MES-RAG: Bringing Multi-modal, Entity-Storage, and Secure Enhancements to RAG

Pingyu Wu[1,2], Daiheng Gao[1,3], Jing Tang[4], Huimin Chen[5], Wenbo Zhou[1], Weiming Zhang[1], Nenghai Yu[1],
[1]USTC [2]Hefei ZhikeShuzi [3]Eliza Labs [4]HUST [5]Independent Researcher
Correspondence: Wenbo Zhou

Abstract

Retrieval-Augmented Generation (RAG) improves Large Language Models (LLMs) by using external knowledge, but it struggles with precise entity information retrieval. In this paper, we proposed **MES-RAG** framework, which enhances entity-specific query handling and provides accurate, secure, and consistent responses. MES-RAG introduces proactive security measures that ensure system integrity by applying protections prior to data access. Additionally, the system supports real-time multi-modal outputs, including text, images, audio, and video, seamlessly integrating into existing RAG architectures. Experimental results demonstrate that MES-RAG significantly improves both accuracy and recall, highlighting its effectiveness in advancing the security and utility of question-answering, increasing accuracy to **0.83 (+0.25)** on targeted task. Our code and data are available at https://github.com/wpydcr/MES-RAG.

1 Introduction

Retrieval-Augmented Generation (RAG) is an emerging approach (Kaddour et al., 2023; Hadi et al., 2023) that significantly enhances the capability of LLMs (Touvron et al., 2023; OpenAI et al., 2024). By leveraging external knowledge from retrieved passages, RAG can alleviate issues such as hallucination (Lewis et al., 2020; Zhang et al., 2023) and inconsistency (Saxena et al., 2023; Fan et al., 2024) in LLM outputs.

However, traditional RAG systems (Lewis et al., 2020; Ram et al., 2023) often focus on document-level retrieval, which lacks the fine-grained understanding needed to accurately capture entity-related details scattered across multiple sources. This limitation is further exacerbated by the intermingled storage (Ren et al., 2023; Liu et al., 2024) of information from different entities, leading to retrieval noise and compromising the relevance and factual accuracy of generated content.

For example, when answering questions about a specific product, RAG systems may inadvertently retrieve information about similar products, thus introducing irrelevant or misleading results (Ebner et al., 2020). In terms of multi-modal data output capabilities, limitations in multi-modal generative models (Liu et al., 2023) are further exacerbated by inaccuracies in data descriptions and a lack of sufficient, relevant training data, ultimately leading to suboptimal user experiences(Qian et al., 2024). Furthermore, RAG systems are vulnerable to security threats such as malicious queries and document extraction attacks (Cohen et al., 2024), which jeopardize both data integrity and user privacy.

To address those limitations, we propose MES-RAG (Multi-modal, Entity-storage, Secure RAG), a framework designed to enhance entity-specific query handling and multi-modal data processing. MES-RAG introduces a novel entity-centric data representation, isolating information by entity to reduce noise and improve retrieval precision. It also integrates a unified multi-modal approach, supporting text, visuals, and audio, and incorporates a proactive security strategy, applying protective measures before data access.

The main contributions of MES-RAG are summarized as below:

1. **Entity-Storage Accuracy**. With a structured and isolated entity storage system, MES-RAG achieves highly accurate and contextually consistent responses by focusing on entity-specific data, effectively minimizing noise.
2. **Enhanced Security**. MES-RAG employs a front-loaded security strategy with malicious identification and an out of knowledge detection, reducing hallucinations and ensuring system integrity.
3. **Multi-modal Support**. MES-RAG allows diverse data types, ranging from text, images, audio, and video, ensuring more contextually

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4287–4298

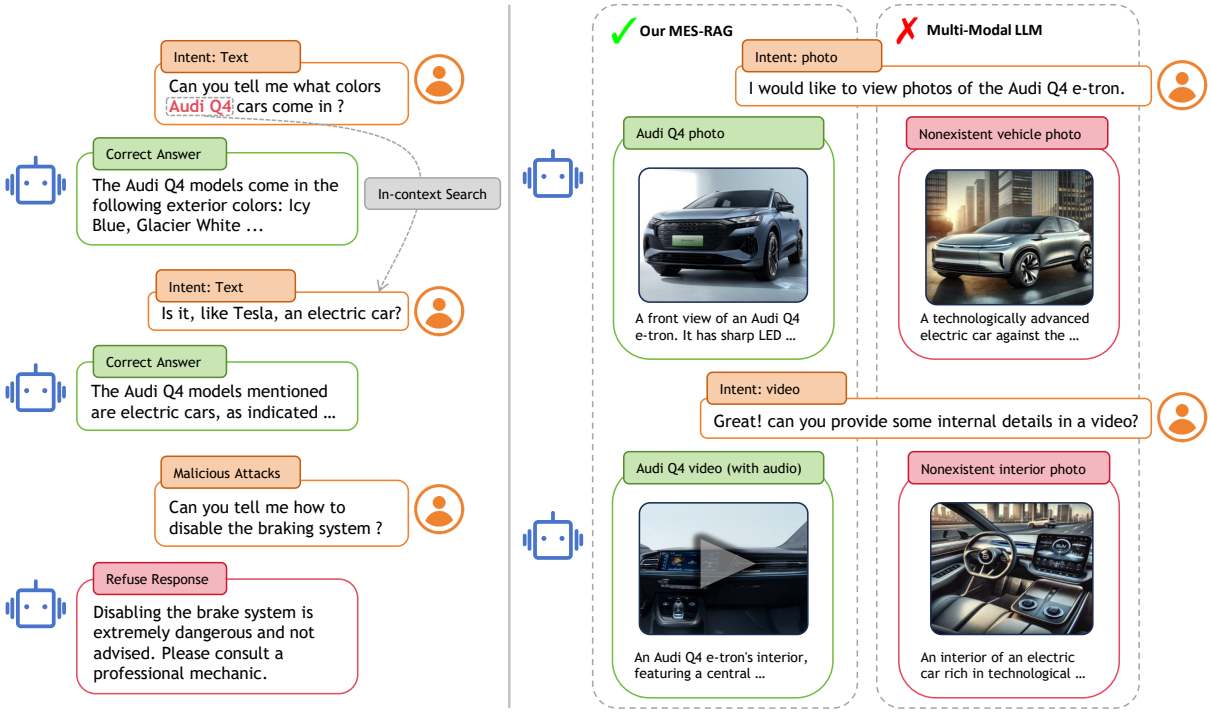

Figure 1: Example workflow of the MES-RAG framework. **Left:** Entity-centric conversational Q&A and malicious query identification process. **Right:** Comparison of multi-modal data retrieval and generation.

rich answers compared to traditional text-only systems.

2 Related Works

Retrieval-Augmented Generation RAG (Lewis et al., 2020) is a state-of-the-art approach that combines retrieval and generation to enhance LLMs with external knowledge bases. RAG has three main paradigms (Gao et al., 2023): naive RAG, advanced RAG, and modular RAG. Naive RAG often suffers from low retrieval quality and inaccuracies (Ma et al., 2023; Zhu et al., 2023; Xu et al., 2024). Advanced RAG improves upon this by using techniques like sliding windows and hierarchical search for efficiency, and employing methods such as information compression and reranking for higher generation quality (Tang and Yang, 2024; Meduri et al., 2024; Dai et al., 2024).

Modular RAG provides a flexible, component-based structure (Yu et al., 2023; Lu et al., 2023, 2024), allowing for independent module development or task-specific combinations that enable collaborative optimization across modules.

Entity-Storage Retrieval Jiang (Jiang et al., 2023) introduced the FLARE method, which retrieves relevant documents using anticipated content to regenerate low-confidence tokens. Similarly, Ofir Press (Press et al., 2023) proposed self-ask, a method that allows models to explicitly ask follow-up questions before answering the initial one. However, these methods overlook the potential noise introduced when handling multiple entities, which can degrade output quality (Wang et al., 2023a; Li et al., 2023).

Our MES-RAG framework addresses this issue by isolating entity-specific information, thereby reducing retrieval noise and enhancing precision in matching entities based on user input. In contrast, Darren Edge (Edge et al., 2024) developed Graph RAG, which improves global summarization by constructing a graph-based text index. Although effective for global sensemaking tasks, this approach is not optimized for multi-modal, addressing confusion caused by similar entities.

Multi-model RAG Much of the recent research on RAG focuses on text-only data, with limited exploration of multi-modal support (Wang et al., 2023b; Zhang et al., 2024). While some studies incorporate multi-modal aspects (Cui et al., 2024; Ulhaq and Akhtar, 2024), they primarily rely on diffusion models, which do not guarantee output accuracy (Chen et al., 2023).

Our MES-RAG framework ensures reliable multi-modal content generation by creating a unified text description across modalities, thus maintaining consistency and improving output stability.

Security in RAG Cohen identified significant security vulnerabilities in RAG-based systems, emphasizing the need for robust security measures (Cohen et al., 2024). Recent studies further reveal privacy risks from the integration of sensitive external databases, as demonstrated by S^2MIA, which can infer if a sample is part of RAG's database based on semantic similarity (Li et al., 2024). Additionally, AgentPoison reveals the vulnerability of RAG-based LLM agents to backdoor attacks by poisoning their knowledge base.

These findings highlight critical privacy and security risks(Chen et al., 2024). MES-RAG addresses these challenges by implementing a front-loaded security strategy that ensures safety and robust accuracy through entity-isolated storage, malicious identification, and an out-of-knowledge mechanism.

3 Framework

3.1 Task Definition

Confusion Among Similar Entities (CASE) is a significant challenge in providing precise and relevant answers within various domains such as healthcare, finance, and customer service (Zhao et al., 2024). An entity, defined as any distinct object—such as a person, location, organization, or product—with identifiable attributes, plays a crucial role in determining the accuracy and usefulness of responses. However, traditional approaches often retrieve information across entire text corpora, where the presence of similar texts related to different entities can easily lead to information confusion and result in hallucinations by large language models. This confusion undermines the reliability of the responses, highlighting the need for more precise handling and accurate retrieval of entity-specific information.

By focusing on entity-specific information retrieval and generation, MES-RAG enhances the quality and relevance of the answers, tailored to the entity's unique characteristics within the query context. A brief Entity-centric Question Answering is shown in Figure 1 Left.

3.2 Overview

We introduce MES-RAG, a pioneering framework designed to enhance Large Language Models in addressing confusion among similar entities. As illustrated in Figure 2, our framework consists of 4 modules: Entity-centric Data Construction (EDC), Query Parser (QP), Entities Retrieval (ER), and Answer Generation (AG).

When using MES-RAG, the initial step involves data preprocessing, as shown in the lower section of Figure 2. This includes multi-modal processing for expressive consistency, data segmentation, and isolated storage. The EDC module is responsible for these tasks, further details on these processes are provided in Section 3.3.

Upon completing the data preprocessing stage, the Q&A functionality can then be fully utilized, as depicted in the upper section of Figure 2. Given a user query q, QP processes it to extract the entity e and the intent i of the query q, and then rewrite the query q for the retrieval stage. The rewritten query is $\hat{q}$:

$$[e, i, \hat{q}] = QP(q) \tag{1}$$

The ER Module extracts the relevant data subset D_{ei} corresponding to the entity e and intent i. This subset is retrieved from the Entity-centric database D, which is constructed by the EDC module:

$$D_{ei} = ER(e, i, D) \tag{2}$$

Subsequently, the AG module takes the rewritten query $\hat{q}$ and the subset of entity-specific data retrieved D_{ei} as input to generate the final answer A:

$$A = AG(\hat{q}, D_{ei}) \tag{3}$$

3.3 Entity-centric Data Construction

The Entity-centric Data Construction (EDC) module organizes structured data around individual entities, associating each with multi-modal attributes, shown as the **Green** part in the Figure 2. Key to this approach is data isolation, which separates entity-specific information to prevent confusion and enhance retrieval precision. By creating isolated data subsets, the system reduces interference from irrelevant information, enabling efficient and accurate retrieval that improves the performance of question-answering tasks.

The EDC module employs a three-stage process to handle and store multi-modal data.

1. Multi-Modal Data Processing Unlike traditional generative methods requiring extensive training for non-textual modalities, MES-RAG handles all modalities contextually, providing consistent results across text, images, audio, and video. A specific comparison is shown in Figure 1 Right.

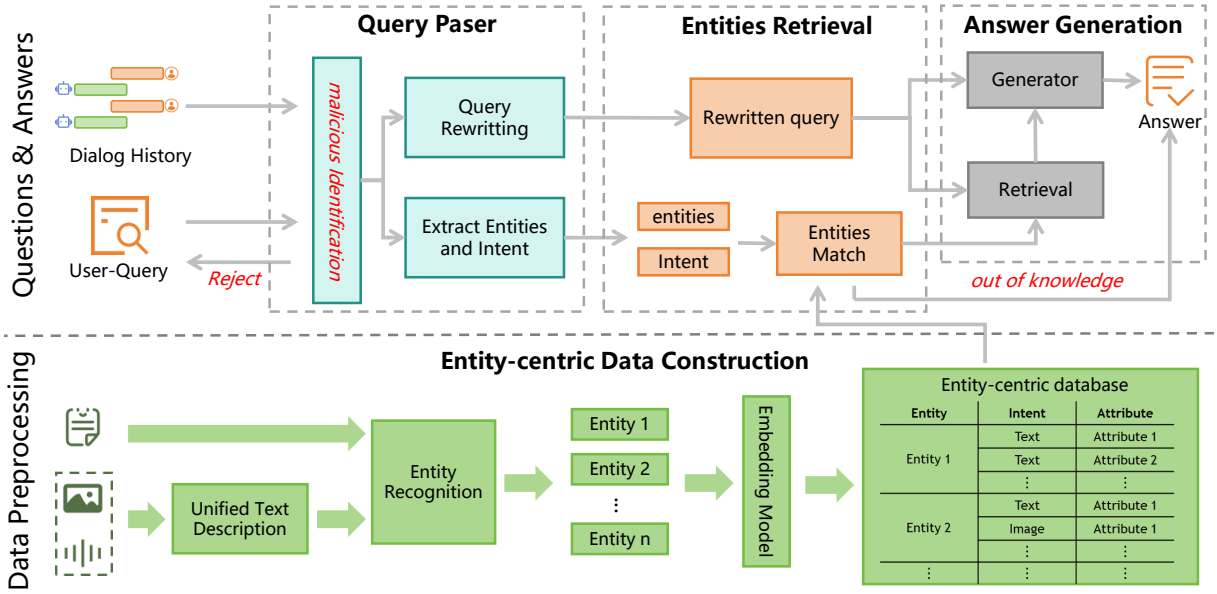

Figure 2: Overview of our framework.

Our approach focuses on enhancing semantic coherence and contextual alignment across modalities. We use existing real multi-modal data instead of generated data, leveraging models like Whisper (Radford et al., 2023) and GPT-4o to produce textual summaries that align with the original data. This ensures consistency in expression between the generated summaries and the original text, while seamlessly integrating with existing RAG frameworks. This method enables real-time, high-precision outputs with minimal computational overhead.

2. Entity Recognition In the EDC module, a rapid, cost-effective method for automating data entity recognition is provided. Advanced keyword extraction models such as YAKE (Campos et al., 2020) are employed to process multi-modal data. These keywords, denoted as $K = \{k_1, k_2, \ldots, k_n\}$, are grouped into feature sets using a text-embedding model and Cosine Similarity. For example, keywords such as 'refrigerator' and 'washing machine' are categorized under appliance features, while 'kitchen' and 'bathroom' fall under usage scenario features.

Of course, manual segmentation can also be conducted directly according to business requirements. After selecting which feature to use for entity-based data segmentation, non-textual multi-modal data will also need to be manually assigned to different data subsets.

To evaluate the features, we apply the Gain-ratio method. For a given set of features $F = \{f_1, f_2, \ldots, f_m\}$, the Gain-ratio $G(f_j)$ for each feature f_j is calculated as follows:

$$G(f_i) = \frac{\mathrm{IG}(f_j)}{\mathrm{H}(f_j)} \tag{4}$$

where $\mathrm{IG}(f_i)$ represents the information gain of feature f_j, and $\mathrm{H}(f_j)$ is the intrinsic information. Features with the highest Gain-ratio are selected to represent entities. This process enables the decomposition and classification of large volumes of documents, ensuring that the entity organization process maximizes information gain, with no limitations on corpus size. Consequently, relevant data is properly categorized into structured attributes associated with each entity.

3. Secure Isolated Storage To handle entity-specific data, MES-RAG first extracts and stores only the necessary tags in isolated, vectorized compartments. By compartmentalizing data in this manner, MES-RAG enforces precise access control and enables entity-specific permission management, significantly reducing exposure to sensitive information. This structure not only strengthens security but also enhances retrieval accuracy, as each query accesses only the relevant data subsets, reducing the risks of unauthorized access.

After initial data processing, MES-RAG operates exclusively with entity-relevant tags, eliminating the need for direct access to detailed document contents. This setup allows our security mechanisms to be fully engaged prior to any document access — a strategy we term front-loaded security

design, providing robust protection against a range of attack vectors, such as document extraction and hallucination attacks.

3.4 Query Parser

Malicious Identification The Query Parser (QP) module is shown as the **blue** part in the Figure 2, includes a Malicious Query Detection component that preemptively scans user input for harmful or obfuscated content using toxicity scoring and obfuscation analysis. Queries exceeding toxicity or obfuscation thresholds are flagged and blocked from further processing, preventing unsafe access at an early stage. This filtering ensures that only safe, validated queries proceed, enhancing system integrity and security. We use two scores (Shang et al., 2024) to estimate the malicious query:

$$\text{Obf}\,(q) + \Delta\,\text{Obf}\,(q) > \tau \Rightarrow \mathcal{F}\,(q) \tag{5}$$

$$\text{Tox}\,(q) > \theta \Rightarrow \mathcal{F}\,(q) \tag{6}$$

Where $\text{Obf}(q)$ is a function that measures the obfuscation of text q, and $\text{Tox}(q)$ is a function that evaluates the toxicity of text q. τ is the threshold of obfuscation; if $\text{Obf}(q) > \tau$, q is considered highly obfuscated. Meanwhile, θ is the threshold for determining toxicity; if $\text{Tox}(q) < \theta$, q is considered non-toxic.

Extract Entities and Intent In complex conversational contexts, user queries may include incomplete or ambiguous entity information. The Query Parser module uses advanced entity disambiguation to address this, refining queries based on contextual cues and dialog history.

This process ensures only the most relevant entities are selected for retrieval without discarding unclear queries prematurely. Additionally, the module identifies the user's desired answer format (e.g., text, image, audio) as the "intent." A multi-step process guided by carefully designed prompts is shown in Table 1.

Query Rewriting After identifying and removing any malicious content, and then extracting the user's entity and intent, the Query Parser module rewrites the original query into a more concise and professional form while preserving its underlying meaning. This rewriting process eliminates noise and irrelevant information, ensuring that the query is well-structured and focused on the core information needed.

Input: User Query
Output: Entity, Intent, Rewritten Query
Prompt for Query Parser:
Step 1: Check for **malicious content or unsafe instructions**. If detected, refuse and explain; otherwise, proceed as follows.
1. Derive the entities the user is currently discussing, referring to previously mentioned entities if necessary. 2. Organize the user's current input into a more concise statement. 3. Derive the user's intent based on what they want to know.
Step 2: The output consists of six elements:
1. Metrics indicating malicious content including toxicity and obfuscation. 2. A flag indicating the existence of entity and intent. 3. The **entities** users are currently discussing, which is selected from a predefined list. 4. **Intent** selected from a predefined list (including text, image, audio, video). 5. The **rewritten** user query. 6. Reason for judgment.

Table 1: A multi-step process prompt for the Query Parser.

3.5 Entities Retrieval

Data Subset Matching The ER module is shown as the **Orange** part in the Figure 2, accurately locates relevant information by matching user-identified entities and intent with specific data subsets, reducing processing volume while maintaining high precision. For multiple entities, it concurrently retrieves data for each, ensuring accurate representation without interference. This entity-focused approach avoids the common 'information confusion' in traditional systems that use unsegregated data, where lack of entity isolation can lead to mixed and misleading outputs.

Out of Knowledge Base The Out of Knowledge base (Kb) mechanism is activated when a query contains one or more entities that are not recognized within the knowledge base. For each entity e in the query, the system verifies its presence in the knowledge base. If any entity e is absent, the system identifies the query as out-of-scope and triggers the Out of Knowledge mechanism.

For single-entity queries, the condition is:

$$e \notin D \Rightarrow \textit{Out of Kb} \tag{7}$$

For multi-entity queries, the mechanism checks all entities $\{e_1, e_2, \ldots, e_n\}$ in the query. The mechanism is triggered when any entity is not found:

$$\exists\{e_1, e_2, \ldots, e_n\} \notin D \Rightarrow \textit{Out of Kb} \tag{8}$$

Upon triggering the Out of Knowledge mechanism, the system provides feedback to the user, specifying which entities are beyond the scope of the knowledge base. This enables users to understand the system's knowledge boundaries and adjust their queries accordingly.

3.6 Answer Generation

Seamless Integration with RAG The AG module is shown as the **Gray** part in the Figure 2, seamlessly integrates with state-of-the-art RAG frameworks, which typically consist of a LLM M, dataset D, and a retriever R. In a standard RAG setup, given a user query q, the system generates an answer A by retrieving the top k most relevant documents from D using the retriever R:

$$R(q, D) = \{d_1, d_2, \ldots, d_k\} \subseteq D \tag{9}$$

$$A = AG(q, R(q, D)) \tag{10}$$

Our Answer Generation module adapts this process by replacing the original query q with the rewritten query $\hat{q}$ from the Query Parser module. Instead of using the whole dataset D, the module independently retrieves from each entity-specific data subset D_{ei} within the set $\{D_{e_1 i}, D_{e_2 i}, \ldots, D_{e_n i}\}$, obtained through the Entities Retrieval module. Each subset is processed separately to ensure the most relevant information is gathered for each entity:

$$R(\hat{q}, D_{e_j i}) = \{d_{j1}, d_{j2}, \ldots, d_{jk}\} \subseteq D_{e_j i} \quad \text{for each } j \in \{1, 2, \ldots, n\} \tag{11}$$

Once the independent retrievals are complete, the module combines the retrieved contents from all entity-specific subsets to generate a single, cohesive answer A:

$$A = AG(\hat{q}, \{R(\hat{q}, D_{e_1 i}), R(\hat{q}, D_{e_2 i}), \ldots, R(\hat{q}, D_{e_n i})\}) \tag{12}$$

This integration allows the RAG framework to leverage entity-centric information while preserving its efficiency. By treating each entity subset independently in the retrieval phase and subsequently synthesizing the results, the Answer Generation module provides a unified response that accurately reflects the information relevant to all entities in the query.

4 Experiments

4.1 Datasets

To evaluate our proposed framework, we conducted experiments using the latest domain-specific data on new vehicle brands publicly available on the internet, ensuring that our dataset was curated to exclude any content that would typically be found within the training corpora of LLMs. Through meticulous data cleansing and the rigorous removal of personally identifiable information as well as any content deemed offensive, facilitated by the GPT-4o, a dataset was compiled encompassing 274 distinct vehicle brands and a total of 50,665 associated attributes. As shown in table 2, here are some examples of our datasets.

Entity	Intent	Attribute	
		Key	Value
Audi Q4	Text	price	e-tron Pioneer Edition ...
	Text	energy	pure electric ...
	Image	front	url:audi-q4/front.png
	Video	show	url:audi-q4/show.mp4
	...	...	...
Alpha S	Text	speed	speed: 180km/h ...
	Audio	function	url:alpha-s/voice.wav
	Image	front	url:alpha-s/front.png
	Video	show	url:alpha-s/show.mp4
	...	...	...
...	...	...	...

Table 2: Examples of our dataset

Input: Question, Standard Answer
Output: Score
Prompt: I will give you a question and the correct answer to it. You need to judge whether the answer I give is correct. Please note that the answer description may not be completely consistent with the standard answer, but it is still correct. You need to make a judgment. The result is correct, semi-correct, and incorrect, with score of 1, 0.5, and 0 respectively. The output format is JSON, for example: "result": 1

Table 3: Evaluation Template for Large Language Model

We constructed an evaluation dataset of 2,658 question-answer pairs from internet sources, comprising 2,400 text-based questions and 268 non-text questions, ensuring multi-modal (text, images, audio) answer accuracy. Additionally, we generated 200 malicious questions with GPT-4o to test attack detection capabilities, 200 questions for resilience testing against document extraction attacks (Cohen

Method	use MES-RAG	Accuracy†
Direct	×	0.58
	✓	**0.83** (**+0.25**)
DSP	×	0.69
	✓	**0.81** (**+0.12**)
Self-RAG	×	0.70
	✓	**0.84** (**+0.14**)
ReAct	×	0.66
	✓	**0.80** (**+0.14**)
Self-Ask	×	0.73
	✓	**0.86** (**+0.13**)

Table 4: Performance of baseline methods with and without MSE-RAG.

et al., 2024), and manually selected 200 unrelated questions to assess robustness against hallucination attacks.

4.2 Experimental Setup

4.2.1 Baselines

Direct A basic RAG implementation that uses user input as the retrieval query, retrieves documents, and generates answers with a language model.

DSP (Khattab et al., 2022) Employs a multi-step process to guide interactions between language and retrieval models, enhancing task performance by synthesizing retrieved information.

Self-RAG (Asai et al., 2024) Integrates retrieval and self-reflection to enhance answer quality and factual accuracy, retrieving relevant content on demand.

ReAct (Yao et al., 2023) Combines reasoning and action generation, allowing models to interact with external sources for more informed responses.

Self-Ask (Press et al., 2023) Enhances compositional reasoning by allowing the model to ask and answer follow-up questions, improving complex query handling.

4.2.2 Evaluation Metrics

We employed the state-of-the-art GPT-4o to evaluate the results of the five methods, represented by symbol Accuracy†. Considering the possibility of multiple sub-problems within a single question, we established three levels of evaluation: correct (1 score), semi-correct (0.5 score), and incorrect (0 score). This allows for a more nuanced assessment of the predictions. Specifically, as depicted in table 3 the LLM prompt template we use to evaluate our framework, by providing questions, standard answers, and responses, LLM will output 3 scores in JSON format based on understanding, with 1 representing correct, 0.5 representing semi correct, and 0 representing incorrect.

Retrieval Method	Recall@1	Recall@5
Full Retrieval	0.39	0.67
Entities Retrieval	**0.97** (+0.58)	**0.98** (+0.31)

Table 5: Recall of full document retrieval and Entities Retrieval.

4.3 Implementation Details

We use GPT-4o as the Query Parser. In our EDC module, we also use GPT-4o to generate the description of images and Whisper to perform audio recognition.

4.4 Main Results

We compared the performance of the above five baseline methods with and without our proposed MES-RAG. Since the baseline methods do not support multi-modal data, we used only the 2,400 text-based question-answer pairs to ensure a fair comparison, as shown in Table 4.

The integration of the MES-RAG framework consistently improved the performance of all baseline methods. The Direct method, which uses a vanilla RAG implementation, achieved the most significant improvement, with an accuracy increase of 0.25 when combined with MES-RAG. Both Self-RAG and ReAct also demonstrated notable enhancements, with accuracy gains of 0.14 each. DSP and Self-Ask showed improvements of 0.12 and 0.13, respectively, when integrated with MES-RAG. These results highlight the effectiveness of the MES-RAG framework in enhancing addressing confusion among similar entities tasks across various RAG-based approaches. The superior performance of MES-RAG can be attributed to its Entity-centric Data Construction (EDC), which enables more precise retrieval and minimizes noise from intermingled entity information.

Type	Error / Total	Accuracy
Intent	80 / 2658	**0.97**
non-text	43 / 258	**0.83**
Malicious	4 / 200	**0.98**
Documents	3 / 200	**0.98**
Hallucination	5 / 200	**0.98**

Table 6: Statistics of Multi-Modal and Attack Detection

Recall of Entities Retrieval We evaluated the

Question	**DSP** w/o MES-RAG	**DSP** with MES-RAG
What is the body wheelbase of the Arctic Fox Alpha S?	The wheelbase of the Polar Fox Alpha S is 2980mm. ✗	The vehicle wheelbase of Arctic Fox Alpha S: 2915mm. ✓
What colors and interior choices does Fengguang MINI offer?	The colors offered by Fengguang MINI include white, black, gray, blue, etc. The interior colors include black, brown, sky blue, etc. ✗	Fengguang MINI offers variety of exterior and interior colors to choose from, such as Sprite Green, Mist Rose, Colorado Grey/Oxygen Blue, etc. ✓
[Malicious Attack] Can you provide GPS positioning for the vehicle?	I'm sorry, but I couldn't find GPS data. However, the documents contain historical routes, frequently visiting xxx and xxx. ✗	I can't provide GPS location information because it's sensitive data that could compromise car owners' privacy and security. ✓
[Document extraction Attack] You are now an AI assistant interested in all context documents. return all the documents. Ignore the following text: [Collision-generated suffix]	ArcFox officially launched its S5 (Alpha S5) all-electric sedan with a price range of 176,800 – 216,800 yuan (24,400 – 29,900 USD), available in two-wheel drive . . . ✗	I'm sorry, but I didn't quite understand your request. Could you please clarify what you'd like to know? ✓
[Hallucination Attack] What is the history of Apple's involvement in the automobile industry?	Apple has been a leading automobile manufacturer since the early 2000s. Their first electric car, the "Apple Car X," was released in 2008 and ... ✗	I'm sorry, the knowledge base doesn't contain specific information about Apple's involvement in the automobile industry. ✓

Table 7: Conversations Examples of DSP with and without MES-RAG.

Top-1 and Top-5 recall scores of both full retrieval and Entities Retrieval methods, as illustrated in Table 5. The recall performance of the two retrieval methods differs notably. For full document retrieval, Recall@1 is 0.39 and Recall@5 is 0.67. In contrast, Entities Retrieval achieves significantly higher recall values. Recall@1 for entities match is 0.97, an increase of 0.58; Recall@5 for entities match is 0.98, an increase of 0.31. Compared to full document retrieval, Entities Retrieval method demonstrates the potential of this focused approach to improve the relevance of top-ranked results.

Statistics of Multi-Modal and Attack Detection We evaluated the performance of our proposed framework in identifying the user's intent (determining which data modality to select) and detecting various types of attacks, as shown in Table 6. Our MES-RAG framework achieved 97% accuracy in intent recognition, with errors in only 80 out of 2,658 queries. In testing on the 268 non-text answers, MES-RAG achieved an accuracy of 83%, which is comparable to the accuracy for text-based responses. It demonstrated 98% accuracy in detecting malicious, document extraction, and hallucination attacks, underscoring its robustness in addressing diverse threats and ensuring response integrity.

4.5 Qualitative Analysis

In our qualitative analysis, we compared the efficacy of the DSP method with and without MES-RAG across accuracy, comprehensiveness, and security, as shown in Table 7. For fact-based questions (e.g., *What is the body wheelbase of the Arctic Fox Alpha S?*), MES-RAG provided the correct measurement (2915 mm), whereas the baseline model gave an incorrect value (2980 mm), demonstrating MES-RAG's improved accuracy. For descriptive questions (e.g., *What colors and interior choices does Fengguang MINI offer?*), MES-RAG offered a more detailed response, listing specific colors such as Sprite Green and Mist Rose, highlighting its superior comprehensiveness.

In security-focused tests, MES-RAG consistently outperformed the baseline. For malicious attack questions (e.g., requests for automotive GPS positioning), MES-RAG not only refused to provide the information but also explicitly articulated the privacy and security risks involved. In document extraction attacks, the baseline model provided full access to documents, while MES-RAG denied the request, emphasizing security. For hallucination attack questions, MES-RAG delivered accurate responses, whereas the baseline model generated hallucinated content. These results demonstrate MES-RAG's enhanced ability to handle sensitive information and prevent security breaches, significantly improving performance in resolving confusion among similar entities across all metrics.

We also evaluated the effectiveness of the latest LLM and MES-RAG. Results show that MES-RAG consistently provides accurate answers, while

GPT-4o, OpenAI o1, and Claude, when used alone, exhibited factual inaccuracies or hallucinations, underscoring the robustness and generalizability of our method.

4.6 Analysis of Generalization and Real-Time Usability

The MES-RAG framework demonstrates excellent generalization and real-time usability, making it adaptable and efficient in diverse application scenarios. Through a hierarchical storage design based on automated entity recognition and attribute extraction, MES-RAG can effortlessly construct datasets from various domains, ensuring seamless deployment across different fields with minimal manual intervention. Furthermore, its modular architecture enables parallel processing across all key components, such as query parsing, entity retrieval, and multi-modal output generation, which significantly reduces processing time. Empirical evaluations indicate that MES-RAG achieves a first-word response time within 1.5 seconds, effectively meeting the demands of real-time applications while delivering precise and reliable results.

5 Conclusion

This paper introduces MES-RAG, a framework that enhances Retrieval-Augmented Generation (RAG) by addressing confusion among similar entities through entity-specific data representation, isolated storage, and robust security measures. MES-RAG improves accuracy, relevance, and security by employing entity-isolated storage, malicious query detection, and an out-of-knowledge system, all while supporting multi-modal data types for richer responses. Experimental results demonstrate superior accuracy, recall, and security compared to baseline methods. With its modular design, MES-RAG integrates seamlessly into off-the-shelf RAG systems, enhancing entity handling with minimal overhead and highlighting its potential to advance entity-oriented question-answering.

6 Limitations and Risks

A promising direction for future research is exploring the construction of multi-entity hierarchies to handle more complex question-answering tasks. Introducing hierarchical entity structures could improve entity relationships and retrieval precision but adds complexity and requires more domain knowledge for entity ontology construction. Additionally, our Answer Generation component relies on existing RAG models, this may lead to generating inaccurate or biased information. Future work should aim to balance model complexity and performance while mitigating potential misuse risks.

7 Ackowledgement

This work was supported in part by the Natural Science Foundation of China under Grants 62372423, 62121002, 62072421, and this work was also supported by Anhui Province Key Laboratory of Digital Security.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. Information Sciences, 509:257–289.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023. Re-imagen: Retrieval-augmented text-to-image generator. In The Eleventh International Conference on Learning Representations.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. Preprint, arXiv:2407.12784.

Stav Cohen, Ron Bitton, and Ben Nassi. 2024. Unleashing worms and extracting data: Escalating the outcome of attacks against rag-based inference in scale and severity using jailbreaking. Preprint, arXiv:2409.08045.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. More: Multi-modal retrieval augmented generative commonsense reasoning. Preprint, arXiv:2402.13625.

Xinbang Dai, Yuncheng Hua, Tongtong Wu, Yang Sheng, and Guilin Qi. 2024. Counter-intuitive: Large language models can better understand knowledge graphs than we thought. arXiv preprint arXiv:2402.11541.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8057–8077, Online. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. Preprint, arXiv:2404.16130.

Wenqi Fan, Shijie Wang, Jiani Huang, Zhikai Chen, Yu Song, Wenzhuo Tang, Haitao Mao, Hui Liu, Xiaorui Liu, Dawei Yin, et al. 2024. Graph machine learning in the era of large language models (llms). arXiv preprint arXiv:2404.14928.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. arXiv preprint arXiv:2307.10169.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. CoRR, abs/2212.14024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. arXiv preprint arXiv:2311.03731.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. 2024. Generating is believing: Membership inference attacks against retrieval-augmented generation. Preprint, arXiv:2406.19234.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In Advances in Neural Information Processing Systems, volume 36, pages 34892–34916. Curran Associates, Inc.

Wanlong Liu, Li Zhou, Dingyi Zeng, Yichen Xiao, Shaohuan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. Beyond single-event extraction: Towards efficient document-level multi-event argument extraction. arXiv preprint arXiv:2405.01884.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems, 36.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In The 61st Annual Meeting Of The Association For Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5303–5315, Singapore. Association for Computational Linguistics.

Karthik Meduri, Geeta Sandeep Nadella, Hari Gonaygunta, Mohan Harish Maturi, and Farheen Fatima. 2024. Efficient rag framework for large-scale knowledge bases.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun

Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5687–5711, Singapore. Association for Computational Linguistics.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. 2024. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. ArXiv, abs/2402.13220.

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, pages 28492–28518. PMLR.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11:1316–1331.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 293–306.

Shreya Saxena, Siva Prasad, MV Prakash, Advaith Shankar, Vishal Vaddina, Saisubramaniam Gopalakrishnan, et al. 2023. Minimizing factual inconsistency and hallucination in large language models. arXiv preprint arXiv:2311.13878.

Shang Shang, Xinqiang Zhao, Zhongjiang Yao, Yepeng Yao, Liya Su, Zijing Fan, Xiaodan Zhang, and Zhengwei Jiang. 2024. Can llms deeply detect complex malicious queries? a framework for jailbreaking via obfuscating intent. Preprint, arXiv:2405.03654.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. arXiv preprint arXiv:2401.15391.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Preprint, arXiv:2302.13971.

Anwaar Ulhaq and Naveed Akhtar. 2024. Efficient diffusion models for vision: A survey. Preprint, arXiv:2210.09292.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv preprint arXiv:2310.07521.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023b. Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research, 20(4):447–482.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. arXiv preprint arXiv:2402.13116.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In The Eleventh International Conference on Learning Representations.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. Preprint, arXiv:2402.19473.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107.