MuRAR: A Simple and Effective Multimodal Retrieval and Answer Refinement Framework for Multimodal Question Answering

Zhengyuan Zhu[1], Daniel Lee[2], Hong Zhang[2],
Sai Sree Harsha[2], Loic Feujio[2], Akash Maharaj[2], Yunyao Li[2]
[1]University of Texas at Arlington, [2]Adobe
zhengyuan.zhu@mavs.uta.edu, dlee1@adobe.com, yunyaol@adobe.com

Abstract

Recent advancements in retrieval-augmented generation have demonstrated impressive performance on the question-answering task. However, most previous work predominantly focuses on text-based answers. Although some studies have explored multimodal data, they still fall short in generating comprehensive multimodal answers, especially step-by-step tutorials for accomplishing specific goals. This capability is especially valuable in application scenarios such as enterprise chatbots, customer service systems, and educational platforms. In this paper, we propose a simple and effective framework, MuRAR (**Mu**ltimodal **R**etrieval and **A**nswer **R**efinement). MuRAR starts by generating an initial text answer based on the user's question. It then retrieves multimodal data relevant to the snippets of the initial text answer. By leveraging the retrieved multimodal data and contextual features, MuRAR refines the initial text answer to create a more comprehensive and informative response. This highly adaptable framework can be easily integrated into an enterprise chatbot to produce multimodal answers with minimal modifications. Human evaluations demonstrate that the multimodal answers generated by MuRAR are significantly more useful and readable than plain text responses. A video demo of MuRAR is available at https://youtu.be/ykGRtyVVQpU.

1 Introduction

The emergence of retrieval-augmented generation (RAG) techniques (Lewis et al., 2020; Gao et al., 2023) and large language models (LLMs), such as GPT models (Brown et al., 2020; OpenAI, 2023), Gemini (Anil et al., 2023), and Llama (Touvron et al., 2023), has significantly transformed the field of question answering (QA) and improved the quality of responses generated by AI assistants. However, the current generation of AI assistants has limitations in delivering comprehensive multimodal answers to user questions, especially step-by-step tutorials on accomplishing specific goals. This capability is particularly valuable in enterprise scenarios, where critical information can often be extracted from product documentation that includes multimodal data. In such cases, images, tables, and videos are often *crucial* for understanding complex, domain-specific topics. Enhancing AI assistants with the ability to incorporate multimodal information can, therefore, significantly improve user comprehension and engagement (Zhang et al., 2024; Singh et al., 2021). This improvement offers several benefits, including increased productivity, reduced barriers to entry, higher product adoption rates, enhanced creativity, and improved user experiences.

Previous work (Talmor et al., 2021; Kumar et al., 2020; Joshi et al., 2024) has primarily focused on leveraging various techniques to better understand multimodal data as input and generate plain text answers to a given query. In another scenario (Singh et al., 2021; Wu et al., 2023), the output may consist of either a text answer or a text answer accompanied by a retrieved image or video appended at the end. However, the existing solutions fail to adequately address the challenges posed by complex questions that require illustrating multiple steps to achieve a goal and integrating various multimodal content within the answer.

In summary, the main challenges are: **a)** How to retrieve the relevant multimodal data that are related and helpful to answer the user questions, and **b)** How to generate a coherent multimodal answer that integrates the retrieved multimodal data. To address these challenges, we present MuRAR (**Mu**ltimodal **R**etrieval and **A**nswer **R**efinement). This simple and effective framework generates coherent multimodal answers containing retrieved multimodal data, such as images, videos, and tables. Our framework comprises three main components: text answer generation, source-based multimodal retrieval, and multimodal answer refinement.

The text answer generation component retrieves relevant text documents based on the user's query and generates an initial text answer using an LLM. The source-based multimodal retrieval component retrieves multimodal data that are relevant to the text answer snippets in the initial text answer. Finally, the multimodal answer refinement component prompts an LLM to generate the final answer by integrating the retrieved multimodal data with the initial text answer.

We implemented and applied this framework on an enterprise-level AI assistant. To verify the framework's effectiveness, we evaluated the quality of the multimodal answers on a human-annotated dataset of 300 questions and answers. The human evaluation results show that the MuRAR framework effectively retrieves useful and relevant multimodal data while maintaining answer readability. Furthermore, the multimodal answers were consistently preferred over plain text answers. Our framework can be adapted to other enterprise-level AI assistants by collecting topic-specific multimodal data and fine-tuning a topic-specific text retrieval model. This work is, to our knowledge, the first to address the problem of generating coherent multimodal answers to user questions.

2 Design of MuRAR

Formally, given a user question q as input and a set of multimodal data $\mathcal{D} = \{D_S, D_I, D_T, D_V\}$, where $\{D_S, D_I, D_T, D_V\}$ denote collections of text document snippets, images, tables, and videos, respectively, the objective is to generate a multimodal answer $A_{mm} = F(S, I, T, V)$. Here, F denotes a function that organizes a set of retrieved multimodal data $(S, I, T, V) \in D$, relevant to the user's question, into a coherent and informative answer.

To achieve a high quality A_{mm}, we propose MuRAR, as illustrated in Figure 1. The text answer generation component uses a RAG-style approach (Lewis et al., 2020; Gao et al., 2023), first retrieving relevant text document snippets $S = \{s_1, s_2, ..., s_n\} \in D_S$ based on user query q and then generating an initial text answer A_t by prompting an LLM. Next, we retrieve multimodal data, namely, $I = \{i_1, i_2, ..., i_m\} \in D_I$, $T = \{t_1, t_2, ..., t_k\} \in D_T$, and $V = \{v_1, v_2, ..., v_l\} \in D_V$ that are relevant to the text answer. Finally, the multimodal answer refinement component generates a final multimodal answer A_{mm} by incorpo-

rating the retrieved multimodal data into the initial text answer.

Notably, directly prompting or using techniques such as chain-of-thought (Wei et al., 2022) with LLMs for tasks involving both text and multimodal data is ineffective for two reasons. First, the complexity of the task overwhelms the LLM as it needs to determine which data to reference, decide whether to display multimodal data and figure out where to place it within the answer. Additionally, this complexity results in low-quality multimodal answers.

2.1 Text Answer Generation

Our text answer generation component follows the RAG-style approach. Specifically, we fine-tuned a pre-trained text embedding model (Reimers and Gurevych, 2019) on an in-house annotated dataset, which includes labels indicating whether a text document snippet s_i is relevant to a user query q. The fine-tuned embedding model is then applied to the text document snippets D_S to create vector indexes using FAISS (Johnson et al., 2019). These vector indexes serve as a database for retrieving relevant text document snippets by calculating the cosine similarity between the user query q and each text document snippet s_i. For each user query q, the top five relevant text snippets are selected. An LLM is then prompted with the user query q and the retrieved top five text snippets to generate initial text answer A_t. The detailed prompt is provided in Appendix A.4.

2.2 Source-Based Multimodal Retrieval

The source-based multimodal retrieval component involves two steps: source attribution and section-level multimodal data retrieval.

Source Attribution. The initial text answer A_t is segmented into multiple sentences, with each sentence representing a continuous text answer snippet $a_{[i,j]}$ spanning from the i-th token to the j-th token in A_t. Each text answer snippet $a_{[i,j]}$ is then compared to every text document snippet in D_S by calculating the cosine similarity between their embeddings. The text document snippet s_i with the highest cosine similarity score is identified as the source for $a_{[i,j]}$. Notably, if the highest similarity score is below 0.6, no source is assigned to $a_{[i,j]}$. These text answer snippets and their sources are the foundation for retrieving multimodal data.
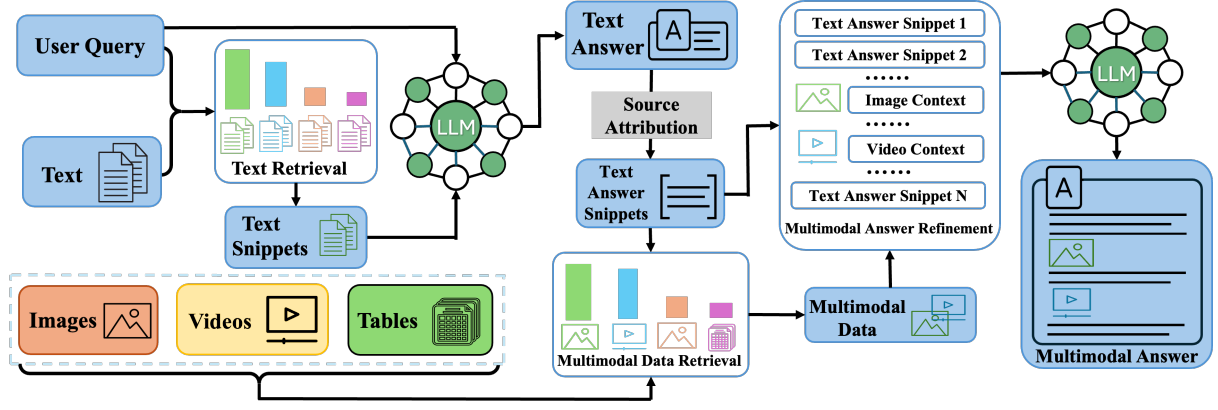
Figure 1: The architecture of the MuRAR framework.

Section-Level Multimodal Data Retrieval. For each text answer snippet $a_{[i,j]}$ and its corresponding source s_i, we first locate the original web document containing s_i. We then identify the section where s_i resides and collect all multimodal data within that section. This approach significantly reduces the search space for multimodal data and improves the precision of the retrieval results. The multimodal data may include images, tables, and videos. Multimodal data are represented using contextual and LLM-generated text features. The contextual text features are derived from the paragraphs before and after the multimodal data. For the LLM-generated text features, we utilize the image captions generated by GPT-4 and the "alt" text attribute extracted from the raw document HTML for images. For tables and videos, we use summarized table content and summarized video transcripts generated by the LLM, respectively.

To efficiently encode these text features, text embeddings are computed using the same fine-tuned embedding model employed for text document retrieval. The multimodal data included in the multimodal answer are retrieved based on the cosine similarity between the text answer snippet $a_{[i,j]}$ and the text embeddings of multimodal data. Only the highest-ranked multimodal data are selected for a text answer snippet. However, the same multimodal data may occasionally be selected for multiple text answer snippets, resulting in duplication within the multimodal answer. To address this, we only keep the multimodal data with the highest similarity score from the retrieval results.

2.3 Multimodal Answer Refinement

After retrieving the multimodal data, an LLM is prompted to refine the initial text answer into a multimodal answer. The prompt includes the user question q, initial text answer A_t, and retrieved multimodal data accompanied by their contextual text features.

To guide the LLM to generate multimodal answers, placeholders are inserted into A_t after the textual answer snippets corresponding to retrieved multimodal data. Each placeholder includes the URL of the multimodal data and its contextual text features, ensuring the LLM incorporates relevant information while minimizing the risk of generating irrelevant details and hallucinations. Additionally, the prompt includes several illustrative question-answering examples to demostrate how multimodal data should be integrated into the final answer. An example prompt is provided in Appendix A.4. The LLM is instructed to replace the placeholders with appropriate descriptions and modify the text answer snippets to produce a coherent and readable multimodal response. Finally, for display in the user interface, the URLs for multimodal data are converted into HTML elements, enabling users to interact with the multimodal content directly.

3 User Interface

We implemented the MuRAR framework and integrated it into a prototype version of the AI assistant within Adobe Experience Platform.[1] As illustrated in Figure 2, when a user queries the AI assistant (A), for example by asking, "What is a good tutorial for creating a schema?", MuRAR retrieves relevant text snippets (B) from documents on Adobe Experience League[2] and generates an initial text answer.
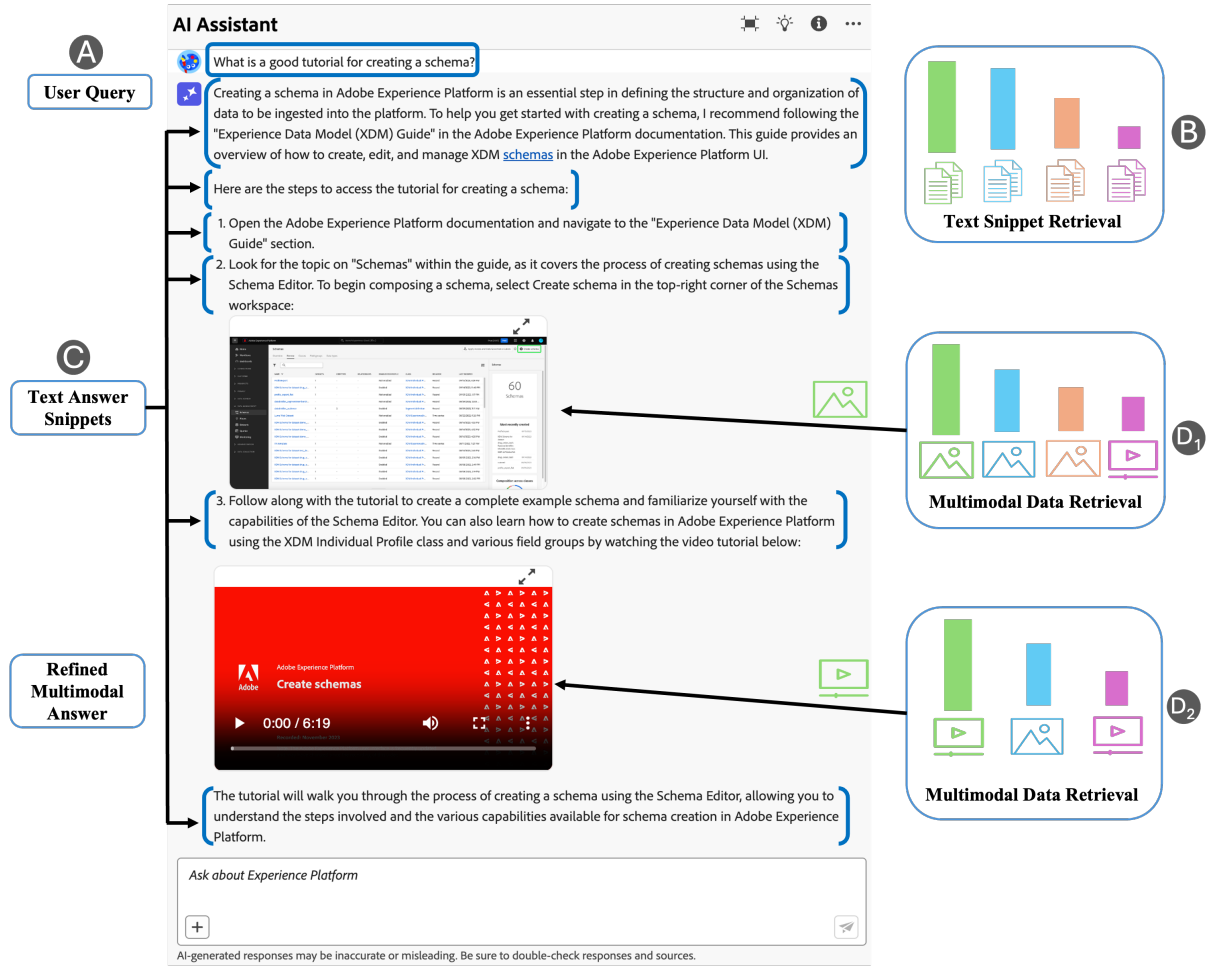

Figure 2: The interface of AI Assistant demonstrating multimodal answers is constructed by combining multimodal data retrieval and answer refinement.

Subsequently, source attribution is applied to each text answer snippet (C) to identify potential multimodal data candidates. Next, MuRAR retrieves the most relevant multimodal data (D1 and D2), which, in this case, include a screenshot of the Schemas Workspace and a video tutorial explaining the schema creation process. Finally, MuRAR integrates the relevant multimodal data into the initial text answer through multimodal answer refinement. Notably, the multimodal elements are interactive: when a user clicks on an element, a pop-up window appears, displaying the multimodal data in full size for detailed viewing.

4 Data Collection

We curated two datasets to support the development of the MuRAR framework and facilitate human evaluation. A multimodal document dataset was created to serve as $\mathcal{D}$ for developing the MuRAR framework. Additionally, a multimodal question-answering dataset was collected and annotated for use in human evaluation.

4.1 Multimodal Document Dataset

The multimodal document dataset was collected from 2,173 web documents from Adobe Experience League [3]. As summarized in Table 1, the dataset contains four modalities: text, image, table, and video. The textual data includes both plain text and tabular content, while the visual data comprises images and videos. Additional details about the data collection process can be found in Appendix A.1.

The text content was scraped from web documents and tokenized using the GPT-2 tokenizer (Radford et al., 2019). Each document was segmented into smaller snippets ranging from 11 to 1,500 tokens, resulting in a total of 18,071 text snippets. For image data, we extracted image URLs along with their surrounding textual context, which includes the text preceding the image (pre-image

---

[3] https://experienceleague.adobe.com/en/docs

Modality	Metric	Value
Text	Count	18,071
	Avg content tokens	192
Image	Count	6,320
	Avg context tokens	238
	Avg caption tokens	94
Video	Count	253
	Avg context tokens	91
	Avg summary tokens	33
Table	Count	2,644
	Avg context tokens	160
	Avg table tokens	223

Table 1: Multimodal dataset statistics.

context) and the text following it (post-image context). To enrich the dataset with more text features for multimodal retrieval, GPT-4 was used to generate captions for the images, resulting in 6,320 annotated image entries. Tables were extracted in JSON format, along with their associated contextual text. As with the image data, the contextual text for tables includes the text preceding and following the table. This process yielded 2,646 table entries. For video content, we collected video URLs, contextual text, and transcripts. In cases where transcripts were unavailable, videos were downloaded, and Whisper (Radford et al., 2023) was used to generate transcripts from the audio. GPT-4 was then employed to summarize these transcripts, resulting in 253 video entries. For all modalities, additional metadata were gathered, including the titles of the web document and the headings of the sections where the multimodal data were located.

4.2 Multimodal Question Answering Dataset

To construct a multimodal question-answer dataset, we collected 764 real customer questions and applied MuRAR to generate answers. Among these, 306 questions were answered with multimodal data. For human evaluation, we randomly sampled 150 questions from this subset of 306 to assess the quality of the generated answers. To analyze the impact of different backbone LLMs on the quality of multimodal answers, we conducted a comparative study using GPT-3.5 and GPT-4. For each model, we generated 150 text-based and multimodal answers, resulting in a total of 300 question-answer pairs. This evaluation enabled us to compare the effectiveness and coherence of multimodal responses across the two models, providing insights into how the choice of LLM influences performance.

5 Human Evaluation

To assess the effectiveness of our multimodal question-answering system, we conducted a human evaluation study in two phases: (1) single-model evaluation and (2) pairwise comparison. The single-model psychometric evaluation was designed to measure three key metrics: usefulness, readability, and relevance of the multimodal answer. For the pairwise comparison, we used a preference-based ranking to determine the overall user preference between text-only and multimodal responses. Notably, we did not assess the quality of the text content itself but rather the added value of integrating multimodal data into the text answers.

5.1 Study Setup

We compiled a dataset of 300 question-answer pairs, evenly distributed between outputs from GPT-3.5 and GPT-4 models. For each question, both text-only and multimodal answers were generated. To evaluate the quality of these outputs, we recruited eight expert annotators, all holding advanced degrees in computer science with substantial experience in natural language processing (NLP). The annotation process was conducted on LabelStudio (Tkachenko et al., 2020-2022), with each question-answer pair evaluated by at least three experts.

5.2 Evaluation Schema

The annotators were asked to rate each multimodal answer on a 5-point Likert scale (1 being lowest, 5 being highest) for the following metrics, which were adapted from (Pradeep et al., 2024):

- **Usefulness:** This metric measures the extent to which multimodal elements contribute to the user's comprehension of the text content. High scores indicate that the multimodal output provides valuable additional information, clarifies complex concepts, or illustrates key points in ways that significantly aid understanding.
- **Readability:** This assesses how well the multimodal elements are integrated with the text, considering factors such as placement, size, and formatting. High scores indicate seamless integration that enhances the overall reading experience.
- **Relevance:** This measures how closely the multimodal elements relate to the content of the

text. High scores indicate that the multimodal output directly supports or illustrates the textual content.

	Model		
Metric	**GPT-3.5**	**GPT-4**	**Average**
Usefulness	3.34	**3.60**	3.47
Readability	3.49	**3.76**	3.63
Relevance	3.66	**3.90**	3.78
Preference Rate	0.82	**0.90**	0.86

Table 2: Evaluation results for multimodal answers generated by GPT-3.5 and GPT-4.

After rating the multimodal answers, annotators were asked to indicate their overall preference between the text-only version, the multimodal version, or if they found them equally effective ("Same").

5.3 Results

Psychometric Evaluation Results. The psychometric evaluation focused on three key aspects: usefulness, readability, and relevance. As shown in Table 2, when examining the answers generated by both GPT-3.5 and GPT-4, the usefulness metric achieved an average score of 3.47, while readability and relevance scored 3.63 and 3.78, respectively. These scores, all above the midpoint of the scale as reflected in Figure 3, suggest that our approach performs well in producing useful, readable, and relevant output. Qualitative feedback from annotators further supports this conclusion, indicating that the multimodal answer provides informative additions to the text, is understandable through its placement, and remains relevant to the associated content. In addition, the average preference rating of 0.86 demonstrates a strong overall preference for our method compared to the text-only alternative. When comparing GPT-3.5 with GPT-4, we found that using GPT-4 as the backbone LLM increased all the metrics, showcasing its superior performance in generating high-quality, multimodal content within our framework.

Inter-Annotator Agreement. To assess annotation reliability, we calculated two inter-annotator agreement (IAA) measures (Table 3). Krippendorff's alpha (Krippendorff, 2011) was 0.4179 overall, indicating moderate agreement across all annotators. However, Cohen's kappa (Cohen, 1960)

	Agreement Metric		
Metric&Model	**K-α_{normal}**	**K-$\alpha_{combined}$**	**C-κ**
Overall	0.4179	0.3437	0.7100
Usefulness$_{GPT-3.5}$	0.4150	0.3468	0.6879
Usefulness$_{GPT-4}$	0.5424	0.4993	0.7900
Usefulness$_{all}$	0.4758	0.4164	0.7383
Readability$_{GPT-3.5}$	0.3418	0.2147	0.6852
Readability$_{GPT-4}$	0.3187	0.2502	0.7048
Readability$_{all}$	0.3424	0.2374	0.6968
Relevance$_{GPT-3.5}$	0.3369	0.2958	0.6459
Relevance$_{GPT-4}$	0.4465	0.3664	0.7291
Relevance$_{all}$	0.3925	0.3323	0.6872

Table 3: The inter-annotator agreement among the eight annotators.

	Model		
Metric	**GPT-3.5**	**GPT-4**	**GPT-3.5 & GPT-4**
Usefulness	0.9741	0.9059	0.9496
Readability	0.7686	0.7059	0.7500
Relevance	0.8576	0.8301	0.8519

Table 4: Standard deviation of evaluation metrics for multimodal answers generated by GPT-3.5 and GPT-4.

between the top two annotators was 0.71, suggesting substantial agreement when excluding outlier data points. To investigate this discrepancy, we conducted further analyses. The standard deviation ranged from 0.8519 to 0.9496, indicating a relatively tight distribution of scores (Table 4). An annotator-specific analysis (Table 5) revealed that Annotators 2 and 4, accounting for 26% and 5.8% of annotations, respectively, had lower average scores compared to others. Annotator 2 averaged 3.0619 (SD = 0.5648), while Annotator 4 averaged 3.3836 (SD = 0.253). These findings suggest that the lower Krippendorff's alpha may be attributed to systematic differences in scoring patterns among a subset of annotators rather than widespread disagreement. It is worth noting that while annotators were provided with in-depth instructions, they did not undergo formal training.

In conclusion, despite some variability in IAA results, the high Cohen's kappa for the top annotators, combined with strong psychometric evaluation scores and preference ratings, supports the overall reliability and effectiveness of our approach. The MuRAR framwork demonstrates clear benefits over text-only alternatives, providing valuable enhancements to textual content.
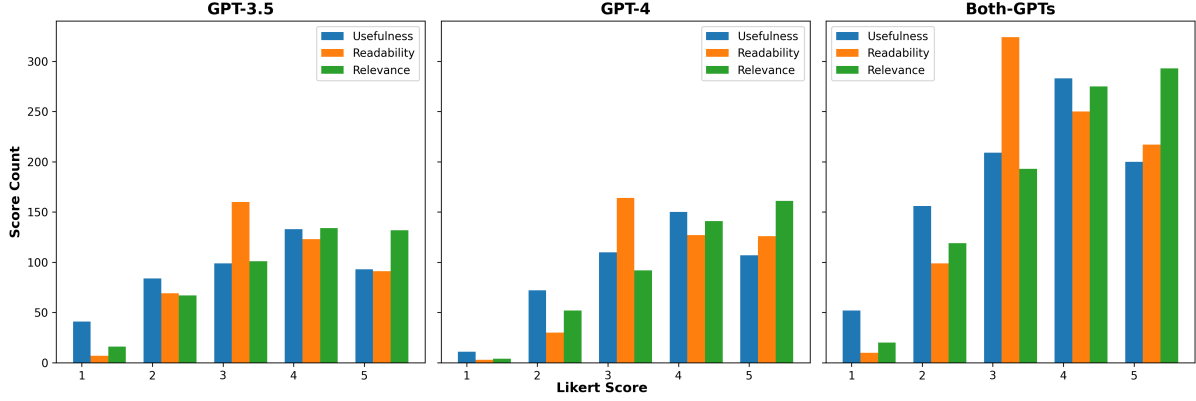
Figure 3: Distribution of score counts for GPT-3.5, GPT-4, and the combined results of GPT-3.5 and GPT-4.

5.4 Limitations

We analyzed the errors and mistakes made by the MuRAR framework during our human evaluation. We identified some issues in the multimodal retrieval component. Although source attribution ensures precision, it can result in low recall, *i.e.*, relevant multimodal data for a text answer snippet may not be in the same section or even the same web document. Additionally, readability can be affected by the multimodal answer refinement component. For instance, multimodal data may contain duplicated information already explained in plain text. This repetition can negatively impact the readability and clarity of the multimodal answer, making it less effective for users.

6 Conclusion and Future Work

We introduced MuRAR, a framework designed to enhance text-based responses by incorporating images, tables, and videos. Human evaluations showed that MuRAR's multimodal answers are more useful, readable, and relevant than text-only responses. The system integrates text answer generation, source-based multimodal retrieval, and answer refinement to produce coherent multimodal answers. Future work will focus on two key areas: improving the quality of multimodal answers and enhancing the user experience. To improve the quality of multimodal answers, we plan to incorporate a broader range of multimodal documents to expand the dataset, train a custom LLM to replace reliance on proprietary models such as GPT-3.5, and address the relatively low recall performance in multimodal data retrieval. To enhance the user experience, future developments could include features such as enabling videos to jump directly to time segments relevant to the user's question and

providing a more intuitive and seamless interaction for users.

7 Ethical Considerations

Content Appropriateness. The MuRAR framework integrates diverse data modalities, such as images and videos, into answers. The multimodal data used in this work is sourced from the Adobe Experience Platform, a business-focused document repository considered free from harmful content. However, our framework is designed for flexibility and can be applied to a wide range of multimodal question-answering systems and data sources. This flexibility introduces the risk of retrieving inappropriate, offensive, or irrelevant material from poorly regulated or curated sources. Such content could harm users or erode trust and safety. To address this risk, it is essential to implement robust mechanisms for filtering, validating, and contextualizing multimodal content.

Privacy Concerns. Multimodal data retrieval could inadvertently expose sensitive information. Although the data source used in this work consists of enterprise-level documents devoid of sensitive content, the flexible nature of the MuRAR framework allows integration with diverse data sources, where privacy concerns may arise. Privacy-preserving techniques are essential to protect sensitive user and organizational data during retrieval and integration, ensuring compliance with regulations and maintaining trust.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. 2024. Robust multi model rag pipeline for documents containing text, table & images. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 993–999. IEEE.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Abhishek Kumar, Trisha Mittal, and Dinesh Manocha. 2020. MCQA: multimodal co-attention based network for question answering. *CoRR*, abs/2004.12238.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ronak Pradeep, Daniel Lee, Ali Mousavi, Jeff Pound, Yisi Sang, Jimmy Lin, Ihab Ilyas, Saloni Potdar, Mostafa Arefiyan, and Yunyao Li. 2024. Convkgyarn: Spinning configurable and scalable conversational knowledge graph qa datasets with large language models.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5317–5332. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix,

Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519.

Lichao Zhang, Jia Yu, Shuai Zhang, Long Li, Yangyang Zhong, Guanbao Liang, Yuming Yan, Qing Ma, Fangsheng Weng, Fayu Pan, Jing Li, Renjun Xu, and Zhenzhong Lan. 2024. Unveiling the impact of multimodal interactions on user engagement: A comprehensive evaluation in ai-driven conversations. *CoRR*, abs/2406.15000.

A Appendix

A.1 Multimodal Scraper Design

Our multimodal scraper design collects various fields and metadata from Adobe Experience League for images, videos, and tables. For images, we gather the link to the image and the surrounding context, specifically the text between the previous and current image and between the current and next image. The metadata collected includes the title of the document, the header of each section containing the image, and the URL of the document. For videos, the fields collected include the URL of the video, the context text before the video, and the video transcript. The metadata gathered is similar, including the document title, section headers, and document URL. For tables, we collect the table content in the form of a JSON string, the context text before the table, and the document URL. Additional metadata includes the document title and the header of each section containing the table.

A.2 Human Evaluation Metrics

Usefulness   Usefulness measures how much the multimodal elements contribute to the user's comprehension of the text content.

• 1 - Not at all useful: Multimodal elements provide no additional understanding or actively confuse the user.

• 2 - Slightly useful: Multimodal elements offer minimal enhancement to understanding.

• 3 - Moderately useful: Multimodal elements provide some additional clarity or information.

• 4 - Very useful: Multimodal elements significantly enhance understanding of the text.

• 5 - Extremely useful: Multimodal elements are crucial for full comprehension of the text.

Readability   Readability assesses how well the multimodal elements are integrated with the text.

• 1 - Severely impairs readability: Multimodal elements are poorly placed, causing significant disruption to reading flow.

• 2 - Somewhat impairs readability: Multimodal elements are not well-integrated, causing minor disruptions.

• 3 - Neutral impact on readability: Multimodal elements neither enhance nor impair the reading experience.

• 4 - Enhances readability: Multimodal elements are well-placed, supporting smooth reading flow.

Annotator	Answers	Usefulness	Readability	Relevance	Preference (Multi-Modal / Text Only / Same)
No.1	294	3.6463	3.8367	4.0170	207 (70.41%) / 41 (13.95%) / 46 (15.65%)
No.2	237	2.8861	3.0295	3.2700	115 (48.52%) / 81 (34.18%) / 41 (17.30%)
No.3	259	3.7452	3.8340	3.8764	205 (79.15%) / 28 (10.81%) / 26 (10.04%)
No.4	53	3.0566	3.5094	3.5849	30 (56.60%) / 16 (30.19%) / 7 (13.21%)
No.5	9	4.1111	4.4444	4.4444	9 (100.0%) / 0 (0.0%) / 0 (0.0%)
No.6	22	4.0000	4.0909	4.1818	19 (86.36%) / 2 (9.09%) / 1 (4.55%)
No.7	24	4.2500	4.2083	4.6250	23 (95.83%) / 0 (0.0%) / 0 (0.0%)
No.8	2	4.0000	4.5000	4.5000	2 (100.0%) / 0 (0.0%) / 0 (0.0%)

Table 5: Per-annotator average scores and preference.

• 5 - Significantly enhances readability: Multimodal elements are perfectly integrated, greatly improving the reading experience.

Relevance Relevance measures how closely the multimodal elements relate to the text content.

• 1 - Completely irrelevant: Multimodal elements have no apparent connection to the text.

• 2 - Mostly irrelevant: Multimodal elements have only a tenuous connection to the text.

• 3 - Somewhat relevant: Multimodal elements relate to the text but not be entirely on-point.

• 4 - Highly relevant: Multimodal elements clearly support and illustrate the text content.

• 5 - Perfectly relevant: Multimodal elements are essential to the text, providing crucial illustrations or data.

Preference Grading Annotators also indicate their overall preference between the text answer and the multimodal answer:

• **Text Only**: Choose this if you believe the text alone would be more effective without the multimodal elements.

• **Multi-Modal**: Select this if you think the combination of text and multimodal elements provides the best experience.

• **Same**: Choose this if you feel text-only and multimodal versions are equally effective.

A.3 Additional Human Evaluation Results

Due to space constraints, we include additional human evaluation results in the Appendix. The average score and preference per annotator are presented in Table 5.

A.4 Prompts

Please note that the actual prompts used in the system development differ from the prompts shown below. These are simplified versions that capture the essence of the prompt design.

Prompt for Text Answer Generation The prompt for text answer generation can be found in Figure 4.

Figure 4: Prompt for text answer generation.

Prompt for Multimodal Answer Refinement The prompt for multimodal answer refinement can be found in Figure 5.

Figure 5: Prompt for multimodal answer refinement.