Esmail Gumaan

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

TECHNICAL SKILL

Python
Deep Learning
machine learning
C/C++
JavaScript
HTML/CSS
Clean Code
Pytroch framework
Numpy
Pandas
MySQL
Git
Linux OS
Data Structures and alogrithms

PROFESSIONAL SKILL

Problem-Solving
Research skills
Solid Principles

EDUCATION

Sana'a University — Dec 2022 - Present

- Bachelor of Computer Science

WORK EXPERIENCE

AI/ML Software Engineer — Jun, 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 on cutting-edge research focused on Large Language Models (LLMs) and their practical 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.

PROJECTS

TransformersFactory

Nov 2024 - Present

- **Project aim:**
  - Develop a Framework for building and fine-tuning Transformer models with the ability of integrating multiple pre-trained models in the same interface, the user can build his own transformer model without writing a single line of code and providing an API to call models easily.

- **Project Outcome:**
  - By the conclusion of this project, we will have a fully developed, user-friendly framework that enables both technical and non-technical users to seamlessly create, integrate, and fine-tune Transformer-based deep learning models without the need for writing code. This framework will include a unified graphical interface to select from multiple pre-trained Transformer architectures—such as BERT, GPT, or RoBERTa—apply domain-specific fine-tuning routines (e.g., adjusting hyperparameters, training on custom data sets, and running validation tests), and then deploy the resulting model via a standardized API endpoint.

Nanograd

June 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 - June 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.**