

Brian Woods

Staff / Principal ML & Systems Engineer

San Francisco, CA (relocating Oct 2026) · me@brian.biz · brian.biz · github.com/brianosaurus · linkedin.com/in/brian-r-woods

SUMMARY

Systems engineer with 30 years building and leading production infrastructure: Yahoo Mail anti-spam systems at 6B+ messages/month (6 U.S. patents), engineering leadership at three startups (teams of 8 to 12), real-time distributed systems, and modern ML infrastructure. Current work spans LLM inference performance, production ML, and agent infrastructure, with merged upstream contributions to vLLM and Pydantic AI and published H100 performance research (brian.biz/h100). Hands-on in Python, C++, Rust, Go, and PyTorch. UC Berkeley M.S., Information & Data Science (AI concentration), 2026.

EXPERIENCE

brian.biz (independent) · ML Systems Engineer & Researcher Nov 2025 – Present · San Francisco, CA

- Built and operate a real-time distributed ML platform processing 15M+ transactions/run with gRPC/Geyser streaming, multi-region Jito execution, Rust/Anchor programs, Python orchestration, and Rust extension modules on latency-sensitive paths.
- Built the full ML lifecycle: model training, deployment, monitoring, drift detection, champion/challenger evaluation, kill rules, and public telemetry; deployed PPO, behavioral cloning, offline CQL, and statistical models.
- Published causal systems research, an agent-benchmark reliability study, and a pre-registered data-quality ablation; open-source human-gated agentic workflow. Live systems and write-ups at brian.biz.

AGEN.AI · Chief Technology Officer Jan 2024 – Nov 2025 · San Francisco Bay Area

- Led a 10-person team at an AI startup; built the initial Python/Svelte product and backend, set architecture and engineering practice, and led AI integration against product requirements.

Sabbatical · Travel in Africa Nov 2022 – Dec 2023

Umee · Senior Software Developer Apr 2022 – Oct 2022 · Go, Cosmos SDK

Sifchain · Senior Cosmos Protocol Engineer Apr 2021 – Apr 2022 · IBC, Cosmos SDK

Augur Project · Core Developer Oct 2017 – Oct 2020

Dispatch Labs · Core Developer Aug 2015 – Oct 2017

HelloBit · Director of Engineering Jun 2013 – Aug 2015 · team of 8

AnythingSocial · Director of Engineering 2010 – Apr 2013 · team of 12

Bebo · Head of Viral Marketing and International Sep 2005 – Dec 2009

Tickle · Software Engineer 2001 – 2005

Yahoo · Technical Yahoo / Senior Engineer, Yahoo Mail 1996 – 2001

- Built Yahoo Mail anti-spam systems in C++ while the service handled 6B+ messages/month; earned 6 U.S. patents from that work. Also built e-commerce systems for Yahoo Mail products.

OPEN SOURCE & INFERENCE RESEARCH

vLLM and **Pydantic AI**, merged contributor; H100 study at brian.biz/h100.

- Found a KV-cache correctness bug in vLLM where stale block hashes produced incorrect prefix-cache hits; authored the scheduler fix and regression test (#49377, #52806). Further merged and review contributions on attention configuration, stop-string and KV-cache correctness, including a decode-path livelock incorporated into the upstream fix.
- Measured a 25% throughput effect from prefix-cache contamination in vLLM's repeated serving benchmarks at concurrency 64; corrected the study and reported the methodology issue upstream (#52884).
- Pydantic AI: filed and fixed correctness issues in embeddings and AG-UI streaming; EmbeddingSettings dimensions provider-support fix merged Aug 2026.
- H100 inference study: measured real ceilings (90% of specified HBM bandwidth, 78% of specified BF16 FLOPs), predicted decode rates from weight traffic, and attributed residual latency, including 7.2 ms/step of kernel-launch overhead recovered through CUDA graphs.
- Characterized memory-bandwidth utilization, concurrency scaling, MoE batching economics, KV-cache behavior, and TTFT/throughput trade-offs across serving configurations.

EDUCATION

University of California, Berkeley · M.S., Information & Data Science (AI concentration) 2024 – 2026

- Capstone VitalFoods.ai: designed the original RotatE knowledge-graph models, built the full site, and ran SageMaker/EC2 infrastructure; a teammate led embedding training.

University of Colorado Boulder · B.S., Computer Science & Economics

SKILLS

Systems & infrastructure: distributed systems, real-time streaming, gRPC, queueing/backpressure, multi-region execution, observability, fail-closed design; Python, C++, Rust, Go, SQL, TypeScript/JavaScript

Inference & performance: vLLM, continuous batching, paged KV cache, prefix caching, CUDA graphs, speculative decoding, TTFT/TPOT/throughput, roofline analysis, MBU/MFU, GPU profiling and benchmark methodology

ML engineering: PyTorch, Hugging Face Transformers/Datasets, PPO, CQL, behavioral cloning, LightGBM/XGBoost, feature pipelines, drift monitoring, champion/challenger, SageMaker/EC2, production MLOps

Research & evaluation: experiment design and preregistration, causal inference (RDD, DiD, fixed effects), benchmark/eval construction, bootstrap uncertainty, power analysis, robustness tests, reproducibility

Leadership & agents: founder/CTO/director experience, hiring and mentoring, research rigor and documentation standards, AI agents in the engineering loop, MCP server development

Evidence: brian.biz for live telemetry, H100 inference research, causal studies, agent evaluation, and technical write-ups; GitHub for upstream OSS work