Leon Chlon

leochlon@gmail.com | linkedin.com/in/leochlon | github.com/leochlon

EDUCATION

Massachusetts Institute of Technology - Postdoctoral Researcher (Machine Learning) **2018**
University of Cambridge - Ph.D. (Machine Learning) 2017
University of Cambridge - M.Phil. (Theoretical Physics) 2014

SELECTED PUBLICATIONS & PREPRINTS

- **Predictable Compression Failures: Why Language Models Actually Hallucinate** (arXiv:2509.11208) — International Conference on Machine Learning (ICML 2026), accepted.
- **LLMs are Bayesian, in Expectation, not in Realization** (arXiv:2507.11768) — Annual Conference on Neural Information Processing Systems (NeurIPS 2026), in submission.
- **Attention Deficits in Language Models: Causal Explanations for Procedural Hallucinations** (arXiv:2602.19239) — Annual Conference on Neural Information Processing Systems (NeurIPS 2026), in submission.

EXPERIENCE

University of Oxford **March 2025 – Present**
Visiting Fellow

- Won **$300K** in backing from **Microsoft, Google, and NVIDIA** to fund an AI research sabbatical at **Oxford's Torr Vision Group**.
- Built and shipped **HallBayes**, a model-agnostic post-training system for suppressing LLM hallucinations; reached **1.6k GitHub stars** and **150+ forks**, integrated into **PyTorch Geometric**, featured at **NVIDIA GTC 2026**.
- Co-designed CUDA kernels for **INT4 compression** with **Prof. Omri Weinstein** (Hebrew University of Jerusalem) and **Unsloth**, cutting quantization penalty by **~9%** in **GPTQ/GGUF vLLMs** including Qwen3.5.
- Sponsored by **HP** & **NVIDIA** to extend fused **Triton** transformer kernels to 2025 **Blackwell-generation GB10 GPUs** (**sm_121**), achieving **6.1×** acceleration vs PyTorch baselines.
- Built and released **Mezzanine**, a post-training toolkit adopted by teams at **DeepMind, CERN, DAMTP, and CRUKCI**; powered the first **pixel-free I-JEPA** for robotics and improved mean retrieval rank by **2.1×** (**114.5 → 54.3**) on **LeRobot ALOHA**.

Selected Industry Engagements — Apple, TikTok, World Bank Group **August 2023 – March 2025**
Applied ML systems across ranking, forecasting, and large-scale inference

- **Apple** — Content Ranking: Architected a production MMoE ranking system in TensorFlow using ONNX Runtime, Redis feature stores, and GPU-accelerated batch inference, reducing p50 latency by 43% for workloads serving 50M+ requests/day.
- **World Bank Group** — Post-training: Fine-tuned LLaMA-2 70B LoRa adapters across 300k socioeconomic indicators, reducing forecast error by 35% sMAPE with immediate adoption across 15 cross-country teams.
- **TikTok** — Post-training: Bayesian-optimised retraining of MMoE content ranking model across a 500M+ DAU ranking system using Ax, BoTorch, MLFlow & Ray, yielding double-digit gains in cross-platform sharing.

Tailor Bio (University of Cambridge Spinout)
0 -> 1 Founding AI Engineer **December 2022 – August 2023**

- Engineered a scalable drug discovery stack on AWS EC2 (p4d instances) using PyTorch, RDKit, and Neo4j; developed custom sparse matrix multiplication routines and gradient checkpointing strategies to train billion-edge graphs on limited startup GPU resources, achieving 3x throughput gains that accelerated the validation of 3 lead compounds for Series B.

Uber Technologies (Careem)
Lead ML Engineer – Dynamic Pricing **January 2022 – August 2022**

Leon Chlon
leochlon@gmail.com | linkedin.com/in/leochlon | github.com/leochlon

- Redesigned marketplace pricing for 100M+ users using algorithmic game theory, achieved 2.7x efficiency improvement over legacy system and $2M+ revenue impact.
- Deployed via Kubernetes stacked with Istio service mesh, enabling shadow testing with production via canary rollouts tracked via MLflow.

Meta
Senior Research Scientist – (AI Safety) **October 2020 – January 2022**
- Introduced Bayesian reinforcement learning tools for crash detection across Instagram, Whatsapp and Facebook deployments using PyMC3 10-15% improvement in cross-platform detection rates.
- Fine-tuned vector embeddings over HNSW-indexed KNN via FAISS; improved hate speech detection recall by 15% while maintaining precision at billion-post scale.
- Built data pipelines and PySpark jobs integrating Hadoop/Hive historical data with Scuba real-time analytics; with health degradation warning systems using Prophet forecasting.

McKinsey & Company
Senior Data Scientist **November 2018 – June 2020**
- Designed gradient-boosting credit risk models for Tier 1 banks covering $100B+ exposures; results included >40% improvement in early warning detection for an Asia-based client.
- Built and presented strategies under Basel III/IFRS 9 compliance for C-suite executives at three major financial institutions.

OPEN SOURCE & LEADERSHIP

Open Source
- Core contributions to DeepMind Optax and HuggingFace; 4x speedup in SAM image processing; awarded the EWOR Fellowship (0.1% acceptance rate).

Community & Talks
- Invited globally to deliver talks on AI safety and production ML systems including Imperial College London, AUB Lebanon, Sharjah University UAE.
- Social outreach (Linkedin 40,500 followers, Instagram 50.5k, TikTok 90k): Translating updates at the intersection of research and deployment to a diverse audience.