Saif Mahmoud

AI Engineer

Abu Dhabi, UAE • contact@saifmb.com • Portfolio • GitHub • LinkedIn

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 • Vision Transformers, Triton, CUDA — Nov 2025 – Present

- Accelerated tree-based speculative decoding attention by ∼16× over PyTorch SDPA (Turing, d=5, b=4), observing better speedups and memory savings at deeper trees and branching
- Optimized for Blackwell architectures (SM120) with ancestor-aware attention, achieving ∼4.5× kernel-level speed-up for tree-based SD against FlashAttention-2, the fastest available dispatch on SM120 where FlashInfer has no support
- Accelerated sparse Vision Transformers throughput by 2.5× on Turing (90% sparsity) over SDPA, reducing kernel latency ∼9× via custom varlen attention. Ported to A100s, showing 1.8× kernel speedup over FlashAttention-2
- Automated screening of 550+ research papers by building a tool to log submissions, flag duplicates, and confirm eligibility

AI Engineering Intern — **Remote**

LuxAI • Triton, CUDA, TensorRT, Nsight, Whisper, SBERT, FastAPI, Triton Inference Server, Playwright — 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
- 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)
- 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 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

Software Engineering Intern — **UAE**

Smart Navigation Systems • Python, C++, TypeScript, Django, OpenCV, IoT, Next.js — 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

Publications

- **Accelerating Vision Transformers with Hardware Aware Triton-based Sparse Attention** – ZU SRCAC 2026 (Poster)
- **Only Ancestors Matter: Sparse Flash Attention for Tree-Structured Speculative Decoding** – Manuscript in preparation
- **Structured pruning in Vision Transformers: A Systematic Literature Review** – Manuscript in preparation

Open Source

- **PyTorch PR#178698** – Fixed two bugs causing `torch.compile()` to silently produce wrong uint8 values from ceil(log2(...)) calls caused by a failing hardware capability check. Added pre-SM100 IEEE 754 bit-manipulation fallback (Under review)
- **PyTorch PR#178098** – Fixed RuntimeError in torch.compile caused by incorrect shape restoration in Inductor's mix-order reduction codegen, resolving gradient shape mismatches during backward pass (Under review)
- **PyTorch PR#178723** – Fixed silent wrong results for `uint8` tensor vs negative signed scalar, identifying type promotion wrapper as root cause. Resolved via `is_comparison_op_` flag and re-promotion through `promoteTypes()` (Under review)
- **vLLM PR#38475** – Fixed OOM on decode instances in the P2P NCCL KV-cache connector under high QPS by popping `recv_store` entries post-injection, freeing pinned-RAM, and refactoring cleanup to use per-request tracking dict (Open)

Projects

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

RAG Email Assistant – GitHub • Python, FastAPI, pgvector, Gemini

- Built a RAG pipeline over Gmail via OAuth2 with hybrid vector and lexical search using pgvector and full-text indexing
- Enforced chunk citation to avoid model hallucinations. Masked PII using session-scoped token replacement.

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