Nils Matteson

Madison, WI | (208) 921-2576 | nilsmatteson@icloud.com | linkedin.com/in/nilsmatteson | github.com/matteso1 | nilsmatteson.com

CS master's student (Northeastern, Silicon Valley) and Data Science & CS senior (UW–Madison). Founder of **thaw** (LLM-inference infrastructure) and **Matteson Systems LLC**, with experience across GPU/CUDA inference, distributed systems, and applied ML.

Education

Northeastern University Sep 2026 – May 2028
M.S. Computer Science, Silicon Valley campus *San Jose, CA*

University of Wisconsin–Madison Expected May 2026
B.S. Data Science, Minor in Computer Science *Madison, WI*

- **Coursework**: Big Data Systems, Artificial Intelligence, Machine Learning, Machine Organization, Programming III (Data Structures), Linear Algebra, Discrete Math, Data Science Modeling.

Technical Skills

Languages: Python, Rust, Go, C++, CUDA, TypeScript/JavaScript, SQL, R, Java
ML & Inference: PyTorch, vLLM, SGLang, Triton, Hugging Face, XGBoost, scikit-learn, RAG, conformal prediction, NumPy/Pandas
Systems & Infra: CUDA/PyO3 FFI, gRPC, Raft, LSM-trees, Kafka, AWS (Bedrock/S3), Docker, PostgreSQL, Redis, CI/CD, Linux
Web & Data: Next.js, React, Node, WebSockets, DeckGL/MapLibre, Drizzle, Flask, DuckDB

Experience

thaw (startup) 2026 – Present
Founder & Lead Engineer *Rust · CUDA · Python*

- Architected **thaw**, a Rust + CUDA library that snapshots and restores live LLM inference state (weights, KV cache, prefix-hash table, scheduler), enabling sub-second forking of vLLM sessions: **0.88s** median fork vs. ~340s cold boot on H100 (committed benchmark receipts).
- Built a double-buffered `O_DIRECT` DMA pipeline overlapping disk reads, PCIe transfer, and an 8-shard parallel CRC32C verifier that matches a serial pass exactly, hitting **14.3 GB/s** weight restore and 3.4× faster 70B load with bit-identical output across **8 models**.
- Eliminated a 60× KV-cache snapshot bottleneck by coalescing ~16K tiny per-block DMAs into one contiguous gather, then reconstructing vLLM's prefix-cache hash table on restore so cold-started requests skip prefill.
- Designed a `CudaBackend` trait (mock + real CUDA behind one contract) running the full pipeline and **172 Rust tests** on macOS with no CUDA toolchain; opened vLLM RFC #34303 (co-founded with M. Yu, K. Kapur).

Matteson Systems LLC 2026 – Present
Founder & Engineer *Next.js · Postgres · Playwright · Claude*

- Built and shipped an **autonomous** SMB cold-outreach product that scraped and scored **10,500+** local businesses (OpenStreetMap) and surfaced 158 high-priority leads in a single run.
- Engineered a two-stage Claude-vision audit pipeline combining real-device Playwright screenshots (Chromium + WebKit/iPhone), live Lighthouse Core Web Vitals, and on-page analysis into per-lead website critiques and personalized cold emails at **~$0.04/lead**.
- Designed an append-only event-sourced Postgres CRM (6 tables, 8 Drizzle migrations) and a threaded Gmail drip on Vercel cron (RFC822 in-thread replies, reputation-aware send caps) with a per-call cost ledger in integer microcents.

Research Cyberinfrastructure, UW–Madison DoIT Jan 2026 – Present
AI Workflows Research Assistant *AWS Bedrock · RAG · LLM eval*

- Built the evaluation and cost-tracking framework benchmarking **9 LLMs** and 3 ensemble strategies on AWS Bedrock across 282 questions; ensemble majority voting (0.840) outperformed every individual model.
- Showed via Pareto analysis that Llama-4 Maverick reaches **98%** of top accuracy at a fraction of cost and latency, informing production model selection; presented at the UW–Madison ML+X Forum.

Selected Projects

Sentinel: Distributed Message Queue | github.com/matteso1/sentinel *Go*

- Built a Kafka-inspired distributed log engine from scratch: custom **LSM-tree** storage (skip-list memtables at 3.9M reads/s, SSTables w/ CRC32, WAL, leveled compaction), a gRPC wire protocol, and topic/partition consumer groups.
- Implemented **Raft consensus** from scratch (leader election, log replication, split-brain prevention), validated by a deterministic in-memory network simulator covering partition + heal; 45 passing tests.

Madison Metro ML: Real-Time Bus Arrival Prediction | madisonbuseta.com *Python, XGBoost, React*

- Shipped a live ML system correcting transit-API ETAs with a **47-feature** XGBoost model and **Mondrian conformal prediction** (calibrated 90%-coverage intervals stratified by route × day-type × horizon).
- Automated nightly retraining (GitHub Actions) behind a hard deployment gate (≥2s MAE gain, no metric regression); React/DeckGL/MapLibre map rendering 200+ live vehicles at 60fps.

Also: Lattice (Rust+PyO3 reactive framework, Cranelift JIT, 10.6× vs. Streamlit) · brain2text (BCI Kaggle: 5-layer GRU + CTC decoder, 40% WER) · gitstare (Rust TUI, on crates.io) · LockBox (AES-256-GCM password manager).
Coursework projects: distributed-systems visualizations (Cassandra/Kafka/HDFS/HBase, CS544) · 1.2M-tweet sentiment pipeline on the UW HTCondor grid (STAT405) · causal-inference propensity-score matching in R (STAT479).