Seifeldin Abdellatif

AI Engineer

+971557871870 • contact@saifmb.com • Portfolio • GitHub • LinkedIn • United Arab Emirates

Education

Al Ain University – BSc. Software Engineering • GPA: 3.68/4.00, 3× Honors List — Sep 2023 – Expected May 2027

Experience

AI Mobile Quality Assurance Intern — **Dubai, UAE**

Incoming @ Samsung Gulf — Jun 15th - Aug 14th

Research Assistant — **Abu Dhabi, UAE**

Al Ain University • Python, C++, Triton, CUDA — Nov 2025 – Present

- Accelerated tree-based attention 8× over FlashInfer on larger trees (N≈680), achieving linear $\mathcal{O}(N \cdot d)$ scaling through a query-guided tree walk. Achieved up to 1.6× Llama-3.1-8B throughput over EAGLE-3 on an A100 (L=10240, N=343)
- Recovered additional +20.7% throughput on Llama-3.1-8B over EAGLE-3 by resolving a redundant draft bug
- Accelerated sparse Vision Transformers throughput 12% over FlashAttention-2 varlen on Lovelace (90% sparsity, BS=128); reducing kernel latency 2.12× by optimizing for short ViT sequences and removing unnecessary backward pass overhead
- Recovered 57% PPL degradation and +3 pt on GSM8K over 3-bit TurboQuant via an attention QK adapter at a 5μs cost
- Provisioned an 8x RTX 4000 cluster to train a custom draft head. Achieved 97.3% scaling efficiency via a 2.3s all-reduce, syncing every 12 steps on 1Gbps inter-connect for a combined training throughput of 1480 tok/s

AI Engineering Intern — **Remote**

LuxAI • TensorRT, Triton, CUDA, Nsight, FastAPI — Jul 2025 – Mar 2026

- Designed a multi-modal inference pipeline with a sub-35 ms VADER triage gate to filter 60% of content pre-inference
- Achieved 1.75× throughput and 2.34× reduced P99 latency on Whisper via Int8 quantization and VAD-gating. Fit single worker into 48MB VRAM, affording dual threading for +60% throughput on constrained memory
- Reduced audio encoder inference latency by 22% over PyTorch by exporting to TensorRT FP16. Profiled the output engine with Nsight, identifying an unfused softmax bottleneck caused by missing FlashAttention support on SM75
- Accelerated SBERT throughput by +71% over SDPA at small workloads (N=1000) using custom fused attention on Turing
- Reduced GPU residency to ∼12% of runtime using event-driven serverless architecture for data-bound GraphQL streams.

Software Engineering Intern — **Abu Dhabi, UAE**

Smart Navigation Systems • Python, C++, TypeScript, Django, OpenCV, IoT (ESP32) — May 2025 – Nov 2025

- Designed the core backend for Himaya71 (UAEU I2P 3rd place winner), a smart campus safety system aggregating occupancy and fire alert states from distributed nodes running YOLOv8s, posting to a Django PostgreSQL database

Research & Publications

- **S. Abdellatif**, 2026. "QAttention: Tree-Sparse Attention and Acceptance Decay in SD." Submitted to NeurIPS 2026 [GitHub]
- **S. Abdellatif**, A. Almasri, 2026. "Dispatch-Aware Ragged Attention for Pruned Vision Transformers." arXiv:2604.15408
- **S. Abdellatif**, 2026. "Acceptance Dynamics in Speculative Decoding Across Cognitive Domains." arXiv:2604.14682
- **S. Abdellatif**, 2026. "Empirical Analysis of KV Cache Compression and Sparsity for Hardware Acceleration." In preparation
- **S. Abdellatif**, 2026. "LoRA Adapter for KV Quantization: Quality Recovery and Lossy No-Dequant." In preparation
- **S. Abdellatif** et al., 2026. "Structured pruning in Vision Transformers: A Systematic Review". In preparation

Open Source

- **PyTorch PR#178847** – Resolved cpp_wrapper backward failure due to lazy backward graphs compilation missing context
- **PyTorch PR#178698** – Fixed uint8 corruption from failing inline-asm capability check, adding IEEE 754 fallback
- **PyTorch PR#178098** – Fixed incorrect shape restoration in Inductor's codegen by resolving gradient shape mismatch

Projects

Local Inference Engine • Python, Ollama, Flask

- Engineered a 1.6T A49B MoE inference pipeline with a 3-tiered cache (HBM-RAM-NVME) on a 20GB RTX 4000 Ada.
- Exported expert routing sequences to distill an 8M T5 predictor from DeepSeek V4 Pro via pre-fill sequences and forward masking. Achieved 3.1min/seq (512 tokens) via an LRFU cache for cross-domain decay, with 42% expert hit rate
- Achieved 99.7% top-4 expert recall, pre-fetching dynamic experts into memory JIT, achieving a 90% cache hit-rate
- Hosted Qwen-3.6-27B IQ3_XS as a sub-agent on an RTX 4000 Ada to afford 128k context, achieving single-batch 21 tok/s

Skills

Inference: Nsight, Triton, CUDA, Triton Inference Server, TensorRT, ONNX Runtime, vLLM

AI/ML: PyTorch, scikit-learn, Hugging Face, Transformers, NumPy, Pandas, OpenCV, RAG, LoRA

Languages & Deployment: Python, Bash, C++, GCP, Linux, Docker, GitHub Actions, FastAPI, PostgreSQL