Gabin Maxime Nguegnang

AI Research Scientist · AI/ML Engineer

LLM Fine-Tuning and Evaluation · Deep Learning Optimization · MLOps

Freising, Bavaria – 85356, Germany | +49-176-875-103-53 | gmnguegnang@gmail.com

LinkedIn: linkedin.com/in/gabin-maxime-nguegnang-06b807162 | GitHub: github.com/mnguegnang

What defines me

AI research scientist and machine learning engineer. Builds AI systems that reach real users, then proves they work with measured numbers. The current project is an AI assistant that takes a business question, splits it into steps, and picks the right tool for each step. It checks its own answer before a person sees it. An earlier project turned a fuel study into software that field technicians ran themselves, securing 84,617 liters of fuel and cutting reporting time from days to seconds. Springer Nature published the research side: how to train deep learning models faster and more reliably. Owns the problem from the first data pull through to the running service. Nothing goes live until it clears a test suite built for it.

Selected achievements

- **An AI Assistant That Plans the Work, Picks the Right Tool, and Checks Itself:** Built an agentic AI system for supply-chain decisions: the language model reads the question, and 7 mathematical solvers run the calculation, so verifiable code produces the answer (LangGraph, MCP, OR-Tools). Deployed on Azure with human sign-off on any decision above $10,000, 113 automated tests, and a security review of 58 attack attempts, 0 of which broke through.
- **Trained a Language Model to Get Finance Answers Right, 49.9% to 74.8%:** Fine-tuned Llama 3 8B on 6,000 finance questions with LoRA, lifting answer accuracy from 49.9% to 74.8%. Trained it on 1,800 comparison pairs to match which answers people prefer, reaching 98.1% agreement with the preferred answer (RLAIF and DPO). A separate model graded the results to keep the scoring honest, and MLflow tracked every run.
- **Secured 84,617 Liters of Fuel and Reduced Reporting Time from Days to Seconds:** Traced the cause of fuel loss at telecom base stations and compared 4 machine learning models (Gradient Boosting led at 98% Nash efficiency). Shipped the winning model as a web application with a live dashboard that the technicians ran themselves. Later built a document question-answering service that cites its sources for every claim, released through automated testing and deployment (FastAPI, Docker, GitHub Actions).
- **Published Research on Making Deep Learning Training Faster and More Reliable:** Proved when deep neural network training is guaranteed to reach a solution and how large a step it can safely take, a result that holds as networks get deeper. Springer Nature published the work in 2024 after peer review, confirmed by large-scale experiments on a supercomputing cluster. 5 papers and 98 citations to date.

What i'm looking for

Seeking a role as an AI Research Scientist, Applied Scientist, Machine Learning Engineer, or Generative AI Engineer. The best fit is a team that needs someone to carry an idea from a pain problem (or research) through to a running product and stand behind the numbers it reports. Based near Munich, Germany, and open to on-site, hybrid, or remote work, full-time.

Education

Sep. 2019 – Sep. 2025 **PhD in Applied Mathematics: Deep Neural Networks Optimization**
LMU Munich and RWTH Aachen University, Germany

Aug. 2017 – Feb. 2019 **Co-operative MSc in Industrial Mathematics: Machine Learning**
African Institute for Mathematical Sciences (AIMS), Limbe, Cameroon

Oct. 2016 – Nov. 2018 **MSc in Applied Mathematics: Dynamical Systems and Modeling**
University of Yaoundé I, Cameroon

Oct. 2014 – Jul. 2016 **BSc Honours in Applied Mathematics**
University of Douala, Cameroon

Professional Experience

Oct. 2025 – Present **AI Research Consultant: LLM Reasoning and Agentic AI Evaluation**
Outlier AI and Mercor, Remote, Germany

- Identify where a language model's step-by-step reasoning goes wrong on hard STEM problems and write the correction, which becomes the training signal (RLHF) that makes the next version of the model more accurate.
- Grade AI agents on whether they pick the right tool and stay on track across long tasks (ReAct-style reasoning, function calling, long-context management), and settle the ambiguous cases against a written rubric.
- Design the scoring rubrics and lead peer review to keep prompt engineering consistent across a distributed group, turning hard mathematical content into training data the team can trust.

Apr. 2024 – Sep. 2025 **Doctoral Research and Teaching Assistant**
Ludwig Maximilian University of Munich, Germany

- Established the conditions under which stochastic gradient descent (SGD), the method behind almost all model training, is guaranteed to converge, showing engineers how aggressively they can train before a run falls apart.
- Delivered an explainable-AI (XAI) project that flags failures early in industrial power plants and gives the reason behind each alert. Reached F1 0.99 with 6 ensemble learners and SHAP, and held fairness across sites (Disparate Impact Ratio 0.95).
- Taught postgraduate courses in optimization, machine learning, deep learning and data science, and defended a disputed proof step before the Springer Nature editorial board through to acceptance.

Sep. 2019 – Mar. 2024 **Doctoral Research and Teaching Assistant**
RWTH Aachen University, Germany

- Derived practical learning-rate conditions for gradient descent on deep networks that do not shrink exponentially as the network gets deeper, removing a limit that kept the earlier theory out of practical use.
- Backed the theory with large-scale PyTorch training runs on high-performance computing infrastructure, spreading the workload across nodes to cut the runtime of each experiment.
- Taught continuous optimization, mathematics of data science and higher mathematics, and built course materials that made advanced theory concrete for engineering students, within a research group of 9 nationalities.

Jun. 2018 – May 2019 **Machine Learning Intern**
Group One Holding Company, Limbe, Cameroon

- Analyzed telecom fuel-consumption data to pinpoint the root cause of fuel loss and compared 4 machine learning models (Gradient Boosting led at 98% Nash efficiency). Deployed the winning model as a Flask web application with a monitoring dashboard, securing 84,617 liters of fuel.
- Automated log ingestion, cutting reporting time from days to seconds, and built a monitoring dashboard that tracked API success rates and processing latency.
- Mapped operational constraints with base-station technicians, then presented the findings, the model results, and the case for rollout to company managers and the operations director.

Projects

Agentic ERP Supply-Chain Copilot | *LangGraph, MCP, OR-Tools, CVXPY, CRAG, DSPy, AKS*

- Split the work between the language model and code: a LangGraph orchestrator sorts each question into 1 of 10 types, the model handles the language, and 7 OR/CVXPY/SciPy solvers run the math behind typed Pydantic contracts. Code with provable guarantees executes the plan the model writes.
- Built the safety net: a live 100-question test on the real classifier (90% accuracy gate), a retrieval pipeline (BGE-large, pgvector and BM25) that rewrites its own query when the documents returned are off-target (Recall@5 above 0.80), and a human sign-off on decisions above $10,000.

LLM Alignment Pipeline | *PyTorch, Hugging Face, LoRA, QLoRA, DPO, MLflow*

- Fine-tuned Llama 3 8B on 6,000 FinQA examples via LoRA, raising token accuracy from 49.9% to 74.8%. Labeled 1,800 preference pairs with Zephyr 7B-β (RLAIF), then applied DPO to push reward accuracy to 98.1%. Graded with Prometheus 2, kept separate from Zephyr so no model marked its own work, tracked every run in MLflow, and published the adapter to Hugging Face (`G-Maxime-N/llama3-8b-finqa-dpo`).

Production RAG Chatbot | *FastAPI, Chainlit, FAISS, LangChain, Docker, GitHub Actions*

- Built an async FastAPI service that streams answers token by token, remembers the last 12 turns of a conversation, and reports its own health. Fixed a thread-safety fault in an asyncio generator and removed a 230 MB cold-start delay behind Docker health checks.
- Automated releases through GitHub Actions to GHCR so every deployed image traces back to the commit that produced it, and reached 0.8373 citation precision (ALCE) on questions spanning multiple scientific documents.

Selected Publications

- Nguegnang, G.M., Rauhut, H., and Terstiege, U. *Convergence of gradient descent for learning linear neural networks.* Advances in Continuous and Discrete Models, Springer Nature, 2024.
- Mulongo, G., Atemkeng, M., Ansah-Narh, T., Rockefeller, R., Nguegnang, G.M., et al. *Anomaly detection in power generation plants using machine learning and neural networks.* Applied Artificial Intelligence, Taylor & Francis, 2020.

Selected Certifications

Advanced Agent Coding, Outlier AI (Jan. 2026)

Advanced Prompt Engineering, Outlier AI (Dec. 2025)

Generative AI with Large Language Models, DeepLearning.AI / Coursera (Oct. 2024)

Machine Learning Engineering for Production (MLOps) Specialization, DeepLearning.AI (May 2024)

Technical Skills

Programming & Data: Python (modular, pandas, NumPy, SciPy, scikit-learn), PyTorch, SQL and vector databases (PostgreSQL, pgvector, FAISS), FastAPI, Uvicorn, Pydantic, Flask, REST APIs, Bash
Agentic AI, LLMs & RAG: AI Agents, Agentic Systems, Multi-Agent Orchestration, Chain-of-Thought, ReAct, Long-Context Management, Tool-Augmented Generation, LangGraph, LangChain, MCP, DSPy, RAG, CRAG, Prompt Engineering, Hugging Face (Transformers, TRL, PEFT), Fine-tuning (LoRA, QLoRA), SFT, RLAIF, DPO, RLHF
Evaluation & Experiment Tracking: Evaluation Framework Design, Live Evals, Failure-Mode Analysis, Root-Cause Analysis, Red-Teaming, Rubric-Based Evaluation, Preference Labeling, Quality Gates, Recall@k, ALCE, RAGAS, Prometheus 2, MLflow, Hallucination Mitigation
Machine Learning & Research: Deep Learning, Statistics, Non-convex Optimization, Convergence Analysis, Bayesian Optimization, Experimental Design, Gradient Boosting, Random Forest, SVM, Logistic Regression, Lasso, Anomaly Detection, Explainable AI (SHAP), Predictive Modeling, Exploratory Data Analysis (EDA), NLP, Operations Research Solvers
MLOps, DevOps & Cloud: Docker, GitHub Actions (CI/CD), GHCR, Kubernetes, OpenTelemetry, LangSmith, Git, pytest, Distributed Training, HPC, AWS, Google Cloud Vertex AI, Azure Kubernetes Service (AKS), Agile (Scrum)

Memberships & Awards

Munich Center for Machine Learning (MCML), Member (Jun. 2024 – Present)

Community of Computational and Mathematical Methods in Data Science, Member (Dec. 2019 – Present)

Mastercard Foundation Scholars Program Award (Aug. 2017): Recognition of academic excellence, commitment to community impact, and dedication to African leadership development.

Languages

French: Mother tongue (C2) | **English**: Proficient (C1) | **German**: Professional working proficiency (B2+)