Zahra Golpayegani

ML Engineer
g.golpa75@gmail.com
✉ ☎
https://zahragolpa.github.io/

About Me

I am a machine learning engineer with 3+ years of experience **scaling and optimizing** large language models for **efficient training and deployment**. I specialize in **hardware-aware performance tuning**, **parallelism strategies**, and building robust ML systems a**cross the full stack**.

Skills

Programming Languages: Python, C/C++, CUDA
Deep Learning Frameworks: PyTorch, **Megatron-LM**, **DeepSpeed**, **HuggingFace**, **LLaMA Factory**, **Accelerate**, **VeRL**, **vLLM**
MLOps & DevOps: Docker, LangChain, MLflow, Git
Other: Linux, LaTeX, Technical Writing

Experience

Huawei Canada – *Senior ML Engineer* Sep 2024 – Present
Tools: PyTorch, DeepSpeed, Huawei Ascend SDK, HuggingFace, Megatron-LM, LLaMA Factory, vLLM, VeRL

- Integrated RL training recipes with multi-turn tool calls on the VeRL platform into the CANN-recipes-train repository (PR #101).
- Implemented Attention-FFN Disaggregation (AFD) for SOTA LLMs to improve serving efficiency.
- Diagnosed hardware bottlenecks on **Huawei Ascend chips** and applied operator-level optimizations to enhance compute efficiency.
- Helped develop an internal simulation library to evaluate parallelism paradigms, reducing trial-and-error in LLM deployment.
- Benchmarked parallelization strategies—including **tensor, expert, context, sequence, and data parallelism**—for models like **DeepSeek-NSA** and **MiniMax-Text-01**, quantifying memory, communication, and throughput impacts.
- Accelerated pre-training of **LLaMA-based LLMs** using conditional computation, cutting training time and cost by **up to 16%** without significant accuracy loss.
- Designed and implemented an end-to-end **Agentic AI system** for **video generation**.
- Conducted long-context evaluations of **hybrid LLMs** to study the effect of architecture on performance.
- Co-authored a paper (ETT) proposing a test-time training algorithm that extends GPT-Large and Phi-2 context up to 32×, improving accuracy by **30%**.

Zetane Systems – *AI Engineer* Dec 2021 – Aug 2024
Tools: Python, LangChain, OpenAI API, PyTorch, Docker, Git, MLflow

- Co-designed and developed ZetaForge, an AI platform automating prompt-to-pipeline generation using LLMs.
- Developed complex end-to-end multi-agent AI pipelines leveraging state-of-the-art libraries such as **LangChain** and **openai-python** (blog post).

- Implemented a robust **MLOps pipeline** (end-to-end machine learning operations workflow) featuring data and model validation, preprocessing, model training with **automated experiment tracking**, and continuous model improvement (open-source project).
- Contributed to flagship products including Protector—a platform for assessing vision model reliability in real-world conditions.
- Delivered product demos and technical talks to prospective clients and international audiences at events including Scale AI, World Summit AI, and ALL IN.

XAI Lab, Concordia University – *Research Assistant* Sep 2021 – Jan 2024
Tools: Python, PyTorch, OpenCV, MATLAB, LaTeX

- Investigated the interplay between **model robustness, accuracy, and shape bias** under Professor Nizar Bouguila's supervision; published findings at **CRV 2023**.
- Proposed a novel image compression algorithm inspired by **Singular Value Decomposition** (SVD) to optimize robustness of vision models trained on compressed data; work published at **ICPRAM 2024**.

Education

Concordia University, MASc Information Systems Engineering 2024
GPA: 4.0 / 4.0
Thesis: Enhancing Deep Learning Model Robustness: Insights from Out of Distribution Data Augmentation and an Innovative Image Compression Technique
Supervisor: Prof. Nizar Bouguila
Amirkabir University of Technology, BSc Computer Science 2020
GPA: 3.7 / 4.0
Final Project: Food9K: Detecting Food in Social Media Images Using YOLOv3
Supervisor: Prof. Mohammad Akbari

Publications

- ETT: Expanding the Long Context Understanding Capability of LLMs at Test-Time. 2025. arXiv.
- PatchSVD: A Non-Uniform SVD-Based Image Compression Algorithm. 2024. ICPRAM.
- Clarifying myths about the relationship between shape bias, accuracy, and robustness. 2023. CRV.

Awards

- **Campaign Hero Award**, Huawei *2025*
  Honored for delivering rapid, high-impact results on a key internal campaign at Huawei HQ.
- **Future Star Award**, Huawei *2025*
  Recognized for outstanding performance and potential, highlighting my contributions.
- **Best Paper Award Candidate**, ICPRAM *2024*
  Nominated for research on model compression and robustness.
- **Conference and Exposition Award**, Concordia University *2023*
  Awarded for excellence in presenting graduate-level research.
- **Accelerate Explore Award**, Mitacs *2021*
  For conducting applied AI research in collaboration with industry partners.
- **Merit Scholarship Entrance Award**, Concordia University *2021*
  Awarded for high academic achievement upon admission to MASc program.
- **2nd Place, Worldwide**, RoboCup Junior Soccer League *2013*
  RoboCup World Competitions – international robotics tournament.