# Ayman Mahfuz

512-705-8897 | ayman.afeef@gmail.com | linkedin.com/in/aymanmahfuz | aymanmahfuzportfolio.com

## Education

**The University of Texas at Austin** — Austin, TX

*Bachelor of Science in Computer Science* — *Aug 2023 - May 2027*

- **Concentration:** AI and ML | **GPA:** 3.5/4.0
- **Courses:** HPC, NLP, Visual Generative Computing, Data Structures, Computer Architecture, Algorithms

## Skills

**Languages**: Python, C++, C, JavaScript, TypeScript, SQL, Java, HTML/CSS
**Systems/Tools**: System Design, Distributed Systems, Multi-threading, Concurrency, Docker, Linux, Git, AWS, GCP
**Frameworks/Libraries**: React.js, Flask, Django, FastAPI, REST APIs, OAuth/JWT, Pandas, Numpy
**ML/AI**: PyTorch, TensorFlow, Transformers, LLMs, Image Generation, Reinforcement Learning (PPO, MAPPO)

## Experience

**Arm** — May 2025 – Present

*ML Research Engineer Intern*

- Engineered a modular, high-performance, ML-driven Bayesian optimization platform (Python) automating 10,000+ hours of CPU/memory stress tests; achieved 99.8th-percentile coverage using less than 1% of test cases and accelerated failure discovery.
- Presented tool to SVP and directors, collaborated across firmware/validation teams, authored 1,400+ lines of documentation, and conducted code reviews to ensure reliability, maintainability, and adoption of the platform.

**The University of Texas at Austin - AI Lab, Texas Robotics** — Jan 2025 – Present

*Undergraduate Research Assistant*

- Optimized a 400K+ LOC C++ simulator and training stack; cutting RL training time  67% via profiling, memory tuning, and early-termination heuristics.
- Scaled curriculum training to 5M+ episodes with reproducible GPU configs and CI/CD testing pipelines; tuned sim-to-real parameters to boost on-robot reliability.
- Deployed hierarchical multi-agent policies (dribble, pass, shoot, defend) on NAO robots, securing a 3rd place finish globally at RoboCup 2025 SPL.

**The University of Texas at Austin - Center for Media Engagement** — Aug 2023 – May 2025

*Software Engineer, Undergraduate Research Assistant*

- Designed from the ground up a scalable full-stack research platform: three React web application games tied to a Flask backend with a round-robin game assigner, Firebase logging of every participant action, and dashboards for rapid data analysis; supported 1,000+ participants with 99.99% uptime.
- Engineered ETL pipelines and cron-based scrapers that ingested and organized a 250M+ row dataset into GCP BigQuery, enabling efficient queries and domain/content-level analysis.
- Built and fine-tuned BERT/LLM classifiers (clickbait, story detection, sentiment, entity recognition) and deployed them behind Flask REST APIs with logging/error handling for reliable study use.

**Lockheed Martin** — Jun 2022 – Oct 2022

*Software Engineering Intern*

- Built JavaScript tooling to deduplicate asset records in internal CRM systems, improving data integrity and streamlining reporting across departments.

## Projects

**Helm — AI-Native Product Management Platform** | *Next.js, FastAPI, PostgreSQL, OpenAI*

- Built a "Cursor for PMs" — a Jira alternative with AI-first workflows: natural language ticket management, auto-generated daily briefings, and live meeting capture that converts discussions into tickets automatically.
- Architected an agentic LLM system with 25+ callable tools via OpenAI function calling, enabling contextual reasoning over tickets, GitHub PRs, commits, and transcripts to answer complex product questions in chat.
- Designed a hybrid RAG layer combining SQL filtering with vector embeddings across all project data; powers briefings that synthesize overnight activity, surface blockers, and recommend prioritized next actions.

**Modern Transformer LLM with RLHF Alignment** | *Pytorch*

- Built decoder-only LLM from scratch with state-of-the-art architecture (RoPE, RMSNorm, SwiGLU, attention sinks); developed Direct Preference Optimization (DPO) alignment pipeline; optimized training for H100 GPU

**Attention-Based Image Gen Model** | *Pytorch*

- Built DDPM from scratch with transformer attention layers and U-Net backbone; implemented both DDPM and DDIM samplers achieving 20× speedup while maintaining generation quality on CIFAR; trained end-to-end noise prediction and denoising pipeline