**Energy-Efficient Generative AI: Optimizing Retrieval-Augmented Generation (RAG) with FAISS, HuggingFace, and Groq API (Llama 3-70B)**

**Summary**

This research proposal focuses on a **retrieval-augmented generation (RAG) based generative text model** utilizing **FAISS vector store, HuggingFace embeddings, and Groq API (Llama 3-70B).** The objective is to improve the **energy efficiency of large language models (LLMs),** aligning with **Green AI principles.** The study will explore methods like reducing model size, using less computing power, and optimizing how the model works.. The final aim is to create a faster, cheaper, and greener AI system**.**

**Introduction**

Generative AI models use a lot of energy, making them expensive and less eco-friendly. This research aims to make a **text-generating model** more efficient using **FAISS, HuggingFace embeddings, and Groq API.** The goal is to **reduce energy use while keeping the model fast and accurate**.

**Related Work**

Researchers have worked on **energy-efficient AI models** like **Evolved Transformer** and **Primer**, designed using **Neural Architecture Search (NAS)**. The concept of **Green AI** was introduced to balance performance and energy use. Studies also emphasize the need to **publish ML energy consumption data** for transparency. This work builds on these ideas to develop **more efficient AI models**

**Methodology**

This research improves a **text-generating AI model** using **FAISS, HuggingFace embeddings, and Groq API (Llama 3-70B).** The model finds answers by searching a document with **FAISS vector search**. To use **less energy**, methods like **making the model smaller and running it more efficiently** are applied. Energy use and computing costs are measured to see improvements. The goal is to make AI **faster, cheaper, and more eco-friendly** while keeping it accurate.

**Dataset**

This research uses **PDF documents** as the dataset, which are processed **using FAISS vector search** for retrieval-based text generation.

**References**

 Schwartz, R., et al., 2020. Green AI. Communications of the ACM, 63(12)​.

 Lacoste, A., et al., 2019. Quantifying the Carbon Emissions of Machine Learning. arXiv:1910.09700​.

 Bender, E.M., et al., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ACM Conference on Fairness, Accountability, and Transparency​.

So, D.R., et al., 2019. **The Evolved Transformer.** International Conference on Machine Learning​

**Performance Measures**

The performance of the generative text model will be measured using the following metrics:

1. Energy Consumption – Measuring the power usage of training and inference to evaluate efficiency.
2. Processing Speed – Checking response time and latency for generating text.
3. Accuracy & Relevance – Assessing the correctness of generated responses using benchmark datasets.
4. Computational Cost – Comparing hardware resource usage before and after optimization.
5. Carbon Footprint Reduction – Estimating the decrease in environmental impact due to model optimizations.