Esmail Gumaan

esmail.agumaan@gmail.com +967-772-945-392 esmail a.gumaan Esmail-Ibraheem
Google Scholar Portfolio

About Me

AI Research Engineer focused on neural networks, large language models, and scalable ML systems. Experienced in PyTorch and CUDA, with strong contributions to open-source research bridging theory and applications.

Education

University of Sana'a *Jan 2023 – Mar 2025*
BS in Computer Science

- GPA: 79.72% ≈ 2.2/5.0 (Gut)
- **Coursework:** Artificial Intelligence, Data Science, Data Mining, Advanced Programming

Experience

ML Engineer *Philippines*
Automa8e *Dec 2024 – Jan 2025*

- Developed AI-powered applications for generating dynamic responses and automating actionable insights extraction from documents using advanced NLP and document processing techniques

AI/ML Software Engineer *Germany*
AI Development Collaborator *Nov 2023 – Aug 2024*

- Worked on AI projects related to Retrieval-Augmented Generation (RAG) systems
- 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

AI Engineer *Sana'a*
Creative Point *Jan 2024 – Nov 2024*

- Worked 12-hours per day to build AI models completely from scratch
- Trained algorithms to meet high-performance standards and solve specific challenges

AI/ML Student Researcher *Sana'a*
Sana'a University *May 2023 – Aug 2023*

- Collaborated with my professor to implement key research papers, including the Transformer (Attention Is All You Need) and diffusion models, focusing on Large Language Models (LLMs) and applications
- Contributed to the development and fine-tuning of diffusion models for generative AI tasks
- Built diverse Retrieval-Augmented Generation (RAG) systems—agentic, graph-based, and self-correcting

Publications

Theoretical Foundations and Mitigation of Hallucination in Large Language Models arXiv: 2507.22915) July 2025

Universal Approximation Theorem for a Single-Layer Transformer arXiv: 2507.10581) July 2025

Mixture of Transformers: Macro-Level Gating for Sparse Activation in Large Language Model Ensembles ResearchGate: RG.2.2.25049.02400) April 2025

ExpertRAG: Efficient RAG with Mixture of Experts – Optimizing Context Retrieval for Adaptive LLM Responses arXiv: 2505.08744) Mar 2025

Galvatron: Automatic Distributed Training for Large Transformer Models arXiv: 2505.03662) Mar 2025

Projects

nanograd: AI Engine *github.com/Esmail/nanograd*

- anograd ML/DL and neural net ecosystem, run models like GPT, llama, stable diffusion, vision transformer, reinforcement learning, autotrainer, your Unreal Engine, but for AI, essentially making it an AI engine or an AI Ecosystem
- Tools Used: Python, PyTorch, Gradio, FastAPI

Axon: AI Research Lab *github.com/Esmail/Axon*

- 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 highquality, reproducible implementations of seminal and contemporary AI models such as InstructGPT, LLaMA, transformers, diffusion models, and Reinforcement Learning from Human Feedback (RLHF)
- Tools Used: Python, PyTorch, JavaScript

TinyLlamas *github.com/Esmail/Tinyllamas*

- Tinyllamas is an advanced language model framework, inspired by the original Llama model but enhanced with additional features such as Grouped Query Attention (GQA), Multi-Head Attention (MHA), and more. This project aims to provide a flexible and extensible platform for experimenting with various attention mechanisms and building state-of-the-art natural language processing models.
- Tools Used: Python, PyTorch, JavaScript, Streamlit, Gradio, HuggingFace, PyTorch-Lightning

Certificates And Rewards

- Machine Learning Specialization by Stanford and Deeplearning.AI
- Introduction to Artificial Intelligence by IBM
- Generative AI by Google Cloud
- AI Exhibition Projects Competitions by Sana'a University
- Introduction to Artificial Intelligence and Data Science by Sana'a University
- SmartX Theses Competitions for undergraduates

Skills and Technologies

Programming: Python, C++, Java, SQL, JavaScript, CUDA, LaTex, PyTorch, TensorFlow, PyTorch Lightning, LangChain, Haystack, Colab, GitHub, Git, HuggingFace, Databases, Linux OS

Mathematics: Good understanding of differential equations, calculus, and linear algebra
Languages: English (B2, IELTS: 6.0/9.0), Arabic (native)