CoMuS-KG: A Collaborative Framework of Multimodal Unstructured Data and Knowledge Graph

Shuhao Hu[1,2], Xin Wang[1,2], Ji Xiang[1,2], Xiaobo Guo[1,2], Lei Wang[1,2], Jiahui Shen[1,2]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{hushuhao24}@mails.ucas.ac.cn, {wangxin, xiangji, guoxiaobo, wanglei, shenjiahui}@iie.ac.cn

***Abstract*—Large language models (LLMs) have demonstrated remarkable capabilities in many fields, especially in complex neural lagnuage processing tasks. Despite their impressive performance, the content generated by LLMs still suffers from the problem of hallucination, particularly in tasks that require real-time data or specialized domain knowledge. Knowledge graphs and multimodal unstructured data serve as important sources of knowledge that can help address the hallucination issues in LLMs. However, existing methods mostly utilize knowledge graphs or multimodal unstructured data in isolation, neglecting the interaction between the two and it is the interaction that contributes to the extraction of deep knowledge in the knowledge base. In this paper, we propose a novel framework called the Collaborative Framework of Multimodal Unstructured Data and Knowledge Graph (CoMuS-KG). This framework enhances the reasoning capabilities of LLMs by enabling interaction between multimodal unstructured data and knowledge graphs, extracting deep knowledge from unstructured data, and completing missing information in knowledge graphs. Specifically, CoMuS-KG first decompose the question posed to the LLMs into multiple sub-questions and convert these sub-questions into knowledge graph triplets with missing head entity, tail entity, or relation. And then the knowledge graph and multimodal unstructured data are used to complete these triplets. Finally, we use the completed triplets to answer the original question and the completed triplets can be updated back into the knowledge graph to assist in other reasoning tasks. Extensive experiments on three KGQA benchmark datasets demonstrate the question-answering performance and reasoning capabilities of CoMuS-KG. Our code is publicly available at: https://github.com/GuChongAn/CoMuS-KG**

***Index Terms*—Knowledge Graph, Multi-modal Data, Large Language Model, Collaborative Framework**

I. Introduction

In recent years, Large Language Models (LLMs) [1]–[3] have achieved remarkable results in various fields, particularly in question-answer applications. However, when it comes to answering questions that require specific domain knowledge or real-time data, LLMs often encounter the issue of hallucination [4], [5]. Fortunately, LLMs can mitigate this problem by leveraging external knowledge, such as documents, images, and knowledge graphs [6], [7]. These external knowledge is typically organized into various knowledge bases according to different domains (e.g., medical knowledge bases, computer knowledge bases, and policy knowledge bases). It is noteworthy that each knowledge base not only contains surface factual information, such as medical records of a pneumonia patient, but also implicitly encompasses structural information of concepts that are not readily detectable by existing models, such as the common symptoms of pneumonia. The latter is particularly crucial for answering complex questions, such as accurately identifying COVID-19 as a novel coronavirus pneumonia during its initial emergence.

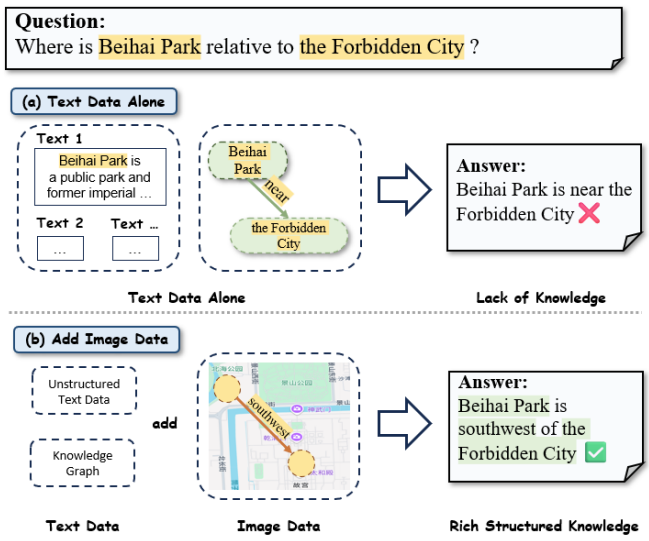

Fig. 1. As shown in Figure(a), text data alone cannot provide accurate knowledge for some problems. And as shown in Figure(b), these problems can be solved when image is added, especially when image structure information is extracted by interaction of knowledge graph and unstructured data.

Recent approaches typically leverage external knowledge bases within the framework of Retrieval-Augmented Generation (RAG). Based on the type of the retrieval knowledge base, current RAG methods can be categorized into two types: unstructured data retrieval and knowledge graph retrieval. unstructured data retrieval methods [8], [9] obtain the necessary knowledge directly from the unstructured data knowledge bases. While these methods can locate knowledge related to the question, they overlook the interconnections between pieces of knowledge, thus failing to capture the implicit conceptual information within the knowledge base.

Recently, methods that organize knowledge bases into knowledge graphs and then perform retrieval, exemplified by GraphRAG [10] and KAG [11], have garnered attention. These approaches first transform the knowledge base into a knowledge graph format and subsequently retrieve relevant knowledge from the graph, which allows for better utilization of structural relationships among data. However, existing knowledge graphs suffer from a lack of multimodal information. Methods like GraphRAG focus solely on text-modal data, treating images merely as text summaries. Additionally, images in common multimodal knowledge graphs are treated as attributes or subordinate entities of text entities [12]. Both summarizing images as text and treating them as entity attributes result in the loss of their rich structural information, which is a crucial component of the implicit conceptual information within the knowledge base. This information loss is particularly serious in images with multiple entities and rich inter-entity relationships (e.g., maps, photographs with multiple people, etc.). As illustrated in Figure 1, knowledge graphs commonly include only the landmark buildings of a place and their relationships, while map images contain a much richer array of locations and their interrelationships. It is the latter that provides sufficient knowledge for answering relevant questions.

In this paper, to enable LLMs to fully utilize the structural information inherent in multimodal knowledge base, we propose a collaborative framework between multimodal unstructured data and knowledge graphs, which called the Collaborative Framework of Multimodal Unstructured Data and Knowledge Graph (CoMuS-KG). For question answering tasks, CoMuS-KG first decompose the input question into multiple sub-questions, each of which should pertain to a single entity or relationship. These sub-questions are then converted into incomplete triplet formats, where the triplets lack either the head node, tail node, or relationship. The next step involves attempting to complete these triplets by knowledge graph retrieval. If the knowledge graph is unable to complete these triplets, the Retrieval-Augmented Generation (RAG) method is employed to utilize external multimodal unstructured data for triplet completion. Finally, the large language model generates the answer to the question based on these completed triplets and completed triplets can updata back to the original Knowledge Graph.

In summary, the contributions of this paper can be summarized into three key points,

- We have constructed the Collaborative Framework of Multimodal Unstructured Data and Knowledge Graph (CoMuS-KG). Through the interaction between knowledge graphs and multimodal unstructured knowledge base, our approach can effectively address complex questions that cannot be answered by the surface knowledge.
- We have addressed the problem of how to capture the implicit conceptual structures within unstructured databases and have developed a resulting method.
- We have conducted extensive experiments to demonstrate the performance of our approach in knowledge base question answer.

II. Related Work

Knowledge Base Question Answering (KBQA) [13] focuses on answering given questions based on external knowledge bases, which typically exist in formats such as text, image and knowledge graph. According to the kinds of external databases, the methods of KBQA can be divided into two categories: methods based on unstructured data and methods based on knowledge graph.

A. Methods Based on Unstructured Data

Methods based on unstructured data mostly fit Retrieval-Augmented Generation(RAG) [8] paradigm: retrieving relevant knowledge from the knowledge base based on the question (retrieval), augmenting the model with the retrieved knowledge (augmentation), and finally generating the answer to the question (generation) [14]. The improved RAG method can be divided into pre-retrieval and post-retrieval strategies according to the improved location [15].

Pre-retrieval and retrieval methods aim to handle the information to be retrieved and enhance retrieval accuracy [16]. The most typical method is dense search and its variants [17], [18]. Dense search encodes the unstructured data and the question into the same vector space, and then calculates the similarity between the two to find the top-k data that is most similar to the question. Although these methods can find relevant knowledge, they overlook the structural relationships within the retrieval information sources.

Post-retrieval methods concentrate on processing the retrieved information; for instance, [9] segments retrieved images into multiple regions for fine-grained analysis, and [19] performs object recognition on retrieved images to search for more comprehensive related information. However, these methods fail to perceive the structural information of the retrieval sources from the limited results obtained.

B. Methods Based on Knowledge Graph

Recent approaches, such as GraphRAG [10] and KAG [11], have garnered attention for their organization of unstructured data into knowledge graphs, which are then queried based on the given questions. Although organizing unstructured data into knowledge graphs reveals their structural relationships, GraphRAG's focus on global information overlooks the detailed connections within the knowledge graph, and KAG struggles with multimodal information, only converting images into text, thereby losing the inherent structural information of the images.

In addition to the KBQA approach, there are other approaches that address Knowledge Graph Question Answering (KGQA) [6] task. Unlike KBQA, KGQA focuses solely on directly retrieving information from knowledge graphs and utilizing the retrieved knowledge to answer questions. Early methods achieved this by embedding knowledge graphs into models during the pre-training or fine-tuning stages [20], [21]. However, embedding knowledge graphs into LLMs inevitably

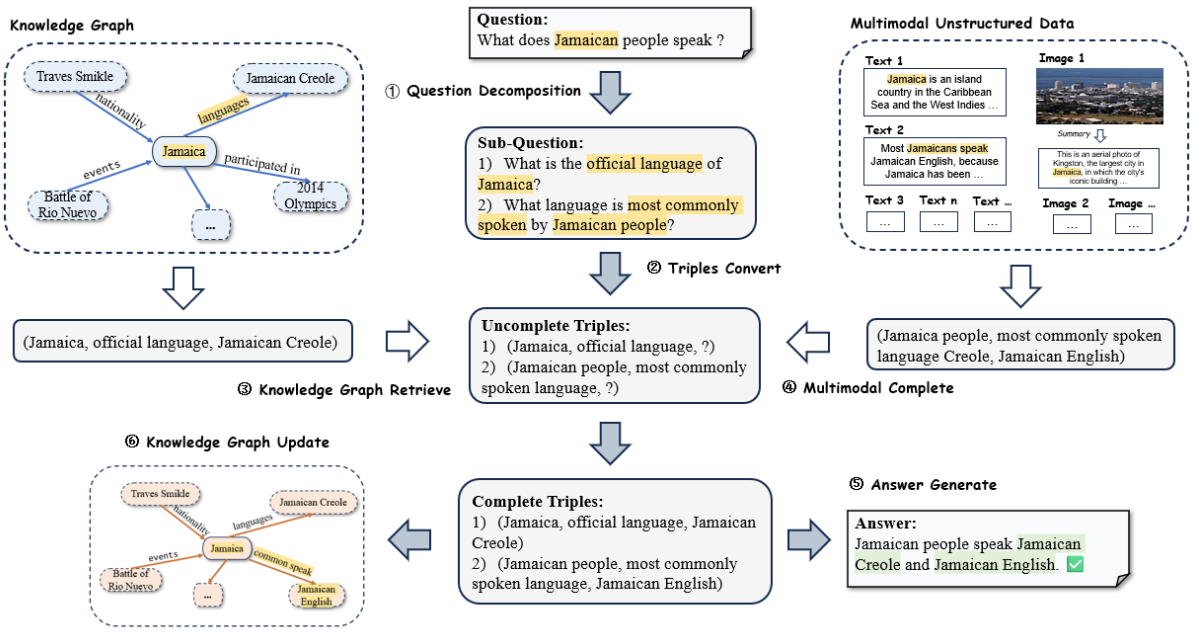

Fig. 2. An example workflow of CoMuS-KG.

compromises the interpretability, updating ability and reasoning ability of the model.

Recent research has employed semantic parsing and dense search methods to retrieve relevant knowledge from knowledge graphs. Semantic parsing methods convert input questions into executable logical format statements on the knowledge graph which are then executed to retrieve relevant knowledge [22], [23]. However, this requires extremely meticulous handling, and any minor error can lead to execution failure. Dense search methods embed entities and relationships from the knowledge graph into vector spaces, embed the question into the same vector space, and calculate the similarity between them to find the most relevant entities and relationships [24]. While dense search methods can identify relevant entities or relationships, they overlook the structural aspects of the knowledge graph.

Among recent methods, only CogMG [25] uses unstructured data and knowledge graph at the same time as we do. However, on the one hand, CogMG ignores the use of multi-modal data, thus missing the rich knowledge contained in multi-modal data. On the other hand, CogMG pays more attention to the update of knowledge graph by using unstructured data, thus the extraction of deep information from unstructured data is neglected.

III. Methodes

In this section, we first introduce the overall framework of CoMuS-KG, as illustrated in Figure 2, and then delve into the specifics, including six steps: problem decomposition, triplet generation, knowledge graph search, multimodal RAG completion, answer generation, and knowledge graph update.

A. Overview of CoMuS-KG

Overall, CoMuS-KG is a planning-retrieval-reasoning KBQA framework. In the planning and retrieval phases, CoMuS-KG is designed to interact with unstructured data and knowledge graphs, enabling the model to leverage implicit structural information from unstructured data sources during the reasoning process. Specifically, for a given question Q, CoMuS-KG first decomposes it into a set of sub-questions $[q_1, q_2, ..., q_n]$. These sub-questions are then individually transformed into incomplete triplets $[t_1, t_2, ..., t_n]$. The knowledge graph is searched to complete these triplets. If these triplets are not successfully completed, multimodal unstructured data sources are utilized, employing the RAG method to complete these triplets, resulting in completed triplets $[\bar{t}_1, \bar{t}_2, ..., \bar{t}_n]$. Finally, the completed triplets are used by the LLMs to answer the original question Q, and those triplets completed by RAG are updated into the knowledge graph.

B. Planning: Problem Decomposition and Triple Generation

Given a question, CoMuS-KG leverages the underlying LLM to identify the knowledge required to address the question and concretizes the operations to acquire this knowledge into corresponding sub-questions. Notably, to ensure the success of subsequent triple conversion, each sub-question is constrained to pertain to only one specific entity or relationship during the generation process. This process can be represented as,

$$[q_1, q_2, ..., q_n] = Decompose(Q) \tag{1}$$

where q_i denotes the i-th sub-question, Q represents the original question that needs to be answered, and $Decompose$

denotes the problem decomposition operation based on the LLM. After obtaining a series of sub-questions, these sub-questions are required to be converted into incomplete triples, such as (Einstein, ?, Tesla) or (Einstein, student, ?), where "?" represents the part to be completed. This process can be represented as:

$$[t_1, t_2, ..., t_n] = Question2Triple([q_1, q_2, ..., q_n]) \quad (2)$$

where t_i denotes the i-th incomplete triple, corresponding to the i-th sub-question, and $Question2Triple$ denotes the conversion operation from a question to a triple based on the LLM.

C. Retrieval: Knowledge Graph Retrieval and Multimodal RAG Completion

Following the problem decomposition and triplet generation operations in the planning phase, we have obtained a set of incomplete triplets. The next natural step is to complete these triplets, which involves two operations: knowledge graph retrieval and multimodal RAG completion.

Before performing knowledge graph retrieval to complete the triplets $[t_1, t_2, ..., t_n]$, we first use an embedding model to generate embeddings for all entities and relationships in the knowledge graph, preparing for dense search. For any incomplete triplet ti, there are two possible scenarios: it either lacks a head or tail node, or it lacks a relationship (e.g., (Canberra, capital of, ?) and (Canberra, ?, Australia)). In the former case, we first locate the existing node and then calculate the similarity between all connections of this node and the connections of the triplet. We obtain the k most similar candidate relationships (where k may be 0) that exceed a certain threshold. The nodes linked by these candidate relationships are the missing head or tail nodes of the triplet. For missing relation triples, we first locate their head nodes and then traverse all other nodes connected to this node, calculating their similarity to the tail node of the triple. The most similar k candidate nodes (where k can also be 0) with similarity exceeding a certain threshold are obtained. The relationship between the candidate nodes and the head node is the missing relation of the triple. This process can be described as:

$$[\bar{t}_1, \bar{t}_2, ..., \bar{t}_n] = KGComplete([t_1, t_2, ..., t_n]) \quad (3)$$

where $\bar{t}_i$ represents the completed triple, and $KGComplete$ denotes the Knowledge Complete operation. If both k values are 0 in the previous context, the knowledge graph retrieval fails, and the completion of the triple cannot be achieved. In this case, the next step is to proceed with multimodal RAG completion. If the knowledge graph retrieval successfully completes all ti, the reasoning step is directly performed to answer the original question.

For triples that failed to be completed through knowledge graph retrieval, we employ multimodal RAG for completion. Specifically, for an incomplete triple ti, we first convert it from the triple format into natural language format, and then compute its corresponding embedding representation. By calculating the similarity between this embedding and the embeddings of the text and images to be retrieved, we identify the most relevant text and images to the triple. Subsequently, we utilize a large model to complete the triple with the support of relevant knowledge.

In particular, we have processed the images in a fine-grained manner to obtain richer entity information. Each image to be retrieved is segmented into multiple sub-images, which are then summarized using a large model. Each sub-image focuses on different positions of the original image, allowing us to capture detailed information that would not be possible with a direct summary of the original image.

D. Reasoning and Knowledge Update

After obtaining the completed triples, we prompt the LLM to generate answers to the original questions using the completed triples. Based on the triples generated through the interaction between the knowledge graph and multimodal unstructured data, the LLM gains hidden, accurate, and structured contextual data. This enables the LLM to answer questions that it could not answer on its own or through simple retrieval from unstructured data sources and the knowledge graph, significantly enhancing the model's reasoning capabilities.

CoMuS-KG also supports updating the completed triples back into the knowledge graph to further facilitate the interaction between the knowledge graph and multimodal unstructured data. Unlike typical knowledge graph construction efforts, which mostly extract entities and relationships directly from unstructured data, completely rebuild a knowledge graph, and then require cumbersome alignment work to be put into practical use, our update operations are real-time and already aligned. Since the interaction between the knowledge graph and multimodal unstructured data has been achieved, the completed triples are actually already aligned with the constraints of the knowledge graph. After being updated into the knowledge graph, they can be used directly.

IV. Experiment

We further designed and conducted experiments to demonstrate the effectiveness of the CoMuS-KG framework in the KGQA task. Our findings illustrate the advantages of integrating multimodal unstructured data with knowledge graph, highlighting its positive impact on the accuracy of question answering

A. Experiment Settings

Datasets. We evaluated the performance of CoMuS-KG on two typical KGQA benchmark datasets: WebQuestionSp (WebQSP) [27] and Complex WebQuestions (CWQ) [26]. Specifically, WebQSP contains 1,628 test question-answer pairs, while CWQ contains 3,531 test question-answer pairs. Both datasets include questions that require multi-hop reasoning and are based on the Freebase knowledge base. To better compare with CogMG [25], we also tested on the KQApro dataset [23], a recently released KGQA dataset for complex tasks. As with

TABLE I
RESULT ON WQSP, CWQ AND KQA PRO DATASET

Method	WQSP		CWQ		KQA pro$_{sub}$	
	EM	Acc	EM	Acc	EM	Acc
gpt-3.5-tubro	0.4519	0.2548	0.7469	0.5491	0.46	0.32
CogMG [25]	-	-	-	-	-	0.86
CoMuS-KG	**0.6966**	**0.5281**	**0.8931**	**0.6744**	**0.92**	**0.92**
CoMuS-KG w/o RAG	0.6454	0.4981	0.8789	0.6701	0.78	0.78
CoMuS-KG w/o KG	0.4681	0.3548	0.7886	0.4889	0.64	0.64
CoMuS-KG w/o Knowledge	0.4316	0.2288	0.7389	0.4496	0.58	0.58

CogMG, we tested 50 questions randomly selected from this dataset, and we call the sub-dataset KQA pro$_{sub}$.

Baselines. To the best of our knowledge, while there are many previous papers that use either knowledge graphs or unstructured databases alone to aid LLMs question answering, only CogMG leverages both modalities of data as we do. So we only compared CoMuS-KG with CogMG in the experiment.

Metrics. Since we leverage LLM to generate the final answers rather than performing retrieval on the knowledge graph, our answers cannot be strictly matched with the reference answers in the test data. Following previous work, we use Exact Match scores and accuracy as evaluation metrics. The Exact Match score determines whether the answer is correct by calculating the similarity between the model's output and the reference answer, while accuracy is calculated based on whether the reference answer appears in the output of the model.

Implementations. During the execution of CoMuS-KG, multiple calls to the LLM are required. Unless specified otherwise, we use the gpt-3.5-turbo API. For image data in multimodal datasets, we use InternVL [28] to summarize it.

B. Main Result

All experimental results are presented in Table I, with the best results on each dataset highlighted in bold.

The comparison between CoMuS-KG and CogMG on the KQApro-sub50 dataset should be given primary attention. CoMuS-KG achieved a 92% accuracy, which is 6 percentage points higher than CogMG's 86% accuracy. From the ablation experiments on the same dataset, it is evident that this performance improvement is clearly due to the interaction between multimodal unstructured data and the knowledge graph, as neither CoMuS-KG w/o RAG, which uses only knowledge graph data, nor CoMuS-KG w/o RAG, which uses only multimodal data, outperformed CogMG.

To further substantiate the importance of the interaction between multimodal unstructured data and the knowledge graph, we conducted extensive ablation experiments on three datasets. Specifically, CoMuS-KG represents the complete framework as described in the method section. CoMuS-KG w/o RAG removes the part of CoMuS-KG that uses the RAG method to complete triples with multimodal unstructured data. CoMuS-KG w/o KG removes the part of CoMuS-KG that uses the knowledge graph to complete triples. CoMuS-KG w/o Knowledge indicates that no external knowledge is used, and only the steps of question decomposition and triple transformation as described earlier are performed, with the completion of triples relying solely on the LLM's own capabilities. In the ablation experiments, we observed two key characteristics. Firstly, across all three datasets, the method that does not incorporate any knowledge yielded the worst results, performing even worse than the direct LLM responses on the CWQ and WebQSP datasets. This indicates that the process of problem decomposition and transformation into triples actually serves the interaction between multimodal unstructured data and knowledge graphs. The absence of this interaction results in a significant loss of method performance when problem decomposition and triple transformation are performed in isolation. Secondly, the introduction of a knowledge graph greatly enhanced the method's performance. For instance, on the CWQ dataset, the accuracy jumped from 35% with CoMuS-KG w/o KG to 52% with CoMuS-KG, and even without unstructured data, CoMuS-KG w/o RAG achieved an accuracy of 49%. This trend underscores the critical importance of knowledge graphs for question answering, primarily because all three datasets are designed for KGQA tasks, where the questions are inherently tailored to knowledge graph data.

Finally, we compared the performance of CoMuS-KG and direct LLM responses across the three datasets. CoMuS-KG consistently achieved substantial improvements, indicating that it can mitigate the hallucination issue of LLMs to a certain extent.

C. Knowledge Graph Update

After completing the question-and-answer process, CoMuS-KG has the option to update triples with multimodal unstructured data completion back into the knowledge graph to help with subsequent inference tasks. We compared the composition elements and reasoning ability of the original knowledge graph and the updated knowledge graph on CWQ and WebQSP datasets. As shown in Table 2, CoMuS-KG

TABLE II
RESULTS OF KNOWLEDGE GRAPH UPDATE FOR WQSP AND CWQ

Dataset	Question Number	Entity	Relation	EM	ACC
CWQ	3531	+761	+104	+0.051	+0.030
WebQSP	1628	+129	+84	+0.015	+0.004

updated relevant entities and relationships needed to answer questions on both datasets. Both EM and correct rate are improved without any modification of the method.

V. CONCLUSION

In this paper, we address the challenge of extracting hidden, deep structural information from unstructured data sources through the interaction of multimodal unstructured data and knowledge graphs. Specifically, we propose CoMuS-KG, a collaborative framework that integrates multimodal unstructured data with knowledge graphs. This framework leverages the structural information of knowledge graphs to achieve a deep understanding of unstructured data, while also using multimodal unstructured data to complement the multimodal information that knowledge graphs may lack. The interaction between multimodal unstructured data and knowledge graphs is realized through the completion of knowledge graph triples, which in turn aids LLMs in question answering. Rich experiments on three KGQA benchmark datasets demonstrate the question answering performance and reasoning capabilities of the CoMuS-KG framework. We believe that the interaction between unstructured data and knowledge graphs offers a new paradigm for knowledge base question answering, enabling more effective utilization of the deep details within knowledge bases.

REFERENCES

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, and P. Mishkin, et al., "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, and F. L. Aleman, et al., "GPT-4 technical report," 2023, *arXiv preprint arXiv:2303.08774*.

[3] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, and H. T. Cheng, et al., "LaMDA: Language models for dialog applications," 2022, *arXiv preprint arXiv:2201.08239*.

[4] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4149–4158.

[5] A. Talmor and J. Berant, "The Web as a Knowledge-base for Answering Complex Questions," 2018, *arXiv preprint arXiv:1803.06643*.

[6] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, 'Unifying large language models and knowledge graphs: A roadmap', *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[7] J. Baek, A. F. Aji, and A. Saffari, 'Knowledge-augmented language model prompting for zero-shot knowledge graph question answering', 2023, *arXiv preprint arXiv:2306.04136*.

[8] P. Lewis et al., 'Retrieval-augmented generation for knowledge-intensive NLP tasks', in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 2020.

[9] W. Lin, J. Chen, J. Mei, A. Coca, and B. Byrne, 'Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering', in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2024.

[10] D. Edge et al., 'From local to global: A graph rag approach to query-focused summarization', 2024, *arXiv preprint arXiv:2404.16130*.

[11] L. Liang et al., 'KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation', 2024, *arXiv preprint arXiv:2409.13731*.

[12] X. Zhu et al., 'Multi-Modal Knowledge Graph Construction and Application: A Survey', *IEEE Transactions on Knowledge & Data Engineering*, vol. 36, no. 02, pp. 715–735, Feb. 2024.

[13] A. Bordes, N. Usunier, S. Chopra, and J. Weston, 'Large-scale Simple Question Answering with Memory Networks', 2015, *arXiv preprint arXiv:1506.02075*.

[14] O. Ram et al., 'In-Context Retrieval-Augmented Language Models', *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1316–1331, 2023.

[15] Y. Gao et al., 'Retrieval-augmented generation for large language models: A survey', 2023, *arXiv preprint arXiv:2312.10997*.

[16] V. Karpukhin et al., 'Dense Passage Retrieval for Open-Domain Question Answering', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[17] W. Lin and B. Byrne, 'Retrieval Augmented Visual Question Answering with Outside Knowledge', in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11238–11254.

[18] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, 'KAT: A Knowledge Augmented Transformer for Vision-and-Language', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 956–968.

[19] J. Qiu et al., 'Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm', 2024, *arXiv preprint arXiv:2403.04735*.

[20] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, 'Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning', in *International Conference on Learning Representations*, 2024.

[21] A. Saxena, A. Kochsiek, and R. Gemulla, 'Sequence-to-Sequence Knowledge Graph Completion and Question Answering', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[22] H. Luo et al., 'ChatKBQA: A Generate-then-Retrieve Framework for Knowledge Base Question Answering with Fine-tuned Large Language Models', in *Findings of the Association for Computational Linguistics*, 2024, pp. 2039–2056.

[23] S. Cao et al., 'KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: ACL 2024*, 2022, pp. 6101–6119.

[24] Y. Wei, Q. Huang, Y. Zhang, and J. Kwok, 'KICGPT: Large Language Model with Knowledge in Context for Knowledge Graph Completion', in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 8667–8683.

[25] T. Zhou, Y. Chen, K. Liu, and J. Zhao, 'CogMG: Collaborative Augmentation Between Large Language Model and Knowledge Graph', in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2024, pp. 365–373.

[26] A. Talmor and J. Berant, 'The Web as a Knowledge-Base for Answering Complex Questions', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 641–651.

[27] W. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh, "The Value of Semantic Parse Labeling for Knowledge Base Question Answering," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, Jan. 2016.

[28] Z. Chen et al., 'Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24185–24198.