DUC-ANH NGUYEN

AI/ML Research Engineer
Munich, Germany | +49 176 7564 4076 | anh.nguyen1@campus.lmu.de
github.com/DucAnhValentinoNguyen | linkedin.com/in/duc-anh-nguyen-ml | ducanhvalentinonguyen.com

PROFILE

M.Sc. Statistics and Data Science student at LMU Munich building machine-learning systems from research through deployment. Focus areas include multimodal vision-language models, self-supervised learning, long-video understanding, LLM agents and reliable data pipelines. Combines rigorous evaluation and statistical reasoning with practical MLOps on GCP and AWS. Current work spans an LMU x ZEISS surgical-video research collaboration, archaeological LiDAR, applied forecasting, open-source ML/agent contributions and production-minded AI projects.

TECHNICAL SKILLS

Programming	Python (PyTorch, Hugging Face Transformers, PEFT/LoRA, scikit-learn, pandas, NumPy, Optuna, LightGBM), SQL, C++/CUDA, R (tidyverse, mlr3), Bash, Git, LaTeX.
LLM and agents	Tool calling, MCP, structured outputs, RAG, embeddings and vector search, prompt/context engineering, LLM/VLM fine-tuning, distillation, quantization, pruning, agent guardrails and adversarial testing of tool-using agents.
Vision and ML	Vision-language/foundation models (SigLIP2, PaliGemma2, Qwen-VL, LLaVA, InternVL3); computer vision, segmentation and long-video understanding; self-supervised learning (MAE, DINO, LeJEPA); domain adaptation (DANN); explainable AI (SHAP); time-series forecasting; anomaly detection.
Data and evaluation	ETL for image, video, point-cloud and tabular data; foundation-model annotation/auto-labeling; label quality control; versioned event contracts; medallion pipelines; DVC; drift monitoring (Evidently); evaluation harnesses for classification, VQA and video; confidence intervals, baseline benchmarking, per-class error analysis, Weights & Biases and Bayesian optimization.
Backend and MLOps	FastAPI, REST APIs (JSON, authentication, webhooks), Streamlit, Docker, Terraform, GitHub Actions CI/CD, Hydra, Prometheus, pytest, ruff, mypy; Linux and SLURM GPU clusters (H100/A100); int4 quantization and pruning.
Cloud	GCP (Vertex AI, Cloud Run, Cloud Build, Cloud Storage, BigQuery, Pub/Sub, Dataflow); AWS (S3, EC2, SageMaker).
Ways of working	AI developer tooling (Claude Code, GitHub Copilot); close work with domain experts in clinical research, retail and archaeology; structured documentation, reproducible analysis and code review.

PROFESSIONAL EXPERIENCE

Research Assistant (HiWi) - Department of Statistics, LMU Munich *01/2026 - present*

LMU x ZEISS: Multimodal vision-language models for surgical video

- Fine-tuned SigLIP2 into a surgical-domain vision encoder for endoscopy VLMs. On 16-class HyperKvasir frame classification, exceeded evaluated zero-shot baselines (MedSigLIP, Qwen3-VL-8B and LLaVA-OneVision-7B) by more than 10 percentage points in balanced accuracy; scaled the pipeline to videos up to 120 minutes.
- Built an ImageNet-initialized ViT-B/16 self-supervised pretraining benchmark comparing DINO v1 with LeJEPA, then evaluated frozen representations on five HyperKvasir tasks (four classification tasks and polyp segmentation). Applied a pre-registered selection rule: no regression in mean balanced accuracy or Dice against the pre-SSL baseline. DINO/ImageNet passed cleanly across all five tasks with no regression in accuracy, balanced accuracy, AUROC or Dice; LeJEPA/ImageNet regressed.
- Designed a checkpoint-resumable streaming pipeline for the raw GastroNet-5M corpus: 4,823,974 images per epoch across all 506 portal archives, 3 epochs and 113,064 optimizer steps per setting. Used deterministic shuffling, bounded cache/prefetch, parallel decode and augmentation, and deterministic blank-sample fallback to preserve exact step counts across SLURM restarts without staging the corpus locally; trained on LRZ H100/A100 GPUs.
- Audited all 10,619 local HyperKvasir evaluation images and a 226,256-image sample from 24 GastroNet zip archives. Used perceptual-hash checks to remove 727 calibration/query frames near-duplicate to reference images and reduce evaluation leakage; documented sampling limitations and corpus integrity findings.

Full-corpus benchmark and data-quality results

VLF-ZEISS project

Self-supervised segmentation of archaeological LiDAR point clouds

- Built an autoencoder pipeline to segment aerial LiDAR spanning 10+ hectares and flag anomalies for targeted expert review, reducing reliance on full-site manual inspection. Second author on a paper in preparation (ISPRS Journal; not yet submitted).

Store sales forecasting - NKD, German retail chain

- Developed daily store-level demand forecasts from 1.5 years of sales and store attributes; LightGBM models tuned with Optuna reached 24% WAPE at a 14-day horizon and were presented to the retail partner.

Wildboar compensation and tolerance - statistical consulting

- Produced a reproducible analysis of a field survey in two Chinese counties, covering compensation applications, agricultural losses, policy support and tolerance. Applied EDA, EFA/CFA/PCA, ordinal models, Bayesian checks, PLS-SEM, mediation, prediction and robustness/stress tests; documented limits including county-period confounding, measurement overlap, structural missingness and unmeasured confounding.

Reproducible analysis repository

IT Team Member - Department of Statistics, LMU Munich *11/2025 - present*

- Administer department access rights and GPU-cluster use with Linux, CUDA, SLURM and shell scripting.
- Built and deployed a GUI for real-time cluster status and job submission, saving about four hours per week of manual checks.

PUBLICATIONS

Test-Time Domain Generalization for Endoscopy via Optimal-Transport Steering *First author*

In preparation; target CVPR 2027.

Self-Supervised Learning for Semantic Segmentation of Archaeological LiDAR Point Cloud Data *Second author*

In preparation for the ISPRS Journal; not yet submitted; manuscript preparation underway.

OPEN-SOURCE CONTRIBUTIONS

gnhf - token-budget correctness fix *08/2026*

Merged upstream in the 3.8k-star autonomous coding-agent orchestrator; shipped in v0.1.46.

- Found a 183x undercount in the --max-tokens safety cap: provider adapters reported four token buckets while the orchestrator stored two, silently dropping cache reads and writes. A reported run appeared to use 22% of a 1,000,000-token budget after billing 41.1M tokens and $48.64.
- Normalized adapter accounting while keeping provider buckets mutually exclusive, and propagated corrected totals through budget enforcement, rendering, exit summaries, telemetry and CLI help. Added regression tests and confirmed they fail without the fix; 760 tests and 6 CI checks passed through three maintainer review rounds.

Merged pull request #213

MONAI - per-class rank in EmbeddingCollapseMetric aggregate *09/2026*

- Fixed aggregate reduction that could omit a class's collapse indicator when its per-class rank was invalid, hiding a collapsed minority class in the summary score. Added regression tests (verified failing without the fix) and clarified docstrings in response to maintainer feedback; pull request merged on 16 Oct 2026.

Merged pull request #9117

HACKATHONS AND CHALLENGES

TraceForge - agentic PCB design, European Hackathon League Munich *08/2026*

- In a five-person team and 20-hour build, developed an agent that converts plain-English circuit descriptions into manufacturable boards, validates electrical/design rules against live component stock, repairs errors and exports Gerbers. Owned the Mouser availability integration (lifecycle, stock depth and minimum order quantity) and web-scraping sub-agents.

TraceForge repository

Cloud Innovation Challenge - LMU x AWS, Velocity Racing case study *2026*

- Designed and presented a cloud architecture for a fictional Formula 1 team across five areas: generative-AI race strategy, sovereign cloud/data security, fan engagement, global broadcasting and real-time telemetry. Delivered as a scored case-study report and interactive demo using synthetic data; an architecture proposal, not a production system.

Case-study report

Interactive demo

Amazon SD Engineering Hackathon - registered participant *2026*

EMEA Early Career Talent Hackathon; registered, with no project outcome claimed.

Hackathon page

TEACHING EXPERIENCE

Teaching Assistant - 2 courses: Logic & Discrete Structures + Calculus for CS and statisticians, LMU Munich *04/2023 - 09/2025*

- Led exercise sessions and graded exams for two undergraduate courses of 100+ students each; routine practice translating abstract technical material for non-expert audiences.

PROJECTS

AI agents and data pipeline applications

Service Desk AI - bounded multi-agent operations desk *08/2026*

- Built a full-stack service desk where an AI router retrieves cited policy evidence and delegates to bounded ERP, CRM, HR and scheduling specialists. Trusted Python validates model output, selects permitted tools and requires human approval for protected writes.
- The model never sees raw records, selects arbitrary tools or executes mutations; unavailable, malformed, unsafe or low-confidence outputs route to human review. FastAPI, Next.js, SQLite, Docker Compose, GitHub Actions and Terraform/AWS scaffold.

Source code

Live demo

Service Desk AI data platform *2026*

- Implemented bronze/silver/quarantine/gold processing behind versioned JSON Schema contracts, runnable locally or through Apache Beam/Dataflow on GCP. Invalid records are quarantined with a reason rather than silently dropped.

Source code

SciPaLI - MLOps pipeline for multimodal reasoning *04/2026*

- Two-person course project fine-tuning PaliGemma2-3B with LoRA/PEFT on ScienceQA: 72.2% exact match; systematic prompt-order ablation yielded +16 points, and int4 quantization reduced GPU memory by 51%. Built DVC versioning, Hydra configuration, cloud training, FastAPI/Cloud Run serving with hot-swappable adapters, drift monitoring and CI/CD.

Coursework repository; not independently publishable.

AI Agent Security - multi-step tool attacks *2026*

- Built Python harnesses and prompt suites for reproducible adversarial multi-step tool-use tests with GPT-OSS and Gemma. Ranked 461 of 4,187 teams on the public Kaggle leaderboard.

Competition leaderboard

ML/AI and statistical research

Bachelor thesis - decision theory for Bayesian optimization *01/2025 - 04/2025*

- Designed a decision-theoretic acquisition policy for Bayesian optimization and benchmarked it on standard test functions.

Thesis code

Additional projects

EdgeLoop - ML from training set to device firmware *2026*

- Built an end-to-end edge-ML path from synthetic building telemetry to a fixed-seed trainer, checksummed JSON artifact, generated C firmware constants and dependency-free device inference, plus audited fleet telemetry/deployment. Implemented matching inference contracts in Python and C; ingestion is idempotent.

Source code

Other competitions, projects and applied analysis *2026*

Kaggriculture (rank 1,162 of 10,148); RSNA Knee Abnormality Detection (rank 1,467 of 6,299); NVIDIA Nemotron Model Reasoning Challenge; BirdCLEF+26 (03/2026); neetcode-gpt; NN-from-scratch; and InstaSHAP benchmarks. Social-data-science PLS-SEM analysis of a wildlife-compensation policy in Giant Panda National Park (R-squared = 0.30), supporting a partner publication.

Kaggriculture
RSNA Knee Abnormality Detection
NVIDIA Nemotron Model Reasoning
BirdCLEF+26
neetcode-gpt
NN-from-scratch
InstaSHAP benchmarks

EDUCATION

M.Sc. Statistics & Data Science - LMU Munich *05/2025 - present (expected graduation: spring 2027)*

Relevant courses: MLOps, Applied Deep Learning, Deep Learning for NLP, Research Seminar in Explainable AI, Statistical Learning, Generative AI and Visual Synthesis, Analysis of High-dimensional Biological Data, Diagnostic Accuracy Studies.

B.Sc. Statistics & Data Science (major), Computer Science (minor) - LMU Munich *10/2021 - 04/2025*

Final grade 1.8 (German scale; 1.0 is best). Relevant courses: Database Systems, Applied Machine Learning, Machine and Deep Learning with Python, Introduction to Machine Learning.

Academic service *LMU Munich*

Administer the Department of Statistics GPU cluster; two and a half years as a teaching assistant for two undergraduate courses with 100+ students each.

LANGUAGES

Languages	Vietnamese (native), German (C1), English (C1), Italian (B1), Chinese (HSK 2), Arabic (beginner), Thai (beginner).

HOBBIES AND SOCIAL ACTIVITIES

Social activities	Blood donor at the German Red Cross Munich (13 donations as of 1 October 2026). Dorm speaker at Olydorf (11/2024 - 05/2026).
Music	Guitarist and vocalist in jam sessions.
Running	Long runs; marathon personal best 4:20:07 (unofficial).
Sports	Basketball (5x5 and 3x3), softball in the university league, skateboarding, martial arts (Judo and Karate), hiking.