

Omar Mohammad Al-Saleh

Ph.D. Student | LLM Systems & AI Infrastructure

LinkedIn | Portfolio | oalsaleh@uwaterloo.ca | GitHub

EDUCATION

University of Waterloo

Jan 2026 – Present

Ph.D. in Computer Science

- **WASL Lab** | Supervisor: Prof. Samer Al-Kiswany | GPA: 95.25/100 (A+)

Research focus: LLM serving systems, scalable inference, KV-cache management, workload reconstruction, and high-performance AI infrastructure.

Jordan University of Science and Technology

Sep 2020 – Jun 2025

B.Sc. in Computer Engineering

- GPA: 3.94/4.00 — Excellent with Honors; top of graduating class.

RESEARCH EXPERIENCE

TraceWeaver — Cache-Faithful LLM Workload Reconstruction

Jan 2026 – Present

- Developing a privacy-preserving trace-reconstruction framework that converts anonymized LLM serving traces into synthetic prompts while preserving workload structure and KV-cache sharing behavior.
- Designed replay modes from semantic reconstruction to exact-token-oracle replay, using deterministic synthetic block identities to reproduce repeated cache blocks without original prompt text or token IDs.
- Built trace-matching/replay pipelines that preserve input/output characteristics, prefix-sharing topology, cache depth, and cached-token fraction across large traces and long-context workloads.
- Evaluate reconstructed workloads on GPU hardware with vLLM using TTFT, inter-token latency, throughput, and prefix-cache behavior; tooling includes Python, Linux, SLURM, Hugging Face tokenizers, and GPU clusters.

TECHNICAL SKILLS

LLM/Systems: vLLM, KV-cache & prefix caching, trace-driven replay, LLM inference evaluation, Linux, SLURM, GPU clusters

Programming: Python, C/C++, C#, JavaScript/TypeScript, Verilog, SQL

Frameworks/Tools: .NET Core, Angular, React, Node.js, SQL Server, MongoDB, Git/GitHub

PUBLICATION

Applying AI and Machine Learning to Enhance Automated Cybersecurity and Network Threat Identification

Oct 2024

The 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2024) — Leuven, Belgium

- Co-authored and presented work on AI/ML techniques for cyber-threat detection in distributed networks and deployment challenges including adversarial attacks, data imbalance, and resource constraints.

EXPERIENCE

Teaching Assistant — CS 350 Operating Systems / CS 106 Intro to Computer Science 2

Jan 2026 – Present

University of Waterloo

- Supported Operating Systems and introductory CS courses through technical consulting, debugging, grading, and explanations of concurrency, synchronization, and core programming concepts.

Software Developer — Anazeem Information Technology

Jun – Dec 2025

- Developed and maintained enterprise ERP functionality across the full stack using Angular, .NET Core REST APIs, and SQL Server, working across frontend workflows, backend services, and database logic.
- Owned production-facing concerns across the application environment, including authentication and role-based authorization, API/data security, performance optimization, debugging, and end-to-end integration.

Software Developer Intern — Optimiza

Jul – Oct 2024

- Developed full-stack services with .NET Core, React, Entity Framework Core, and SQL Server, including REST APIs, authentication/security, database logic, and Python workflow automation.

PROJECTS & HONORS

FlowBot — No-Code Workflow-Based Chatbot Builder

2025

Best Graduation Project Award — The 16th International Conference on Information and Communication Systems (ICICS 2025)

- Built an end-to-end drag-and-drop chatbot platform using React, .NET Core, SQL Server, authentication/session management, and Levenshtein-distance-based response matching.

MIPS Pipelined CPU with Stack Memory

2023

Top 5 Finalist — National Semiconductor Design Competition (JoSDC'23)

- Designed a five-stage MIPS CPU in Verilog with stack memory, forwarding/stalling, branch prediction, and test benches; ranked among top teams in a 100+ participant national competition.