

Xavier Theimer-Lienhard

ML Research Engineer – [[website](#)] · [[Scholar](#)] · [[GitHub](#)] · [[HuggingFace](#)] · [[email](#)]

SUMMARY

I am a ML Research Engineer specializing in Medical LLMs. 2+ years' experience on the end-to-end LLM stack: pre/post-training on 1000+ GPU clusters, RLHF/DPO/GRPO, scalable inference, model evaluation, and orchestration. Led open-weight LLM releases, evaluation, and cross-functional teams across tech and academia.

EXPERIENCE

LiGHT Lab, EPFL & Harvard - Research Engineer

Apr 2025-present

- Finetuning of Medical LLMs, SFT; DPO; distributed trainings on up to 128 nodes.
- Clinical alignment of LLM using evaluations of medical doctors in 22+ organizations ([Moove project](#)).
- Supervised 4 Master students for their theses on reinforcement learning and reasoning.

Yale School of Medicine - Short Term Scholar & Postgraduate Associate

Sep 2024-Apr 2025

- Finetuned a Medical reasoning model: +30% on general reasoning and +9% on medical accuracy.
- Built multimodal synthetic data pipeline generating 10k+ clinical notes; direct feedback from clinicians.
- Supervised 2 Master students for their theses on synthetic data generation.

MLO Lab, EPFL - Research Assistant

Sep 2023-Sep 2024

- Created Meditree: an inference pipeline boosting accuracy +5% on medical exams.
- Meditron: a medical finetune of Llama-3 reaching 81% on medical exams, GPT-4 performance ([Meta Blog](#)).

Irbis Consulting SA - ML Engineer

Sep 2023-Sep 2024

- Built electron app; Pytorch Deeplearning captcha solver; webscraper.

EDUCATION

- **Master, MSc, Data Science, (EPFL)** - Sep 2022 - Apr 2025
- **Bachelor, BSc, Computer Science, (EPFL)** - Sep 2018 - Sep 2022

PUBLICATIONS & TALKS

- [Llama-3-Meditron: An Open-Weight Suite of Medical LLMs](#), AAAI 2025 (GenAI4Health)
- [GPoT: A language Model for Rhyme Generation on Synthetic Data](#), ACL 2024
- **Invited Speaker**, [Meta Open-Source AI Summit 2024](#).

SKILLS

ML: PyTorch, RL (DPO/GRPO), Transformers, Finetuning, Quantization, Prompt Engineering, Experimental Design.

Engineering: HPC/Slurm, Distributed Training (FSDP/DeepSpeed), Docker/Containers, Linux, Git, vLLM/Inference Optimization.

Languages: Python, C++, Java, SQL, French, English