RV College of Engineering®

Go, change the world

VI Semester Minor Project Phase 1 (18CS64)

Department of Computer Science,
R V College of Engineering,
Bengaluru.

DOCSONA

NAMES	USN
ANISH FELIX MATHIAS	1RV20CS022
ARINDAM THAKUR	1RV20CS027
JOEL SHAJI MATHEW	1RV20CS061

Internal Guide : Dr. Vishalakshi Prabhu

PROBLEM STATEMENT

This project aims to address the challenge of extracting key insights from a large volume of textual content by developing an interactive document summarizer similar to Blinkist. The system should allow users to upload documents, search for specific books or documents, and engage in chat-like interactions with the content. The goal is to provide users with concise summaries and facilitate efficient access to essential information within diverse textual materials

LITERATURE REVIEW

Sl	Paper Title and Authors	Abstract Key points	Approach used	Limitations
1	Author(s): Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N. and Kaiser, Łukasz and Polosukhin, Illia Title: "**Attention Is All You Need**" Year: 2017	1. The paper proposes the Transformer model, a novel architecture that relies solely on self-attention mechanisms for sequence processing tasks. 2. The Transformer model avoids the use of recurrent or convolutional neural networks, making it computationally efficient and highly parallelizable. 3. The self-attention mechanism allows the model to capture dependencies between different positions in a sequence by attending to all positions at once. 4. The paper introduces positional encoding to provide the model with information about the position of words or tokens in the sequence, enabling it to process sequences of variable length. 5. The Transformer model exhibits superior performance compared to previous models while maintaining the ability to model long-range dependencies effectively.	The paper suggested a move to start using the Transformer model for NLP and sequence processing tasks. This paper gave the Neural Networks a way to take care of positions of tokens and their relationship with other word. Also helped us take care previous issues with long term dependencies and all this could be parallelized.	1. The transformer model had Increased memory requirements. 2. Interpretability of the learned attention patterns is still difficult. 3. Requires substantial amounts of training data. 4. Fine tuning of the transformer model is probably the hardest task involved in using it. 5. The computation complexity grew exponentially and it demanded huge amounts of computing resources.

LITERATURE REVIEW

Sl	Paper Title and Authors	Abstract Key points	Approach used	Limitations
2	Author(s): Tim Dettmers, Mike Lewis, Younes Belkada, Luke Zettlemoyer Title: "**LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale**" Year: 2022	1. The paper introduces LLM.int8(), a method for 8-bit matrix multiplication in transformers that preserves the performance of 16-bit precision models. 2. LLM.int8() combines vector-wise quantization and mixed-precision decomposition to achieve memory reduction while maintaining model performance. 3. The proposed method demonstrates that LLM.int8() can handle transformers with up to 175B parameters without performance degradation, while reducing memory consumption by approximately 50%. 4. The process involves scaling inputs using row and column-wise maximum values, performing 8-bit vector-wise multiplication, de-quantizing the outputs, and accumulating the results in 16-bit floating point.	The approach used is a clever and effective method for reducing GPU memory requirements while maintaining the performance of LLMs. By leveraging vector-wise quantization and mixed-precision decomposition, the authors achieve a significant reduction in memory usage without sacrificing model accuracy. It can make LLMs more accessible for deployment on consumer GPUs.	1. The paper focuses solely on the Int8 data type and does not explore the potential of 8-bit floating-point (FP8) data types. Since current GPUs and TPUs do not support FP8, further investigation into this data type is left for future work. 2. The scalability and effectiveness of LLM.int8() for even larger models remain uncertain. 3. The paper primarily focuses on the inference phase and does not extensively study the training or fine-tuning of models using Int8 quantization.

LITERATURE REVIEW

Sl	Paper Title and Authors	Abstract Key points	Approach used	Limitations
3	Author(s): Elias Frantar, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh Title: "**GPTQ : Accurate post training Quantization for Generative pre-trained transformers**" Year: 2023	1. The paper introduces GPTQ, a one-shot weight quantization method based on approximate second-order information, capable of quantizing GPT models with 175 billion parameters. 2. GPTQ achieves significant compression gains, reducing the bitwidth to 3 or 4 bits per weight with negligible accuracy degradation. 3. The method enables the execution of a 175 billion-parameter model on a single GPU, allowing for generative inference. 4. GPTQ demonstrates reasonable accuracy even in extreme quantization scenarios, such as 2-bit or ternary quantization levels. 5. The method presents potential for making large language models more accessible	The authors find that quantizing weights in a greedy order, which minimizes additional quantization error, performs well. The authors propose "lazy batch updates." Instead of updating weights column-wise, the updates are performed in batches, which enhances GPU utilization and leads to better performance for large models. To address numerical inaccuracies that arise when dealing with large models, especially during block updates, the authors introduce a Cholesky reformulation.	1. The method focuses on optimizing memory utilization and GPU efficiency but does not address computational efficiency explicitly. 2. The effectiveness and performance of GPTQ may depend on specific optimizations tailored to the hardware architecture, which could restrict its generalizability across different platforms. 3. The paper does not provide a detailed analysis of the specific accuracy loss incurred by the quantization process, and the extent of the trade-off may vary depending on the specific model and task.

LITERATURE REVIEW

Sl	Paper Title and Authors	Abstract Key points	Approach used	Limitations
4	Author(s): Rohan Taori* and Ishaan Gulrajani* and Tianyi Zhang* and Yann Dubois* and Xuechen Li* and Carlos Guestrin and Percy Liang and Tatsunori B. Hashimoto Title: "**Alpaca: A Strong, Replicable Instruction-Following Model**" Year: 2023	1. Alpaca is trained on 52K instruction-following demonstrations generated in the style of self-instruct using OpenAI's text-davinci-003. 2. Alpaca behaves qualitatively similar to text-davinci-003, while being small, easy, and inexpensive to reproduce. 3. The paper provides the training recipe, data, and plans to release the model weights, emphasizing that Alpaca is intended for academic research only. 4. Challenges in training an instruction-following model under an academic budget are addressed by utilizing a strong pretrained language model (LLaMA) and high-quality instruction-following data. 5. Alpaca's training pipeline involves fine-tuning LLaMA models using Hugging Face's training framework, leveraging techniques like Fully Sharded Data Parallel and mixed precision training	Alpaca is fine-tuned from Meta's LLaMA 7B model using supervised learning on 52K instruction - following demonstrations generated from OpenAI's text-davinci-003. The training process involves utilizing Hugging Face's training framework, applying techniques such as Fully Sharded Data Parallel and mixed precision training. A preliminary evaluation is conducted, including a blind pairwise comparison between Alpaca and text-davinci-003, as well as interactive testing to assess Alpaca's behavior.	1. Limited evaluation scale and diversity: The evaluation of Alpaca's performance may be limited due to the scale and diversity of the evaluation set used. 2. Lack of safety measures: The paper mentions that Alpaca is not ready to be deployed for general use due to the absence of adequate safety measures. 3. Prohibited commercial use: Alpaca's usage is restricted to academic research, and commercial use is prohibited due to licensing restrictions and the terms of use of the underlying models.

LITERATURE REVIEW

Sl	Paper Title and Authors	Abstract Key points	Approach used	Limitations
5	Author(s):Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, Julien Launay Title: "**The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only**" Year: 2023	1. The paper introduces Falcon LLM, a model independent of popular frameworks trained on curated data. 2. Large language models can achieve high performance using filtered and deduplicated web data alone, outperforming curated models. 3. The authors extracted five trillion tokens from the web, releasing a subset of 600 billion tokens as the RefinedWeb dataset. 4. Language models with 1.3/7.5 billion parameters were trained on the RefinedWeb dataset.	The paper demonstrates the effectiveness of training language models solely on filtered and deduplicated web data, challenging the need for curation. The authors utilize the CommonCrawl dataset to extract a significant amount of high-quality web data. Specific details regarding the training techniques, hyperparameters, and model architecture are not mentioned in the abstract.	1. The generalization abilities of web-only models compared to curated models for different domains or tasks are not discussed. 2. The representativeness of the RefinedWeb dataset in relation to the entire web and potential biases are not addressed. 3. Detailed information about training techniques, hyperparameters, and model architecture is missing, hindering reproducibility and comparison to existing methods.

LITERATURE REVIEW

Sl	Paper Title and Authors	Abstract Key points	Approach used	Limitations
6	Author(s): Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, Ahmed Awadallah Title: "**Orca: Progressive Learning from Complex Explanation Traces of GPT-4"** Year: 2023	1. Research focuses on improving smaller models through imitation learning from large foundation models (LFMs). 2. Orca, a 13-billion parameter model, learns to imitate LFMs' reasoning process using rich signals from GPT-4. 3. Orca surpasses conventional models on complex zero-shot reasoning benchmarks and shows competitive performance on professional/academic examinations.	Orca utilizes imitation learning from LFMs and learns from GPT-4's explanation traces and step-by-step thought processes. Teacher assistance from ChatGPT guides the learning process. Large-scale and diverse imitation data are employed through judicious sampling and selection techniques.	1. Lack of rigorous evaluation in previous research overestimates the capabilities of smaller models. 2. Details about the learning process and the role of ChatGPT are not provided in the abstract. 3. Evaluation in domains beyond complex reasoning and limited performance comparison with GPT-4 are not mentioned.

CURRENT STATE

PRIVATE GPT

CHAT PDF

GENIUS PDF

LANG CHAIN

Feasibility of the project

The difficulty of our project is adjustable, allowing us flexibility in its execution. Instead of training a multi-billion parameter model from scratch, we opted to use an existing model that meets our RAM requirements and has performed well in the hugging face open LLM rankings.

Performance Evaluation: The hugging face open LLM rankings serve as a benchmarking platform to assess the model's performance and compare it with other models.

Our research indicates that fine tuning the selected model for our summarization task is feasible, although there may not be a significant advantage since summarization is a generic task. However, we plan to investigate further if time permits. To overcome the limitation of a limited context window in the 13 billion parameter model, we employ a vector datastore and contextual compression techniques.

Project Viability: Considering the available resources, time, and objectives, our project is viable and can be successfully executed.

OBJECTIVES

- Enable users to upload documents in various formats, such as PDF, Word, or plain text.
- Develop a search functionality that allows users to search for existing documents by title, author, or keywords.
- Develop algorithms that can process and extract key information from documents to create concise and informative summaries.
- Design a user-friendly interface to display the generated summaries, including the title, author, and main points of the document.
- Develop an interface to communicate with a document as if it was a person.

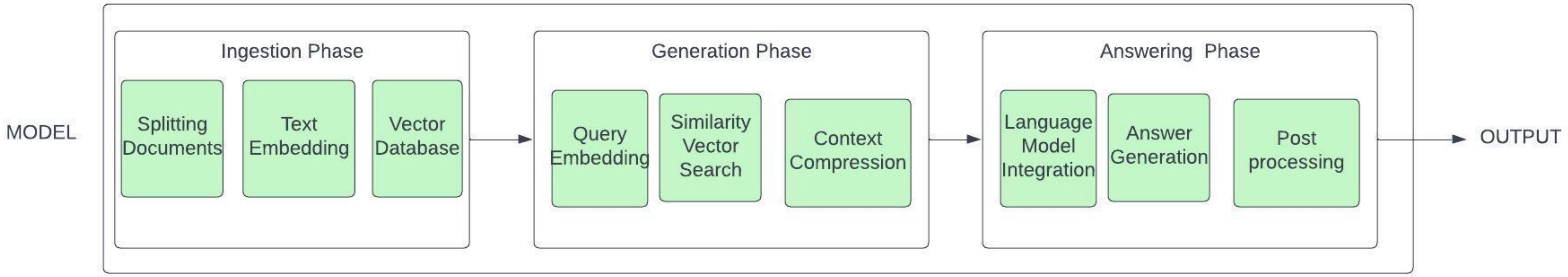
Formulation of objectives and methodology

Our model consists of two distinct phases: **ingestion** and **generation**

1. **Ingestion Phase**: During ingestion, the model receives the entire document collection and employs a semantic and context-preserving splitting mechanism. The resulting text chunks are then transformed into embeddings, which are stored in a vector database for efficient retrieval.
2. **Generation Phase**: In the generation phase, the user's query is embedded using a similar technique. The model performs a similarity vector search to fetch the relevant text chunks from the stored document embeddings. A context compressor is applied to filter out any irrelevant information, ensuring a more focused input for subsequent processing.

The distilled information, obtained through the ingestion and generation phases, is combined with the user query and fed into our Language Model (LLM). The LLM leverages this refined input to generate accurate and contextually appropriate answers.

Formulation of objectives and methodology



Formulation of objectives and methodology

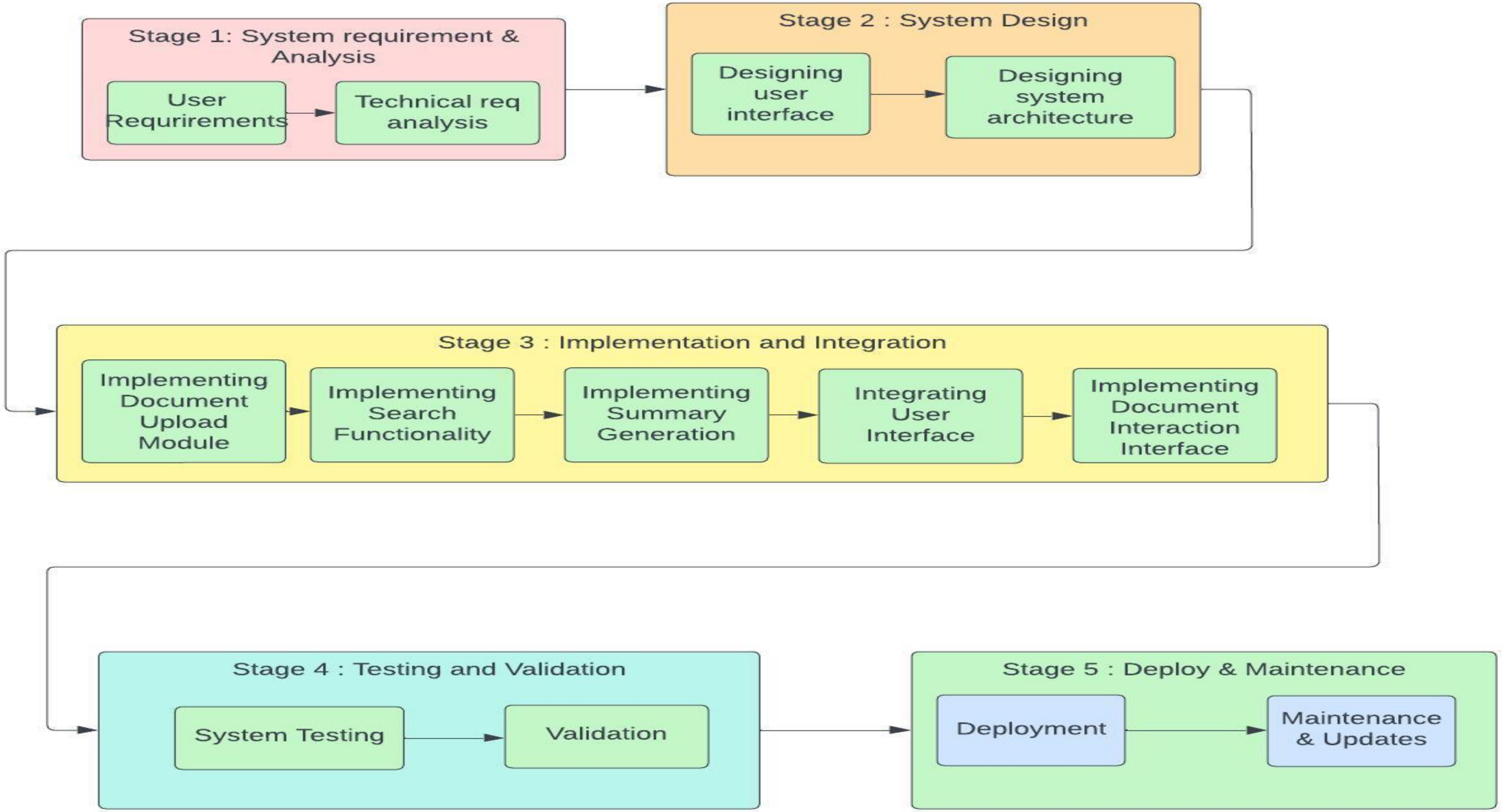
Benefits of the Methodology:

Our methodology offers several advantages over a simplistic approach of passing the entire document to the LLM :

- It significantly reduces latency
- enhances accuracy by utilizing relevant information
- allows for effective utilization of smaller models without sacrificing performance.

Formulation of objectives and methodology

Methodology Flow Diagram


Stage 1: System requirement & Analysis
User Requirements
Technical req analysis
Stage 2 : System Design
Designing user interface
Designing system architecture
Stage 3 : Implementation and Integration
Implementing Document Upload Module
Implementing Search Functionality
Implementing Summary Generation
Integrating User Interface
Implementing Document Interaction Interface
Stage 4 : Testing and Validation
System Testing
Validation
Stage 5 : Deploy & Maintenance
Deployment
Maintenance & Updates

EXPECTED OUTCOMES

The expected outcome of our project is to develop a tool that can efficiently summarise books and documents, providing users with condensed versions of the content. By offering a catalogue of pre-summarized books and allowing users to submit their own texts for summarising, the AI aims to save time and enhance information accessibility. The ability to ask questions about the documents further improves user interaction, enabling specific information retrieval. The project aims to create a valuable resource for quickly understanding and extracting key insights from a variety of texts.

References

Research Papers:

1. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
2. https://openreview.net/pdf?id=dXiGWqBoxaD
3. https://arxiv.org/abs/2210.17323
4. https://crfm.stanford.edu/2023/03/13/alpaca.html
5. https://arxiv.org/abs/2306.01116
6. https://arxiv.org/abs/2306.02707

Tools used

Tools Being used for documentation

Obsidian

Kanban

Git & Github

Google docs & lucid chart

Web links:

1. https://huggingface.co/
2. https://python.langchain.com/docs/get_started/introduction.html
3. https://github.com/oobabooga