Franz Louis Cesista

helped bring down NanoGPT training to 3 minutes ($0.40)

🌐 leloykun.github.io
✉️ franzlouiscesista@gmail.com

Employment

Expedock — **machine learning research/engineer**, logistics automation, raised $20M funding — 2021–2024
- (re-)built our entire AI pipeline, from data collection, data warehousing and streaming, to model training, deployment, inference optimization, & continuous monitoring
- eliminated 99% of distributed system faults via defensive/chaos engineering
- built a DSL in TypeScript for a fully-programmable Metabase-like UX for data viz

Exora — research intern on forecasting energy supply-&-demand in the Philippines — summer 2020

Honours & awards

World Finalist (2x)	International Collegiate Programming Contest **(ICPC)**	Russia, 2021 & Bangladesh, 2022
Regional Finalist (3x) & 🥉 medalist | ICPC		Singapore, 2018 & Malaysia, 2019 & Jakarta, 2020
World Finalist (2x)	International Olympiad in Informatics **(IOI)**	Iran, 2018 & Japan, 2019
🥈🥉🥉 medalist	NOI Singapore Invitational	Singapore, 2016, 2017, 2018
Merit Scholarship	San Ignacio de Loyola Scholarship Program, Ateneo	2018
Merit Scholarship	Department of Science & Technology, Philippines	2018

Education

BS Mathematics from Ateneo de Manila University — 2018-2021; 2024-2025
(also took advanced graduate classes in preparation for further studies)
Philippine Science High School — 2012-2018

Selected publications :: see Google Scholar for the full list

"*Training Transformers with Enforced Lipschitz Bounds*"
NeurIPS 2025, under review — L Newhouse⋆, R Hess⋆, F Cesista⋆, A Zahorodnii, J Bernstein, P Isola

"*Retrieval Augmented Structured Generation: Business Document Information Extraction as Tool Use*"
IEEE MIPR 2024 — F Cesista⋆, R Aguiar, J Kim, P Acilo

"*Multimodal Structured Generation: CVPR's 2nd MMFM Challenge Technical Report*"
CVPR 2024, 2nd Multimodal Foundation Models Challenge — F Cesista

Selected blog posts (some w/ citations from published papers) :: see Ponder for the full list

"*Steepest Descent on Finsler-Structured (Matrix) Geometries via Dual Ascent*"
"*Rethinking Maximal Update Parametrization: Steepest Descent on the Spectral Ball*"
"*Squeezing 1-2% Efficiency Gains Out of Muon by Optimizing the Newton-Schulz Coefficients*"

Selected side projects

"*Modded-NanoGPT*"	NanoGPT (124M) in 3 minutes, NanoGPT (350M) in ~25 minutes
"*Multimodal Structured Generation*"	Interleaved, multimodal in-&-out structured outputs
"*Flash Attention Minimal*"	A 300-line C++ CUDA implementation of Flash Attention 1 & 2
"*Llama2.cpp*"	A C++ implementation of Meta's Llama2

Selected open source contributions

Google Deepmind's "Optax"	JAX implementation of the Muon optimizer
DottxtAI's "Outlines"	Interleaved, multimodal in-&-out structured outputs
Huggingface's "Transformers"	Helped simplify & unify API of multimodal models

Leadership

Cofounder and Former CTO of Google Developer Student Clubs (GDSC) - Loyola Branch

PyTorch • JAX • CUDA • C++ • Python • PostgreSQL • Snowflake • DBT • React • ReactRedux • TypeScript
GraphQL • Keras • Scikit-Learn • AWS SageMaker • Nvidia Triton • vLLM • Docker

optimizer-architecture codesign • multimodal ml • structured generation • information retrieval