# Omer Sajid

San Francisco, CA | +1 765 301 1753 | omersajid9@outlook.com | iomersajid.pw | in/omersajid9 | git/omersajid9

## Education

**University of Illinois Urbana-Champaign** — Champaign, IL
Master's in Computer Science; Cumulative GPA: 4.00/4.00 — Jun 2027
**Courses**: Machine Learning, ML Systems, CUDA Programming, ML & Compilers, High Performance Computing
**Extracurricular:** Member of Technical Staff at Agentic AI Club

**DePauw University** — Greencastle, IN
Bachelor's in Computer Science; Cumulative GPA: 3.8/4.00 — Jan 2023

## Professional Experience

**Amazon** — Urbana, IL
**Machine Learning Systems Researcher** — Aug 2026 – Present

- Co-developed large-scale, end-to-end agentic RL training platform on Trainium, improving SWE-bench resolve rate from 19% to 55% on Qwen3.5-9B.
- Increased rollout throughput 23× using autobucketization and continuous batching, eliminating per-step recompilation.
- Reduced memory pressure 73% on long-horizon rollouts via parallelism strategies and custom NKI kernels.

**Samsung** — Jersey City, NJ
**Machine Learning Systems Engineer** — May 2026 – Aug 2026

- Profiled KV-cache I/O across long-horizon SWE-agent workloads and designed a three-tier offloading strategy.
- Derived SSD bandwidth, latency, and capacity targets for disaggregated prefill/decode serving, incorporated into next-gen product lineup.

**Yabla** — Jersey City, NJ
**Machine Learning Engineer** — Oct 2023 – May 2026

- Fine-tuned LLM (Qwen3-9B) using PEFT to increase multilingual translation accuracy by 23%, reducing educator workflow from 4 hours to 6 minutes per video.
- Deployed the model via SGLang, implementing inference optimizations including prefix caching, continuous batching and KV cache management, serving 1,000 requests/day.
- Optimized video search infrastructure by migrating from SQL-based queries to an inverted-index engine, eliminating user complaints by 98% and reducing latency from 1.5s to 150ms for 800,000+ daily requests.
- Built end-to-end data pipelines encompassing both real-time streaming and batch ETL architectures with visualization dashboards, monitoring 12 KPIs across user engagement, retention and revenue metrics.

**Fermi National Accelerator Laboratory** — Batavia, IL
**Quantum Algorithm Researcher** — May 2022 – Oct 2023

- Developed novel quantum optimization algorithm in Qiskit to train geometric Quantum Neural Networks with 95% accuracy across 8 to 85 qubit systems; published in Physical Review D.
- Reduced simulation costs via tensor network compression and dimensional reduction, cutting memory 9x and runtime from 48+ hours to 5 hours.

## Publications

- **M. O. Sajid**, *"Agentic Reinforcement Learning on Trainium,"* MLSys, *manuscript in preparation*.
- Y.-Y. Li, **M. O. Sajid**, J. Unmuth-Yockey. *"Lattice holography on a quantum computer"* Physical Review D, 110(3), 034507.

## Technical Project

**GPT-2 training and CUDA Inference Engine** — Jan 2026 – May 2026
Trained GPT-2 (124M) from scratch in PyTorch and wrote CUDA inference engine on NVIDIA A40 with Flash Attention, KV-cache, and Tensor Core GEMM kernels, achieving 556× matmul, 84.5× prefill and 10.9× decode speedup vs. baseline.

## Technical Skills

**Languages**: Python, C++, JavaScript/TypeScript, PHP, Rust, Swift, Java, R
**Technologies**: AWS, Docker, Vagrant, Git, SQL, PostgreSQL, Laravel, React.js, Node.js, Slurm
**ML**: CUDA, PyTorch, NumPy, verl, Transformers, PEFT, TRL, Gymnasium, Pandas, Scikit-learn