Zohaib Khan

Ann Arbor, MI | (571) 524-4448 | zohaib.khan3502@gmail.com | Personal Website/Blog | Google Scholar | GitHub

EDUCATION

University of Michigan, Ann Arbor *Ann Arbor, MI, USA*

Master of Science in Information - CGPA: 4.00/4.00 Expected: June 2027

- **UMSI Achievement Fellowship Award Recipient (100% Funded)**

Lahore University of Management Sciences *Lahore, Pakistan*

Bachelor of Science in Computer Science – CGPA: 3.82/4.00 June 2024

- **Head Teaching Assistant** for Machine Learning, Speech & Language Processing, Advanced Topics in ML

EXPERIENCE

Research Engineer — University of Michigan, Ann Arbor *July 2025 – Present*

- Designed a new coding RL environment around the Countdown game to investigate LLM misalignment, adapting to the standard Prime Intellect environments and verifiers frameworks.
- Executed GRPO and SFT fine-tuning runs for Qwen and LLaMA LLMs using verl and llama-factory, integrating LoRA adapters to cut training cost while boosting **code reasoning accuracy by 90%** over baseline.
- Automated fast creation of **20000+ instruction-response pairs** using the asynchronous OpenAI API for distillation.
- Developed AST- and regex-based validators for automatic code extraction, sandboxed execution, and reward scoring.
- Published at **ICLR SPOT 2026.**
- Finetuned LLMs to model human belief data mined from large-scale surveys (120000+ respondents), **boosting accuracy scores by 17% on MCQ-style questions**.
- Implemented a custom finetuning engine with HuggingFace accelerate and peft to extract and optimize first-token probabilities for MCQ answering tasks, logging metrics remotely with wandb.
- Currently investigating **multi-agent systems with memory** for long-term planning.

ML Research Engineer — Fatima Fellowship *Sep 2024 – Present*

- Optimized multi-GPU inference stack for LLMs (up to 72B parameters) using vllm, achieving **4000+ tokens/sec** throughput and enabling high-throughput safety evaluation pipelines.
- Designed asynchronous RAG pipelines with Tavily Search and OpenAI APIs, **reducing redundant completions by 63%** through adaptive retrieval and context filtering.
- Integrated Llama-Guard models for automated safety filtering of LLM outputs and benchmarked detection accuracy via A/B tests against fine-tuned BERT classifiers.

Machine Learning Engineer — ISSM.ai *May 2023 – Sep 2023*

- Built a high-throughput OCR pipeline combining text detection and QR-code extraction to process 10+ invoice formats, **reducing manual data entry by 80%**.
- Trained custom PyTorch text-detection and recognition models on **15K+ annotated images**, increasing recognition **robustness across new invoice templates by 3x**.
- Optimized model inference on Intel CPUs with ONNX + OpenVINO, **achieving 5x lower latency** with no loss in accuracy, enabling real-time processing on client systems.

Machine Learning Intern — CodeSlash *June 2022 – March 2023*

- Built an object-tracking pipeline for satellite imagery using YOLOv8 + StrongSORT, **increasing multi-object tracking accuracy by 20%** for large herds in wide-area scenes.
- Deployed an optimized face-recognition system with OpenVINO, FastAPI, Docker, and MongoDB, **cutting CPU inference latency by 3x**.

PROJECTS

Accelerated and Distributed Transformer Training | *Python, PyTorch, Triton*

- Reproduced LLaMA-3 from scratch in PyTorch, building a custom distributed training stack with DDP, Flash-Attention-2 Triton kernels, mixed-precision casting and checkpointing utilities to optimize efficiency.

Multilingual Fact-checking in LLMs | *Python, vllm,*

- Built a large-scale data pipeline that processed 300K+ multilingual claims using Google Fact Check and asynchronous OpenAI APIs to evaluate LLM accuracy, safety, and consistency.
- Automated scraping, preprocessing, and inference with async batching, caching, and rate-limiting, enabling high-throughput analysis and significantly reducing API cost.
- Published work at **Nature Scientific Reports**.

Reasoning on a Budget | *Python, verl, vllm*

- Conducted ablative tests across multiple LLMs and LoRA ranks under strict compute constraints (24h on a single A40) for mathematical reasoning, boosting performance on competition benchmarks for under $7 of compute.
- Published work at **EACL SRW 2026**.

TECHNICAL SKILLS

Languages and Developer Tools: Python, C/C++, CUDA, Golang, Linux, Docker, Git, AWS, GCP, Slurm/HPC, nsight
Machine Learning: PyTorch, TensorFlow/Keras, JAX, Keras, Transformers, Triton, verl, trl, LlamaFactory, vllm, peft
Web Development: React, React Native, Node, Express, FastAPI, Flask, Django