcher without touching core code and allowing researchers to iterate at the hypothesis level rather than writing boilerplate code.
- The framework became the team standard with version-controlled hyperparameters and artifact management adopted across all research workflows.

Scaling Data Loading Infrastructure by 15x through Decoupling

- Scaled evaluation capacity from 4 datasets to 60+ datasets (15x growth from Q2 to Q3) by decoupling data ingestion logic from data loading infrastructure through a robust TensorFlow Datasets (TFDS) builder.
- Eliminated an unsustainable pattern where researchers wrote custom data loaders for each dataset to handle variable sizes and memory constraints. The TFDS builder solved data loading once, while allowing flexible ingestion pipelines (custom ingestion for non-COCO datasets, unified pipeline for COCO datasets).
- Enabled zero-code integration of new COCO-format datasets and standardized loading infrastructure, removing data as a research bottleneck and allowing the team to onboard 50+ datasets in Q3.
- The TFDS builder was further extended by the team and now serves as primary data loading infrastructure for 60+ datasets across inference and fine-tuning for object detection and instance segmentation.

Gemini Multilinguality (Sept 2023 - March 2025)
Shipped synthetic data generation pipeline to production for Gemini's multilingual instruction-following across 25 languages, improving win rates by 0.04 on average against production baselines.

- Built quality control mechanisms ensuring generated data promoted helpful, unbiased behaviors aligned with human intent: systematically analyzed 1000+ responses (50 per language across 20 languages) to develop a taxonomy of problematic patterns, then implemented filters that removed biased responses while preserving quality.
- Achieved quality improvements through 50+ fine-tuning experiments with systematic hypothesis testing, discovering and validating interventions using advanced prompting techniques and more powerful generation models.
- Architected end-to-end scalable pipeline handling data generation, prompt engineering, model ablations, inference optimization, and rigorous evaluation infrastructure across the complete MLOps lifecycle.
- Adapted pipeline for next-generation model architectures, ensuring continued production readiness as Gemini evolved.

Research Resident

Multilingual Self-Instruction (March 2023 - Sept 2023)
Closed the performance gap between English and other languages in LLMs by extending Self Instruct methodology to generate multilingual instruction-response pairs for fine-tuning.

- 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).