Ahmed Muharram

muharram.dev | linkedin.com/in/ahmed-muharram | github.com/ahmedomuharram
Philadelphia, PA (Relocating May 2026 to Cairo, Egypt) | ahmed.o.muharram@gmail.com

SUMMARY

UPenn MSE/BSE applied AI engineer shipping reliability-focused LLM systems: TB-scale ingestion, retrieval-first question answering, and grounded generation with measurable quality controls. Deployed production enterprise RAG pipelines and an on-prem, secure meeting intelligence platform. Seeking SWE/LLM roles in Egypt with end-to-end ownership across backend, data, and infra.

Egyptian citizen. Military service: Exempt.

EDUCATION

University of Pennsylvania, School of Engineering and Applied Science. Philadelphia, PA

- **Master of Science in Engineering (MSE) in Computer Science** (GPA: 3.82) *Accelerated, Expected May 2026*
- **Bachelor of Science in Engineering (BSE) in Computer Science** (GPA: 3.75) *Expected May 2026*
  ↪ **Minors** in Math, Data Science, and Engineering Entrepreneurship

Onsi Sawiris Scholar: Merit-based full-ride scholarship for Egyptian students to study at top universities in the United States
***Relevant Coursework**: Multiple AI/ML courses, Big Data Analysis, Scalable and Cloud Computing, and Advanced Algorithms*

RESEARCH

Master's Thesis: Linguistic Vagueness in Tool-Calling LLM Agents *Expected May 2026*

Advisor: Chris Callison-Burch • Committee: David S. Roos, Delip Rao, Eric Wong

Case studies in materials science and eukaryotic pathogen informatics, focused on how vague language shapes agent planning, tool selection, and failure modes. *Draft available upon request.*

- Introduced LAMINA, a **tool-augmented LLM agent** that turns open-ended scientific questions into executable workflows over Materials Project + CALPHAD (TDBs) and simulations (DFT/NEB, CHGNet) via Kani tool routing.
- Created a reproducible evaluation harness for materials science claims using **Quadratic Weighted Kappa with bootstrap CIs** to quantify reliability and failure modes.
- Currently building **PathFinder**: A conversational planner that converts natural-language intent into step-by-step strategy graphs for pathogen data exploration, improving transparency and reproducibility.

EXPERIENCE

AI/ML Engineer (Temporary Hire)

Fidelity Investments *Dec 2025 - Jan 2026*

- Rejoined to productionize and scale an internal R&D LLM/RAG system from my internship; expanded it to support **generic data sources** via a standardized ingestion template (connector → ingestion → indexing).
- Reworked retrieval to meet fast-update requirements by replacing the BM25S index with a **PostgreSQL Full-Text Search (inverted index) + `pgvector` hybrid**, cutting **index refresh latency from 10 minutes per 1M chunks to <1 second** while maintaining retrieval quality under evolving corpora.
- Implemented advanced retrieval (**multivector / late-interaction**) and improved vector-store scalability via **Qdrant sharding and load balancing**, increasing corpus capacity and expanding document coverage across sources.
- Backfilled historical embeddings into `pgvector` to unify legacy and new data; presented the system to the R&D department, fielded questions, and supported rollout. **Now used in production by the R&D team (∼ 200 users daily)**.

Data Science and Artificial Intelligence Intern

Fidelity Investments *Jun 2025 - Aug 2025*

- Architected a high-throughput ingestion pipeline processing **terabytes of internal data** with **99.99%+ ingestion success**, using validation, retries, and idempotent writes to support scalable downstream analytics and search.
- Built an LLM/RAG workflow using **Qdrant and `pgvector`** to ground answers in enterprise documents; added hallucination/grounding checks and evaluation metrics to track answer quality and failure modes.
- Containerized and deployed the pipeline and services with **Docker and Kubernetes**, enabling reproducible environments and reliable handoff for internal users and downstream teams.

Software Engineering Intern (AI Integration)

Aydi *Jun 2024 - Aug 2024*

- Built a chatbot for agricultural business owners in **low-connectivity environments** to access real-time farm analytics using Haystack, MongoDB Atlas, Redis, and GPT-4o family models, achieving **85%+ top-1 retrieval accuracy**.
- Implemented four Jest test suites for JavaScript services and added token/latency monitoring (tiktoken `o200k_base`) to guide performance and cost optimizations.
- Reduced LLM spend by **96.25%** by recommending a model migration from GPT-4o to GPT-4o mini based on cost/quality analysis and usage patterns.

Ahmed Muharram muharram.dev | ahmed.o.muharram@gmail.com

PROJECTS

ScenIQ – Enterprise AI Meeting Intelligence (Senior Project)

- Built an **on-prem**, **security-first** meeting intelligence product that turns meeting video into transcripts, topics, and structured action items via a multi-stage pipeline (FFmpeg → diarization → Whisper transcription → segmentation → extraction).
- Implemented a modular Python pipeline and FastAPI backend with PostgreSQL and `pgvector` to store transcripts, topics, and action items plus speaker/org-chart embeddings; exposed REST APIs / webhooks and integrated **Zoom/Teams** inputs with **ClickUp/Jira** outputs.
- Delivered a React + TypeScript (Vite) frontend for enterprise workflows; supported air-gapped and private-cloud deployment and reproducible evaluation tooling (runner/metrics/visualization), with runtime targets (**5-min video in ∼1–2 min**).

ByteNet – Distributed Web Search Engine (Crawler, Indexer, Ranking)

- Built an end-to-end **distributed search engine** on a custom-built Java stack (Webserver + distributed KVS + distributed RDD data-parallel framework), packaging services/jobs into JARs and orchestrating a **coordinator/worker cluster across AWS EC2 instances**.
- Implemented a production-grade **web crawler** at scale (∼ **500k URLs crawled**) with **robots.txt + crawl-delay compliance**, per-host **politeness/rate limiting**, **HEAD/GET** gating + content-type filtering, URL normalization, and **content-hash deduplication**; persisted crawl artifacts and failures in KVS tables to support recovery and iterative reruns.
- Designed the indexing + ranking pipeline as distributed RDD jobs: built an **inverted index** with token normalization (**stemming, stopword filtering, dictionary constraints**) and bounded document processing; computed **TF / IDF** tables and ran **iterative PageRank** over the link graph with convergence checks and dangling-link handling.
- Shipped a query backend + lightweight frontend with a JSON **/search** API and **infinite-scroll pagination (20 results/page)**; blended relevance via **TF-IDF + PageRank** scoring and served ranked results from precomputed KVS tables for fast query-time retrieval.

PennOS – UNIX-like OS Simulator (Kernel, Scheduler, FAT Filesystem, and Shell)

- Implemented a UNIX-style process model on top of **spthreads**, including a **PCB** design (PID/PPID, state machine, per-process FD tables) and a system-call interface (**s_** layer) backed by kernel helpers (**k_** layer) to maintain user/kernel abstraction boundaries.
- Built a **SIGALRM (setitimer) preemptive scheduler** with **100ms quanta** and **3-level priority queues** (weighted scheduling, round-robin within queues, starvation avoidance) plus proper **idling via sigsuspend** when all runnable queues are empty; added tick-level **structured logging** for scheduling + lifecycle events.
- Implemented **PennFAT (FAT16-style)** inside a host file using a memory-mapped **FAT region (mmap)** and disk-backed data blocks (lseek/read/write), supporting open/read/write/seek/unlink semantics with global FD table constraints and a shell with built-ins (`cat`, `echo`, `touch`...) + redirection.

InstaLite – Instagram Clone with Semantic Search

- Deployed a full-stack social platform on **AWS** using **EC2** and **RDS** with a relational schema (**12 tables**); implemented secure connectivity via EC2 tunneling.
- Built core product surfaces end-to-end: posts/images, likes/comments/replies, and hashtag feeds with **infinite scrolling**; added **semantic search** for profile discovery beyond simple keyword matching.
- Implemented real-time chat features with **WebSockets** (invites, messaging, notifications) and shipped a production deployment workflow focused on reliability and cost control.

AWARDS

- NASA Space Apps Cairo Special Mentorship Award from Nilepreneurs (**out of 115 teams and 859 participants**)

SKILLS

Technical Skills	Python, TypeScript/JavaScript, SQL, FastAPI, React, PostgreSQL/`pgvector`, Qdrant, Redis, MongoDB, Pandas, NumPy, C, Java, SQLite, MySQL/MariaDB
Industry Skills	Docker, Kubernetes, AWS (EC2, RDS), Linux/Shell, Git, REST APIs, Webhooks, WebSockets, Apache Spark, Haystack, LlamaIndex, FFmpeg, Whisper
Languages	Native proficiency in Arabic and English; limited German and Italian