WILLIAMS ZANGA

Staff AI/ML Engineer | Founding Engineer
Montreal, Canada | +1 438-979-8319 | zanga.williams@gmail.com | LinkedIn | Github | Research | Blog

AI/ML Engineer & Tech Lead with 6+ years of experience building 0-to-1 production ML systems and leading cross-functional R&D teams. Proven track record of translating applied ML research into scalable, low-latency products with measurable business impact. Deep expertise in recommender systems, multimodal ML, and optimization under real-world constraints, across AdTech and Defense environments.

TECHNICAL SKILLS

AI/ML & Research: Recommender Systems (Ranking & Forecasting), Multimodal ML (image, video, text, audio), Generative AI (LLM-based), Computer Vision, Audio Signal Processing, Vector Quantization, Computational Advertising
Systems & Architecture: ML Systems (0-to-1), Multimodal Pipelines, RAG Systems, Agentic & Workflow Automation, Model Compression & Quantization, Low-latency Inference, Evaluation & A/B Testing
Tools & Frameworks: PyTorch, XGBoost, LightGBM, OpenVINO, FastAPI, FastMCP, Streamlit, Faiss, Docker
Languages, Cloud & Data: Python, AWS (primary), GCP, MongoDB, SQL (projects), OpenAI & Gemini APIs

PROFESSIONAL EXPERIENCE

Staff AI/ML Engineer (Founding Team / Head of AI) | HIPPOC (Startup) **| Canada | Oct. 2022 – Dec. 2025**
(promoted from Lead to Staff to lead full-lifecycle full R&D and AI Architecture)

- Designed and deployed an AWS-native ML stack for multimodal campaign forecasting (FFM, Wide & Deep, Two-Tower, MMM) using millions of campaign logs resulting in 6-30% ROAS uplift across customers in controlled A/B-tests
- Led R&D roadmap and built a brand-compliant ad creative generation engine, reducing time-to-launch from weeks to minutes and enabling rapid creative iteration and ad-fatigue mitigation
- Applied model compression and quantization to reduce inference latency, cutting infra-cost by 40%
- Mentored and led a cross-functional ML team (MLOps, Data Eng, AI/DS), translating high-level R&D goals into executable sprints and shipping production features on predictable timelines
- Built autoscaling ETL pipeline on AWS (S3, SQS, EC2, Lambda) processing 1k+ campaign logs per customer per day, ensuring high-integrity data collection, transformation and storage for downstream models
- Authored SR&ED proposals securing $150k non-dilutive Canadian government startups R&D funding

Founding AI/ML Engineer (Advisory / Part-Time) | HIPPOC (Startup) **| France | May 2020 - Oct. 2022**

- Co-led the end-to-end design of the initial product's full-stack infrastructure and defined early-stage R&D roadmap
- Developed and productionized core ML features (text/object detection, website/image/video analysis) and neuroscience-inspired models (memorability, attention), deployed across backend services and a browser extension
- Optimized runtime by ~50% and memory use by ~20% through surgical model ablations and compression

R&D Software Engineer | CRITEO (AdTech) **| France | Sept. 2021 - March 2022**

- Conducted high-volume campaign analysis to improve CTR predictors, A/B testing and feedback loop calibration
- Identified and resolved critical eCPM misestimations, improving bidding accuracy and campaign return modeling

Spaceflight Dynamics Engineer | THALES (Defense & Space) **| France | *April 2019 - Sept. 2021***

- Ported cutting-edge ML algorithms into production code for satellite image processing (+10% throughput)
- Designed high-precision numerical methods to control error propagation in proprietary automatic differential algebra software, reducing forecasting error by ~35% on benchmarks
- Initiated and developed a semi-Lagrangian debris prediction solver adopted by the French Space Agency

SELECTED RESEARCH CONTRIBUTIONS

Continuous Learning via Liquid Memory Framework| 2025 (ongoing)

- Designing end-to-end (perception, cognition, action) learning framework extending the core principles of Neural Turing Machines and Memory Augmented Neural Networks (MANNs)
- Experimenting with compact modular architecture: MiniLM-L6-v2 (encoder), a VQ-like memory module, GPT-2 (decoder) and Hierarchical Recurrent Memory (inspired by HRM paper) and the AGNews dataset

Efficient Music Source Separation | 2025 (ongoing)

- Developed compact source separation network leveraging hybrid band-split and variational autoencoding
- Achieved competitive SDR performance with 94% fewer parameters than current compact SOTA further highlighting feasibility for efficient edge/low-resource deployment

Loss State-Rate ODE framework for Gradient Flow | 2025

- Proposed a continuous-time ODE for learning dynamics, isolating decay rate, topological and optimizers effects for deeper insights over the contribution of architectural choices and training stability analysis
- Numerically validated a phase transition for CCE loss near loss = log(2) responsible for brake/acceleration effects
- Repository: https://github.com/wkzng/loss-state-rate

iSincNet: Invertible Lightweight Audio Representation | 2024 - 2025

- Designed a reversible, SincNet-inspired transform allowing exact reconstruction from spectral features
- Demonstrated controllable bin/framerate and differentiable inversion in both linear and mel scales
- Repository (model, training code, weights): github.com/wkzng/iSincNet

PhaseFlow: Conservative Transport Equation Simulation Engine | 2020

- Reduced a brute force *n*-particle simulation complexity from O(n^2) to O(n) using a hierarchical spatial graph
- Achieved near C/C++ performance via targeted Cython-optimized kernels
- Repository: https://gitlab.com/wzanga/phaseflow

FORMAL RESEARCH & PUBLICATIONS

- Astrocyte-Inspired Hierarchical Routing for Enhanced Expert Specialization in Mixture-of-Exports Models
- Irreducible Loss Floors in Gradient Descent Convergence — Preprint (submitted to NeurIPS 2025)
- Stable Numerical Laplace Transform Inversion without Over-and-Undershoot — Imperial College 2018 [Report]
- Strategies for space rendezvous on Lunar Distant Retrograde Orbits — ISAE-SUPAERO 2018 [Report]

EDUCATION

MSc. Computer Science — ***Imperial College London*** **| England| 2017 - 2018**
Distinguished Master's Thesis — Ranked top 3% of the Department of Computing

BSc. & MSc. Aerospace Engineering — ***ISAE–SUPAERO*** **| France | 2015 - 2019**
Honors include Excellence Scholarship, Amicale ISAE Scholarship, Award for Social Engagement