Volume 9, Issue 12, December – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

AI-Enhanced Q&A-Summary System Using Advanced Prompt Techniques

Avikal Chauhan[1]; CH Pawan[2]; C Vishwash[3]; Aditya Dillon[4]; Bharani Kumar Depuru[5]

[1,2,3,4,5]AISPRY

Publication Date: 2025/05/15

Abstract: The swift progress of large language models [1] in recent times has profoundly influenced the trajectory of natural language processing. This evolution has been propelled by exponential growth in computational resources, the increasing availability of expansive data, and refinements in algorithmic methodologies. Transitioning from the rudimentary rule-based frameworks to today's complex architectures, LLMs have undergone substantial transformation. Early models showcased the capacity for generating coherent and contextually applicable content but recent enhancements have significantly augmented both comprehension and content generation capabilities marking a pivotal leap in language model sophistication.

The expansion of open-source LLMs [2] has transformed the sphere of advanced linguistic innovations, contributing unprecedented access for experimentation and implementation across sectors, such as education recent breakthroughs in prompt enhancing have redefined how these models are harnessed, producing very accurate contextually aware outputs without requiring exhaustive retraining processes.

Within the educational domain, large language models LLMs present a paradigm shift by streamlining text automation significantly mitigating the laborious, and resource-heavy demands of traditional manual tasks. This advancement empowers educators to devote more attention to pedagogy and direct student interaction. On top of that, the fusion of sophisticated LLMs accompanied by optimized prompting strategies [3] in scholastic platforms elevates the educational experience, delivering tailored high-caliber, and contextually pertinent material. This approach fosters a more adaptive and systematic learning ecosystem enhancing the overall instructional framework.

Keywords: *AI-Enhanced Content Generation, Large Language Models, RAG, Prompt Techniques, Q&A-Summary system, Lang Chain Framework, Plagiarism.*

How to Cite: Avikal Chauhan; CH Pawan; C Vishwash; Aditya Dillon; Bharani Kumar Depuru. (2024). AI-Enhanced Q&A-Summary System Using Advanced Prompt Techniques. *International Journal of Innovative Science and Research Technology*, 9(12), 3170-3179. https://doi.org/10.38124/ijisrt/24dec299.

I. INTRODUCTION

The increasing adoption of cutting-edge technology in academic settings has driven a pronounced rise in the need for superior on-demand educational materials [4]. Educational institutions are relentlessly searching for advanced approaches to efficiently generate and disseminate content that addresses the diverse needs of their learners. Established methods of content creation notably the painstaking manual process of drafting questions and answers are not only laborious but also prone to errors and issues of intellectual property infringement.

To surmount these challenges, this study unveils a pioneering ai-empowered text creation engine precisely crafted to automate the content generation workflow. By harnessing advanced machine learning frameworks through the Gemini Flash model [4], this engine orchestrates the automated extraction and composition of interrogative and responsive content from PDF documents. Central to its prowess are sophisticated natural language processing NLP [5] techniques and cutting-edge large language models LLMs, which confer the capability to interpret and manipulate text with exceptional granularity and sophistication. These technological advancements facilitate the generation of contextually pertinent and highly accurate questions and answers thereby ensuring that the content produced is both profoundly insightful and pedagogically valuable.

A pivotal element of the Q&A-Summary system is the implementation of prompt engineering [6] which orchestrates the large language models LLMs to create contextually precise and pertinent content. By employing meticulously crafted prompts, the system directs the LLMs to produce educational resources that conform to defined pedagogical

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

goals and standards. Through iterative evaluation and optimization the most efficacious prompt configuration was discerned, guaranteeing that the output was both relevant and of superior quality. This methodology augments the system's versatility across various subjects and levels of complexity, ensuring that the produced content remains both exemplary and finely attuned to the requirements of educators.

This manuscript elucidates the advancement and deployment of an AI-augmented content generation apparatus, underscoring its transformative potential within academic environments. Additionally, the system's capability to significantly truncate content generation duration while upholding high-quality benchmarks renders it an invaluable resource for educators. Through this pioneering methodology, the tool aspires to furnish educators with a dependable, effective, and scalable solution for content development.

II. TECHNOLOGY INTEGRATION AND METHODOLOGIES

A. Data Dimension

In this Q&A-Summary System, the data dimension encompasses the management of a substantial corpus of documents, emphasizing their dimensional attributes and classifications. The architecture of the system is optimized to adeptly process diverse input formats such as PDFs and text files, while adhering to a maximum file size constraint of 400mb. This constraint guarantees that even voluminous and intricate documents can be processed efficiently without undermining system efficacy. The system's capability to accommodate various document types endows it with broad applicability rendering it suitable for a multitude of academic scenarios.

B. Methodology

Our initiative to tackle the problem of proficiently deriving questions and answers from PDF documents for pedagogical objectives was directed by the crisp-ML(Q) framework [Fig.1]. This systematic schema [7] offered a strategic blueprint that guaranteed a meticulous and intentional approach throughout our research process spanning from data acquisition to model implementation with the ultimate goal of augmenting educational understanding and mitigating learner exasperation.

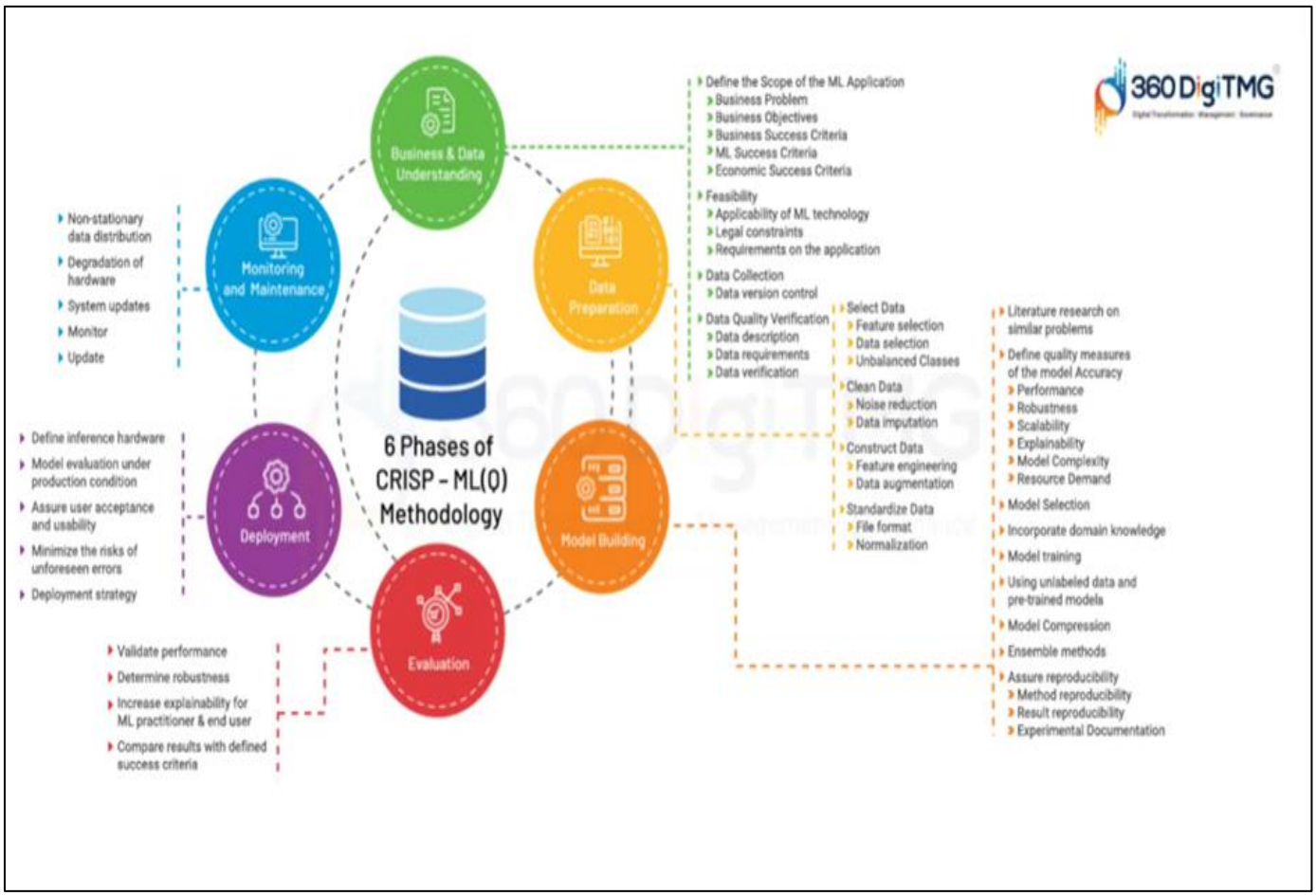

Fig 1: CRISP-ML (Q) Methodological Framework, Outlining its Key Components and Steps Visually. (Source:-Mind Map - 360DigiTMG)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

C. Data Preparation

The data curation protocol for the qa-summary system entails the conversion of PDF documents into a format conducive to model training and analytical querying. This procedure is segmented into several principal phases:

➢ *PDF to Text Extraction*

The initial objective demands the conversion of PDF files into text. This process is executed using the pdfreader class provided by the pypdf2 library [8]. This class interprets the content within the PDF and outputs it as raw unformatted text. This initial stage ensures that the data is formatted appropriately for further computational analysis.

➢ *Text to Chunks*

Upon the extraction of textual data, it is systematically partitioned into discrete manageable segments utilizing the recursive character text splitter [9] class from langchain. The gettextchunks function performs this division by segmenting the text into units of 10000 characters with a 1000-character overlap between segments. This method preserves sufficient contextual continuity within each segment facilitating thorough and coherent analysis.

➢ *Tokens to Vectors*

After the text segmentation process, the subsequent step involves transforming these text segments into vector representations [10]. This is achieved through the googlegenerativeaiembeddings class available in langchain which encodes the text segments into embeddings. These embeddings are numerical vectors that appear for the semantic layers of the text. A vector repository is then prepared using these embeddings and is preserved locally to allow for swift and efficient retrieval. This preparatory framework ensures that text extracted from PDF documents is methodically converted into a format that supports advanced language model functionalities such as enhanced search operations and sophisticated question generation.

D. Framework

LangChain works as the advanced open-source correlation structured framework specifically crafted to enhance the design of applications that utilized large LLMs. It delivers an integrated platform that seamlessly connects LLM-powered solutions with a variety of data inputs and workflow architectures. Supporting a wide range of LLMs, LangChain [11] equips developers with the flexibility to choose models tailored for query interpretation and response formulation. Its efficient abstractions reduce the complexity of development by enabling the effortless composition of components such as prompt generation tools and data access mechanisms with minimal code. Both Python and JavaScript libraries are available to facilitate its deployment.

The architecture comprises a suite of integral factors including large language model (LLM) modules, prompt blueprints, chains, document ingestion mechanisms, text segmentation utilities, memory handlers, and agent systems. These elements optimize processes like summarization, query resolution, and data augmentation by incorporating LLMs with external datasets and orchestrating conversational context. LangChain further incorporates vector databases for optimized data retrieval and introduces advanced memory management to preserve dialogue continuity. Moreover, its agent modules autonomously discern and execute subsequent workflow actions thereby amplifying the system's adaptability.

In the domain of this Q&A-Summary System, LangChain emerges as a cornerstone for the sophisticated manipulation and management of text data derived from PDFs. The system adeptly translates textual content into numerical vectors, which are meticulously archived in a vector-centric database to support advanced querying and intricate question formulation. By weaving together prompt orchestration, expansive language models LLMs, and meticulous output interpretation into a unified procedural framework, LangChain underpins the creation of high-caliber question-and-answer mechanisms. This methodical integration ensures that the qa-summary system proficiently navigates complex data interactions and refines its processing capabilities, thus elevating overall functionality and effectiveness.

E. Prompt Engineering

Prompt engineering, whereby the set of high-level heuristic prompts are built to ensure that the model produces responses in accordance with certain structures and informs these decisions using information known a priori. You do so by drafting prompts incorporating three primary types of information: input, context and examples. The input describes the type of data required for producing an input, while context provides directions on how you expect your model to act, and examples illustrate what kind of response will be output by the model. Prompt creation can sometimes be a trial-and-error process as every model is trained on very specific types of keywords part of the dataset which might cause them to interpret prompts differently. With that, advanced prompting can take LLMs to next level — execute some of the more challenging tasks like reasoning and problem-solving with significantly greater ease. Check out these methods to get great prompts.

➢ *Chain-of-Thought (CoT) Prompting*

This method falls under the Chain-of-Thought (CoT) prompting [12] technique, which is used for improving the cognitional performance of language models because they are forced to explain the processes employed in coming up with an answer. This method is found to be highly effective for complex tasks that involve logical thought processes like number calculation and problem solving. Indirect problems questions make it easier to get a more accurate response by breaking down the steps involved in getting an answer and encourage better understanding of how a certain conclusion was drawn.

- *Example:*

✓ Prompt: "John has 10 apples, gives away 3, and buys 5 more. How many apples does he have now? Explain your reasoning."

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

✓ Output: "Starting with 10 apples, after giving away 3, he has 10 - 3 = 7 apples. Then, adding 5 more, he has 7 + 5 = 12 apples."

➢ *Self-Consistency*

Self-Consistency [13] is a prompting technique that enhances the reliability of model outputs by generating multiple responses for the same query and selecting the most consistent answer. This method is particularly useful for tasks where uncertainty may lead to varied outputs, helping to mitigate errors and biases.

- *Example:*

✓ Prompt: "What is the capital of India?"
✓ Responses:

- "The capital of india is New Delhi."
- "New Delhi is the capital of India."
- "India's capital is New Delhi."

✓ Final Output: "The capital of India is New Delhi." (selected based on consistency).

➢ *Tree-of-Thoughts (ToT) Prompting*

Tree-of-Thoughts (ToT) prompting allows models to explore multiple reasoning paths before arriving at a conclusion. This technique is beneficial for complex decision-making scenarios where different outcomes can arise based on varying assumptions.

- *Example:*

✓ Prompt: "You are at a crossroads. If you turn left, you will find a river; if you turn right, you will find a mountain. What do you choose and why?"
✓ Output: "If I turn left, I can relax by the river and enjoy nature. If I turn right, I could hike the mountain for exercise. I choose to turn left for a peaceful day."

➢ *Active Prompting*

Active prompting [14] enhances LLMs by zeroing in on questions where the model is uncertain or inconsistent. This works by first coming up with multiple answers to find questions containing big disagreements and then selecting these uncertain ones for human annotation that can provide more knowledge about what's being asked and why it's being asked. Finally, using this refined information, the model can improve its responses.

- *Example:*

✓ Prompt: "What were the key causes of the French Revolution?"
✓ Response: If a model gives varying answers to the above question, active prompting would identify this uncertainty, have experts provide detailed explanations, and then use these explanations to train the model to give more accurate and consistent answers in the future.

➢ *Reasoning Without Observation (ReWOO)*

ReWOO (Reasoning WithOut Observation) is a technique that makes LLMs more efficient by separating the reasoning process from real-world data retrieval. Instead of using real-time information, ReWOO divides the workflow into three parts: a Planner that breaks down questions into steps, a Worker that retrieves necessary information, and a Solver that combines the steps and information to produce an answer.

- *Example:*

✓ Prompt: "If a train leaves the station at 5 PM traveling at 60 mph, how far will it travel by 7 PM?"
✓ Output: "The train travels for 2 hours. At 60 mph, it will cover 60 * 2 = 120 miles."

➢ *ReAct*

ReAct combines reasoning with action, allowing models to generate reasoning traces alongside their actions, making it suitable for dynamic decision-making tasks.

- *Example:*

✓ Prompt: "You are at a grocery store. If you see apples on sale for $1 each and oranges for $2 each, what do you buy?"
✓ Output: "I will buy 5 apples because they are cheaper. I can save $5 compared to buying oranges."

➢ *Expert Prompting*

Expert Prompting involves creating specialized personas for specific tasks, allowing language models to respond based on the expertise of the imagined agent. This technique generates more relevant and accurate responses in specialized domains.

- *Example:*

✓ Prompt: "As a financial advisor, what investment strategies would you recommend for a beginner?"
✓ Output: "I recommend starting with low-cost index funds, diversifying your portfolio, and considering a retirement account like a 401(k) or IRA."

➢ *Automatic Prompt Engineering (APE)*

Automatic Prompt Engineering (APE) optimizes prompts by treating them as programs and selecting the best candidates from a pool generated by the model itself. This technique is effective in zero-shot learning scenarios.

- *Example:*

✓ Prompt: "Generate a summary of the benefits of renewable energy."
✓ Output: "Renewable energy reduces greenhouse gas emissions, decreases dependence on fossil fuels, and promotes sustainable development." (The model generates multiple summaries and selects the most coherent one.)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

This separation of explanation and example provides a clearer understanding of each technique and its application.

These techniques illustrate the importance of structured and thoughtful prompts design in maximizing the effectiveness of LLM. Out of these prompt engineering techniques, active prompting and chain of thought (CoT) reasoning were selected for their effectiveness in customized content generation. These were employed in the prompt template in our Q&A-Summary System.

F. Best Model Selection

- *Gemini 1.5 Flash*

While selecting a model for a specific application several key factors must be considered to ensure optimal performance, efficiency, and cost-effectiveness. In comparing Gemini 1.5 Flash to open-source models, it becomes evident that the pre-trained model Gemini 1.5 Flash [15] excels in areas crucial for high-demand environments. Its remarkable output speed of 214 tokens per second and low latency of 038 seconds makes it an ideal choice for applications requiring real-time interaction, such as chatbots and live content generation. Additionally, its ability to handle a context window of up to two million tokens ensures it can manage extended interactions without losing context, a critical feature for tasks involving long-form content or complex dialogues.

Moreover, it presents a highly economical alternative compared to numerous other proprietary and open-source models, rendering it particularly appealing for scenarios involving substantial data volumes. This cost-efficiency, combined with its impressive performance indicators, and a notable MMLU score of 0789 establishes it as a superior choice relative to many open-source counterparts. The model's inherent multimodal abilities further enhance its adaptability allowing for seamless integration and processing of text, images, and audio consequently. Gemini 1.5 flash emerges as an enticing option for developers seeking a robust efficient, and economically advantageous model suitable for a wide range of AI applications.

- *Model Architecture*

Gemini 15 flash represents a sophisticated Transformer decoder model [16] meticulously crafted to achieve both operational efficiency and superior performance. This model is distinguished by its expansive context window, which facilitates the processing of lengthy input sequences with remarkable proficiency. It has been optimized for seamless integration with tensor processing units (TPUs), strategically minimizing latency during deployment by exploiting parallelized computation across both attention mechanisms and feedforward layers..

The model is meticulously synthesized through online distillation from the expansive Gemini 1.5 pro model, refining its operational prowess while upholding formidable multimodal integration. The Gemini 1.5 Flash iteration employs cutting-edge higher-order preconditioning techniques, which substantially elevate its performance and quality metrics. These nuanced architectural innovations ensure that Gemini 1.5 Flash strikes an optimal balance between computational efficiency and adeptness in managing complex multimodal data streams with notably reduced latency.

- *Model Integration and Working*

The incorporation of the Gemini 1.5 Flash model within this Q&A-Summary System leverages its advanced capabilities to enhance the processing of text data and generate high-quality responses. After the text data is converted into numerical embeddings, the model is employed to efficiently handle and analyze these embeddings. The integration process involves querying the vector database, where the text embeddings are stored, allowing the model to retrieve relevant information swiftly. This capability is crucial for maintaining responsiveness in real-time applications, such as interactive question-answering systems [17].

Model efficacy is maximized by leveraging customized prompt methodologies. Through the building of targeted prompt frameworks, the input data is structured in a context-sensitive format, facilitating the generation of accurate and cohesive outputs. These prompt frameworks direct the models attention towards the most relevant features within the embeddings, ensuring that the resultant content is aligned with the user's inquiries. This strategy capitalizes on the model's sophisticated context-processing abilities, enabling it to interpret and generate precise outputs in response to the contextual cues embedded in the input data.

Furthermore, the Gemini 1.5 Flash model's architecture and prompt engineering are instrumental in achieving high efficiency and quality in data processing. The model's robust context window and low latency enable it to handle extensive input sequences with minimal delay, making it well-suited for applications requiring rapid and reliable responses. By integrating the model's capabilities with effective prompt design, the Q&A-Summary System can deliver a sophisticated solution for complex data interactions, ensuring both high performance and relevance in the generated content.

The structural design of the Gemini 1.5 Flash model combined with sophisticated prompt optimization serves as a cornerstone for enhancing data processing effectiveness and speed. The model's expansive contextual capacity and minimal response time allow it to process lengthy input streams with remarkable agility, making it ideal for scenarios that require prompt and precise feedback. By harnessing the models advanced features through deliberate prompt configuration, the Q&A-Summary system can offer an advanced approach to navigating complex data interactions, ensuring optimal output quality and contextual accuracy in the generated results.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

III. RESULTS AND DISCUSSIONS

The results demonstrate the effectiveness of the implemented Q&A-Summary System in transforming PDF documents into meaningful question-and-answer pairs. By utilizing advanced prompt engineering techniques, the system was able to generate contextually relevant questions that align well with the source material. The application of the LangChain framework for chunking and embedding text data ensured that the extracted information from the PDFs was adequately represented and processed. The generated questions and answers [Fig.2a, 2b, 2c, 2d, 2e] maintained high relevance and accuracy, which was evident in the context of educational materials tested during the project.

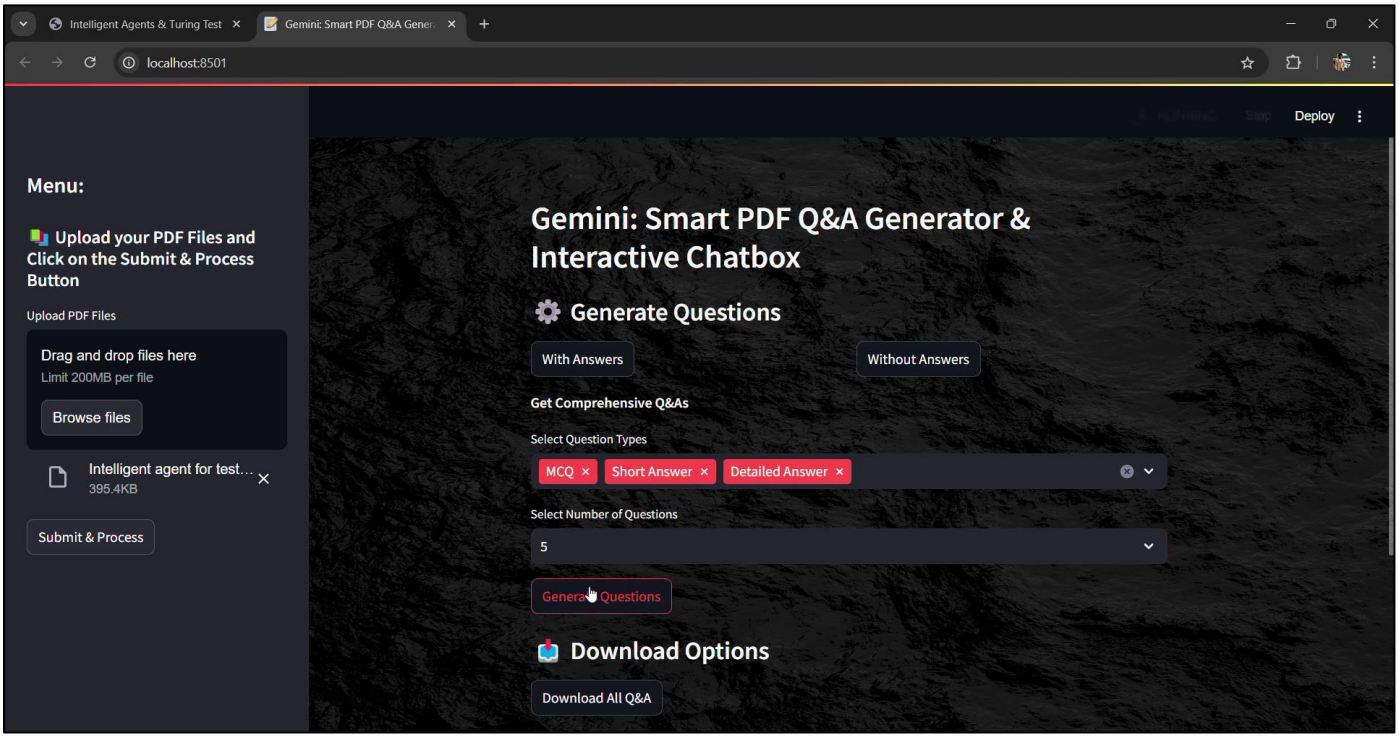

Fig 2(a): Q&A-Summary System: Selection of Q&A Types, Answers, Topic and Number of Questions

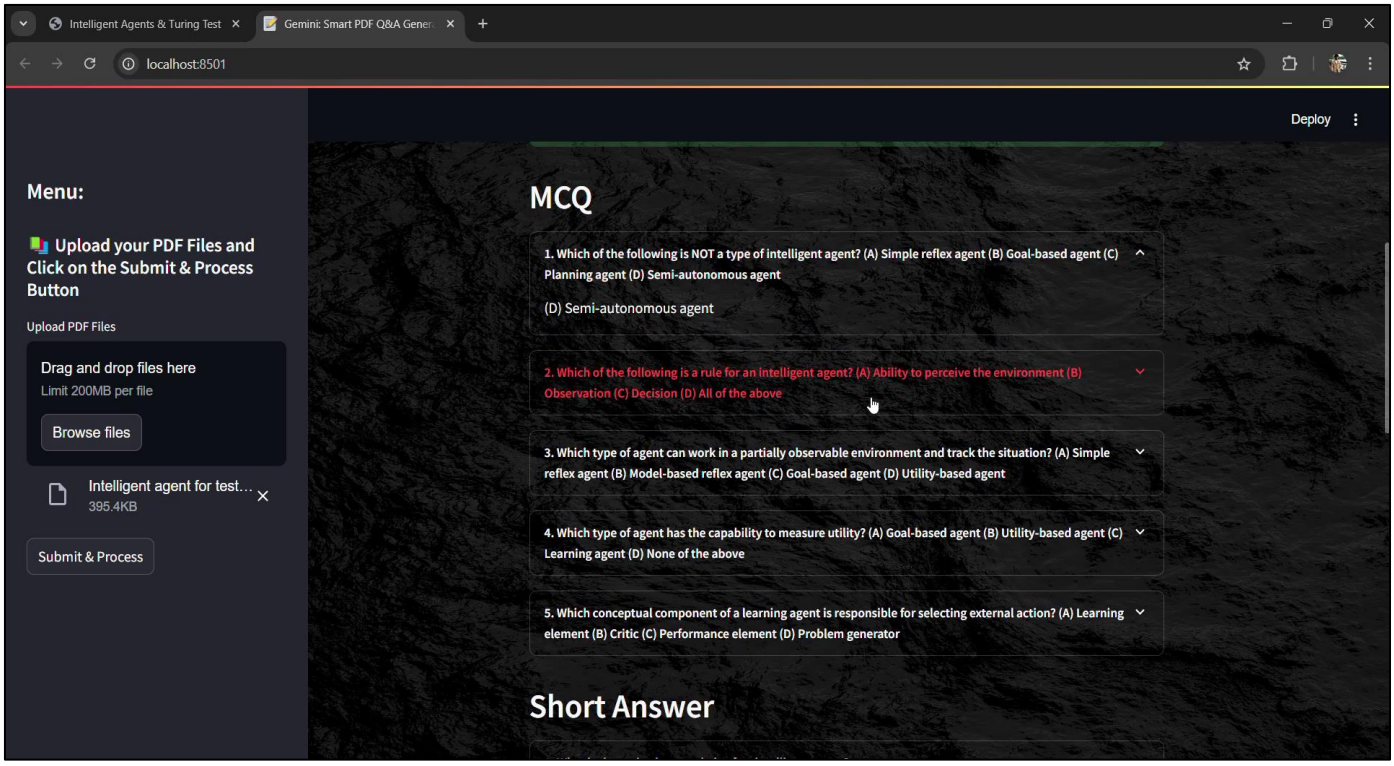

Fig 2(b): Q&A-Summary System: Generation of MCQs

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

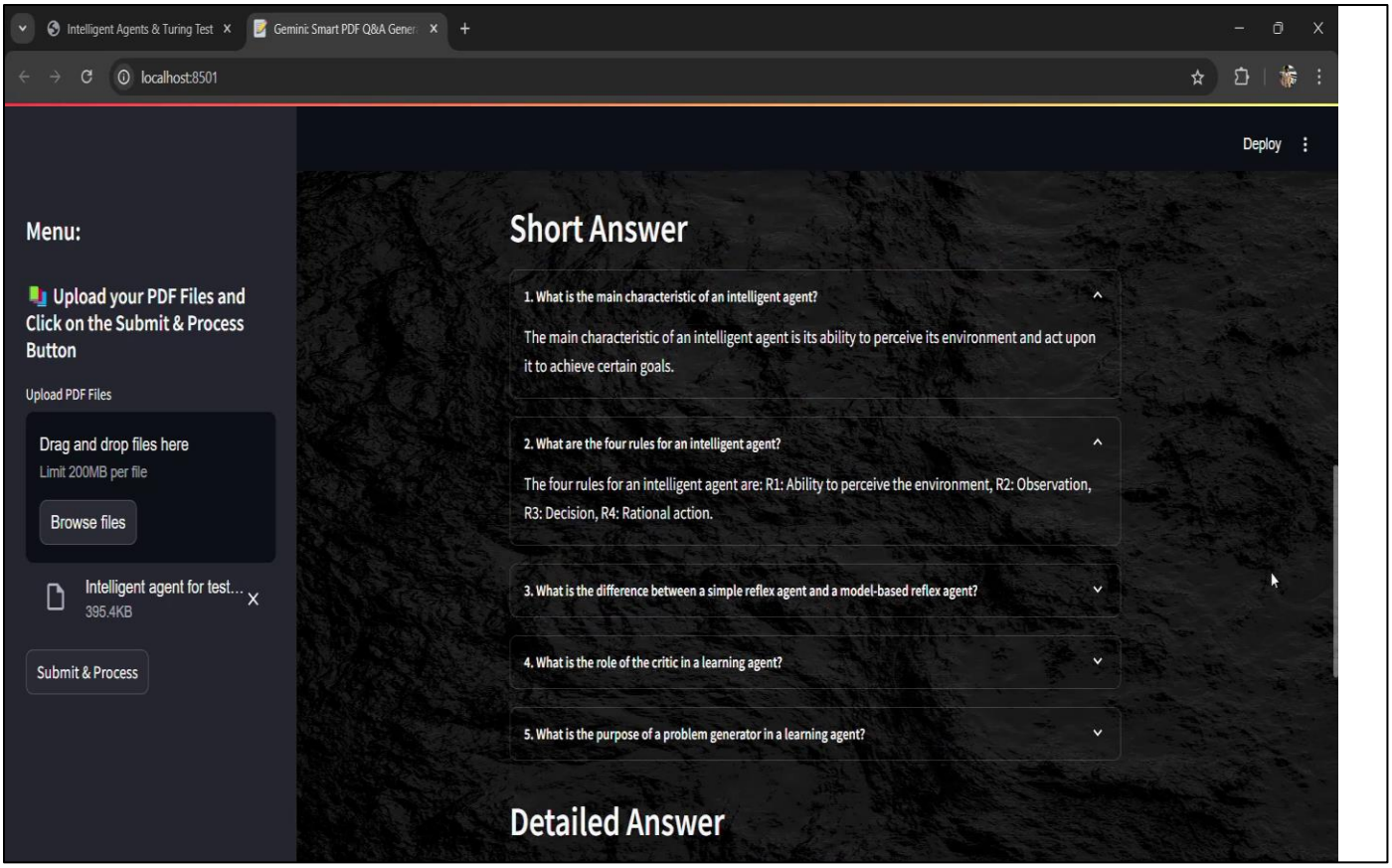

Fig 2(c): Q&A-Summary System: Generation of Short Answers

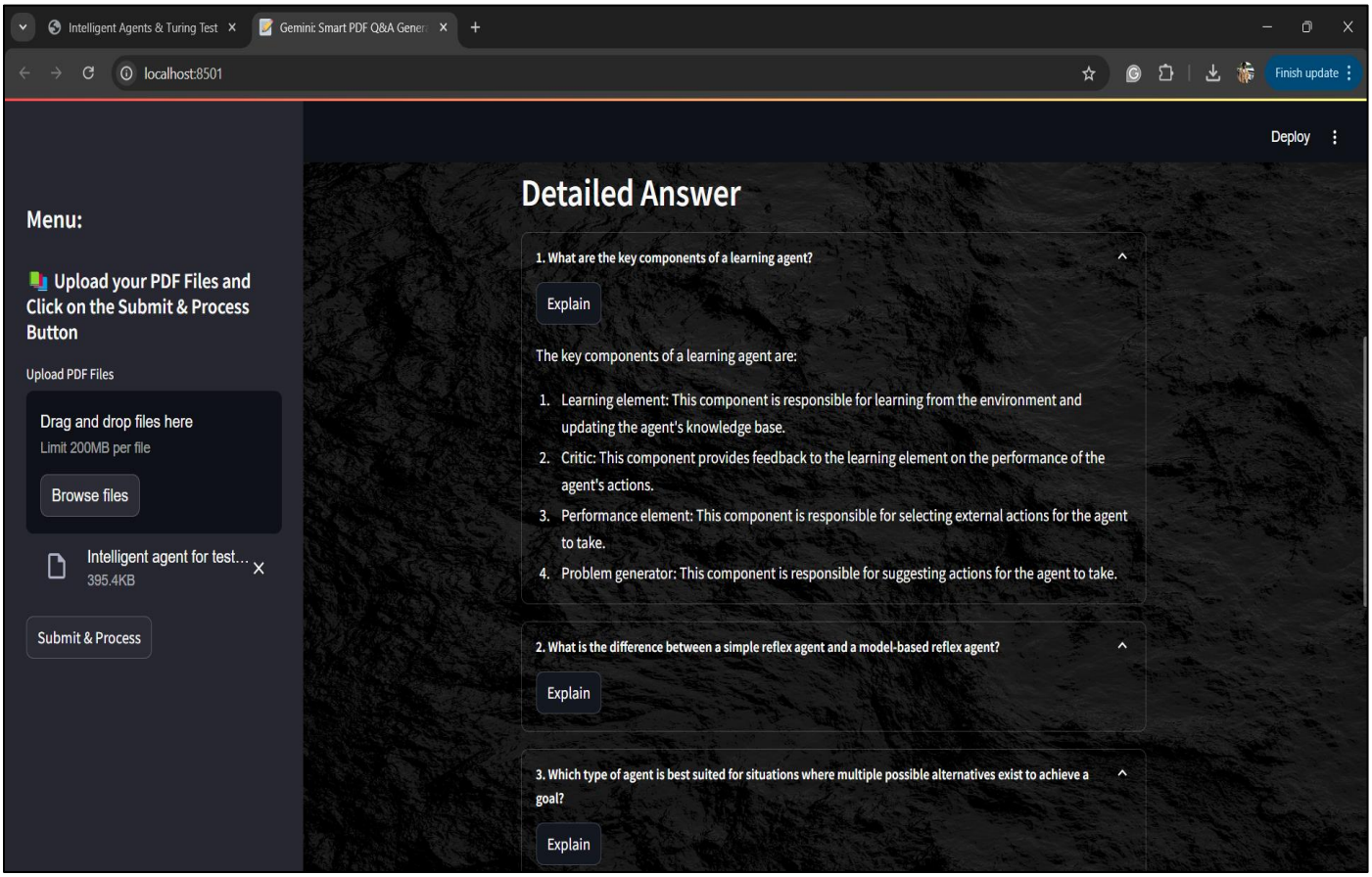

Fig 2(d): Q&A-Summary System: Generation of Detailed Answers

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

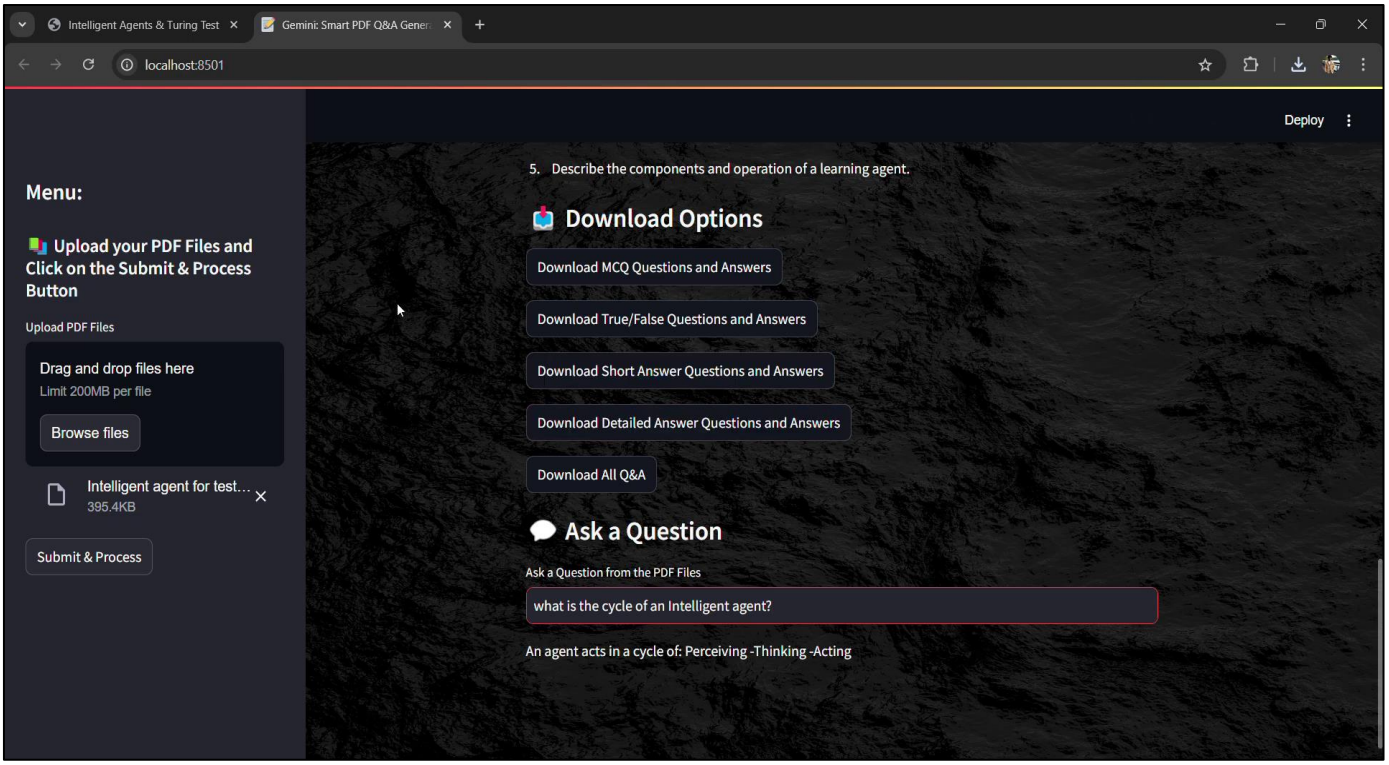

Fig 2(e): Additional Feature: Chatbot

The incorporation of the Gemini 1.5 Flash model markedly elevated the performance and precision of the content generation process. Its proficiency in managing extensive textual data, coupled with its swift output generation, proved especially advantageous for addressing intricate and voluminous documents. This synergy of rapid question generation and high response accuracy underscores the model's adaptability for dynamic, real-time scenarios where prompt content creation is paramount. The applied prompt engineering strategies, including active prompting and chain-of-thought (CoT) reasoning, were instrumental in fine-tuning the question generation mechanism, thereby optimizing the system's overall efficacy.

The evaluation also revealed that the Q&A-Summary System is capable of adapting to a variety of topics and contexts, making it a versatile tool for different domains. The generated content was consistent in quality across diverse subject areas, indicating the robustness of the underlying model and framework. User feedback highlighted the app's potential to reduce the time and effort needed for content creation, further validating its practicality and effectiveness in real-world scenarios. Overall, the results affirm the system's capability to generate high-quality, relevant content efficiently, making it a valuable tool in an educational environment.

IV. CONCLUSION

In summary, the Q&A-Summary System represents a notable advancement in the application of large language models (LLMs) for educational and corporate purposes. By transforming text from PDFs into vectorized formats and employing sophisticated models like Gemini 1.5 Flash, the system effectively generates relevant and contextually accurate questions. This paper demonstrates the power of combining LLMs with frameworks like LangChain, which simplifies the integration of various components and workflows, resulting in a robust and scalable solution.

The deployment of cutting-edge prompt engineering frameworks has notably elevated the app's proficiency in producing superior-quality questions. These methodologies guarantee that the model interprets input efficiently and generates outputs that are contextually precise and intellectually stimulating, establishing the system as a critical tool for a broad spectrum of use cases. As the Q&A-Summary System progresses, it is primed to respond to shifting requirements and deliver impactful solutions within educational and business contexts alike.

FUTURE SCOPE

The Q&A-Summary System offers a broad spectrum of applicability, positioning itself as a multifaceted asset for different industries. Enterprises can harness this system to facilitate worker skill development and training initiatives, by producing tailored quizzes and evaluations that enrich educational outcomes. In the realm of recruitment and talent acquisition, the system's capacity to craft role-specific interview queries significantly optimizes the applicant screening and selection workflow.

The application's prowess in generating FAQ-style content markedly elevates customer education and support, thereby amplifying user engagement and satisfaction. It further contributes to market research by formulating strategic questions that extract important consumer insights. The system proves invaluable for internal knowledge management, ensuring the efficient transfer and retention of

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

information within the enterprise. Additionally, in the realm of content marketing, it excels at producing dynamic, question-centric content that stimulates audience interaction.

Envisioning future advancements, the system holds the prospects for expanded capabilities in adaptive learning. This enhancement would involve the generation of queries specifically designed to align with the personalized learning trajectories of each student, thereby optimizing individualized educational experiences. Another pivotal area of advancement is the continuous refinement of the model. This process entails the continuous calibration of the system through the incorporation of user input and the assimilation of novel datasets, ensuring that the content delivered remains both pertinent and of superior quality.

Furthermore, the system's functionality can be augmented beyond mere question generation to encompass the development of assignments, quizzes, and other pedagogical resources, thus evolving it into an all-encompassing instrument for educational content creation. These enhancements will not only amplify the application's efficacy within academic contexts but also extend its applicability across diverse corporate and educational milieus.

REFERENCES

[1]. Kamath, U., Keenan, K., Somers, G., Sorenson, S. (2024). LLMs: Evolution and New Frontiers. In: Large Language Models: A Deep Dive. Springer, Cham. https://doi.org/10.1007/978-3-031-65647-7_10

[2]. Sanjay Kukreja, Tarun Kumar, Amit Purohit, Abhijit Dasgupta, and Debashis Guha. 2024. A Literature Survey on Open Source Large Language Models. In Proceedings of the 2024 7th International Conference on Computers in Management and Business (ICCMB '24). Association for Computing Machinery, New York, NY, USA, 133–143. https://doi.org/10.1145/3647782.3647803

[3]. Marvin, G., Hellen, N., Jjingo, D., Nakatumba-Nabende, J. (2024). Prompt Engineering in Large Language Models. In: Jacob, I.J., Piramuthu, S., Falkowski-Gilski, P. (eds) Data Intelligence and Cognitive Informatics. ICDICI 2023. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-99-7962-2_30

[4]. Gemini's big upgrade: Faster responses with 1.5 Flash, expanded access and more. https://blog.google/products/gemini/google-gemini-new-features-july-2024/

[5]. Mihalcea, R., Liu, H., Lieberman, H. (2006). NLP (Natural Language Processing) for NLP (Natural Language Programming). In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2006. Lecture Notes in Computer Science, vol 3878. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11671299_34

[6]. Leo S. Lo, The CLEAR path: A framework for enhancing information literacy through prompt engineering,The Journal of Academic Librarianship, Volume 49, Issue 4, 2023, 102720, ISSN 0099-1333, https://doi.org/10.1016/j.acalib.2023.102720

[7]. Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.-R. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. Mach. Learn. Knowl. Extr. 2021, 3, 392-413. https://doi.org/10.3390/make3020020

[8]. Jing Li, Peizhang Wang, Lu Jia, Run Mao, Qian Li, Yongle He, Yi Sun, Pinwang Zhao Design and implementation of an automated PDF drawing statistics tool based on Python. Proceedings Volume 12800, Sixth International Conference on Computer Information Science and Application Technology (CISAT 2023); 128006R (2023) https://doi.org/10.1117/12.3003927

[9]. Bhaskarjit Sarmah, Dhagash Mehta, Stefano Pasquali, and Tianjie Zhu. 2024. Towards reducing hallucination in extracting information from financial reports using Large Language Models. In Proceedings of the Third International Conference on AI-ML Systems (AIMLSystems '23). Association for Computing Machinery, New York, NY, USA, Article 39, 1–5. https://doi.org/10.1145/3639856.3639895

[10]. [10] Selva Birunda, S., Kanniga Devi, R. (2021). A Review on Word Embedding Techniques for Text Classification. In: Raj, J.S., Iliyasu, A.M., Bestak, R., Baig, Z.A. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore. https://doi.org/10.1007/978-981-15-9651-3_23

[11]. A. Singh, A. Ehtesham, S. Mahmud and J. -H. Kim, "Revolutionizing Mental Health Care through LangChain: A Journey with a Large Language Model," 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2024, pp. 0073-0078, https://doi.org/10.1109/CCWC60891.2024.10427865

[12]. Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What Makes Chain-of-Thought Prompting Effective? A Counterfactual Study. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1448–1535, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.101

[13]. T. Ahmed and P. Devanbu, "Better Patching Using LLM Prompting, via Self-Consistency," 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), Luxembourg, Luxembourg, 2023, pp. 1742-1746, https://doi.org/10.1109/ASE56229.2023.00065

[14]. Diao, S., Wang, P., Lin, Y., Pan, R., Liu, X., & Zhang, T. (2023). Active Prompting with Chain-of-Thought for Large Language Models. [Submitted on 23 Feb 2023 (v1), last revised 21 Jul 2024 (this version, v5)] ArXiv. /abs/2302.12246 https://doi.org/10.48550/arXiv.2302.12246

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24dec299

[15]. Mondillo, G., Frattolillo, V., Colosimo, S. et al. Basal knowledge in the field of pediatric nephrology and its enhancement following specific training of ChatGPT-4 “omni” and Gemini 1.5 Flash. Pediatr Nephrol (2024). https://doi.org/10.1007/s00467-024-06486-3

[16]. Multimodality with Gemini-1.5-Flash: Technical Details and Use Cases https://medium.com/google-cloud/multimodality-with-gemini-1-5-flash-technical-details-and-use-cases-84e8440625b6

[17]. Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz, Hong Yu; PaniniQA: Enhancing Patient Education Through Interactive Question Answering. Transactions of the Association for Computational Linguistics 2023; 11 1518–1536. https://doi.org/10.1162/tacl_a_00616