

Gabin Maxime Nguegnang

KI-/ML-Ingenieur · Agentic AI

RAG · LLM-Alignment · API-Integration · Deep Learning · MLOps

Freising, Bayern – 85356, Deutschland | +49-176-875-103-53 | gmaxime@nguegnang-ai.com

Website: nguegnang-ai.com | linkedin.com/in/maxime-nguegnang | github.com/mnguegnang

PROFIL

KI-Forscher und Machine-Learning-Ingenieur. Baut KI-Systeme, die echte Nutzer erreichen, und belegt ihre Wirkung mit gemessenen Zahlen. Das aktuelle Projekt ist ein KI-Assistent, der eine Geschäftsfrage in Schritte zerlegt und für jeden Schritt das passende Werkzeug wählt. Er prüft seine eigene Antwort, bevor ein Mensch sie sieht. Ein früheres Projekt machte aus einer Kraftstoffstudie eine Software, die Techniker vor Ort selbst bedienen: 84.617 Liter Kraftstoff gesichert, Berichtszeit von Tagen auf Sekunden gesenkt. Springer Nature veröffentlichte die Forschungsseite dazu: wie sich Deep-Learning-Modelle schneller und verlässlicher trainieren lassen. Verantwortet den Weg vom ersten Datenabzug bis zum laufenden Dienst. Nichts geht live, bevor es die eigens dafür gebaute Testsuite besteht.

AUSGEWÄHLTE ERFOLGE

- **KI-Assistent, der die Arbeit plant, das passende Werkzeug wählt und sich selbst prüft:** Entwickelte ein Agentic-AI-System für Supply-Chain-Entscheidungen: das Sprachmodell liest die Frage, 7 mathematische Solver führen die Berechnung aus, sodass überprüfbarer Code die Antwort erzeugt (LangGraph, MCP, OR-Tools). Deployte den Stack auf Azure, mit menschlicher Freigabe ab 10.000 USD, 113 automatisierten Tests und einer Sicherheitsprüfung mit 58 Angriffsversuchen, von denen 0 durchkamen.
- **Sprachmodell auf Finanzfragen trainiert, von 49,9 % auf 74,8 %:** Fine-tunte Llama 3 8B mit LoRA auf 6.000 Finanzfragen und hob die Antwortgenauigkeit von 49,9 % auf 74,8 %. Trainierte das Modell auf 1.800 Präferenzpaaren (RLAIF und DPO) bis zu 98,1 % Übereinstimmung mit der bevorzugten Antwort. Ein separates Modell bewertete die Ergebnisse, damit kein Modell die eigene Arbeit benotet, und MLflow protokollierte jeden Lauf.
- **84.617 Liter Kraftstoff gesichert, Berichtszeit von Tagen auf Sekunden gesenkt:** Ermittelte die Ursache der Kraftstoffverluste an Telekommunikations-Basisstationen und verglich 4 Machine-Learning-Modelle (bestes Modell Gradient Boosting, 98 % Nash-Effizienz). Lieferte das beste Modell als Webanwendung mit Live-Dashboard aus, das die Techniker selbst bedienen. Baute später einen Dienst für Dokumentenfragen, der jede Aussage mit Quellen belegt, ausgeliefert über automatisierte Tests und Deployment (FastAPI, Docker, GitHub Actions).
- **Publizierte Forschung zu schnellerem und verlässlicherem Deep-Learning-Training:** Bewies, wann das Training tiefer neuronaler Netze garantiert konvergiert und wie groß die Schrittweite dabei ausfallen darf, ein Ergebnis, das auch mit wachsender Netztiefe gilt. Springer Nature veröffentlichte die Arbeit 2024 nach Peer Review, bestätigt durch groß angelegte Experimente auf einem HPC-Cluster. 5 Publikationen und 98 Zitationen.

ANGESTREBTE POSITION

Angestrebt wird eine Position als AI Engineer, Machine Learning Engineer, Generative AI Engineer oder Data Scientist. Am besten passt ein Team, das eine Idee von einem echten Problem (oder aus der Forschung) bis zum laufenden Produkt tragen soll und für die berichteten Zahlen einsteht. Wohnort im Raum München, verfügbar vor Ort, hybrid oder remote, in Vollzeit.

AUSBILDUNG

- Sep. 2019 – Sep. 2025* **Promotion in angewandter Mathematik: Optimierung neuronaler Netze**
LMU München und RWTH Aachen, Deutschland
- Aug. 2017 – Feb. 2019* **Kooperativer M.Sc. in Industriemathematik: Machine Learning**
African Institute for Mathematical Sciences (AIMS), Limbe, Kamerun
- Okt. 2016 – Nov. 2018* **M.Sc. in angewandter Mathematik: Dynamische Systeme, Modellierung**
Universität Yaunde I, Kamerun
- Okt. 2014 – Jul. 2016* **B.Sc. mit Auszeichnung in angewandter Mathematik**
Universität Duala, Kamerun

BERUFSERFAHRUNG

- Okt. 2025 – heute* **AI Research Consultant: LLM-Reasoning und Evaluierung agentischer KI**
Outlier AI und Mercor, remote, Deutschland
- Analysiere, an welcher Stelle das schrittweise Reasoning eines Sprachmodells bei schweren MINT-Aufgaben scheitert, und schreibe die Korrektur, die als Trainingssignal (RLHF) die nächste Modellversion genauer macht.
 - Bewerte KI-Agenten daraufhin, ob sie das richtige Werkzeug wählen und über lange Aufgaben in der Spur bleiben (ReAct-Reasoning, Function Calling, Long-Context-Management), und entscheide strittige Fälle anhand einer schriftlichen Rubrik.
 - Entwerfe die Bewertungsrubriken und leite das Peer Review, damit Prompt Engineering in einem verteilten Team einheitlich bleibt, und überführe schwierige mathematische Inhalte in belastbare Trainingsdaten.
- Apr. 2024 – Sep. 2025* **Wissenschaftlicher Mitarbeiter in Forschung und Lehre**
Ludwig-Maximilians-Universität München, Deutschland
- Bestimmte die Bedingungen, unter denen der stochastische Gradientenabstieg (SGD) garantiert konvergiert, und zeigte damit, wie aggressiv sich ein Training fahren lässt.
 - Lieferte ein Projekt zur erklärbaren KI (XAI), das Ausfälle in Industriekraftwerken früh meldet und den Grund für jede Warnung nennt. Erreichte F1 0,99 mit 6 Ensemble-Verfahren und SHAP und hielt die Fairness über alle Standorte (Disparate Impact Ratio 0,95).
 - Unterrichtete Optimierung, Machine Learning und Data Science im Masterstudium und verteidigte einen strittigen Beweisschritt vor der Redaktion von Springer Nature bis zur Annahme.
- Sep. 2019 – Mär. 2024* **Wissenschaftlicher Mitarbeiter in Forschung und Lehre**
RWTH Aachen, Deutschland
- Leitete praxistaugliche Bedingungen an die Lernrate für den Gradientenabstieg in tiefen Netzen her, die nicht exponentiell mit der Netztiefe abnehmen, und hob damit eine Schranke auf, die die frühere Theorie für die Praxis unbrauchbar machte.
 - Untermauerte die Theorie mit groß angelegten PyTorch-Trainingsläufen auf HPC-Infrastruktur und verteilte die Last über mehrere Knoten, um die Laufzeit jedes Experiments zu senken.
 - Unterrichtete kontinuierliche Optimierung und Mathematik der Data Science und erstellte Lehrmaterial, das fortgeschrittene Theorie für Ingenieurstudierende greifbar machte.
- Jun. 2018 – Mai 2019* **Praktikant Machine Learning**
Group One Holding Company, Limbe, Kamerun
- Analysierte Kraftstoffdaten von Telekommunikations-Basisstationen zur Bestimmung der Verlustursache und verglich 4 Machine-Learning-Modelle (bestes Modell Gradient Boosting, 98 % Nash-Effizienz). Deployte das beste Modell als Flask-Webanwendung mit Monitoring-Dashboard und sicherte 84.617 Liter Kraftstoff.
 - Automatisierte die Protokollverarbeitung, senkte die Berichtszeit von Tagen auf Sekunden und stellte Ergebnisse und Rollout-Vorschlag der Geschäftsleitung vor.

PROJEKTE

- Agentic-AI-Copilot für ERP-Lieferketten** | *LangGraph, MCP, OR-Tools, CVXPY, CRAG, DSPy, AKS*
- Teilte die Arbeit zwischen Sprachmodell und Code auf: ein LangGraph-Orchestrator ordnet jede Frage 1 von 10 Intents zu, das Modell übernimmt die Sprache, und 7 Solver (OR-Tools, CVXPY, SciPy) rechnen hinter typisierten Pydantic-Verträgen. Code mit beweisbaren Garantien führt den Plan aus, den das Modell schreibt.
 - Baute das Sicherheitsnetz: einen Live-Test mit 100 Fragen am echten Klassifikator (Schwelle 90 % Genauigkeit), eine Retrieval-Pipeline (BGE-large, pgvector und BM25), die ihre Anfrage neu formuliert, wenn die Treffer am Thema vorbeigehen (Recall@5 über 0,80), und eine menschliche Freigabe ab 10.000 USD.
- Pipeline für LLM-Alignment** | *PyTorch, Hugging Face, LoRA, QLoRA, DPO, MLflow*

- Fine-tunte Llama 3 8B mit LoRA auf 6.000 FinQA-Beispielen und steigerte die Token-Genauigkeit von 49,9 % auf 74,8 %. Labelte 1.800 Präferenzpaare mit Zephyr 7B- β (RLAIF) und hob per DPO die Reward-Genauigkeit auf 98,1 %. Bewertete mit Prometheus 2, getrennt von Zephyr, sodass kein Modell die eigene Arbeit benotet, protokollierte jeden Lauf in MLflow und veröffentlichte den Adapter auf Hugging Face (G-Maxime-N/llama3-8b-finqa-dpo).

Multi-Agenten-Pipeline für F&E (3 eigene Agenten) | Claude-Code-Subagenten, MCP, Pydantic V2, bandit

- Entwarf und lieferte 3 spezialisierte Agenten aus, die die Arbeit weiterreichen: ein Rechercheagent schreibt eine belegte Implementierungsspezifikation, ein Coding-Agent baut den modularen Code samt Tests, ein Review-Agent sortiert die Befunde nach Schwere und setzt die Korrekturen um. Jede Übergabe läuft über ein strukturiertes Dokument, sodass keine Stufe vom Kontextfenster der vorherigen abhängt.
- Band den Rechercheagenten an einen MCP-Server für akademische Recherche an (Semantic Scholar, ArXiv, Hugging Face Hub, Zitations-Snowballing), sodass jede Quelle auf ein Werkzeugergebnis zurückgeht statt auf das Modell. Sicherte Commits über ruff, mypy, pytest und bandit ab, verlangte menschliche Freigabe vor Installationen und Commits und speicherte den Sitzungsstand persistent gegen Kontextverlust.

RAG-Chatbot im Produktivbetrieb | FastAPI, Chainlit, FAISS, LangChain, Docker, GitHub Actions

- Baute einen asynchronen FastAPI-Dienst, der Antworten Token für Token streamt, die letzten 12 Gesprächsrunden behält und den eigenen Zustand meldet. Behob einen Thread-Safety-Fehler in einem asyncio-Generator und beseitigte eine Kaltstartverzögerung von 230 MB hinter Docker-Health-Checks.
- Automatisierte die Releases über GitHub Actions in die GHCR, sodass jedes ausgelieferte Image auf den erzeugenden Commit zurückführt, und erreichte 0,8373 Zitationsgenauigkeit (ALCE) bei Fragen über mehrere wissenschaftliche Dokumente hinweg.

AUSGEWÄHLTE PUBLIKATIONEN

- Nguegnang, G.M., Rauhut, H., und 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., u. a. *Anomaly detection in power generation plants using machine learning and neural networks*. Applied Artificial Intelligence, Taylor & Francis, 2020.

AUSGEWÄHLTE ZERTIFIKATE

Advanced Agent Coding, Outlier AI (Jan. 2026) | Model Parallelism: Building and Deploying Large Neural Networks, NVIDIA (Oct. 2025) | Generative AI with Large Language Models, DeepLearning.AI and AWS (Oct. 2024) | Machine Learning Engineering for Production (MLOps) Specialization, DeepLearning.AI (2024) | Structuring Machine Learning Projects, DeepLearning.AI (2020)

TECHNISCHE KENNTNISSE

Programmierung & Daten: Python (modular, pandas, NumPy, SciPy, scikit-learn), PyTorch, SQL und Vektordatenbanken (PostgreSQL, pgvector, FAISS), FastAPI, Pydantic V2, Flask, REST-Schnittstellen, Bash

Agentic AI, LLMs & RAG: KI-Agenten, agentische Systeme, Multi-Agenten-Orchestrierung, Chain-of-Thought, ReAct, Long-Context-Management, Tool-Augmented Generation, Subagenten-Orchestrierung, MCP-Server-Anbindung, Context Engineering, Human-in-the-Loop-Freigaben, LangGraph, LangChain, MCP, DSPy, RAG, CRAG, Prompt Engineering, Hugging Face (Transformers, TRL, PEFT), Fine-Tuning (LoRA, QLoRA), SFT, RLAIF, DPO, RLHF

Evaluierung & Experimentverfolgung: Entwurf von Evaluationsframeworks, Live-Evaluierung, Fehlermuster- und Ursachenanalyse, Red Teaming, rubrikbasierte Bewertung, Präferenzlabeling, Qualitätsprüfungen, Recall@k, ALCE, RAGAS, Prometheus 2, MLflow, Vermeidung von Halluzinationen

Machine Learning & Forschung: Deep Learning, Statistik, nichtkonvexe Optimierung, Konvergenzanalyse, bayessche Optimierung, Versuchsplanung, Gradient Boosting, Random Forest, SVM, Anomalieerkennung, erklärbare KI (SHAP), prädiktive Modellierung, explorative Datenanalyse (EDA), NLP, Operations-Research-Solver

MLOps, DevOps & Cloud: Docker, GitHub Actions (CI/CD), GHCR, Kubernetes, OpenTelemetry, LangSmith, Git, pytest, verteiltes Training, HPC, AWS, Google Cloud Vertex AI, Azure Kubernetes Service (AKS), agiles Arbeiten (Scrum)

MITGLIEDSCHAFTEN & AUSZEICHNUNGEN

Munich Center for Machine Learning (MCML), Mitglied (Jun. 2024 – heute)

Community of Computational and Mathematical Methods in Data Science, Mitglied (Dez. 2019 – heute)

Mastercard Foundation Scholars Program Award (Aug. 2017): Auszeichnung für akademische Leistung und gesellschaftliches Engagement.

SPRACHEN

Französisch: Muttersprache (C2) | **Englisch:** verhandlungssicher (C1) | **Deutsch:** gute Kenntnisse (B2+)