AHMAD TAWIL

Sakhnin, Israel · ahmadtawil.se@gmail.com · github.com/AhmadTawil1 · linkedin.com/in/ahmadtawil1

Software Engineer — GPU Inference Performance, Applied GenAI & Backend Systems

SUMMARY

Final-year B.Sc. Software Engineering student who measures and builds GPU-accelerated AI systems: vLLM inference benchmarking across A100 and L4 tiers, multimodal retrieval pipelines on PyTorch and Hugging Face models, and containerized Python services (FastAPI, Airflow, PostgreSQL, Docker).

EDUCATION

Braude College of Engineering, Karmiel — B.Sc. Software Engineering Expected Feb 2027

- Coursework: Artificial Intelligence, Reinforcement Learning, Data Mining, Cloud Computing, Software Systems Architecture, Databases.

GPU & AI SYSTEMS PROJECTS

LLM Serving Performance — vLLM Prefix Caching Across GPU Tiers 2026

- Benchmarked vLLM prefix caching (Qwen2.5-7B) over 8 concurrency levels $\times$ 2 GPU tiers (A100, L4) $\times$ cache on/off — 80 records.
- Throughput transferred across tiers (2.56$\times$ vs 2.70$\times$ at concurrency 128); latency SLO compliance did not (100% A100 vs 0% L4) — tuning transfers, hardware viability does not.
- Built a closed-loop asyncio load generator with client-side TTFT/TPOT instrumentation; 88 GPU-free tests.

RAG Configuration Transfer Across GPU Tiers 2026

- Swept 96 frozen RAG configurations (chunk, overlap, embedding model, top-k, reranker) identically on A100 and L4 — 192 records; retrieval quality byte-identical, isolating latency.
- Per-stage degradation: rerank 1.403$\times$, generate 1.169$\times$, embed 1.043$\times$; Kendall's τ-b = 0.80 across latency rankings. Verdict *partial* — the embedding model moved the Pareto frontier, not the reranker.
- Provenance on every record (git SHA, corpus hashes, model and CUDA/torch versions); 77 tests.

Video Ingestion Resolution Threshold — SigLIP 2 Embedding Drift 2026

- Measured drift across 3,600 records (600 frames $\times$ 6 resolutions): under 0.1% above the model's 384 px input but 1.4% at 360p — a 13.4$\times$ jump, unanimous across 600/600 frames.
- Validated the instrument with negative and positive controls; label-free design removed annotator bias.

DriftWatch — LLM Regression & Cost Watchdog (Airflow, PostgreSQL) 2026

- Scheduled re-benchmarking of 4 models on a versioned 40-task suite: 480/480 records, 0 failures, $0.1134 per run; rolling-window z-score detection at a measured 3.23% false-positive rate.
- Async dispatch (asyncio/httpx) with concurrency caps and retry/backoff, Airflow dynamic task mapping, database-enforced idempotency; `EXPLAIN`-driven index cut a query 2,069 $\rightarrow$ 424 buffer reads.

OmniSight — Semantic Video Intelligence Platform (Capstone, two-person) 2026 – 2027

- Dual-stream video indexing (SigLIP 2 + InternVideo2) with reciprocal-rank fusion, a four-tier Qwen2.5-VL verification cascade, and a multimodal RAG stage producing cited reports.
- Containerized microservices (Docker Compose): React, FastAPI, RabbitMQ, Qdrant, PostgreSQL, MCP.

TECHNICAL SKILLS

Languages: Python (3.12), SQL, TypeScript, JavaScript, Bash, LaTeX

GPU & Inference: PyTorch, Hugging Face Transformers, vLLM, KV cache / PagedAttention / prefix caching, TTFT / TPOT / goodput, SLO analysis, GPU tiering (A100 / L4 / T4), CUDA runtime debugging, load generation

GenAI & Evaluation: RAG design, agentic pipelines (LangChain, LangGraph), vector databases (Qdrant, FAISS, ChromaDB), reranking, multimodal models (SigLIP 2, InternVideo2, Qwen2.5-VL), MCP, eval suites, drift detection

Backend & Infrastructure: FastAPI, asyncio + httpx, Apache Airflow 3.x, PostgreSQL 16, SQLAlchemy, RabbitMQ, REST APIs, Docker & Docker Compose, Linux, Git

Method: design of experiments, percentile latency (p50 / p95), pre-registered hypotheses, provenance & reproducibility, pytest (unit, real-server integration, live-database)