Saif Mahmoud
AI Engineer

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

EDUCATION

Al Ain University — **Abu Dhabi, UAE**
Bachelor of Science in Software Engineering • GPA: 3.81/4.00, 3x Honors List — **Sep 2023 – Expected May 2027**

EXPERIENCE

Research Assistant — **UAE**
Al Ain University • Vision Transformers, Generative Adversarial Networks (GANs), Triton — **Nov 2025 – Present**

- Conducting a systematic review on structured pruning in Vision Transformers across 90+ papers, evaluating cost-accuracy and compute-throughput tradeoffs
- Improved sparse Vision Transformers' throughput on Turing by 1.5x over PyTorch SDPA by writing a Triton ragged attention kernel eliminating redundant padding on dropped tokens, and reducing kernel latency by 89%
- Evaluating a 3× parameter-reduced U-Net backbone as both GAN Generator and Discriminator, analyzing convergence stability and generative fidelity under reduced capacity constraints
- Automated the screening of 550+ papers by building an internal tool to evaluate against acceptance criteria

AI Engineering Intern — **Remote**
LuxAI • Triton, CUDA, TensorRT, Nsight, Triton Inference Server, Whisper, Sentence-Transformers, FastAPI, Playwright — **Jul 2025 – Mar 2026**

- Architected a multi-model video inference pipeline on Triton IS, enabling independent model scaling
- Accelerated SBERT cold-start inference by +71% over PyTorch baseline by writing a custom FlashAttention Triton kernel with online softmax and tiled QKV accumulation, at a <5% sustained throughput overhead
- Engineered an Int8-quantized Whisper workflow with VAD-gating, maintaining ~48MB VRAM per video. Scaled to 2 concurrent workers, achieving +60% throughput at no memory cost within a 4GB budget
- Reduced audio encoder inference latency by 22% over PyTorch by exporting to TensorRT FP16. Used Nsight to profile output engine, identifying unfused softmax as primary bottleneck due to absent FA support on SM75
- Designed a lightweight triage gate with sub-35ms latency, reducing total inference load by 60%

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 devices
- Built an edge inference pipeline running YOLOv8s, posting to Django endpoints for PostgreSQL aggregation

PROJECTS

Edge Diffusion Engine – GitHub • PyTorch, Diffusers, CUDA

- Engineered a text-to-image inference pipeline to execute a 2.5B parameter Stable Diffusion model on a constrained 4GB GTX 1650, replacing large T5 text encoders with smaller CLIP counterparts
- Implemented 4-bit NF4 quantization, strict component flushing, and dynamic VAE payload interception (FP32 upcasting) to prevent Turing-architecture float16 hardware overflows

Resume Optimizer (On-Device + API) – GitHub • TypeScript, Gemini, WebLLM, Llama-3.2-3B

- Built a hybrid inference router across client-side SLM and Gemini API. Implemented 2-phase prompt chaining with Server-Sent Events (SSE), achieving a 72% reduction in Time-To-First-Token via response streaming

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

- Architected a bilingual GEO/SEO engine via K-Means clustering and hybrid retrieval. Designed a perplexity-based validation gate using Sentence-Transformers and an LLM-as-a-judge to sanitize ingested data
- 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 (TIS), TensorRT, Quantization, LoRA
- AI/ML: PyTorch, scikit-learn, Hugging Face, Transformers, NumPy, OpenCV
- Languages & Frameworks: Python, Bash, C++, TypeScript, Java, Django, FastAPI, Node.js
- DevOps/Deployment: Linux, Docker, Git, GitHub Actions, Playwright, PostgreSQL, pgvector

CERTIFICATIONS

HarvardX CS50AI – Introduction to Artificial Intelligence	**Jul 2025**
DeepLearning.AI & Stanford – Machine Learning Specialization	**Jan 2026**