Stefan Webb, PhD

stefanwebb.me github.com/stefanwebb
linkedin.com/in/stefan-webb youtube.com/@stefan-webb

- Generative-modeling researcher turned developer advocate: PhD from the University of Oxford, now leading developer relations at Oumi, with purpose-built agents scaling discovery, monitoring, and more.
- Impact in both developer reach and go-to-market: an article on Hacker News front page and 50,000 Reddit views; launches driving 10× marketing-site traffic and new pipeline; research behind NYTimes article.
- Advocacy backed by working code: open-source libraries for building low-latency agentic tools with local open-weight models, local real-time AI voice agents, and agentic video production.
- Six years across ML research (Meta, Twitter) and developer relations (Oumi, Zilliz). Publications at NeurIPS, ICML, ICLR, UAI, and JMLR. Based in San Francisco.

SELECTED PORTFOLIO

Video (all produced end-to-end)
- Product demo, *"A Tour of the Oumi Platform: Own Your AI, Compound Your Edge"*
- Vlog, *"AI Voice Agents as a Video Game Mechanic?"*
- Tutorial, *"Produce Videos with AI in Premiere Pro for Free using Local, Open-Weight Models"*

Short-form (likewise)
- *"Custom AI in Hours Not Months"*
- *"How do I create a 3D object from an image using Microsoft's TRELLIS.2 (for free)?"*
- *"The two-and-a-half 3D asset formats you need to know for GameDev"*

Writing (all written without AI assistance)
- Oumi Blog, *"Small Fine-tuned Models are All You Need"*
- Basis for a highly-cited NYTimes article: *"Oumi's Study Finds 50% of AI Overviews Untrustworthy"*
- Milvus Blog, *"I Built a Deep Research with Open Source—and So Can You!"*
- gamedev.blog, *"The Curious Case of the LeetCode Problem that Was Useful in the Real World"*

Events
- AI Dev 26 SF, *"VibeML: Build your AI model in hours, not months"*
- Webinar, *"Build custom AI in hours not months with Oumi"*
- Meetup host and speaker (twice a month), *"Unstructured Data Bay Area"*

Open Source (AI-assisted, using Claude Code)
- named-pipes: alternative to MCP and CLI for local agentic tool servers (LLM inference, TTS, STT, vector search) built on low-latency interprocess communication; ships a TUI and Claude Code skill
- unity-voice-agents: Unity package for fully local, real-time AI voice agents built on open-weight models and named-pipes, for non-player characters and in-game interfaces
- premiere-cli, premiere-ai, premiere-ai-skills: toolchain for automating video production in Adobe Premiere Pro — CLI tool and Premiere plugin, local open-weight speech transcription, word-level timestamps, diarization, silence removal, AI cut selection, and Claude Code skills

Docs & Tutorials
- Product documentation, *"Oumi Platform: Custom AI in hours not months"*
- Tutorial, *"Building a web browsing agent from scratch"*

EXPERIENCE

Oumi, *Lead Developer Relations Engineer* May 2025–Present

- Own the developer-facing surface end to end: product and technical videos, marketing website, webinars, industry booths, product documentation, blog, newsletter, and a hosted agent for scaling developer relations processes.
- Led two go-to-market product launches driving 10× marketing-site traffic and new top-of-funnel pipeline, the most recent reaching 50,000 views, 1,000 reactions, and 250 comments on LinkedIn.
- Performed original research that formed the basis for a highly-cited NYTimes article, syndicated in Newsweek, NewsNation, NYPost, PCMag, Ars Technica, and a dozen other media outlets.

Zilliz, *Developer Advocate* Sep 2024–May 2025

- Chief evangelist in the Bay Area for Milvus, the leading open-source vector database.
- Organized and spoke at twice-monthly meetups, industry conferences, webinars, and hackathons.
- Wrote technical articles on RAG and agent methodology, multimodal semantic search, model evaluation, and research agents. A Deep Research piece reached the Hacker News front page and 50,000 Reddit views.

Startup Founder and Generative AI Advisor Jun 2023–Mar 2024

- Prototyped a web-app for generation and delivery of personalized educational content using fine-tuned LLMs.
- Advised startup founders and investors in hedge funds, mutual funds, and sovereign wealth funds on Generative AI trends: training and inference costs, benchmark performance, open versus closed source, and rate of industry adoption.

Meta, *Machine Learning Researcher* Aug 2021–Nov 2022

- Spearheaded use of Bayesian statistical testing to improve the reliability of online experimentation.
- End-to-end development of ML pipelines including large-scale data processing with Trino.

Twitter, *Machine Learning Researcher* Oct 2019–Aug 2021

- Applied ML Researcher collaborating with product teams in Revenue Science, Health, Experimental Data Science, and Interest Discovery to solve long-term ML challenges, pioneering graph-based neural networks.

EDUCATION

University of Oxford, *Doctor of Philosophy* 2014–2019

- Awarded full funding by EPSRC for dissertation research at intersection of Bayesian statistics and deep learning, supervised by Prof M. Pawan Kumar and Prof Yee Whye Teh.

SKILLS

- **Frameworks:** PyTorch, Hugging Face (Transformers, Diffusers, PEFT, TRL), vLLM, MLX LM, Ollama.
- **Architectures:** Transformers, diffusion models, flow matching, state-space models, mixture-of-experts.
- **Fine-tuning and efficiency:** RLHF, DPO, GRPO, LoRA and its variants, quantization like NF4, GPTQ, and AWQ.

PUBLICATIONS

1. Ben Chamberlain, James Rowbottom, Emanuelle Rossi, **Stefan Webb**, Maria Gorinova, and Michael Bronstein. *GRAND: Graph Neural Diffusion.* Thirty-eighth International Conference on Machine Learning (ICML2021). https://arxiv.org/abs/2106.10934.
2. Benjie Wang, **Stefan Webb**, and Tom Rainforth. *Statistically Robust Neural Network Classification.* Thirty-seventh Conference on Uncertainty in Artificial Intelligence (UAI2021). https://arxiv.org/abs/1912.04884.
3. **Stefan Webb**, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. *A Statistical Approach to Assessing Neural Network Robustness.* Seventh International Conference on Learning Representations (ICLR2019). https://arxiv.org/abs/1811.07209.
4. **Stefan Webb**, Adam Golinski, Robert Zinkov, N. Siddarth, Tom Rainforth, Yee Whye Teh, Frank Wood. *Faithful Inversion of Generative Models for Effective Amortized Inference.* Thirty-second Conference on Neural Information Processing Systems (NeurIPS2018). https://arxiv.org/abs/1712.00287.
5. Leonard Hasenclever, **Stefan Webb**, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. *Distributed Bayesian Learning with Stochastic Natural-gradient Expectation Propagation and the Posterior Server.* JMLR 18, 1–37 (2017). http://arxiv.org/abs/1512.09327.