Multi-Modal Summarization for Scientfic Papers

INFO 5731, Computational Methods for Information Systems – Section 020

Fardeen Ali Mohammed - 11687259
FardeenaliMohammed@my.unt.edu

Contents

Abstract

Advanced summarizing methods must be developed due to the increasing amount of scientific articles to assist researchers effectively in understanding articles without reading them fully. This study investigates a multi-modal approach to articles through the integration of textual and visual data, including figures, tables, and images into the summary process. Moreover, this study attempts to figure out how adding visual information improves the quality and informativeness of summaries using a retrieval-augmented generation (RAG) framework. Preliminary findings indicate that multi-modal summarization provides richer and more contextual accurate insights than a text-based approach. Further, we want to evaluate 100+ articles with human-based summarization.The code, data, analysis, and results can be accessed on GitHub at: https://github.com/Fardeen210/Text-Summarization

Keywords:Summarization, RAG, Multi-Modal RAG,Vectors

1 Introduction

Text summarization is key to finding solutions for the issues involved in handling large amounts of data in practical use these techniques improve efficiency, decision-making, and availability of necessary details when cumbersome information is simplified into small packages. In research, for example, summarization tools assist scholars in identifying the most relevant documents from the excessive number of articles, reports, and experimental results; methods aiding multi-vector retrieval enable easy organizing of tables, tables, and images for easy retrieval by the search system. A text summary is performed more generally and indicates the problem's directions for further study. In contrast, a table summary shows the most significant statistic values that help researchers avoid trivial work and look at the situation across different disciplines.

Since it combines textual and visual summaries, multi modal summaries contribute to accurate and engaging storytelling. For example, image summaries make intricate charts more straightforward to understand, and text summaries draw out significant highlights of the interviews or reports. Media outlets using multi-vector retrievers can effectively facilitate the enhanced and efficient categorization of multimedia databases to release breaking news quickly.

Text and table summaries condense financial reports or operational data into valuable information, and image summaries help managers and employees who need more background knowledge of the topic understand graphically presented information. For example, the retail industry can apply the summarization model in producing daily sales data summaries for different regions to help discover patterns and resolve discrepancies on the same day. A good integration of summarization into the business processes means that new trends can be responded to quickly and enhance strategic management.

The approach of utilizing multi vector retrievers, which was explained in the methodology, is beneficial for improving summary since summarization involves indexing of summaries in addition to preserving raw data. This approach increases the large datasets' scalability, specificity, and efficiency with multi modal features. Summaries help retrieve information while enabling easy integration of modalities,

such as adding image summaries to text, making the result more contextually rich and better.

1.1 Background

Accoding to El-Kassas et al. (2020b) [5], the earliest development of the summarization field came from Luhn's effort in 1958, which proposed extractive summarization by pointing out the potential text segments for abstracts. Extraction approaches prevailed until the middle of the twentieth century, and the core approach was about choosing and joining significant sentences. In the 1990s, there was a move to abstractive summarization with the intent to paraphrase information, but this brought about problems in NLP. The maturity of the 2010s with RNNs and transformers brought more structured, contextual summaries, thus influencing policy decisions. To get the best of both worlds, new hybrid approaches appeared to reconcile exactness from the extraction with adaptability from abstraction. There is still a concern about how to assess the summaries effectively, and the evaluation is now supported by both the measurable automated and the global qualitative assessments by humans. More recent research intentions are to increase summary quality, increase aids for context recognition, and explore domain-specific implementations because of language analysis and system sophistication developments.

Existing studies on long document summarization rely on text information only while ignoring visual information, such as figures, diagrams, tables, images, and others, which could be particularly useful. This visual information has experimental or various data points that assist users in understanding the article more elegantly. This project aims to enhance the quality of text summarization by combining text and visual information using multi-modal retrieval augmented generative.

1.2 Purpose and Research Questions

This paper purpose a new approach to multi-modal summarisation, focusing on constructing joint schemes for textual and visual data to generate coherent summaries of scientific articles. To achieve better results from big data, the methodology targets meeting the needs of academic researchers, professionals, and practitioners. In addition to enhancing access to scientific information for people with disabilities, this study provides the theoretical background for further enhancements to the multi-modal summarization approach to be used in other areas such as education, business intelligence, and knowledge management. Finally, this work aims to redefine the way to extend document summarization and how to combine different forms of data to improve usability and relevance.

The key research question focuses on how the integration of visual and textual data can improve the quality of text summarization. The purpose of the project is to evaluate the efficacy of combining text and visual elements to create more informative summaries for long documents, particularly scientific papers.

1.3 Overview of the Report

In this research, an inquiry on utilizing large language models (LLMs) and incorporating multimodal data to improve text summarization quality is presented. This paper addresses critical issues such as contextual relevance, synthesis strategies as well as the impact of multimodal fusion focusing on two LLMs, GPT-4o-mini and Mistral-Large-Latest. Critical elements of the study are highlighted in each section of the report, as follows:

- **Introduction** This section provides a breif introduction about our thought process, a short history, background, significance, and purpose of this research. It introduces the research questions and purpose of experiment.
- **Related Work**This section provides an overview of text summarization techniques, methods for the fusion of multi modal data, and applications of LLMs. It welcomes earlier works in the categories of abstractive and extractive summarization, multi modal fusion, and context optimization before relating them to the goals of this research.
- **Data Collection**This section also discusses text extraction and pre-processing, where semantic segmentation techniques are used with OpenAI embeddings to store. It also describes how vector stores were set to store and recover content to enhance the summaries' production.
- **Methodology** describes the methods, techniques, and frameworks used to overcome summarization difficulties: It comprises adding contextual image summaries to the text chunks, using vector stores to optimize search, and input refining for LLMs for summarizing.
- **Evaluation Metric** this section explain in detail about evaluation metric used.
- **Experiment** This work elaborates on the experiment procedure and uses some evaluation metrics, including the ROUGE, BLEU, and BERT scores. It evaluates two LLMs in different input conditions: text—only conditions and text incorporating images of the summaries.
- **Results** analyses the findings arrived at by integrating the multi modal data and restructuring the learning context. Other performance visualizations and comparisons indicate the effectiveness of the presented methodology. The issues concerning multi modal summarization and the impact of these conclusions on LLM are also considered.
- **Conclusion and Limitations** This section wraps up the study's findings, the gains made in the opportunity to provide better summary quality, and the proposal to

optimize vector stores perfectly. It also suggests possible future research areas, such as QA-based summarization and extension of the method to other common types of media, such as auditory and video

2 Related Work

Text summarization has become a hot topic in the natural language process, and its system can be divided into several classes. Below, we discuss three subtopics closely related to this project: methods of summarization, handling of multi modal data, and context effects in large LLMs.

In recent years, immersive work has taken place in the field of text summarization. Some of the related works are mentioned here:

2.1 Summarization Approaches

Based up on methods Text summarization methods are broadly classified into abstractive, extractive, and hybrid approaches. Abstractive summarization aims to paraphrase the original text by generating new sentences while retaining the core meaning. This method relies heavily on natural language generation and semantic representation, which makes it challenging but essential for generating human-like summaries. Balaji et al. (2016) [3]

Based on Input Size: Number documents used to generate the summarization is a vital part. for e.g.: if there is a single document as source, then it is a Single-Document Summarization (SDS) model which has some concerns related to redundancy, coverage, etc. And when there is more than one document used as source, then it is a Multi Document Summarization (MDS) model which is more complex than SDS because it uses various sources (El-Kassas et al., 2020)[5]

2.2 Multi modal Summarization

Li et al. [9] focused on multi modal text summarization (MMTS) where they integrated both text and images features to produce informative summaries. They developed a multi-modal fusion using ResNet-50 for image feature extraction and seq2seq with LSTM layers for text pre-processing. The ROUGE-1 score achieved was 52.70.

Argade et al. [1] (2024) introduces a multi modal abstractive summarization approach using BERT with an attention layer mechanism. This study aims to identify challenges associated with video summarization by embodying multiple attributes such as images, text, audio and video. They used BERT-based attention to combine these attributes. To encode text, Bidirectional Gated Recurrent Unit (BI-GRU) is used and for images and videos LSTM networks. Results obtained by this model are ROGUE-1 score of 60.20.

Jing et al. (2023) [7] introduced a Vison-Enhanced Generative Pre-Trained Language Model which integrates textual data and visual data. This model uses BART and Swin Transformer for text and visual feature extraction respectively. To fuse these two attributes, they used a multi-head attention mechanism and acquired 53.3 of ROGUE-1 score which surpasses all baseline models.

Chen and Zhuge (2019) [4] present an extractive MMTS method using multi modal Recurrent Neural Network(RNN), which deals with images and text in a document. These modal combines both attributes as a classification problem, using a logistic classifier which calculates probability of an image associated with available text. They used bi-directional RNN to encode text features and VGGNet for image feature extraction.

2.3 Context Effects in Large Language Model

Liu et al. (2024b) [11] analyzed the performance of large language models when long context is given as input. Through controlled experiments, they revealed that the performance of language model is degraded depending upon position of relevant information within the context. And models show improved performance when relevant information is at beginning and ending of the input context.

Joshi et al. (2022) [8] explains a new innovative approach using a variety of phrase variables including topic information, semantic content, keywords, and sentence location. RankSum framework uses a graph-based method to rank keywords while considering sentence location into consideration. Siamese network and probabilistic topic models are used to determine sentence embeddings and global importance respectively. It achieved ROUGE-1 score of 53.25

Baek er al. (2024) [2] proposes a interesting VATMAN(Video-Audio-Text Multimodal Abstractive Summarization. An innovative generative modal that uses a trimodal hierarchical multi head attention system to encompass text, audio and video mobilities. This method has improved summarization and achieved ROUGE-1 score of 52.53

3 Data Collection

3.1 Steps in Data Collection

As we considering Scientific articles as our corpus. Most of the articles are easily accessible in PDF format. There are many libraries present which parse PDF and extract Text and Visual content such as PDFPlumber, PDFMinner.six, Pypdf2, PyMuPDF and etc are for Text only. For our experiment we used unstructred.io. which is open source tool-kit, designed to make it easier to import and pre-process a variety of data types, including Images, Tables from various text-based documents. Further, text data extracted and stored in vector store and all figures tables are exported in JPEG format and collected there summaries through GPT-4o-mini.

3.2 Data Processing and Chunking

According to gkamradt. [6] to improve performance language model application there are five level's chunking strategies, which are discussed below:

1. Level 1: Character Splitting: It is most straightforward method to divide the entire text corpus into certain chunks with N-characters in each. Irrespective of document structure we divide out content by predetermined character count.
2. Level 2: Recursive Character Text Splitting: The issue in Level 1, is solved in here. To maintain document structure a series of separators such as [""] - Double new line, or most commonly paragraph breaks "" - New lines, " " - Spaces, "" - Characters.
3. Level 3: Document Specific Splitting: This method is suitable for different document formats such as Markdown, python/Js, tables and images, Each document has its own separators
4. Level 4: Semantic Chunking: The aforementioned methods focuses on structure and content of the document but this method aims to determine semantic relationship between each chunks with extracting semantic meaning from embeddings.
5. Level 5: Agentic Chunking: In this method, with the help of LLM model we determine how much can be included in each chunk based on the context.

4 Evaluation Metric

As with most models in the field, the effectiveness of the summarization models can be measured through ROUGE and BERTScore. ROUGE is a recall-based method that measures the extent to which n-grams from the generated summary overlap with the reference summary.In precise ROUGE-1, ROUGE-2, and ROUGE-L focus on unigrams, bigrams, and longest common subsequences, respectively. BERTScore relies on contextual embedding from BERT to rate the likeness of summaries produced to that of the reference. These measures are precise where necessary, recall-oriented where needed, and preserve semantic equivalence as well.

4.1 ROUGE score

According to Lin (2004) [10] presented ROUGE-1, ROUGE-2 and ROUGE-L are influential measures to assess the quality of summaries in relation to the reference summaries. ROUGE-1 concentrates on overlapping unigrams, which refers to one of the simplest ways of comparing two texts for similarity. On the other hand, ROUGE-2 expands on this by considering bigrams (two successive words) and introducing a two-gram overlap to capture more about the content relation. For that purpose, ROUGE-L computes the longest common subsequence (the sequence of words in the references that provides the highest number of matches in the candidate summary) to capture the order of words and, thus, the content's coherence. Together, these measures provide a more detailed view of the contracts between the original content and the summary and the repetition of the summary structure.

4.2 BERT score

Zhang et al. (2020) [12] BERTSCORE is a wholly automated text evaluation approach used in text generation tasks, where it measures the BERT-based contextual embeddings computed between the generated candidate and reference sentences. Unlike BLEU, which only pays attention to surface-level matching, BERTSCORE evaluates meaning by measuring token similarity based on the cosine similarity of the contextual embedding. It is resistant to adversarial examples, closely correlates with human evaluation, and does not depend on the specific task of NLG. As a result, owing to multiple language support and weight assignment capabilities, BERTSCORE has shown increased efficacy in functions such as machine translation and image captioning; it performs at a higher level than conventional metrics while identifying semantic meaning.

Summary evaluation using these measures offers a complete view of the summarization quality while taking into account the summary's quantitative measure, the extent of the coverage of the input documents, and the summary's semantic correctness.

5 Methodology

Initially, we extracted raw text and images from each article. We sliced the entire document using the semantic segmentation. This approach has been used in various areas including natural language processing, computer vision. When applied to the text slicing to meaningful fragments and using vector storage for integrating the visual content, the envisioned ability to generate good summaries with contextual depth is highly promoted. In terms of approach, this approach can be applied to different industries including the auto-generation of reports to the improvement of the assistive tools for visually impaired users.

Second, all the graphs, flow charts, and tables are saved as JPEGs for flexibility in integrating auxiliary visuals. Converting these visual components into a selected image format is uniform and easy to process for further use. These are then converted into a base64 string, and the base64 string acts as input for the GPT-4o-mini. Base64 is used to translate the image data into text, which is helpful for text-based systems and to match language models capable of multiple modal data processing. Specifically, with the help of a specific question used in the application GPT-4o-mini, it is possible to receive a relatively brief yet informative summary for every picture. These summaries are applied to descriptions for retrieval and guarantee that the key points of each graphical data from graphs, tables, and flowcharts are, to the extent, reproduced in the text.

Third, image summaries are assembled using the respective textual portions based on a similarity search paradigm. The image summaries are compared to each chunk's text to determine the most coherent associations. The similarity

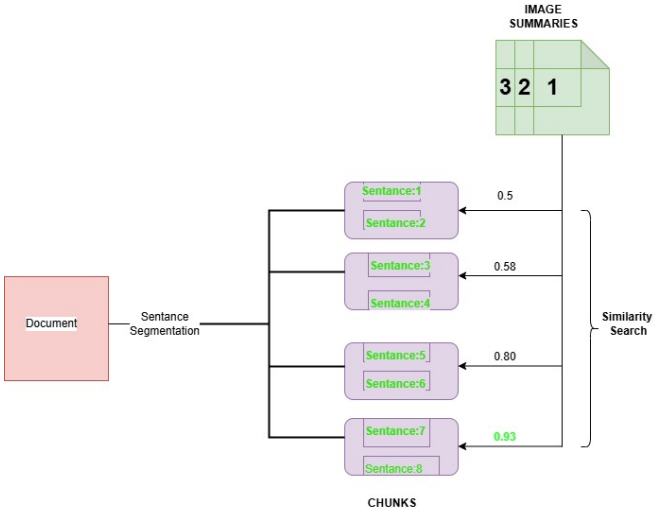

Figure 1. Integrating Image Summaries with Text Chunks

search guarantees that every image summary is added to the most relevant chunk regarding the content's context. Each chunk has been chunked via semantic strategy and will be richer in context. Once the matching is complete, each document chunk from the vector store is fetched and updated with the corresponding image summary. The block of text and the part of the image with enriched information are returned to the vector store to have them in the proper format.

6 Experiment and Data Analysis

First, we conducted our testing phase by identifying five documents to use as a sample. The top 5 most viewed documents from arXiv are selected to determine the effectiveness and efficiency of our proposed approach to summarization.

The first process was to extract the textual content of the documents by using a semantic segmentation method. The segmentation process categorizes the document into relevant categories, which aids in retaining the document's logical form and cohesiveness. Spontaneously after that, the segmented text will convert with OpenAI embeddings, which makes it possible to transform the textual information into vectors of higher dimensions.

6.1 Experiment 1: RAW TEXT ONLY

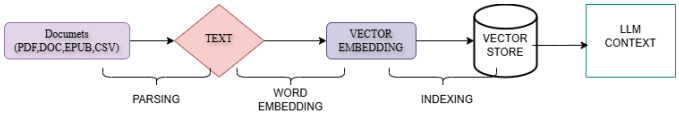

Figure 2. Experiment 1: flow chart

The generated embeddings were then passed to two different LLMs for summarization processes, as discussed next. For this task, the chosen models were GPT-4o-mini by OpenAI and the "mistral-large-latest" of Mistral. The two architectures are expected to produce summaries out of the encoded embeddings. This step was implemented to evaluate these models' summarization capacities based on text inputs only. For Experiment 1, we presented only the text embeddings as input context in each language model.

	RAW TEXT ONLY				
Document	**Model**	**Metric**	**Precision**	**Recall**	**F-Measure**
Article_1	GPT-4o-mini	**rouge1**	0.5486	0.4389	0.4877
		rouge2	0.2448	0.1955	0.2174
		rougeL	0.2986	0.2389	0.2654
	Mistral AI	**rouge1**	0.5948	0.5056	0.5465
		rouge2	0.25	0.2123	0.2296
		rougeL	0.3399	0.2889	0.3123
Article_2	GPT-4o-mini	**rouge1**	0.35430	0.50000	0.41470
		rouge2	0.11490	0.16260	0.13470
		rougeL	0.19430	0.27420	0.22740
	Mistral AI	**rouge1**	0.25190	0.53230	0.34200
		rouge2	0.08050	0.17070	0.10940
		rougeL	0.10690	0.22580	0.14510
Article_3	GPT-4o-mini	**rouge1**	0.35430	0.50000	0.41470
		rouge2	0.11490	0.16260	0.13470
		rougeL	0.19430	0.27420	0.22740
	Mistral AI	**rouge1**	0.25190	0.53230	0.34200
		rouge2	0.08050	0.17070	0.10940
		rougeL	0.10690	0.22580	0.14510
Article_4	GPT-4o-mini	**rouge1**	0.52340	0.32210	0.39880
		rouge2	0.14170	0.08700	0.10780
		rougeL	0.36720	0.22600	0.27980
	Mistral AI	**rouge1**	0.39290	0.58170	0.46900
		rouge2	0.16940	0.25120	0.20230
		rougeL	0.24030	0.35580	0.28680
Article_5	GPT-4o-mini	**rouge1**	0.5143	0.5217	0.5180
		rouge2	0.1942	0.1971	0.1957
		rougeL	0.3714	0.3768	0.3741

Figure 3. Experiment:1 ROUGE Results

I finally tested the performance of the models using Length, ROUGE-1, ROUGE-2, ROUGE-L, BERT, and BLEU scores. All these metrics were compared with the abstract of each document, which, in turn, was used to evaluate the quality of the generated summaries. The above experiment was performed independently on both models to clearly understand how the two models perform in terms of summarization when they're incorporated into text-only data.

6.2 Experiment: 1 Analysis

Mistral AI performs slightly better than GPT-4o-mini with recall values, mainly in **ROUGE-1** scores concerning most articles. For instance, in **Article 1**, the percentage of relevant content captured by Mistral AI is higher than in GPT-4o-mini: recall =0.5056 vs 0.4389. This tendency is noticed in other articles: compared with Mistral, the proposed model has better **ROUGE 1 ** and **2 ROUGE – L **results. Still,

	RAW TEXT ONLY	
	Model	BERT Score
Article 1		
	GPT	0.88869
	Mistral AI	0.86661
Article 2		
	GPT	0.8800
	Mistral AI	0.8666
Article 3		
	GPT	0.8787
	Mistral AI	0.8415
Article 4		
	GPT	0.87318
	Mistral AI	0.85995
Article 5		
	GPT	0.89410
	Mistral AI	0.86458

Figure 4. Experiment:1 BERT Score

it displays overall higher accuracy, including **Article 5** where the **ROUGE-1** accuracy of GPT-4o-mini (0.5143) is higher compared to Mistral AI (0.4565) clearly showing that GPT-4o-mini has higher accuracy in the selection of its terms and the focus of the outputs.

Nevertheless, on the specific **BERT scores**, which are metrics of semantic closeness, GPT-4o-mini attains higher results. For example, as for Article 5**, GPT-4o-mini achieved 0.89410, while the Mistral AI obtained 0.86458, which means more of GPT-4o-mini's output fits the context in which the reference text was written. Altogether, Mistral AI proves to be more effective in the aspect of recall and coverage so that it is more adequate for covering different facets of information; at the same time, GPT-4o-mini adapts better to providing more accurate and semantically similar responses, more congruent with the original text.

Furthermore, these experimental results were compared with the results from Experiment 2, which was performed on text and image summaries.

6.3 Experiment 2: RAW TEXT AND IMAGES

In the next phase of the study, the second phase, we added a multi-modal aspect to the summarization procedure. To achieve this, we first created a vector store that would store all the textual raw data of the documents. In other words, the vector store stores the segmented text chunks that can be retrieved for different content requirements. We then turned our attention to how to incorporate visual data into the summarization flow to support the improved summaries.

More precisely, we created summaries for relatively complex images like graphs, flowcharts, and tables extracted from the documents. These image descriptions were derived from

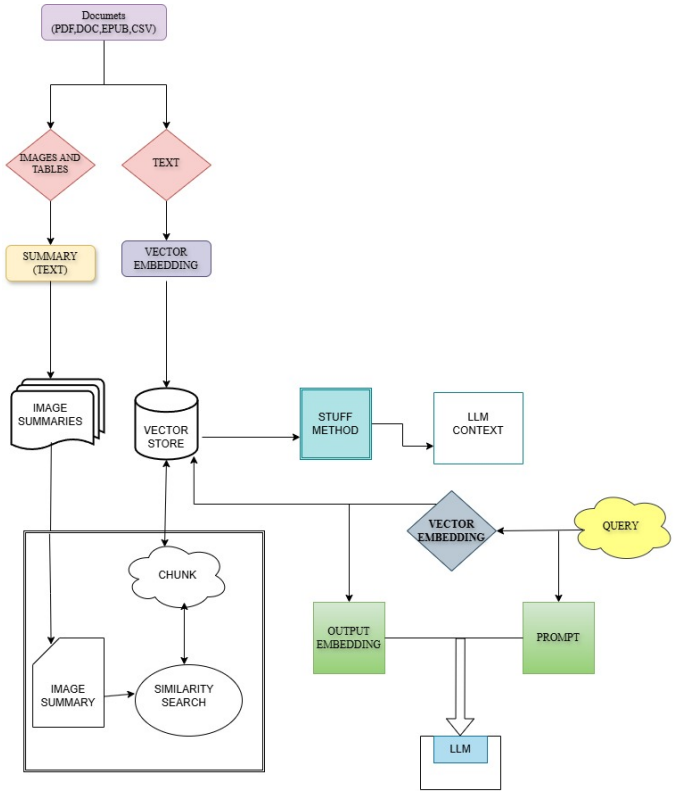

Figure 5. Experiment 2: flow chart

post-image-IDs with a new GPT-4o-mini prompt describing the photos briefly. Applying all these strategies, we used a similarity search mechanism to identify relevant text chunks for every image summary. This mechanism compares the features of the image summary with the features of the text chunks in the vector store, and retrieves the most similar text chunks, once the image summary is in its complete form and stored in the vector store.

The rationale for performing this step was to generate a representation of the document likely to contain richer context information. In this way, by adding summaries derived from images to the endings of the text chunks, we attempted to give more comprehensive information to the language model of the document. This multi-modal input, which combines textual and visual information, was designed to enhance the language model's understanding of the document's content. The combined information was then stored in the vector store for further experimentation.

6.4 Experiment 2: Analysis

For Experiment 2, we retrieved the enriched content from the vector store and provided it as context to the same two LLMs: GPT-4o-mini and Mistrals mistral-large-latest. This experiment was conducted to identify the impact of including visual summaries in producing text summaries.

	RAW TEXT AND IMAGES				
Document	Model	Metric	Precision	Recall	F-Measure
Article_1	GPT-4o-mini	rouge1	0.3578	0.6778	0.4683
		rouge2	0.1206	0.2291	0.158
		rougeL	0.1584	0.3	0.2073
	Mistral AI	rouge1	0.287	0.7111	0.4089
		rouge2	0.1258	0.3128	0.1795
		rougeL	0.1614	0.4	0.23
Article_2	GPT-4o-mini	rouge1	0.181	0.629	0.2811
		rouge2	0.0581	0.2033	0.0904
		rougeL	0.0951	0.3306	0.1477
	Mistral AI	rouge1	0.1542	0.5806	0.2437
		rouge2	0.0472	0.1789	0.0747
		rougeL	0.0707	0.2661	0.1117
Article_3	GPT-4o-mini	rouge1	0.3088	0.4746	0.3742
		rouge2	0.1144	0.1761	0.1387
		rougeL	0.1618	0.2486	0.196
	Mistral AI	rouge1	0.2799	0.661	0.3933
		rouge2	0.1367	0.3239	0.1922
		rougeL	0.1722	0.4068	0.242
Article_4	GPT-4o-mini	rouge1	0.3627	0.5337	0.4319
		rouge2	0.1115	0.1643	0.1328
		rougeL	0.2157	0.3173	0.2568
	Mistral AI	rouge1	0.2648	0.5817	0.3639
		rouge2	0.0811	0.1787	0.1116
		rougeL	0.1422	0.3125	0.1955
Article_5	GPT-4o-mini	rouge1	0.2537	0.6159	0.3594
		rouge2	0.0868	0.2117	0.1231
		rougeL	0.1433	0.3478	0.203

Figure 6. Experiment:2 Results

	RAW TEXT AND IMAGES	
	Model	BERT Score
Article 1		
	GPT	0.8887
	Mistral AI	0.8666
Article 2		
	GPT	0.8967
	Mistral AI	0.8610
Article 3		
	GPT	0.8481
	Mistral AI	0.8323
Article 4		
	GPT	0.8617
	Mistral AI	0.8399
Article 5		
	GPT	0.8891
	Mistral AI	0.8455

Figure 7. Experiment:2 Bert Score Results

The evaluation reveals a tradeoff between GPT-4o-mini and Mistral AI; while it outperforms in one, it lags in other areas. F-measure is higher, and the BERT score is higher in all cases with GPT-4o-mini. For instance, in Article 1, concerning Rouge-1, GPT-4o-mini has an F-score of 0.4683 and a BERT score of 0.8887, while for Mistral AI, the Rouge-1 F-score is 0.4089, and the BERT score is 0.8666. However, Mistral AI has a better recall, precisely 0.7111 in the first article, in contrast to 0.6778 in the GPT-4o-mini. The same general patterns repeat in other articles; in the fourth article, GPT-4o-mini yields 0.4319 in Rouge-1 F-measure and 0.8617 in BERT score vs. Mistral AI scoring 0.3639 in Rouge-1 F-measure and 0.8399 in BERT score; however, Mistral AI has a slightly higher recall of 0.5817 as compared to As it can be seen from these results, GPT-4o-mini performs better in terms of accuracy and lower similarity in semantics. At the same time, Mistral AI prefers scoring higher in the recall area with the parameters lowered to sacrifice precision and alignment.

The evaluations of GPT-4o-mini and Mistral AI show the variability of precision, recall, and the F-measure; the F-measure of ROUGE-1 ranges from 0.2437 up to 4683. However, it is worth mentioning that GPT-4o-mini is slightly more consistent, but the difference could be more pronounced and depend on the context. The evidence of using the ROUGE-2 and the ROUGE-L also proves that the assessment of text similarity is much more complicated and that only some models can dominate the others across all the articles.

As with semantic similarity, the overall scores for both model types are high, as presented by the BERT Score results. GPT remains between 0.8481 and 0.8967, whereas Mistral AI varies between 0.8323 and 0.8666. These high scores indicate that both models produce text that retains a high degree of semantic similarity to reference documents; this is in concert with the differences in the micro-level ROUGE metrics that were previously observed.

7 Results and Discussion

The observations reveal a clear trend: The graph-based results are better than those when using the raw text and images with all the tested metrics. For example, in [Article 1], the BERT score in the raw text-only condition is 0.88869 with GPT-4o-mini, a tad higher than 0.8887 with raw text and images. Similar to the previous case, for Article 5, raw text only gives GPT-4o-mini a score of 0.89410 while the inclusion of images decreases to 0.8891. The same is done for Mistral AI, and it becomes clear that adding images does not improve performance based on standard measures.

Further, to evaluate summarization we have compared the length of the result as one of the category.Surprisingly, length of each model output has tripled. Since the overall length of the summaries has gradually extended, particularly in text-with-image conditions, it can be inferred that the integration

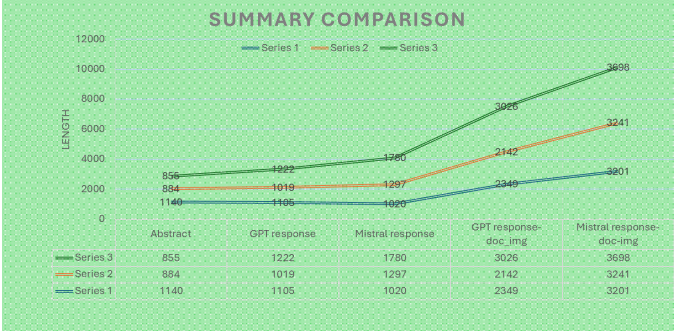

Figure 8. Output length

of picture information causes more comprehensive outputs to be generated. But it does not necessarily mean higher ROUGE or BERT scores To do that, let us switch to a different approach.

One has to say that this can be discussed in terms of the evaluation approach since here the abstract was considered to be the reference summary. As abstracts are limited to the textual representation and do and not directly related to images, incorporation of images could also add noise, rather than helping in terms of semantic matching. Thus, integration of images does not increase standard for better quantity measurements such as BERT and ROUGE scores.

8 Conclusion and Limitations

8.1 Limitations

There are few limitations should be taken into account:

1. Dependence solely on the abstracts as the benchmark for assessment could yield a less accurate measure of the quality of the summaries produced by the models because the models make more elaborate summaries than the mere abstracts in longer forms.
2. Current methods of evaluation (ROUGE and BERT scores) may need to be revised for the summative quality of image-inclusive summaries because they are text-driven.These measures may even need to capture the generated summaries' coherence, relevance, or informativeness, especially when graphics are involved.
3. The text and image extraction method can be only partially accurate as some arrows and texts may have been misplaced, leaving inconsistencies in the summarized input and result.
4. The lack of human contribution to evaluation prevents an indication of the qualitative aspects of the summaries, such as their quality, ease of reading, organization, and ability to apply to real-world scenarios. The human evaluation could concentrate on specific aspects of the summary quality that have yet to be addressed by the most popular metrics, such as the correctness of facts and coherency of the summary.

8.2 Conclusion

The result of our experiment shows that adding a summary of images of an article to the text, greatly lengthened the generated outputs, especially with GPT and Mistral models. This implies that the models made good use of the extra visual information, producing summaries that were more thorough. However, even with this improvement, evaluation metrics like ROUGE and BERT decreased when the abstract was used as a reference to compare with machine response. Further, this research will be focused to evaluate more articles and manually compare each response quality to get a final conclusion.

References

[1] Dakshata Argade, Vaishali Khairnar, Deepali Vora, Shruti Patil, Ketan Kotecha, and Sultan Alfarhood. 2024. Multimodal abstractive summarization using bidirectional encoder representations from transformer with attention mechanism. *Heliyon* 10, 4 (2 2024), e26162. https://doi.org/10.1016/j.heliyon.2024.e26162

[2] Doosan Baek, Jiho Kim, and Hongchul Lee. 2024. VATMAN: integrating Video-Audio-Text for Multimodal Abstractive summarizatioN via Crossmodal Multi-head Attention Fusion. *IEEE Access* 12 (1 2024), 119174–119184. https://doi.org/10.1109/access.2024.3447737

[3] J. Balaji, T.V. Geetha, and Ranjani Parthasarathi. 2016. Abstractive summarization. *International Journal on Semantic Web and Information Systems* 12, 2 (4 2016), 76–99. https://doi.org/10.4018/ijswis.2016040104

[4] Jingqiang Chen and Hai Zhuge. 2019. Extractive summarization of documents with images based on multi-modal RNN. *Future Generation Computer Systems* 99 (4 2019), 186–196. https://doi.org/10.1016/j.future.2019.04.045

[5] Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2020. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications* 165 (7 2020), 113679. https://doi.org/10.1016/j.eswa.2020.113679

[6] gkamradt. [n. d.]. 5 Levels Of Text Splitting at main · FullStackRetrieval com/RetrievalTutorials. https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb

[7] Liqiang Jing, Yiren Li, Junhao Xu, Yongcan Yu, Pei Shen, and Xuemeng Song. 2023. Vision Enhanced generative pre-trained language model for multimodal sentence summarization. *Deleted Journal* 20, 2 (1 2023), 289–298. https://doi.org/10.1007/s11633-022-1372-x

[8] Akanksha Joshi, Eduardo Fidalgo, Enrique Alegre, and Rocio Alaiz-Rodriguez. 2022. RankSum—An unsupervised extractive text summarization based on rank fusion. *Expert Systems with Applications* 200 (3 2022), 116846. https://doi.org/10.1016/j.eswa.2022.116846

[9] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018. Read, watch, listen, and summarize: Multi-Modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering* 31, 5 (6 2018), 996–1009. https://doi.org/10.1109/tkde.2018.2848260

[10] Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. *Meeting of the Association for Computational Linguistics* (7 2004), 74–81. http://anthology.aclweb.org/W/W04/W04-1013.pdf

[11] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (1 2024), 157–173. a{00638

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv (Cornell University)* (4 2020). https://arxiv.org/pdf/1904.09675.pdf