I GUSTI BAGUS RAMADHA SAVERIAN RANUH

Jakarta | +62 81393131312 | i.ranuh001@binus.ac.id | https://www.linkedin.com/in/rama-ranuh/ | github.com/RamadhaRanuh | ramadharanuh.github.io/Portfolio

SUMMARY

AI Engineer with a strong Computer Science background specializing in Large Language Models (LLMs) and high-performance inference optimization. Proficient in deploying scalable AI solutions using Python, JavaScript, and C++. Applied concepts such as MCP, A2A, and model fine-tuning. Expertise includes leveraging frameworks such as vLLM and SGLang for high-performance inference, as well as LangChain, LlamaIndex, and HuggingFace for orchestrating LLM applications. Capable of technical problem-solving and collaborative initiatives. Dedicated to contributing my capabilities in predictive modeling, system optimization, and data-driven decision-making to help organizations achieve their strategic objectives.

EDUCATION

BINUS UNIVERSITY – Jakarta, Indonesia September 2022 – March 2026
Bachelor of Computer Science - GPA 3.93/4.00

ORGANIZATION

Data Science Club – Jakarta, Indonesia September 2024 – July 2025
Member at BINUS University
- Actively collaborate with fellow members, participate in workshops, and engage in discussions to enhance skills in machine learning, data analysis, and research methodologies.
- Lead technical discussions on emerging AI trends, breaking down complex research papers into actionable insights for fellow members.

UREEKA BINUS – Jakarta, Indonesia April 2024 – April 2025
Member at UREEKA BINUS University
- Engages in data mining training through UREEKA, collaborating in teams to compete in various data competitions.
- Develops end-to-end data pipelines for competition submissions using classic ML solutions, utilizing feature engineering and hyperparameter tuning to improve model performance metrics.

EXPERIENCE

AI Engineer Intern at GDP Labs – Jakarta, Indonesia February 2025 – February 2026
AI Engineer Internship at GDP Labs
- Collaborates in a rapid environment, developing prototype emerging State-of-the-Art (SOTA) technologies, applying critical problem-solving to resolve architectural blockers alongside cross-functional teams.
- Focuses on optimizing LLM inference to support high-concurrency user workloads, achieving a 1.8× increase in throughput and up to 70% faster latency by migrating to the SGLang framework.
- Streamline deployment workflows by packaging inference services with Docker Compose, enabling efficient integration with production systems and simplifying environment reproducibility.
- Integrates a new relevancy metric in the internal evaluation SDK, solving the issue of subjective queries in production, automating evaluation workflows, and enabling structured and repeatable assessments across multiple datasets.

Data Science Dicoding Bootcamp – Online October 2024 – February 2025
Enrolled in Dicoding Bootcamp Online
- Complete 600+ hours of an intensive data science and ML program for 6 months, designed to equip digital talents with industry-standard skills.
- Develop multiple hands-on projects with final capstone projects that apply data science and machine learning techniques to real-world problems.

Freshmen Partner – Semarang, Indonesia September 2023 – July 2024
Freshmen Partner at BINUS @Semarang
- Guides freshmen by sharing essential experiences and materials to help them navigate first-year university life with focus and enjoyment.
- Facilitate weekly discussion and sharing sessions, fostering an inclusive environment that encourages open communication, peer support, and collaborative problem-solving.

Tutor – Semarang, Indonesia September 2023 – January 2024
Tutor at BINUS @Semarang
- Taught two subjects: Algorithm & Programming and Discrete Math to first-semester students.
- Voluntarily conducted two classes weekly, each session consistently attracted over 10-12 students, to assist students struggling with course material, helping them improve their understanding and achieve their academic goals in the subjects.

Applied Mathematics Exchange Student at CYCU – Online February 2023 – July 2023
Grade: A-
- Opportunity to participate in an online semester exchange program at Chung Yuan University (CYCU), studying foundational mathematical models and applied mathematics, strengthening quantitative analysis and critical thinking.

- Gain exposure to an international academic environment early in my academic career.

ACHIEVEMENTS

Researcher at ICCSCI – Jakarta, Indonesia August 2024

Researcher at ICCSCI held by Procedia Computer Science and Elsevier

- Conduct a research paper presented at the International Conference on Computer Science and Computational Intelligence (ICCSCI) with a Q2 index, focusing on a comparative analysis of deep learning algorithms for the classification of hyperpigmented skin diseases. The study explored performance metrics across various models to optimize diagnostic accuracy.

10th National Finalist ADSE Competition – Jakarta, Indonesia April 2024 – August 2024

National Finalist ASEAN Data Scientist Explorers Competition Issued by ASEAN Foundation and SAP

- Developing a pioneering data-driven application and solution for the 2024 ASEAN Data Scientist Explorers Competition, which focuses on “Enhancing Waste Management through Integrated Waste Banks and Gamification Incentives,” aims to advance sustainable waste management practices across ASEAN countries.

20th Finalist SOCS AI For Accessibility Hackathon – Jakarta, Indonesia April 2024

Finalist School of Computer Science Hackathon 2024 BINUS x Microsoft (AI4A)

- Builds a mobile web application designed to assist those with vision impairments in identifying objects and text from a distance or from areas inaccessible with a white cane (upper body). making use of Microsoft Azure features, including Speech Service, Translator, and Computer Vision. Everything will be spoken using text-to-speech.

SKILLS

- **Programming Language :** Python, JavaScript, C/C++
- **Frameworks:** SGLang, HuggingFace, LangChain, Streamlit, DeepEvals
- **Database Management System:** MySQL, MongoDB, Supabase
- **Data Visualization:** Tableau, SAP Analytics Cloud
- **Deep Learning:** PyTorch, TensorFlow
- **DevOps:** Git/GitHub, Docker
- **Data Science & Miscellaneous Technology:** A/B Testing, ETL, Data Pipeline, Statistics, Time series, Experimental design, Hypothesis testing, OOP, APIs
- Natural Language Processing (NLP)
- Computer Vision

PROJECTS

NLP Multimodal RAG-Based Chatbot Model

Python, LLM, RAG, Llama CCP, Llama Index, Gradio

Implemented an AI chatbot using Large Language Models (LLM) and Retrieval Augmented Generation (RAG) techniques to provide detailed answers based on knowledge extracted from PDF documents, integrating LlamaCPP for LLM functionality and llama_index for effective document management. The system leverages medical knowledge from PDF documents to provide enhanced natural language responses and advanced information retrieval.

Optimize LLM Inference using SGLang

Python, SGLang, CUDA, TensorFlow, Keras, Librosa

Evaluated and optimized LLM inference performance using SGLang as the primary inference server, conducting a comparative study against vLLM under high concurrent workloads. The scope includes systematic experimentation and hyperparameter tuning across optimization strategies such as quantization, data parallelism, and tensor parallelism. The resulting analysis provided actionable insights on production trade-offs between throughput, latency, memory efficiency, and model fidelity, enabling multiple internal AI teams to make informed deployment decisions.

SafeSight

Bootstrap, JavaScript, Node.Js, Express.Js, Docker, Cloudflare, Microsoft Azure

A mobile web application designed to assist visually impaired person in detecting objects and recognizing text in areas beyond their reach, where a white cane cannot be used for the upper body. It uses Microsoft Azure services, including Computer Vision, Translator, and Speech Services, to provide real-time verbal feedback through text-to-speech, ensuring seamless accessibility and navigation.

Deep Learning Algorithms for Image-Based Classification of Hyperpigmented Skin Disease

Python, YOLO, DenseNet201, GoogleNet, InceptionResNetV2, MobileNet

Analyze and compare hyperpigmented skin disease using machine learning and deep learning techniques. The study analyzed pretrained models like YOLO, DenseNet201, GoogLeNet, InceptionResNetV2, and MobileNet. The study highlights DenseNet201 as the best-performing model for the accurate classification of hyperpigmented skin conditions, based on both accuracy and AUC. However, YOLO was ultimately chosen due to its effective object detection capability.

Speech Emotion Recognition System

Python, CNN, TensorFlow, Keras, Librosa

Developed a Speech Emotion Recognition system using Convolutional Neural Networks (CNN) with TensorFlow. The project involved training a model to recognize emotions from speech input, utilizing audio features for accurate emotion classification.

Multi-Agent Synthetic Data Pipeline & LLM Benchmarking

Python, LangChain, OpenAI, HuggingFace

Engineered an autonomous multi-agent orchestration framework designed to manage the end-to-end lifecycle of synthetic data generation, from raw text retrieval and processing to validation and annotation. Develop a synthetic dataset that focuses on a novel causal reasoning benchmark for the Indonesian language, utilizing LLM agents to generate culturally nuanced samples across standard and colloquial dialects to address low-resource language scarcity. Executed benchmarking, revealing significant performance degradation compared to human-curated baselines and validating the dataset's rigor.

Agentic AI with automated evaluator

MongoDB, Express.JS, React, Node.JS

This project will provide businesses with a more effective means of gaining insight into their financial data. Users can track and analyze, on this dashboard developed with MongoDB, Express.js, React, and Node.js, key financial indicators on profit, revenue, operational and non-operational expenses, and product pricing against their respective targets. Further, predictive analytics through regression modeling enables one to predict revenue with accuracy so that businesses can make strategic decisions and optimize their financial results accordingly.

Patient's Condition Classification Using Drug Reviews

Python, XGBoost, PAC, Logistic Regression, KNN, SVM, Random Forest, Streamlit

Analyze the disease description to predict the specific medical condition or diagnosis using multiple advanced machine learning algorithms (XGBoost, PAC, LR, and more). The project also applies text processing techniques, including CountVectorizer, TFIDFVectorizer, and Word2Vec, to analyze user reviews from a dataset. The project then deploys in Streamlit.

Youtube Scraping Sentiment Analysis

Python, Logistic Regression, LSTM, Google Cloud

Developed a sentiment analysis project by scraping YouTube comments with the Google Cloud YouTube API v3 and applying three models: Random Forest with TF-IDF, Logistic Regression with TF-IDF, and LSTM with Word2Vec. Achieved an average train-test accuracy of 85-91%.

CERTIFICATES

Dicoding Data Science Bootcamp 2025 | Dicoding Indonesia | 2025

ASEAN Data Science Explorers 2024 National Final Indonesia | ASEAN Foundation | 2024

Oral Presentation | International Conference on Computer Science and Computational Intelligence (ISSCSI) | 2024

Certificate of Completion Samsung Innovation Campus Batch 5 Stage 1 | Skilvul | 2024

Certificate of Completion Samsung Innovation Campus Batch 5 Stage 2 | Skilvul | 2024

Shell Nxplorers Training Certificate | Shell | 2024