Esmail Gumaan

Sana'a-Yemen esmail.agumaan@gmail.com +967-772-945-392 esmail-ibraheem.github.io
esmail a.gumaan Esmail-Ibraheem

About Me

AI Research Engineer with a deep passion for neural networks, large language models, and cutting-edge machine learning techniques. Skilled in designing and optimizing scalable systems with a strong foundation in PyTorch and CUDA. Committed to open-source development, multilingual AI accessibility, and building intuitive tools that bridge the gap between research and real-world applications.

Education

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

- GPA: 2.7/4.0 (Very Good)
- **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
- optimizing embedding models for enhanced accuracy and efficiency
- Spearheaded the development of agentic Retrieval-Augmented Generation (RAG) systems

AI/ML Software Engineer *Germany*
AI Research 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, with the goal of creating an AI-powered assistant for medical professionals
- Developed and tested models for enhancing document retrieval
- Created a test case generation tool that creates random XML docs from XML Schema
- Automated the extraction and processing of large datasets from legacy systems using SQL and Perl scripts

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
- Designed and implemented Chatbots tailored to meet the unique needs of clients or projects
- Independently designed and executed end-to-end AI projects, demonstrating strong self-direction and problem-solving skills in solo environments

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

- 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
- 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

Publications

ExpertRAG: Efficient RAG with Mixture of Experts – Optimizing Context Retrieval for Adaptive LLM Responses arXiv: 2504.08744) — 23 Mar 2025

Galvatron: Automatic Distributed Training for Large Transformer Models arXiv: 2504.03662) — 13 Mar 2025

MoT: Mixture of Transformers – An Adaptive Gating Framework for Efficient LLM Ensembles — Present 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

Technologies

Languages: Python, C++, C, Java, HTML/CSS, C#, SQL, JavaScript, CUDA/C++

Technologies: PyTorch, TensorFlow, PyTorch Lightning, LangChain, Haystack, Colab
Tools, Platforms: GitHub, Git, HuggingFace, Databases, Linux OS