SMART INDIA HACKATHON 2024

Transformo Docs Application

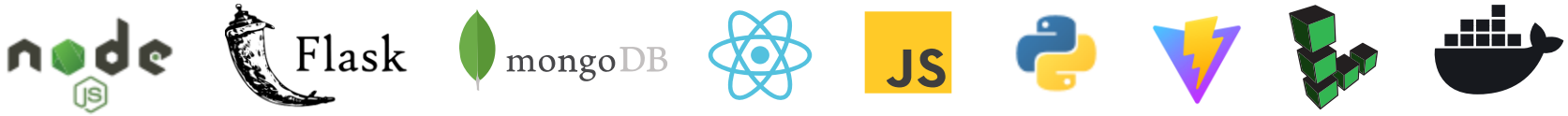
- **Problem Statement ID – SIH 1669**
- **Problem Statement Title- Transformo Docs App**
- **Theme- Smart Automation**
- **PS Category- Software**
- **Team ID- 15388**
- **Team Name - Cross Validators**

Cross Validators

Cognitive Document Engine

SMART INDIA HACKATHON 2024

Proposed Solution



- The proposed solution is a comprehensive **cloud-based** platform that integrates multiple Artificial Intelligence driven tools to **analyze**, **summarize**, **translate**, **listen**(Text-To-Speech) and users can interactively **ask follow-up questions** about their uploaded documents.

- Users can upload various document formats (e.g., **PDFs, DOCX, TXT**) to the platform (e.g., AWS S3, Google Cloud Storage, or Azure Blob Storage) with appropriate **encryption** via a web interface. State of the art **Large Language Models** (e.g., Mistral-7B, GPT-3.5) and Qdrant vector database will be utilized.

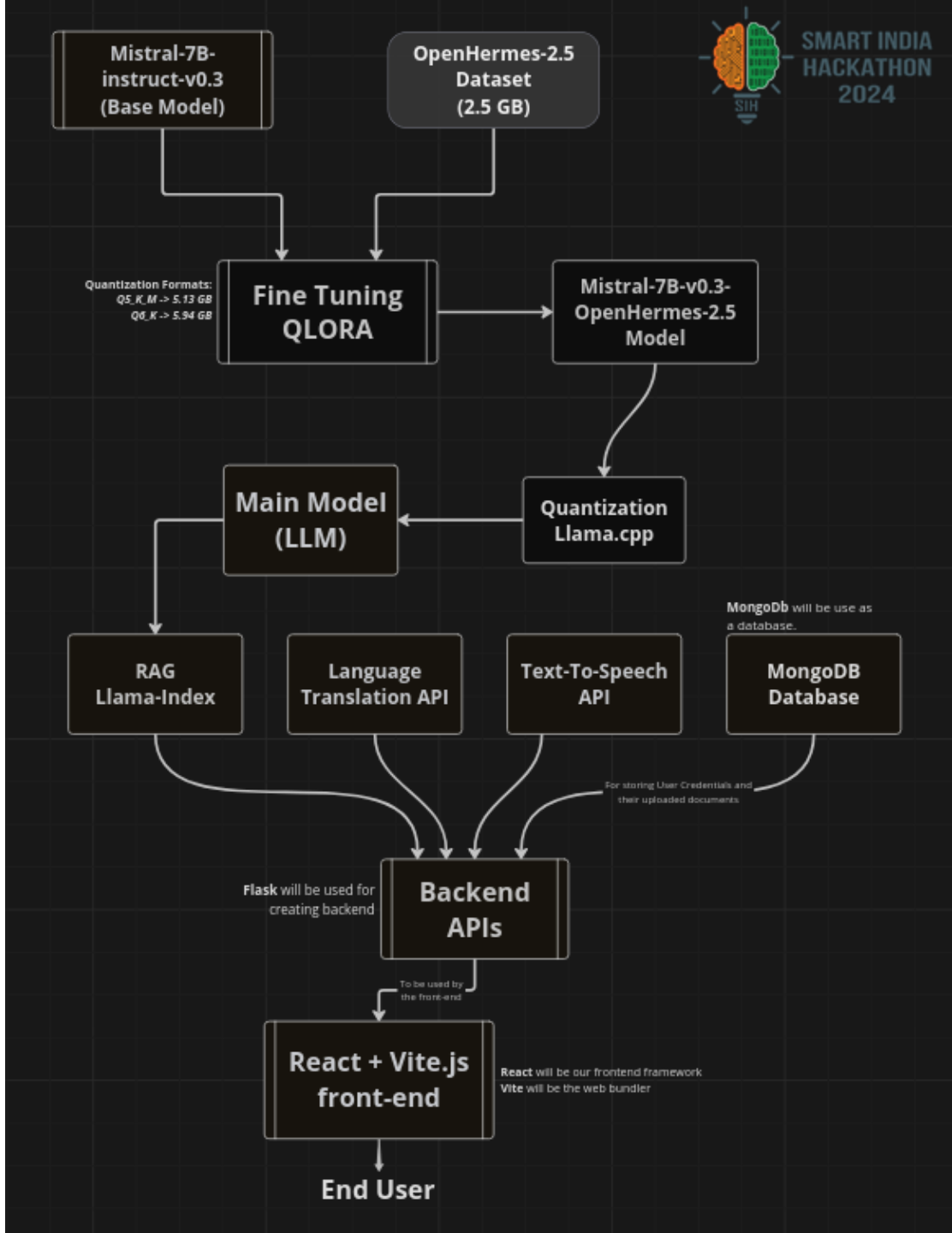
- **Addressing the problem:** It addresses the problem using a state of the art Mistral Instruct 7-billion-parameter open-source model, which is **quantized** for **memory efficiency** and **fine-tuned** with the **Open Hermes dataset**. This dataset contains 1 million samples with questions **covering diverse fields** like Maths, Biology, Science, Law, History, Geography, Politics, Economics, Finance, Business and more.

- **Uniqueness:** The TTS and translation capabilities ensure that the platform is accessible to a wider audience, including **visually impaired** users and speakers of various languages promoting **broader adoption** and usability across diverse user groups. The platform offers **chapter-wise summarization** for books allowing users to generate concise summaries for each chapter separately.

Cross Validators

TECHNICAL APPROACH

- **Llama-Index**: A tool used to build, manage, and query large language models (LLMs) efficiently by providing a way to retrieve relevant information from documents.
- **Llama.cpp**: A C++ implementation for running LLaMA models locally. It's optimized for CPU performance.
- **Quantization**: A technique that reduces the precision of the numbers used to represent a model's weights, making the model smaller and faster.
- **QLoRA**: Quantized Low-Rank Adaptation is a technique where we first quantize the model and freeze some of the weights during fine tuning the remaining weights are trained using low rank matrices.
- **Fine-tuning**: The process of taking a pre-trained machine learning model and training it further on a specific dataset to adapt it to a particular task.
- **Retrieval-Augmented Generation (RAG)**: A hybrid approach that combines retrieval-based methods with generative models.



Cross Validators

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2024

Potential Challenges and Risks:-

- **High Computational Requirements:** AI models for summarization, QnA and TTS are computationally expensive and require significant processing power.
- **Data Privacy:** Handling sensitive documents involves stringent data privacy.

Feasibility and Solutions :-

Business Potential:

- **Enterprise Document Management:** Large organizations need effective document management, summarization, and retrieval tools for internal knowledge management. Can be useful for Academic institutions, research organizations, healthcare sector and legal firms.
- **Subscription and Tiered Pricing Model:** The platform can be monetized through a subscription-based model, with tiered pricing based on usage.

Offline Solution:

For highly sensitive or private documents, an **offline version** of the platform could be developed:

- **Local Deployment:** A version of the platform could be offered as a **local software installation** (on-premises) or as a **Docker container** that enterprises can run on their secure servers.

Quantizing a model's parameters from 32 bits to 5 bits is a highly effective strategy for reducing computational requirements and enabling deployment in more resource-constrained environments without significantly sacrificing accuracy.

Cross Validators

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2024

Potential Impact on the Target Audience

- **Increased Productivity:** Employees can quickly summarize and retrieve information from lengthy documents.
- **Cost Efficiency:** Automated document management reduces the need for manual labor in organizing, categorizing, and summarizing documents.
- **Interactive Learning:** The QnA system can serve as a powerful educational tool, enabling students to engage more deeply with study material and providing instant answers to their questions.
- **Streamlined Information Access:** Quick summarization and QnA capabilities allow healthcare professionals to access relevant patient information, research papers, and guidelines

Benefits of the Solution

- **Inclusive Digital Transformation:** The platform promotes inclusivity by providing tools that cater to diverse user needs, fostering a more inclusive digital environment.
- **Job Creation in AI and Tech:** Developing and maintaining such a platform would create jobs in AI development, data science, customer support, and cybersecurity, contributing to the growth of the tech sector.
- **Remote Accessibility and Reduced Commuting:** By enabling remote access to summarized documents and information, the platform reduces the need for commuting to access physical files, thereby contributing to lower carbon footprints.
- **Reduced Paper Usage:** The platform encourages digital workflows over paper-based processes, reducing the demand for paper and the environmental impact of deforestation and waste.

Cross Validators

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2024

- Transformers Attention is all you need
https://arxiv.org/pdf/1706.03762

- Mistral 7B
https://arxiv.org/pdf/2310.06825

- QLORA: Efficient Finetuning of Quantized LLMs
https://arxiv.org/pdf/2305.14314

- Mistral-7B-Instruct-v0.3 base model
https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

- OpenHermes-2.5 dataset
https://huggingface.co/datasets/teknium/OpenHermes-2.5

- Llama.cpp github repository
https://github.com/ggerganov/llama.cpp

- Llama-Index documentation
https://docs.llamaindex.ai/en/stable/module_guides/

- Our project link
https://github.com/Harshroxnox/Transformo-docs