Owen Li Murphy

Machine Learning Engineer

Shanghai, Shanghai | +86 17321219810 | owen261@icloud.com | Owen Murphy | olmurphy

PROFESSIONAL EXPERIENCE

Software Engineer 08/2026 – Present | Shanghai, China
Global Business AI

- **Engineered a self-hosted CPU inference engine** running an optimized DeBERTa-v3 model via ONNX Runtime for asynchronous conversation satisfaction classification; implemented atomic memory pointer swapping for zero-downtime model hot-reloading and graph-level optimizations.
- **Built an asynchronous, event-driven multimodal RAG pipeline** across Kafka/Redpanda and OpenSearch; orchestrated contextual document chunking, distributed Redis token-bucket rate limiting (300k TPM), and hybrid lexical/vector retrieval.

Software Engineer 06/2022 – 09/2025 | New York, U.S.
IBM

- Architected a high-throughput static analysis pipeline processing 100+ architectural code bases to enforce automated semantic compliance, reduced audit lifecycle from **336 hours to 8 hours** (42x speedup) while ensuring optimal policy enforcement across the enterprise.
- Engineered a **fault-tolerant event-processing framework** featuring exponential backoff retries and tiered DLQ management; achieved **99.9% data durability** for heterogeneous metadata payloads and eliminated manual intervention for transient system failures.
- Engineered a high-concurrency GenAI microservices architecture supporting **50k concurrent requests**; leveraged a distributed Redis/Postgres cache-aside pattern to reduce p99 latency by **20%** and optimize database throughput by **3x**.
- Optimized a **RAG pipeline** for a fleet of 50k users, implementing **semantic caching and vector-search optimization** to drive a 10% uplift in user retention and engagement.
- Directed an **infrastructure-as-code (IaC) modernization** from legacy EC2 to **OpenShift/Kubernetes**, implementing horizontal pod autoscaling (HPA) and right-sizing clusters to capture **$5M in annual OpEx savings**.

RESEARCH & PUBLICATIONS

Governing AI Agent Swarms via Cost-Aware Multi-Task Peer Prediction
Proposed a novel mechanism for governing AI economies where agents strategically choose their inference effort based on token costs and peer rewards. This work established theoretical guarantees for truthful equilibria and achieved state-of-the-art aggregation accuracy on heterogeneous LLM swarms by weighting agents based on their peer-prediction scores.

EDUCATION

M.S. Computer Science 2025 – 2026
Georgia Institute of Technology
Computer Architecture, ML, DL, RL, NLP

B.S. Math and Computer Science 2018 – 2022 | Waco, U.S.
Baylor University
Magna Cum Laude

SKILLS

PyTorch | Scikit-learn | GPU Architecture | Distributed Systems | HPC | Kubernetes | Spark | Python | C++ | Java | SQL/NoSQL | FlashAttention (1/2/3)

CERTIFICATIONS

AWS Certified Solutions Architect / Developer / SysOps Associate

PROJECTS

PyTorch Optimization & Regularization Harness 05/2026 – 07/2026

- Conducted controlled optimizer ablations (SGD vs. Nesterov vs. custom AdamW variants) , analyzing gradient evaluations, convergence thresholds, and generalization gaps.
- Evaluated regularization methods (L2 weight decay, dropout, label smoothing) to mitigate overfitting and preserve minority-class recall.

Unsupervised Feature Geometry & Downstream Neural Learning

- Analyzed the high-dimensional geometric structure of Covertype using clustering (K-Means, GMM/EM) and dimensionality reduction (PCA, ICA, Randomized Projections).
- Evaluated distance preservation and projection distortion across random seeds to optimize downstream PyTorch neural network runtime and generalization.

High-Performance Systems & IPC

- Developed high-performance, multi-threaded systems in C/C++ leveraging custom thread pools, gRPC, and socket-based IPC to minimize latency and memory overhead.
- Optimized shared-memory concurrency barriers and virtual memory allocation patterns, mirroring techniques used in distributed GPU training setups.