Esmail Gumaan

LinkedIn | +967 772945392 | esm.agumaan@gmail.com | Github | Website

TECHNICAL SKILL

Programming: Python, CUDA/C++, JavaScript, Java, HTML/CSS

AI, Machine Learning: Deep Learning, Machine Learning, LLMs, RAG Systems

Frameworks & Libraries: PyTorch, TensorFlow, HuggingFace, PyTorch Lightning, LangChain, Haystack

Data Science: Numpy, Pandas, Colab

Tools, Platforms: GitHub, Git, Databases, Linux OS

PROFESSIONAL SKILL

Problem Solving, Research & Implementation, Software Design Principles (SOLID), Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) Systems

EDUCATION

Sana'a University — Dec 2023 - Feb 2025

- Bachelor of Computer Science

WORK EXPERIENCE

AI/ML Software Engineer — Nov 2024 - Mar 2025

AI Research Collaborator-Remote , Germany

- Worked on AI projects related to Retrieval-Augmented Generation (RAG) systems.
- Contributed to designing and optimizing retrieval mechanisms for improving LLM responses.
- Developed and tested models for enhancing document retrieval
- Developed a Retrieval-Augmented Generation (RAG) system for medical applications, integrating Optical Character Recognition (OCR) and Large Language Models (LLMs) to enhance MRI detection capabilities, with the goal of creating an AI-powered assistant for medical professionals.

ML Engineer — Dec 2024 - Mar 2025

Automa8e-Remote , Philippines

- Developed AI-powered applications for generating dynamic responses and automating actionable insights extraction from documents using advanced NLP and document-processing techniques.
- Spearheaded the development of agentic Retrieval-Augmented Generation (RAG) systems,
- optimizing embedding models for enhanced accuracy and efficiency.
- Independently designed and executed end-to-end AI projects, demonstrating strong self-direction and problem-solving skills in solo environments.

AI/ML Software Engineer Jan, 2024 - Nov, 2024

Creative-Point , Sana'a

- Worked 12-hours per day to build **AI models** completely from scratch.
- Trained algorithms to meet high-performance standards and solve specific challenges.
- Played a key role in **developing RAG (Retrieval-Augmented Generation) systems** to enhance AI-powered information retrieval and response generation.
- Designed and implemented **chatbots** tailored to meet the unique needs of clients or projects.

AI/ML Researcher May, 2023 - Agu, 2023

Sana'a University, Sana'a

- Collaborated with my professor to implement foundational research papers from scratch, including the 'Attention Is All You Need' transformer architecture and diffusion models, with a focus on advancing Large Language Models (LLMs) and their real-world applications.
- Contributed to the development and fine-tuning of diffusion models, exploring their potential in generative AI tasks.
- Worked extensively on distributed training techniques, optimizing the performance of machine learning models across multiple systems.
- Conducted in-depth research to push the boundaries of LLM capabilities, from architecture enhancements to dataset curation strategies.
- Built diverse Retrieval-Augmented Generation (RAG) architectures, including agentic (autonomous agent-driven), graph-based, self-correcting, and adaptive RAG systems, designed to optimize dynamic knowledge retrieval and context-aware decision-making.

PROJECTS

Galvatron: Automatic Distributed Training for Large Transformer Models Jan 2025 - Mar 2025

- **Project aim:**
  - Engineered Galvatron, a distributed training framework for optimizing large transformer models (e.g., BERT, GPT) across multi-GPU environments. The project aimed to address scalability and efficiency challenges in training massive neural networks by integrating tensor parallelism, pipeline parallelism, and heterogeneous memory management.

- **Project Outcome:**
  - Achieved ~30% faster training speeds compared to baseline distributed setups by implementing hybrid parallelism strategies (tensor + pipeline parallelism) and optimizing GPU communication.
  - Reduced GPU memory usage by 25% through dynamic memory allocation and gradient checkpointing, enabling training of models with 1B+ parameters on commodity hardware.
  - Designed a modular pipeline for seamless integration with PyTorch and Hugging Face Transformers, supporting flexible configurations for data, model, and pipeline parallelism.
  - Validated scalability by training BERT-Large on 8 GPUs with near-linear speedup, achieving 90% hardware utilization efficiency.
  - Open-sourced the framework, providing detailed documentation and benchmarks for researchers and engineers working on large-scale NLP/LLM training.

Nanograd: Graudation Project

Jan 2024 -Nov 2024

- **Project aim:**
  - Develop nanograd, a comprehensive Engine or toolkit for building and fine-tuning Generative Pre-trained Transformer (GPT) models from scratch using Python. Implement cutting-edge Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA and adapters, along with optimizations like quantization and FlashAttention. Gain expertise in state-of-the-art methodologies for enhancing performance, including sentiment analysis with Proximal Policy Optimization (PPO). Achieve a deep understanding of the entire pipeline, from tokenization to model training, optimized for CUDA.

- **Project outcome:**
  - Successfully built and fine-tuned a GPT model using various PEFT techniques, achieving high performance and efficiency. Conducted detailed performance evaluations using relevant metrics to assess model quality. Experimented with hyperparameters and architectural modifications to optimize the model further. Produced a well-documented repository with comprehensive code and documentation, demonstrating proficiency in Python and advanced deep learning techniques for natural language processing. The project serves as a valuable resource for developing and fine-tuning large language models.

Axon: AI reseach Lab

May 2024 - Jun 2024

- **Project aim:**
  - Establish Axon as an AI research lab and collaborative platform for implementing cutting-edge AI research papers and conducting novel research across various AI domains. Focus on bridging the gap between theoretical research and practical applications by providing high-quality, reproducible implementations of seminal and contemporary AI models such as InstructGPT, LLaMA, transformers, diffusion models, and Reinforcement Learning from Human Feedback (RLHF).

- **Project Outcome:**
  - Successfully developed and shared high-quality, reproducible implementations of advanced AI models. Conducted extensive evaluations to ensure the fidelity and performance of these implementations. Fostered a collaborative environment for ongoing research and development in AI. Produced well-documented code and comprehensive research papers, demonstrating proficiency in state-of-the-art AI techniques. The project serves as a valuable resource for both researchers and practitioners, advancing the practical application of AI research.

LITRA model Oct 2023 - Dec 2023

- **Project aim:**
  - Develop the LITRA-Model, a specialized language-to-language transformer model designed for translation tasks, using PyTorch. Implement the architecture based on the "Attention is All You Need" paper, focusing on accurate and contextually rich translations between Arabic and English. Gain a deep understanding of transformer models, attention mechanisms, and their application in machine translation.

- **Project Outcome:**
  - Successfully built and trained a language-to-language transformer model capable of translating between Arabic and English with high semantic fidelity and contextual awareness. Conducted comprehensive evaluations, achieving notable BLEU scores. Produced a well-documented repository with detailed code and clear documentation, demonstrating proficiency in PyTorch and transformer-based model implementation. The project advances Arabic-to-English translation capabilities and provides a robust framework for further research and development in machine translation.

CERTIFICATES

- **Machine Learning Specialization from Stanford University.**
- **Introduction to Artificial Intelligence from IBM.**