Saif Mahmoud

AI Engineer

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

Education

Al Ain University – BSc. Software Engineering • GPA: 3.81/4.00, 3× Honors List — Sep 2023 – Expected May 2027

Experience

Research Assistant — **UAE**

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

- Accelerated tree-based attention 8× over FlashInfer on larger trees (N≈680) by achieving linear $\mathcal{O}(N \cdot d)$ scaling through a query-guided tree walk, enabling larger sequence verification per target model forward pass
- Achieved up to 1.6× Llama-3.1-8B throughput over EAGLE-3 with QAttention on a single A100 (L=10240, N=343)
- Recovered additional +20.7% throughput on Llama-3.1-8B by resolving a missed exit in EAGLE-3's draft loop
- 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
- Evaluated KV growth forecasting to avoid pre-emption, finding re-prefill (4000 tok/s) to be faster at all tested configs

AI Engineering Intern — **Remote**

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

- Architected a multi-modal inference pipeline with a sub-35 ms VADER triage gate to filter 60% of content pre-inference
- Engineered an Int8-quantized Whisper workflow with VAD-gating, maintaining ∼48 MB VRAM per worker. Scaled to 2 parallel workers, achieving +60% throughput at no additional memory cost within a 4 GB budget
- 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
- Aggregated previous optimizations for 1.75× E2E throughput and 2.34× reduced P99 latency on Whisper, reducing GPU residency to ∼12% of runtime using event-driven architecture for data-bound GraphQL streams.
- Accelerated SBERT cold-start throughput by +71% over SDPA by implementing a fused attention kernel with online softmax and tiled accumulation, at under 5% sustained throughput overhead (Turing)

Software Engineering Intern — **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
- Handled concurrent state updates via pessimistic locking, purging stale records under high-frequency IoT event streams

Research & Publications

- **S. Mahmoud**, A. Almasri, 2026. "Dispatch-Aware Ragged Attention for Pruned Vision Transformers." arXiv:2604.15408
- **S. Mahmoud**, 2026. "Acceptance Dynamics in Speculative Decoding Across Cognitive Domains." arXiv:2604.14682
- **S. Mahmoud**, 2026. "QAttention: Tree-Sparse Attention and Acceptance Decay in Speculative Decoding." Under review
- **S. Mahmoud**, 2026. "MemForecaster-R: Trajectory-Aware Memory Scheduling for Reasoning LLMs." In-preparation
- **S. Mahmoud** 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 Agent • Python, Ollama, Flask

- Hosted a 27B Dense Qwen-3.6 with IQ3-XS quantization (∼3.4 bit) on an RTX 4000 Ada with 4-bit KV quantization to afford ∼128k context in 20GB VRAM. Achieved 20 tok/s at BS=1, scaling to ∼13 tok/s at >90k context tokens
- Integrated into Qwen Code to handle long agentic tasks, offsetting a daily ∼20M/100k input/output tokens in-house

Search Intelligence Engine – GitHub • Python, SpaCy, Scikit-Learn, GitHub Actions

- Designed a multilingual ingestion-based GEO/SEO engine gated with SBERT and LLM-as-a-judge to avoid hallucinations
- Drove +288% increase in traffic and #1 SERP by finding content gaps using TF-IDF and dense embeddings

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