

TAdaRAG: Task Adaptive Retrieval-Augmented Generation via On-the-Fly Knowledge Graph Construction

Jie Zhang^{1*}, Bo Tang^{2,3*}, Wanzi Shao¹, Wenqiang Wei³, Jihao Zhao^{3,4}, Jianqing Zhu⁵, Zhiyu Li³, Wen Xi⁵, Zehao Lin³, Feiyu Xiong³, Yanchao Tan^{1†}

¹College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China

²AIDS and SIAR, University of Science and Technology of China, Suzhou 215123, China

³MemTensor (Shanghai) Technology Co., Ltd., Shanghai 201306, China

⁴School of Information, Renmin University of China, Beijing 100086, China

⁵China Haisum Engineering Co., Ltd., Shanghai 200031, China

ee6r1c7@gmail.com, tangbo@mail.ustc.edu.cn, yctan@fzu.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) improves large language models by retrieving external knowledge, often truncated into smaller chunks due to the input context window, which leads to information loss, resulting in response hallucinations and broken reasoning chains. Moreover, traditional RAG retrieves unstructured knowledge, introducing irrelevant details that hinder accurate reasoning. To address these issues, we propose TAdaRAG, a novel RAG framework for on-the-fly task-adaptive knowledge graph construction from external sources. Specifically, we design an intent-driven routing mechanism to a domain-specific extraction template, followed by supervised fine-tuning and a reinforcement learning-based implicit extraction mechanism, ensuring concise, coherent, and non-redundant knowledge integration. Evaluations on six public benchmarks and a real-world business benchmark (NowNewsQA) across three backbone models demonstrate that TAdaRAG outperforms existing methods across diverse domains and long-text tasks, highlighting its strong generalization and practical effectiveness.

Code — <https://github.com/IAAR-Shanghai/TAdaRAG>

Introduction

In recent years, large language models (LLMs) (OpenAI 2024; Guo et al. 2025; Team, Q. 2024) have achieved significant breakthroughs in natural language processing, particularly in tasks such as text generation and question-answering systems (Yang et al. 2024; Laban et al. 2024). Despite these advances, LLMs often generate plausible but factually incorrect responses, known as hallucinations (Maynez et al. 2020; Zhou et al. 2020; Feng et al. 2024; Sun et al. 2024). To mitigate hallucinations, retrieval-augmented generation (RAG) integrates external knowledge sources into LLMs, providing enriched and contextually grounded inputs for more accurate responses (Jiang et al. 2023b).

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

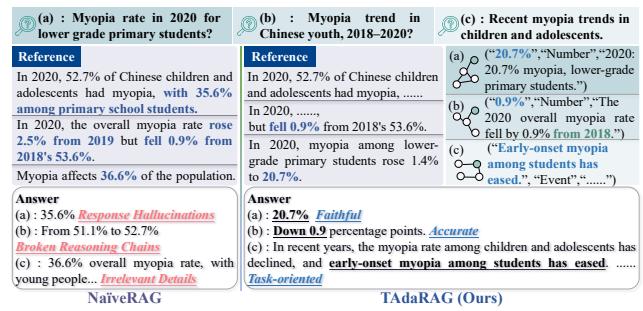


Figure 1: An illustrative example where the RAG system fails to generate correct responses due to truncation, leading to hallucinations, broken reasoning and irrelevant details. Our proposed TAdaRAG addresses these issues by integrating the task-adaptive knowledge graph dynamically.

However, current RAG approaches face several critical limitations in practical scenarios, retrieving large amounts of relevant information, such as related documents, which are split into smaller chunks due to input size constraints (Finardi et al. 2024; Yepes et al. 2024). This approach truncates complete knowledge, leading to information loss in chunks, which can cause hallucinations in responses, as shown in Case (a) of Figure 1. Additionally, discrete chunks fail to capture the inherent logical relationships within the corpora, disrupting reasoning chains and affecting accuracy in complex tasks, as seen in Case (b). Moreover, traditional RAG models input unorganized knowledge, recalling irrelevant details that hinder key information extraction and may impact practical usability, as illustrated in Case (c).

Recent graph-based RAG methods leverage knowledge graphs (KGs) to organize information through structured relationships, enhancing reasoning with summary-like responses (Peng et al. 2024; Guo et al. 2024). However, these methods rely on preconstructed KGs, which require manual maintenance, lack scalability, and introduce redundant or incomplete information, limiting retrieval accuracy. To this end, we propose TAdaRAG, a novel framework that diverges from traditional graph-based RAG by integrating task-oriented KG construction directly into the reasoning

process, rather than during the retrieval phase. This dynamic structure mitigates text fragmentation hallucinations, enhances complex task reasoning, and enables more precise knowledge extraction without requiring predefined graphs.

Specifically, we first employ intent detection to route input texts to tailored extraction templates, ensuring precise initial graph construction across diverse domains. Then we refine the process through supervised fine-tuning, transforming fragmented external knowledge into concise, logically organized, and non-redundant structures. Finally, to achieve automatic KG extraction and self-optimization, we introduce an instruction-level implicit extraction mechanism optimized via reinforcement learning, which significantly enhances the relevance and accuracy of the extracted KGs.

We comprehensively evaluated the performance of TAdaRAG and existing RAG baselines on six public benchmarks and NowNewsQA, a business-scenario benchmark for Chinese current affairs news question answering (QA). Meanwhile, the incremental results at each stage shown in the ablation results demonstrate the effectiveness of our proposed method. In addition, we conducted a human evaluation of the answers generated by different methods, and the results further confirm that TAdaRAG consistently outperforms other advanced baseline methods in terms of answer quality. We also examined the consistency between LLMs ratings and human ratings, as well as the consistency among expert ratings, further demonstrating the reliability of LLM scoring and TAdaRAG. To sum up, our main contributions are as follows:

- We propose TAdaRAG, a task-adaptive RAG framework that integrates structured KG representations into the reasoning process. By dynamically constructing domain-relevant subgraphs, TAdaRAG effectively addresses the hallucination problem caused by chunked input in long-text tasks, enhances reasoning capabilities in complex scenarios, and ensures more accurate extraction and utilization of external knowledge.
- We evaluate TAdaRAG on both public benchmarks and real-world business scenarios, covering tasks across various domains and long-text settings. The results show that TAdaRAG outperforms existing methods in multiple fields and long-context RAG tasks, demonstrating strong cross-domain generalization.
- TAdaRAG has been successfully deployed in commercial applications, with trial accounts now available for user access and testing.

Related Work

Retrieval-Augmented Generation (RAG). RAG (Borgeaud et al. 2022; Guu et al. 2020) enhances language models by retrieving information from external knowledge bases to improve text accuracy and credibility. Early methods encoded documents into vectors for fast retrieval (Chen et al. 2024b; Karpukhin et al. 2020a), while later approaches introduced multi-step retrieval mechanisms to iteratively refine results and enhance long-text comprehension (Jiang et al. 2023b; Su et al. 2024; Trivedi et al. 2023). RAG has been widely used in tasks such as

Multi-Document QA (Karpukhin et al. 2020b; Trivedi et al. 2023), summarization (Edge et al. 2024; Laban et al. 2024), and Open-domain QA (Siriwardhana et al. 2023; Yang et al. 2024; Zhang, Fang, and Chen 2024). Recent advancements also explore adaptive retrieval strategies to better align retrieved content with query intent (Mo et al. 2024). However, existing methods often process input text coarsely, introducing irrelevant information and omitting key details, which limits reasoning accuracy..

Graph-enhanced Retrieval-Augmented Generation.

Graph-enhanced RAG (Peng et al. 2024) models complex knowledge relationships through graph structures, enhancing retrieval comprehensiveness and reasoning capabilities. Early efforts extracted facts from predefined KGs (e.g., Wikidata (Vrandečić and Krötzsch 2014)) but struggled with dynamic task adaptation. Recent studies shift toward building task-specific graphs from raw text. For example, GraphRAG (Edge et al. 2024) constructs KGs from textual data and generates community-based summaries; PathRAG (Chen et al. 2025) significantly improves retrieval efficiency and reduces redundancy by identifying key relation paths; HippoRAG (Jimenez Gutierrez et al. 2024) uses LLMs to convert document corpora into open KGs, serving as its artificial hippocampal index; and Chain of Knowledge (Wang et al. 2024) helps mitigate hallucinations through structured evidence generation and rigorous verification. These advancements highlight the potential of graph-based methods to dynamically adapt to varied tasks and improve contextual relevance. These studies underscore the role of graph structures in improving generation logic and reasoning accuracy. However, most existing methods rely on manually crafted graph models or static corpora, limiting their generalization ability.

Handling Long-Context Tasks in LLMs. In long-text tasks, LLM enhancement methods can be broadly divided into three categories. The first category focuses on expanding the context window through direct extension techniques (Jin et al. 2024) and key-value (KV) cache pruning (Zhang et al. 2023), where THINK (Xu et al. 2024b) reduces memory overhead by pruning redundant KV channels based on low-rank attention patterns. Neurocache (Safaya and Yuret 2024) introduces an external vector cache that stores compressed past states and employs efficient k-Nearest Neighbors retrieval to extend effective context lengths without full model retraining. The second category improves long-text response quality through model fine-tuning, employing strategies such as supervised fine-tuning (SFT) (Chen et al. 2023) and reinforcement learning from human feedback (RLHF) (Zhao et al. 2024). The third category introduces auxiliary structures to guide reasoning. For example, Quiet-STaR (Zelikman et al. 2024) enhances CoT reasoning by generating token-level predictions; MEMORAG (Qian et al. 2024) utilizes a memory module to generate retrieval cues; and INFO-RAG (Xu et al. 2024a) refines retrieved content through a document optimizer. These methods help reduce noise, redundancy, and coherence issues in text generation but often require manually designed schemas or costly retraining.

Methodology

The TAdaRAG framework for language generation task (e.g., summarization) is formulated as follows. Given a database $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{|\mathbb{D}|}$, where each pair (x, y) represents a document and its summary, the framework consists of training and inference stages. For training (Figure 2), the **Supervised Knowledge Extraction Fine-tuning Stage** leverages strong LLMs (OpenAI 2024; Guo et al. 2025) and domain-specific templates to generate KGs from \mathbb{D} , enabling SFT for cold-start. The **Task-Adaptive Knowledge Graph Construction Stage** then trains the model to dynamically extract task-adaptive KG using the REINFORCE algorithm. During inference, the adaptive KG supports LLMs in generating accurate responses.

Supervised Knowledge Extraction Fine-tuning

Intention Detection Pretrained language models often struggle with precise entity extraction, introducing irrelevant or redundant entities, which harms downstream performance, especially in real-world industry scenarios where capturing relevant entities is crucial for reliable responses.

To address this challenge, we introduce carefully designed external templates to guide entity extraction, reducing redundancy and improving entity relevance. Considering the diverse industry-specific requirements, we first identify key application domains and manually select high-impact entity types that most effectively support answer generation. Additionally, we design tailored extraction templates for general as well as specialized scenarios, enabling the model to determine the required entity types, entity description specifications, and relationships among entities for each domain. This significantly enhances the model’s cross-domain generalization capability.

In practice, given a user query q and external knowledge r , we use prompts to perform intent detection and select the appropriate template t , enabling the model to accurately identify the required node types and relation patterns for the current task. The model then extracts a typed knowledge graph that directly aligns with the detected intent and maximizes downstream answer quality. In the RAG setting, this dynamically constructed, domain-aware knowledge graph is integrated into the generation pipeline to produce more accurate and contextually appropriate responses.

Fine-tuning on High-Quality Corpus Based on the obtained template t , we construct an instruction set $I = \{q, r, t\}$ by integrating the question q and external knowledge r , serving as data samples for supervised knowledge extraction fine-tuning (Friel, Belyi, and Sanyal 2024). Then, we use strong LLMs to perform the knowledge extraction task on the instruction set, generating high-quality knowledge extraction results G . Based on these instruction-graph pairs, we create a high-quality dataset for supervised knowledge extraction fine-tuning, covering four question domains and seven sub-datasets, totaling 9,548 fine-tuning samples. Finally, leveraging LoRA (Hu et al. 2022) for supervised fine-tuning on a pretrained LLM with the dataset, we train the model to achieve excellent knowledge extraction capa-

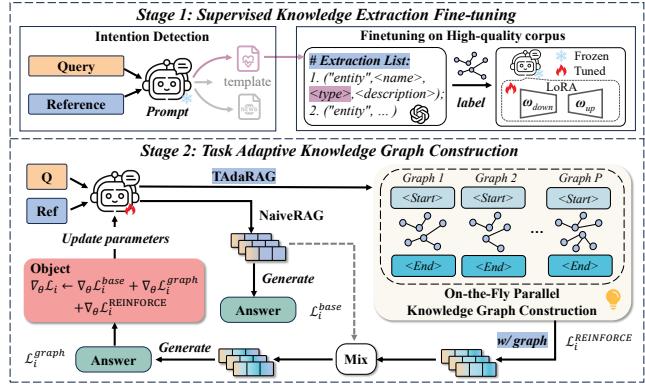


Figure 2: An illustration of our proposed TAdaRAG framework and its two-stage training: (1) Supervised Knowledge Extraction Fine-tuning and (2) Task-Adaptive Knowledge Graph Construction.

bilities, laying a crucial foundation for on-the-fly construction of high-quality KGs.

Task-Adaptive Knowledge Graph Construction

Parallel Construction LLMs’ inability to assess single KG quality can lead to suboptimal outcomes when relying on only one KG for next-token generation. Sampling multiple candidate KGs helps mitigate this issue. Therefore, based on the training set $\mathbb{D}_{train} = \{(x_i, y_i)\}_{i=1}^N$, the model performs knowledge extraction on the input text and constructs p parallel subgraphs $g_i = \{g_i^1, g_i^2, \dots, g_i^p\}$. Learnable tokens $<\text{startextraction}>$ and $<\text{endextraction}>$ are used to indicate the start and end positions of knowledge extraction, enabling the model to naturally embed the knowledge graph during generation and achieve implicit knowledge extraction.

Mixing Network To enhance the model’s understanding and utilization of structured information in knowledge graphs, we propose a graph-structured fusion network. Given an instruction-response pair (x_i, y_i) and a subgraph g_i^k , we first obtain the hidden state $H_{i,j}^{\text{base}} = \text{hs}[x_i; y_i^{(j-1)}]$ of the j -th token in the response under the original RAG method. In our proposed TAdaRAG, which incorporates the subgraph g_i^k , the corresponding hidden state is $H_{i,j,k}^{\text{graph}} = \text{hs}[x_i; y_i^{(j-1)}; g_i^k]$.

To better measure the importance of knowledge graph compared to direct responses, we concatenate the two obtained hidden states and input them into a three-layer MLP with ReLU activation to compute the weight for each token in the response:

$$\omega_{i,j,k} = \text{MLP}(\text{concat}(H_{i,j}^{\text{base}}, H_{i,j,k}^{\text{graph}})). \quad (1)$$

where $\text{concat}(\cdot)$ denotes the concatenation.

Meanwhile, we compute the log-likelihoods under the conditions of without and with the knowledge graph:

$$l_{i,j,k}^{\text{w/o graph}} = \log p_\theta(y_i^j | x_i, y_i^{(j-1)}) = f(H_{i,j}^{\text{base}}), \quad (2)$$

$$l_{i,j,k}^{\text{w/graph}} = \log p_\theta(y_i^j | x_i, y_i^{(j-1)}, g_i^k) = f(H_{i,j,k}^{\text{graph}}), \quad (3)$$

where $f(\cdot) = \text{softmax}(\text{Imhead}(\cdot))$, $\text{Imhead}(\cdot)$ maps the hidden state to the vocabulary space to obtain logits, and $\text{softmax}(\cdot)$ converts them into a probability distribution.

Finally, we compute the weighted sum of the j -th token to obtain the log-likelihood incorporating the graph structure for generation and optimization:

$$l_{i,j,k}^{\text{mix}} = \omega_{i,j,k} \cdot l_{i,j,k}^{\text{w/graph}} + (1 - \omega_{i,j,k}) \cdot l_{i,j,k}^{\text{w/o graph}}. \quad (4)$$

Optimizing Graph Construction Given that the extracted KGs vary in length and content, we aim to determine the most beneficial subgraph for optimizing the response. Specifically, for all instruction-answer pairs (x_i, y_i) in the training set $\mathbb{D}_{\text{train}}$, our objective is to find an optimal subgraph $\tilde{g}^{(i)}$ that maximizes $\pi_\theta(y_i | x_i, \tilde{g}^{(i)})$.

To achieve this, we design a reward function R based on the REINFORCE algorithm to quantify the impact of introducing subgraphs on response generation. We first consider the loss of model direct response to ensure its ability to answer without external knowledge:

$$\mathcal{L}^{\text{base}} = -\mathbb{E}_{(x,y) \sim \mathbb{D}} [\log \pi_\theta(y | x)]. \quad (5)$$

Next, we consider the loss when incorporating the knowledge graph into the response, aiming to train the model to learn how to integrate the input instruction and subgraph to generate more accurate answers:

$$\mathcal{L}^{\text{graph}} = -\mathbb{E}_{(x,y,g) \sim \mathbb{D} \cup g} [\log \pi_\theta(y | x, g)]. \quad (6)$$

At this point, for the case where the i -th instruction references the k -th subgraph, we design the reward function R as follows:

$$R_{i,k} = \max(0, \mathcal{L}_i^{\text{base}} - \mathcal{L}_{i,k}^{\text{graph}} - \bar{R}_i). \quad (7)$$

This reward function increases the likelihood of selecting knowledge graphs that perform better than the average \bar{R}_i . Finally, the REINFORCE loss term is defined as:

$$\mathcal{L}^{\text{REINFORCE}} = -R_{i,k} \cdot \log \pi_\theta(g_i^k | x_i). \quad (8)$$

Thus, the total loss function for model training is defined as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}^{\text{base}} + (1 - \alpha) \cdot \mathcal{L}^{\text{graph}} + \beta \cdot \mathcal{L}^{\text{REINFORCE}}, \quad (9)$$

where α and β are hyperparameters.

Experiments

In the experiments, we investigated the following research questions (**RQs**): **RQ1:** How does the effectiveness of our proposed TAdaRAG framework compare to the state-of-the-art RAG baselines? **RQ2:** How effectively does the TAdaRAG framework handle long-context tasks? **RQ3:** How does the TAdaRAG framework generalize to real-world task scenarios? **RQ4:** Has each training stage of our framework contributed effectively? **RQ5:** How do different hyperparameter values in TAdaRAG influence the performance?

In addition to addressing these questions, we conducted further experiments including statistical significance testing, latency analysis, KG refinement evaluation, evidence verification and case studies in the extend version.

Evaluation of Public Datasets

Datasets & Metrics. (1) **For Q&A Tasks:** We conducted experiments on Health, Biology, and Legal datasets for

open-domain QA (Qian et al. 2024), and HotpotQA and 2WikiMQA for restricted QA (Yang et al. 2018; Ho et al. 2020; Bai et al. 2024), evaluated with F1. (2) **For Summarization:** We used GovReport dataset to evaluate summary generation (Huang et al. 2021), evaluated with ROUGE-L.

Baseline. We compare TAdaRAG against seven representative RAG approaches. NaïveRAG (Gao et al. 2023) implements the standard RAG paradigm. BGE-M3 (Chen et al. 2024a) is a versatile text embedding model with multi-lingual, multi-task, and multi-granularity capabilities. RQ-RAG (Chan et al. 2024) extends this framework by jointly learning to reformulate queries before retrieval, thereby improving the relevance of retrieved passages. GraphRAG (Edge et al. 2024) adopts LLM-extracted knowledge graphs and community detection to enable scalable, comprehensive summarization for complex queries. HippoRAG (Jimenez Gutierrez et al. 2024) adopts hippocampal indexing theory and KGs for reasoning. MEMORAG (Qian et al. 2024) effectively expands the application scope of the RAG system by introducing memory modules. Finally, PathRAG (Chen et al. 2025) retrieves nodes, prunes key paths, and converts them to text to guide LLM generation.

Implementation details. In this study, we adopt Mistral-7B-Instruct (Jiang et al. 2023a), Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Team 2024) as the backbone model. For the two-stage training strategy, we perform SFT in **Stage 1** for 5 epochs, with a maximum input sequence length of 20,480 tokens, a batch size of 1, gradient accumulation over 8 steps, and a cosine learning rate scheduler initialized at 5e-5. In **Stage 2**, we train the model using ZeRO stage-2 optimization with the AdamW optimizer, a per-GPU batch size of 1, and bfloat16 precision, for 3 epochs at a learning rate of 5e-7. During training, we generate multiple KGs per instruction using a sampling temperature $T = 0.6$; for evaluation, we apply greedy decoding. The maximum KG length is set to 2048 tokens, with longer outputs truncated. The entire training process takes approximately 16 hours using 8 NVIDIA A100 (80 GB) GPUs, with 4 hours for Stage 1 and 12 hours for Stage 2.

Main Results (RQ1) Table 1 and Figure 4 highlights several key insights. Taking Mistral-7B results as an example:

Hallucination Mitigation. TAdaRAG mitigates information loss in chunked text through on-the-fly knowledge graph construction with two-stage training, outperforming state-of-the-art (SOTA) RAG baselines MEMORAG in factual domains (Health: 37.40 → 40.77; Biology: 35.70 → 39.31). On the Legal dataset, TAdaRAG significantly outperforms NaiveRAG (35.80 → 49.88), demonstrating strong factuality in answering questions that require the integration of lengthy legal clauses. Its close performance to MEMORAG (49.88 vs. 51.20) further suggests potential for domain-specific enhancement. These results demonstrate that the structured knowledge integration of TAdaRAG not only improves answer accuracy but also enhances model robustness across domains.

Reasoning Enhancement. TAdaRAG enhances reasoning chain completeness by dynamically organizing knowledge hierarchies. Compared to MEMORAG, it

Methods		ULTRADOMAIN			LongBench		
		Health	Biology	Legal	HotpotQA	2WikiMQA	GovReport
Based on Mistral-7B-Instruct							
Standard RAG	NaïveRAG	34.80	34.10	35.80	37.60	20.60	27.40
	BGE-M3	33.20	32.20	42.00	36.20	20.30	26.10
Advanced RAG	RQ-RAG	33.37	33.42	42.60	37.00	21.50	18.60
	GraphRAG	35.60	34.80	37.65	38.00	36.50	25.60
	HippoRAG	34.54	34.23	35.36	39.30	33.10	25.22
	PathRAG	21.67	20.10	18.57	24.22	18.71	15.66
	MEMORAG	37.40	35.70	51.20	42.90	30.30	31.60
TAdaRAG	w/ graph	38.19	36.87	32.92	38.30	38.48	33.72
	w/ sft	<u>40.00</u>	<u>38.92</u>	39.32	41.60	<u>38.86</u>	<u>35.39</u>
	w/ reinforce	40.77*	39.31*	49.88	44.83*	39.31*	36.41*
Based on Qwen2.5-7B-Instruct							
Standard RAG	NaïveRAG	35.25	35.28	36.55	45.78	32.28	20.68
	BGE-M3	30.20	33.20	40.60	36.28	33.30	20.1
Advanced RAG	RQ-RAG	31.50	31.90	38.80	37.40	34.10	21
	GraphRAG	36.82	34.67	40.62	43.33	37.52	28.46
	HippoRAG	35.73	35.53	40.31	45.89	36.16	27.23
	PathRAG	32.65	30.46	32.34	33.79	29.09	24.55
	MEMORAG	36.87	36.00	47.60	37.99	35.32	31.13
TAdaRAG	w/ graph	40.77	38.31	41.76	48.74	42.84	33.88
	w/ sft	<u>41.35</u>	<u>39.62</u>	44.55	<u>49.03</u>	<u>43.37</u>	<u>35.47</u>
	w/ reinforce	42.38*	40.75*	46.83	49.23*	43.79*	36.95*

Table 1: Experiments based on Mistral-7B-Instruct and Qwen2.5-7B-Instruct. GovReport uses the ROUGE-L metric, and other datasets use the F1 metric. The w/ graph variant leverages KGs via prompting, w/ sft enhances knowledge extraction through SFT, and w/ reinforce optimizes task-adaptive KG construction with reinforcement learning. * indicates statistically significant improvements ($p < 0.01$) over SOTA RAG baselines.

achieves notable improvements on complex reasoning tasks (2WikiMQA: $30.30 \rightarrow 39.31$) and multi-hop question answering (HotpotQA: $42.90 \rightarrow 44.83$), demonstrating its superior ability to support structured, multi-step inference.

Task-Oriented Extraction. TAdaRAG outperforms the SOTA MEMORAG in summarization tasks (GovReport: $31.60 \rightarrow 36.41$), indicating improved scalability and precision in integrating task-oriented knowledge for long-text summarization.

Long-Context Task Analysis (RQ2) We compared TAdaRAG with three long-context mechanisms on the same base model: Self-Extend (Jin et al. 2024), H2O+THINK, and SnapKV+THINK (Xu et al. 2024b) on six datasets, using the same foundation model, as shown in Figure 3. On HotpotQA and 2WikiMQA, TAdaRAG effectively handles multi-document QA, rivaling dedicated long-context models without modifying storage. In GovReport, it excels in long-document summarization, demonstrating the benefits of task-adaptive KG construction for long-context task.

Evaluation of Business Scenarios

Dataset Construction. Based on real-world business scenarios from Xinyu AI Search (Wu et al. 2024), we crafted NowNewsQA, a multi-document QA dataset focused on the news domain, consisting of 3,150 examples (3,000 for training and 150 for testing). The questions were derived from diverse real-user queries covering trending topics, political

developments, economic shifts, and major societal events, reflecting the dynamic and multifaceted nature of news consumption. Reference documents were retrieved using the Xinyu AI Search engine, which employs hybrid retrieval techniques to aggregate and rank documents based on their relevance to each query. Each document set mirrors actual search outputs from a production-level engine, preserving realistic characteristics such as redundancy, noise, and partial relevance. As such, NowNewsQA serves as a challenging benchmark for evaluating model performance in reasoning over fragmented news content within real-world business contexts.

Multi-faceted Evaluation Criteria. Because it is difficult to establish a gold-standard answer for RAG tasks—especially given the inherent diversity and subjectivity of answers in real-world industry scenarios—we adopted a rating-based evaluation framework instead of relying on exact match metrics. To ensure the robustness and relevance of the evaluation, we invited domain experts with journalism backgrounds and master’s degrees to design a set of multi-dimensional criteria that comprehensively assess the quality of generated answers. The criteria include: (1) Relevance, (2) Numerical Precision, (3) Conciseness, (4) Factuality, (5) Timeliness, (6) Comprehensiveness, (7) Clarity, (8) Coherence, and (9) Insightfulness.

LLM Evaluation. Conducting comprehensive human evaluations using multi-faceted criteria for all experiments

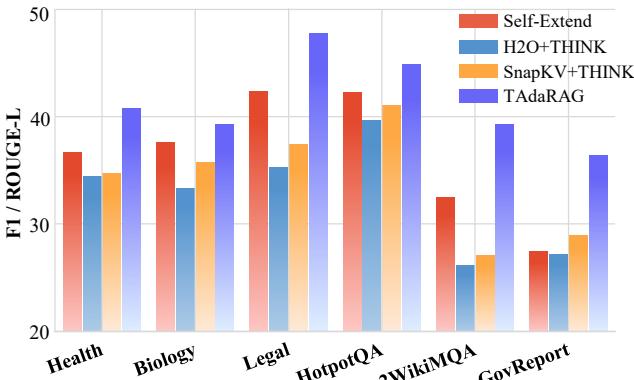


Figure 3: Long-context experiments on Mistral-7B.

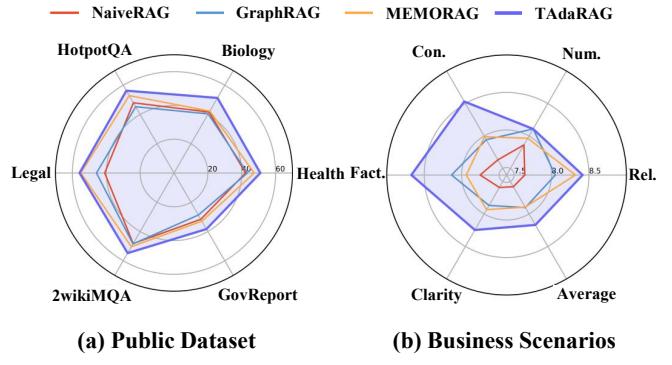


Figure 4: Experiments on Qwen2.5-14B.

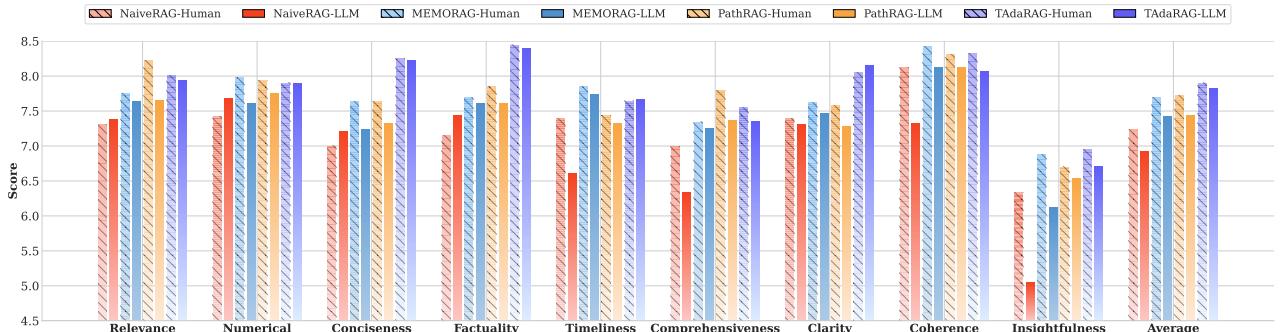


Figure 5: Multi-faceted comparison of different methods based on human and GPT-4o. Higher values indicate better performance, with a maximum value of 10.

is cost-prohibitive. To address this, we adopted an LLM to assess generated answers. This automated evaluation strategy has been widely applied in prior work (Chen et al. 2025; Peng et al. 2024; Qian et al. 2024). Specifically, we utilized GPT-4o, prompting it to generate point-by-point justifications and assign a final score based on the defined criteria. The temperature was set to 0 to ensure deterministic outputs. As shown in Table 2, the LLM-based scores exhibit strong correlation with human evaluations, as evidenced by the comparative results in Figure 5.

Metric	Value	Metric	Value
Relevance	0.706	Comprehensiveness	0.850
Numerical Precision	0.755	Clarity	0.867
Conciseness	0.847	Coherence	0.925
Factuality	0.842	Insightfulness	0.881
Timeliness	0.828	-	-

Table 2: Pearson correlation coefficients between human and LLM scores for different evaluation criteria.

Baseline. We test TAdaRAG against NaïveRAG (Gao et al. 2023), MEMORAG (Qian et al. 2024), and PathRAG (Chen et al. 2025) on real-world business scenarios based on Mistral-7B-Instruct, with performance assessed under multi-faceted evaluation criteria.

Comparative Evaluation (RQ3) Multi-Faceted Evaluation of the Generated Answer. We compare TAdaRAG with existing RAG systems by inviting human experts to assess the generated answers using the

multi-faceted evaluation criteria. As shown in Figure 5, TAdaRAG achieves the highest overall average score (7.904 vs. 7.720), demonstrating superior performance across multiple dimensions. Notably, it significantly outperforms other methods in conciseness (8.251 vs. 7.637) and factuality (8.449 vs. 7.850). These results underscore the effectiveness of our task-adaptive knowledge graph construction in mitigating hallucinations and enhancing response accuracy. Furthermore, TAdaRAG’s strong performance on real-world business queries highlights its practical feasibility and broad generalizability in applied settings.

LLM-Based Multi-Faceted Evaluation. Figure 5 reports the results of the multi-faceted evaluation conducted by GPT-4o. Although the absolute scores differ from those given by human annotators (see Figure 5, the relative rankings are largely consistent, indicating a strong correlation between the two (see Table 2). These findings suggest that LLM-based evaluation is a reliable proxy for human judgment and can effectively reflect model performance across key dimensions.

Human Scoring Consistency Verification. To verify the consistency of human judgments, we computed the inter-rater agreement among three expert annotators on 150 test samples using Pearson correlation across nine evaluation dimensions. The average correlation scores for each metric are reported in Table 3. As shown, all models exhibit consistently high agreement, indicating that the evaluation criteria are well-defined and easy to apply. These strong correlations further suggest that human ratings are stable and reliable,

Model	Rel.	Num.	Concise	Fact.	Time.	Comp.	Clarity	Coh.	Insight
NaïveRAG	0.813	0.727	0.777	0.814	0.742	0.663	0.774	0.764	0.676
MEMORAG	0.764	0.701	0.763	0.841	0.747	0.667	0.815	0.722	0.691
PathRAG	0.797	0.674	0.807	0.806	0.759	0.716	0.712	0.691	0.654
TAdaRAG	0.791	0.691	0.856	0.821	0.719	0.683	0.806	0.68	0.694

Table 3: Pearson correlation coefficients for human scoring consistency verification.

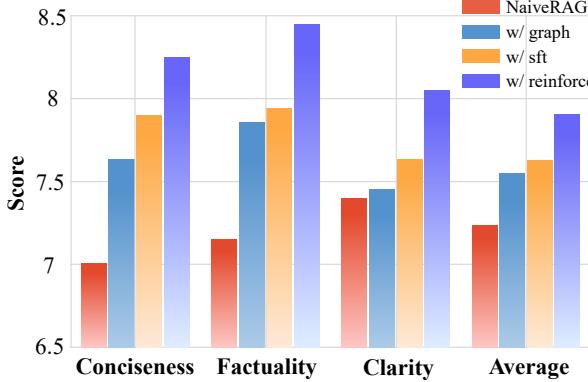


Figure 6: Ablation study on NowNewsQA.

thereby reinforcing the credibility of the evaluation results presented in this study.

Ablation Study (RQ4)

This section investigates the impact of structured KGs in our setting and validates the necessity of the two-stage training design in TAdaRAG. The results, summarized in Table 1 and Figure 6, reveal that:

The introduction of prompt-based KG, which utilizes KGs extracted by the model itself through domain-specific templates, significantly improves model performance across various datasets. For example, it outperforms NaïveRAG, with 2WikiMQA improving from 20.60 to 38.48 and GovReport from 27.40 to 33.72, indicating that KGs effectively enhance model reasoning and summarization capabilities. Moreover, the results of direct prompt-based integration validate the feasibility of enhancing the model through KG incorporation, providing a solid basis for the subsequent optimization of KG construction.

Supervised Knowledge Extraction Fine-tuning enables the model to achieve ideal KGs extraction, bridges prompt-based KG construction and task adaptation. For instance, it improves accuracy by 19.44% on Legal dataset and by 5.56% on Biology dataset, demonstrating its effectiveness in optimizing knowledge extraction and boosting answer precision. Results outperform the prompt-based version, indicating that higher-quality KGs lead to better performance.

Task-Adaptive KG Construction further improves the model across all datasets and tasks, validating the effectiveness of our REINFORCE algorithm. For example, it significantly enhances complex question answering in the Legal domain, with performance improving by 26.86% compared to Stage 1. Moreover, it boosts results across other datasets to achieve SOTA performance, showcasing the indispensability of this stage.

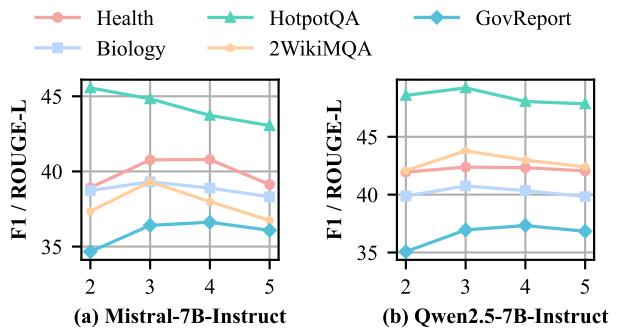


Figure 7: Parallel Subgraphs vs. Performance

Hyperparameter Analysis (RQ5)

We investigate the impact of varying the number of parallel subgraphs sampled during training on model performance. As shown in Figure 7, both Mistral-7B and Qwen2.5-7B achieve optimal performance when the number of subgraphs is set to 3 across most datasets. Using a subgraph number of 2 restricts the model’s reasoning capacity and impedes the identification of optimal substructures. Increasing the number beyond 3 (e.g., 4 or 5) may introduce noise into the optimization process for models with 7B parameters, thereby degrading performance. Specifically, HotpotQA benefits from 2 subgraphs, as this suffices for focused reasoning, whereas GovReport performs better with 4 subgraphs due to its need for more diverse and informative subgraphs to support evidence aggregation in summarization. Moreover, experiments with Qwen show stable performance across varying subgraph numbers, suggesting that stronger models with higher performance exhibit greater training stability.

Conclusion

In this paper, we present TAdaRAG, a task-adaptive RAG framework that dynamically constructs structured knowledge graphs to mitigate hallucinations, strengthen reasoning, and enhance knowledge extraction. Extensive experiments on six public benchmarks and a real-world business dataset demonstrate consistent improvements across factual QA, multi-hop reasoning, long-text summarization, and industry scenarios, highlighting strong generalization and practical utility. The current design involves dynamic KG construction and multi-stage training, adding computational overhead and relying partly on manually crafted templates, which may constrain efficiency and adaptability in more complex scenarios. Future work will further improve the efficiency of KG construction, reduce computation cost, and enhance adaptability, aiming to strengthen the framework’s scalability and reliability in real-world settings.

Acknowledgments

This work was supported in part by the Fujian Provincial Artificial Intelligence Industry Development Technology Project under Grant (2025H0042) and Talent Foundation of Fuzhou University (No. XRC-23027).

References

- Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3119–3137. Bangkok, Thailand: Association for Computational Linguistics.
- Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.
- Chan, C.-M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; and Fu, J. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.
- Chen, B.; Guo, Z.; Yang, Z.; Chen, Y.; Chen, J.; Liu, Z.; Shi, C.; and Yang, C. 2025. PathRAG: Pruning Graph-based Retrieval Augmented Generation with Relational Paths. *arXiv preprint arXiv:2502.14902*.
- Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024b. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Ku, L.-W.; Martins, A.; and Srikanth, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 2318–2335. Bangkok, Thailand: Association for Computational Linguistics.
- Chen, Y.; Qian, S.; Tang, H.; Lai, X.; Liu, Z.; Han, S.; and Jia, J. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.
- Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Balachandran, V.; and Tsvetkov, Y. 2024. Don’t Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14664–14690.
- Finardi, P.; Avila, L.; Castaldoni, R.; Gengo, P.; Larcher, C.; Piau, M.; Costa, P.; and Caridá, V. 2024. The chronicles of rag: The retriever, the chunk and the generator. *arXiv preprint arXiv:2401.07883*.
- Friel, R.; Belyi, M.; and Sanyal, A. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.
- Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; and Huang, C. 2024. LightRAG: Simple and Fast Retrieval-Augmented Generation.
- Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.
- Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Huang, L.; Cao, S.; Parulian, N.; Ji, H.; and Wang, L. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023a. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Jiang, Z.; Xu, F. F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; and Neubig, G. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7969–7992.
- Jimenez Gutierrez, B.; Shu, Y.; Gu, Y.; Yasunaga, M.; and Su, Y. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37: 59532–59569.
- Jin, H.; Han, X.; Yang, J.; Jiang, Z.; Liu, Z.; Chang, C.-Y.; Chen, H.; and Hu, X. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020a. Dense Passage Retrieval for Open-Domain Question Answering. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020b. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP (1)*, 6769–6781.

- Laban, P.; Fabbri, A.; Xiong, C.; and Wu, C.-S. 2024. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9885–9903. Miami, Florida, USA: Association for Computational Linguistics.
- Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Mo, F.; Qu, C.; Mao, K.; Wu, Y.; Su, Z.; Huang, K.; and Nie, J.-Y. 2024. Aligning query representation with rewritten query and relevance judgments in conversational search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1700–1710.
- OpenAI. 2024. Hello GPT - 4O. <https://openai.com/index/hello-gpt-4o/>.
- Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; and Tang, S. 2024. Graph Retrieval-Augmented Generation: A Survey. *ArXiv*, abs/2408.08921.
- Qian, H.; Zhang, P.; Liu, Z.; Mao, K.; and Dou, Z. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.
- Safaya, A.; and Yuret, D. 2024. Neurocache: Efficient Vector Retrieval for Long-range Language Modeling. *arXiv preprint arXiv:2407.02486*.
- Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; and Nanayakkara, S. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11: 1–17.
- Su, W.; Tang, Y.; Ai, Q.; Wu, Z.; and Liu, Y. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12991–13013. Bangkok, Thailand: Association for Computational Linguistics.
- Sun, Z.; Zang, X.; Zheng, K.; Song, Y.; Xu, J.; Zhang, X.; Yu, W.; and Li, H. 2024. Redep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.
- Team, Q. 2024. Qwen2.5: A Party of Foundation Models.
- Team, Q. 2024. Qwq: Reflect deeply on the boundaries of the unknown.
- Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9–14, 2023, 10014–10037. Association for Computational Linguistics.
- Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.
- Wang, J.; Sun, Q.; Li, X.; and Gao, M. 2024. Boosting Language Models Reasoning with Chain-of-Knowledge Prompting. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4958–4981. Bangkok, Thailand: Association for Computational Linguistics.
- Wu, Y.; Tang, B.; Xi, C.; Yu, Y.; Wang, P.; Liu, Y.; Kuang, K.; Deng, H.; Li, Z.; Xiong, F.; et al. 2024. Xinyu: An Efficient LLM-based System for Commentary Generation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6003–6014.
- Xu, S.; Pang, L.; Yu, M.; Meng, F.; Shen, H.; Cheng, X.; and Zhou, J. 2024a. Unsupervised information refinement training of large language models for retrieval-augmented generation. *arXiv preprint arXiv:2402.18150*.
- Xu, Y.; Jie, Z.; Dong, H.; Wang, L.; Lu, X.; Zhou, A.; Saha, A.; Xiong, C.; and Sahoo, D. 2024b. Think: Thinner key cache by query-driven pruning. *arXiv preprint arXiv:2407.21018*.
- Yang, X.; Sun, K.; Xin, H.; Sun, Y.; Bhalla, N.; Chen, X.; Choudhary, S.; Gui, R.; Jiang, Z.; Jiang, Z.; et al. 2024. CRAG-comprehensive RAG benchmark. *Advances in Neural Information Processing Systems*, 37: 10470–10490.
- Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
- Yepes, A. J.; You, Y.; Milczek, J.; Laverde, S.; and Li, R. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*.
- Zelikman, E.; Harik, G. R.; Shao, Y.; Jayasiri, V.; Haber, N.; and Goodman, N. 2024. Quiet-star: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*.
- Zhang, Z.; Fang, M.; and Chen, L. 2024. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 6963–6975. Bangkok, Thailand: Association for Computational Linguistics.
- Zhang, Z.; Sheng, Y.; Zhou, T.; Chen, T.; Zheng, L.; Cai, R.; Song, Z.; Tian, Y.; Ré, C.; Barrett, C.; et al. 2023. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36: 34661–34710.
- Zhao, H.; Andriushchenko, M.; Croce, F.; and Flammarion, N. 2024. Is in-context learning sufficient for instruction following in LLMs? *arXiv preprint arXiv:2405.19874*.
- Zhou, C.; Neubig, G.; Gu, J.; Diab, M.; Guzman, P.; Zettlemoyer, L.; and Ghazvininejad, M. 2020. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.

Appendix A: Data Resources

A1: Dataset Descriptions

We conducted experiments on six public datasets and a real-world business benchmark (denoted as NowNewsQA), with dataset statistics summarized in Table 4. For our task, which leverages KGs to enhance answer generation, we additionally report the typical range of graph sizes and the number of entities used across different datasets.

- **Health dataset:** The Health dataset is derived from medical-related texts such as books, articles, and reports, covering various domains including health, medicine, diseases, drugs, and treatments. It serves as an out-of-domain dataset within the ULTRADOMAIN benchmark.
- **Biology dataset:** The Biology dataset is sourced from biology-related texts such as books, articles, and reports, covering various fields of biology, including cell biology, genetics, ecology, and evolutionary theory. It is an out-of-domain dataset within the ULTRADOMAIN benchmark.
- **Legal dataset:** The Legal dataset primarily consists of legal contracts, focusing on the legal domain and encompassing content such as contract clauses, legal terminology, and legal concepts. It is designed to test a model’s ability to understand and process the complex and nuanced language of legal documents and is one of the in-domain datasets within the ULTRADOMAIN benchmark.
- **HotpotQA dataset:** HotpotQA is a dataset designed to evaluate the multi-hop reasoning capabilities of question-answering systems, collected through crowdsourcing as question-answer pairs based on Wikipedia. It encompasses various types of questions and answers, such as those involving entities, dates, numbers, descriptive attributes, and “yes/no” questions comparing two entities.
- **2WikiMQA dataset:** 2WikiMQA is a large-scale, high-quality multi-hop question-answering dataset constructed from Wikipedia and Wikidata, designed to evaluate a model’s multi-hop reasoning capabilities. It includes four types of questions: Comparison questions, Inference questions, Compositional questions, and Bridge-comparison questions. The answer types are diverse, encompassing “yes/no,” dates, movies, people, and more.
- **GovReport dataset:** GovReport is a large-scale dataset comprising reports published by the U.S. Government Accountability Office and the Congressional Research Service, along with summaries written by experts, focusing primarily on the research and analysis of various policy issues. The key information in GovReport is distributed throughout the entire document, meaning that to generate accurate summaries, a model must comprehensively understand the full content of the document.
- **NowNewsQA dataset:** NowNewsQA is a real-world benchmark derived from Xinyu AI Search, featuring 3,150 multi-document QA examples in the news domain. Questions reflect real user queries on dynamic topics, with reference documents retrieved via a hybrid search engine. The dataset preserves real-world noise and redundancy, making it a challenging testbed for reasoning over fragmented, partially relevant content.

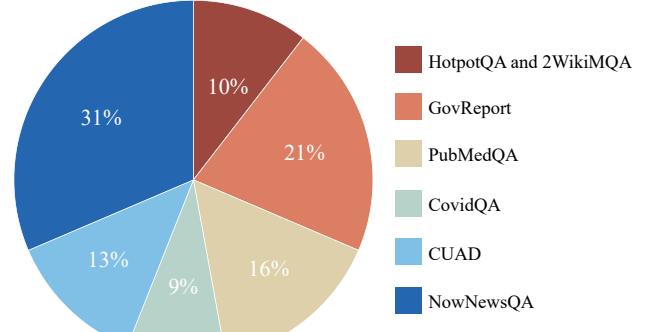


Figure 8: The distribution of SFT dataset.

A2: Corpus for SFT

We allocated the SFT dataset based on the difficulty of downstream tasks and the volume of test data, with a total data size of 9,548, as shown in Figure 8. For general domain tasks, some tasks (e.g., multi-document QA) require less entity information, resulting in a smaller data allocation, while others (e.g., summarization) require comprehensive information integration and high-quality entity lists, leading to a larger data allocation. In specialized domains (e.g., biomedical, legal, and news), open-ended tasks often require rich entity information from materials to improve answer accuracy and quality, leading to a higher proportion of SFT data in these domains. The introduction of each dataset can be found in Appendix A1 and Appendix A3 respectively.

A3: Additional Datasets for SFT

- **PubMedQA dataset:** The PubMedQA dataset originates from the PubMed database, where questions are extracted from research article titles using a rule-based heuristic approach. Each entry includes a research abstract and an automatically generated yes/no/maybe question based on the abstract’s title. This dataset is highly suitable for training and evaluating text-based QA systems, particularly for question-answering applications in the biomedical domain.
- **CovidQA dataset:** The CovidQA dataset is built upon the rich collection of research papers in the CORD-19 dataset, containing approximately 2,000 questions and tens of thousands of relevant background document passages. These questions and documents are not only extensive in number but also detailed in content, focusing closely on various COVID-19-related topics. This makes the dataset highly suitable for training and evaluating text-based QA systems.
- **CUAD dataset:** The CUAD dataset consists of contract texts extracted from publicly available legal documents in the EDGAR database, along with questions designed by domain experts based on these contracts. It includes 21,000 questions and 510 distinct legal contract texts. This dataset is an ideal choice for training and evaluating text-based QA systems, particularly for enhancing text comprehension and question-answering capabilities in the legal contract domain.

Appendix B: Baseline Descriptions

The descriptions of the baselines are detailed as follows:

Datasets	Health	Biology	Legal	HotpotQA	2WikiMQA	GovReport	NowNewsQA
Number of Samples	180	220	438	200	200	200	150
Metric	F1	F1	F1	F1	F1	ROUGE-L	Multi-Faceted
Avg. Length	135,901	125,284	51,413	9,149	4,885	8,169	10,296
Avg. Entities	44.23	42.32	15.64	10.09	8.23	18.82	24.95
Avg. Graph Size	2005.87	2043.98	1877.84	256.67	298.21	1240.87	1887.01

Table 4: Statistics of six public datasets and one real-world business benchmark.

Model	Health	Biology	Legal	HotpotQA	2WikiMQA	GovReport
Based on Mistral-7B-Instruct						
MEMORAG	37.40 ± 1.23	35.70 ± 1.57	51.20 ± 2.45	42.90 ± 1.82	30.30 ± 1.09	31.60 ± 0.87
TAdaRAG	40.77 ± 0.85	39.31 ± 1.13	49.88 ± 2.63	44.83 ± 1.55	39.31 ± 0.66	36.41 ± 0.55
Based on Qwen2.5-7B-Instruct						
MEMORAG	36.87 ± 1.85	36.00 ± 1.85	47.60 ± 4.25	37.99 ± 3.45	35.32 ± 3.68	31.13 ± 2.15
TAdaRAG	42.38 ± 1.33	40.75 ± 2.15	46.83 ± 3.38	49.23 ± 2.95	43.79 ± 2.95	36.95 ± 1.55
Based on Qwen2.5-14B-Instruct						
MEMORAG	47.37 ± 1.05	42.33 ± 2.05	55.04 ± 3.82	52.91 ± 2.05	50.42 ± 3.31	32.99 ± 3.99
TAdaRAG	51.00 ± 1.45	51.32 ± 2.09	55.83 ± 3.33	56.28 ± 2.87	54.79 ± 3.43	38.36 ± 2.73

Table 5: Statistical significance analysis across six datasets

- **NaïveRAG:** This method extracts information from text databases by breaking text into smaller segments for storage, converting them into vector representations using text embeddings. When a user queries, the question becomes a vector, and the system finds the most relevant content by comparing it with stored segment embeddings based on similarity, enabling quick, direct answers as a practical RAG solution.
- **BGE-M3:** This method employs the novel M3-Embedding model to implement standard RAG. The M3-Embedding model uses relevance scores from various retrieval functions as teacher signals, enhancing training quality through knowledge distillation to enable joint learning and mutual reinforcement of multiple retrieval capabilities. It also optimizes batch processing strategies to support large-scale training and high throughput, thereby improving the distinctiveness of the embeddings. Furthermore, it gathers a substantial and diverse set of multilingual data from unsupervised corpora, supervised corpora, and synthetic data, providing a robust foundation for model training.
- **RQ-RAG:** This method can rewrite, decompose, or disambiguate user queries based on different scenarios, thereby improving the accuracy and efficiency of retrieval. It also introduces three distinct sampling strategies—based on perplexity, confidence, and ensemble approaches—to select the optimal retrieval path and answer. As a robust retrieval-augmented generation framework, it significantly enhances the knowledge retrieval and response capabilities of LLMs.
- **GraphRAG:** This is a graph-based RAG. It uses an LLM to extract entities and relationships from the text, representing them as nodes and edges, with descriptions from the original text attached as features to reduce information loss. For each question, a community detection algorithm is applied to summarize and generalize the information contained in the nodes from the bottom up, forming new community

descriptions. Finally, the results of the community detection are used to answer global summarization questions.

- **HippoRAG:** This method adopts the hippocampal indexing theory from cognitive neuroscience and leverages KG techniques to enhance the retrieval capabilities of LLMs. It first converts a text corpus into a schemaless KG, where nodes represent concepts and edges denote semantic relations. For a given query, HippoRAG identifies key concepts and executes the Personalized PageRank algorithm over the KG to perform multi-hop reasoning in a single retrieval step.
- **MEMORAG:** This is a novel RAG method that leverages a lightweight LLM as a memory module to compress an entire database into a compact representation known as “memory”. When a task is received, the memory module generates a series of “clues” based on this memory—rough descriptions of the answer—that guide the retrieval module in locating relevant information.
- **PathRAG:** This is a novel RAG approach that retrieves relevant nodes from the index graph based on keywords in the query. It then extracts key relational paths using a flow-based pruning algorithm and converts them into textual form to guide the LLM in generating more logical and coherent responses.

Appendix C: Simplified Prompt Template

- **Simplified Extraction Prompt Template.** We have carefully crafted a detailed set of extraction prompt templates. First, we guide LLMs to effectively classify questions. This framework includes six clearly defined question types specific to professional domains, each associated with a pre-established list of highly relevant entity types to ensure that the extracted entities are closely aligned with the questions. Additionally, for each specific entity type, we have customized dedicated extraction prompt templates to improve the precision and efficiency of LLMs in identifying and ex-

Metric	TAdaRAG	PathRAG	t	p	d [95% CI]
Conciseness	8.25 ± 0.62	7.63 ± 0.73	7.93	< 0.0001	0.915 [0.466, 0.774]
Factuality	8.45 ± 0.53	7.85 ± 0.55	9.62	< 0.0001	1.111 [0.477, 0.723]

Table 6: Statistical significance analysis of TAdaRAGin business scenarios with Mistral-7B-Instruct as the base model.

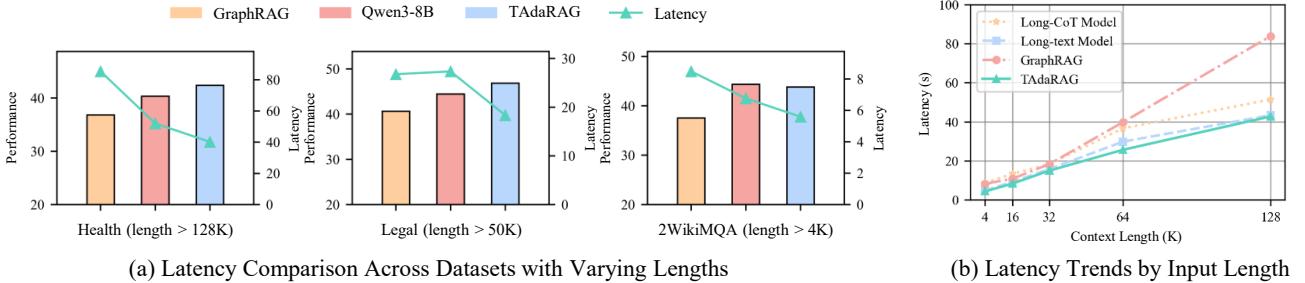


Figure 9: Efficiency analysis results.

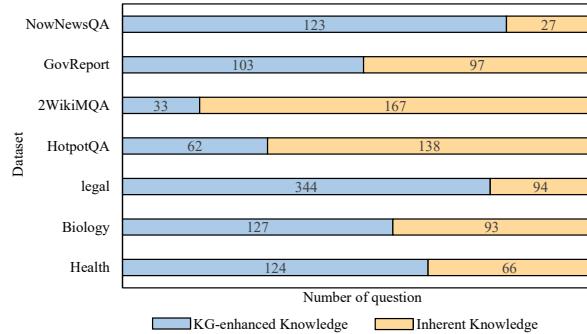


Figure 10: Proportions of TAdaRAG evidence of answers.

tracting relevant entities. For general-domain questions, we have also designed corresponding prompt templates. The simplified extraction prompt template is shown below.

- **Multi-Faceted Evaluation Criteria.** The detailed definitions of the multi-faceted evaluation criteria are provided below. The original text is in Chinese, and we translate it into English to align with the language of this paper.

Appendix D: Additional Results

D1: Statistical Significance Analysis

For Public Dataset. We performed statistical significance tests on TAdaRAG and MEMORAG across six public benchmarks (Health: $n = 220$; Biology: $n = 180$; Legal: $n = 438$; HotpotQA: $n = 200$; 2WikiMQA: $n = 200$; GovReport: $n = 200$). As shown in Table 5, take mistral-7B as backbone, for Health, Biology, 2WikiMQA, and GovReport, the lower bound of TAdaRAG’s 95% CI exceeds the upper bound of MEMORAG’s 95% CI, indicating significant improvements ($p < 0.001$). For Legal and HotpotQA, MEMORAG’s means are slightly higher but their CIs overlap, indicating non-significant differences ($p > 0.05$).

For Business Scenarios. We conducted independent t-tests on TAdaRAG and PathRAG using 300 samples per metric to assess performance differences, as shown in Ta-

ble 6. For *conciseness*: TAdaRAG ($M=8.25$, $SD=0.62$) significantly outperformed PathRAG ($M=7.63$, $SD=0.73$), $t(298) = 7.93$, $p < 0.0001$, $d = 0.916$, 95% CI = [0.667, 1.165]. For *factuality*: TAdaRAG ($M=8.45$, $SD=0.53$) also showed a significant advantage over PathRAG ($M=7.85$, $SD=0.55$), $t(298) = 9.62$, $p < 0.0001$, $d = 1.111$, 95% CI = [0.877, 1.344]. These results highlight the superior performance of TAdaRAGin business contexts, particularly in generating concise and factually accurate outputs.

D2: Efficiency Analysis

We conducted an efficiency analysis on three datasets with significant length differences, as shown in Figure 9(a). The results indicate that compared to the Long-CoT model and the classic GraphRAG method, our model, considering both dynamic graph construction time and first-token generation time, demonstrates superior latency performance. Additionally, it exhibits stable and outstanding performance across datasets of varying scales. In particular, compared to GraphRAG, which relies on pre-built index graphs and subgraph retrieval, our approach achieves significant improvements in both efficiency and effectiveness.

Furthermore, we analyzed the latency trends of different models as the input document length varies in practical tasks, shown in Figure 9(b). The experimental results demonstrate that compared with various long-context answering methods, our model consistently maintains the lowest latency across all input lengths, further validating its strong adaptability to diverse input scales and its acceptable response time cost in real-world applications.

D3: KG Optimization Results

To evaluate the effectiveness of reinforcement learning-based knowledge graph optimization (RL-based KG optimization), we analyzed the knowledge graph size (Size) and entity number (Ent.) before and after training across seven datasets, as shown in Table 7. The results indicate a significant reduction in graph size and entity count after training. For example, the Health dataset’s graph size decreased from approximately 7,300 to around 2,000, and the entity number

Stage	Health		Biology		Legal		HotpotQA	
	Size	Ent.	Size	Ent.	Size	Ent.	Size	Ent.
Base	7303.12	58.34	7864.86	64.37	7287.99	19.97	1894.77	16.09
SFT	5146.43	50.03	5538.78	55.67	5232.32	16.34	573.43	12.49
Reinforce	2005.87	44.23	2043.98	42.32	1877.84	15.64	256.67	10.09
Stage	2WikiMQA		GovReport		NowNewsQA		-	
	Size	Ent.	Size	Ent.	Size	Ent.	Size	Ent.
Base	1524.40	13.32	6684.84	48.66	4433.23	31.44	-	-
SFT	665.45	10.29	3077.34	23.73	2343.23	30.21	-	-
Reinforce	298.21	8.23	1240.87	18.82	1887.01	24.95	-	-

Table 7: Average KG graph size and number of entities across datasets and training stages (Mistral-7B).

dropped from 58 to 44, demonstrating substantial compression.

These changes suggest that the model successfully learned to control the graph size by effectively removing redundant and unnecessary information, focusing on core entities and relations, thereby improving the quality and efficiency of knowledge graph representation and reasoning. The optimization effect is especially notable in large-scale datasets such as Health and GovReport, providing more accurate and efficient knowledge support for downstream tasks.

D4: Evidence Analysis

We analyzed the answers across seven datasets to investigate the sources of evidence supporting TAdaRAGin generating answers, as illustrated in Figure 10. The results indicate that in the general domain, which includes a considerable amount of commonsense questions, a significant proportion of the answers rely exclusively on the intrinsic knowledge embedded within LLM’s parameters. However, in specialized domains such as the medical domain and legal domain, accurate answer generation requires specific domain knowledge. Consequently, the experimental results demonstrate that a substantial proportion of these queries rely on a combination of knowledge from both retrieved KG and the LLM’s inherent knowledge for answer generation.

Appendix E: Case Study

We additionally provide two case studies, as shown in Figure 11 and Figure 12, demonstrating the different responses generated by NaïveRAG, MEMORAG, and TAdaRAGto the same question using the Health and NowNewsQA datasets with Mistral-7B-Instruct as the base model. These responses were evaluated by GPT-4o across four dimensions: relevance, accuracy, hallucination-free performance, and completeness. The results indicate that the charts extracted through our method significantly enhance the quality of the responses. Compared to other approaches, our method exhibits a clear advantage in response performance.

Simplified Extraction Prompt Template

List of Problem Domains and Entity Types:

bio_medical_research: person, organism, symptom, disease, drug, technique, number, device, operation

general_knowledge: person, organism, organization, location, event, time, diet, number, product

legal_contracts: person, organization, location, event, time, number, contract, clause, judgment

customer_support: person, technique, operation, event, time, number, device, product

finance: person, organization, event, time, number, operation, product, policy

news: person, organization, location, event, time, number, product, policy

Intent Analysis: Given a question and a set of documents, classify the question into one of the predefined domains in domain by evaluating the content of the question and its relevance to the provided documents. The classification should be based on the primary focus of the question (e.g., medical, legal, financial) and the domain context found in the documents. Ensure that the answer is based on a careful assessment of the subject matter, keywords, and overall context of both the question and the document(s).

!! The output can only contain one domain of the problem, do not output the cause and other information !!

Entities Extraction: Given a question, documents, and a list of entity types {entity_type}, the task is to extract each type of entity step by step, and finally compile a complete list of entities. The extraction process for each entity type must strictly follow the corresponding extraction rules to identify all relevant entities in the documents that will significantly help answer the question. The output format must strictly follow the example, without any additional text or output in other formats like json. Additionally, do not output the details of entity extraction for each step, only the final list of entities.

Person Entity Extraction: Identify all person entities relevant to answering the question. For each identified person entity, extract the following information:

- Entity Name: The name of the entity.
- Entity Type: person.
- Entity Description: A summary of the information related to the specific person in the context of the question. Based on the information, provide the most relevant and helpful description for answering the question. Optionally, include other relevant information such as the person's identity, position, major life events, significant achievements or awards, involvement in important historical events, published works, contributions, and relationships.
- Each entity's format should be ("Entity", <Entity Name>, <Entity Type>, <Entity Description>).

Organization Entity Extraction: Identify all organizations relevant to answering the question. Note that the identified organizations should not include person entities! For each identified organization, extract the following information:

- Entity Name: The name of the entity.
- Entity Type: organization.
- Entity Description: A summary of the information related to the specific organization in the context of the question. Based on the information, provide the most relevant and helpful description for answering the question. For example, the extracted entity description may include the organization's name, function, goals, leadership, historical background, culture, and its role or influence in the related events. For corporate organizations, in addition to extracting the company name and function, also focus on financial data, market share, annual revenue, number of employees, and other numerical information.
- Each entity's format should be ("Entity", <Entity Name>, <Entity Type>, <Entity Description>).

.....

Multi-Faceted Evaluation Criteria

"Relevance": "The response should directly address the user's question, avoiding irrelevant content, unnecessary elaboration, or roundabout explanations. Each irrelevant statement deducts one point.",

"Numerical Precision": "For questions asking for specific numbers, avoid vague responses like 'several' or 'many.' Answers should be precise and specific. Each ambiguous response deducts one point.",

"Conciseness": "If the user specifies a time period, location, person, or event, the answer must meet these requirements. Each statement that fails to address the question properly deducts one point.",

"Factuality": "Information in the response must be accurate, especially for factual questions. Avoid incorrect numerical data or factual errors. Each numerical or factual mistake deducts one point.",

"Timeliness": "For ongoing news events or urgent reports, the provided information should reflect the latest updates. Note that today's date is xxxx-xx-xx. If the question is not time-sensitive, no points will be deducted. For time-sensitive questions, evaluate the response's timeliness; the less timely it is, the more points should be deducted accordingly.",

"Comprehensiveness": "The response should comprehensively cover all aspects of the user's inquiry, providing detailed information. The answer should be self-contained, requiring no further search to understand the full context. Each missing essential point deducts one point.",

"Clarity": "The response should be clear and easy to understand, with a well-structured format for readability. It should quickly convey the necessary information. (Example of unclear expression: Using a parallel or topic-based structure instead of a chronological one when a time-sequence approach is clearly needed.) Each unclear expression deducts one point.",

"Coherence": "The response should be logically coherent, with smooth transitions between sentences and appropriate word choices. Each instance of incoherence deducts one point.",

"Insightfulness": "The response should provide unique insights and depth. The base score for this criterion is 6 points, with each innovative viewpoint or expression earning an additional 0.5–1 point."

Prompt for multi-faceted evaluation

Assume you are an article quality inspector. Please evaluate the response based on { Metric Title }. I will provide the user's question and the final response. The maximum score is 10 points, and the scoring rules are as follows:

{ Metric Definition }

Please strictly follow the scoring rules. Example output format:

```
{  
  "Issues Identified": "X",  
  "Calculation Process": "10 -1.0 -1.0 -1.0 = 7.0",  
  "Score": 7  
}
```

{ Few - Shot Examples }

Your final score:

""

Question: What role did the Americans with Disabilities Act (ADA) play in the lives of patients with chronic illness?

NaïveRAG response

The article does not mention the Americans with Disabilities Act (ADA) in relation to the lives of patients with chronic illness.

MEMORAG response

The Americans with Disabilities Act (ADA) played a crucial role in advancing the rights of patients with chronic illness by ensuring equal access to health care services, employment opportunities, and public accommodations, thereby improving the quality of life for many individuals living with chronic conditions.

TAdaRAG response

<|startextraction|>

("Entity", "1990", "date and time", "The year the Americans with Disabilities Act was enacted.");
("Entity", "Americans with Disabilities Act (ADA)", "event", "The ADA prohibits discrimination against individuals with disabilities in all areas of public life, including");
.....

<|endextraction|>

The Americans with Disabilities Act (ADA), enacted in 1990, prohibits discrimination against individuals with disabilities in all areas of public life, including jobs, schools, transportation, and all public and private places that are open to the general public. It played a role in improving the care and rights of patients with chronic illness by ensuring equal access and opportunities for them.

LLM Decision

Relevance: Answer 1 is off-topic, failing to answer the question. In contrast, Answer 2 and Answer 3 are highly relevant, summarizing ADA's protection of patients' rights (Answer 2) and its anti-discrimination role in improving care (Answer 3).

Accuracy: Answer 1 lacks substance. Answer 2 correctly describes the ADA's core function, while Answer 3 provides more details, including the enactment year and broader coverage.

Hallucination-Free: Answer 1 is irrelevant. Answer 2 and Answer 3 are fact-based.

Completeness: Answer 1 provides no useful information. Answer 2 offers a general overview but lacks details, while Answer 3 is more comprehensive, covering background and broader protections.

Figure 11: The figure displays the answers provided by different methods to the same question in the Health dataset, along with the LLM's analysis of these answers, visually demonstrating the differences in answer quality across the various methods.

Question: What are the reports in China in 2024 regarding the refutation of the "China overcapacity theory" from the United States and Western countries?

NaïveRAG response

In 2024, China's reports refuting the "China overcapacity theory" from the United States and Western countries mainly cover the following aspects:

1. New Energy Vehicle Industry [...] 2. Capacity Utilization Rate [...] 3. Government Subsidies [...] 4. International Cooperation [...] 5. Green Development [...] 6. Economic Analysis [...] 7. Media and Public Opinion [...]

Conclusion: Through the multifaceted reports outlined above, Chinese media organizations have comprehensively rebutted the U.S. and Western claims of "China overcapacity." They showcase the real development status and market demand in China's NEV industry and other sectors. Backed by authoritative data and expert analyses, these reports effectively counter the accusations from the U.S. and Western countries, safeguarding China's economic image and development rights.

MEMORAG response

In 2024, facing the "China overcapacity theory" put forward by the U.S. and Western countries, China mounted a robust counterattack through media reports and data analysis. These reports are grounded in facts and logically sound. Below is a summary of their core content:

1. Industrial Competitiveness [...] 2. Capacity Utilization Rate [...] 3. Government Subsidies [...] 4. International Cooperation [...] 5. Green Development [...] 6. Economic Analysis [...] 7. Media and Public Opinion [...]

Conclusion: Backed by data and facts, these 2024 reports expose the protectionist nature of the "China overcapacity theory." They not only safeguard China's economic image but also underscore its critical role in global green development and international cooperation.

TAdaRAG response

<|startextraction|>

("Entity", "45 million", "number", "The International Energy Agency predicts that global demand for new energy vehicles will reach 45 million in 2030."); ("Entity", "1.203 million", "number", "Data from the China Association of Automobile Manufacturers show that China will export 1.203 million new energy vehicles in 2023."); ("Entity", "Xing Ziqiang", "person", "Morgan Stanley China Chief Economist Xing Ziqiang questions the theory of 'overcapacity'."); ("Entity", "Nicholas Lardy", "person", "Nicholas Lardy, a senior fellow at the Peterson Institute for International Economics, a US think tank, questions the theory of 'overcapacity'.");

<|endextraction|>

In 2024, China's reports refuting the "China overcapacity theory" from the U.S. and Western countries primarily cover the following aspects:

1. Industrial Logic and Market Reality: Price Adjustments and Market "Clearing": [...]

2. International Data and Market Demand: Global Market Demand: The International Energy Agency estimates that global demand for new energy vehicles (NEVs) will reach 45 million units by 2030, more than four times the figure in 2022. In 2023, China exported 1.203 million NEVs, indicating a substantial latent demand for new energy products in the global market. Subsidies in Other Countries: [...]

3. Political Motives and Protectionism: Political Tool: [...] Double Standards: [...]

4. Expert Opinions and Market Ecosystem: Expert Views: Multiple economists and executives of multinational corporations believe that China holds a significant position in the global supply chain, a result of autonomous decisions by global businesses and consumers. The so-called "overcapacity" in China's new energy vehicle (NEV) sector is, to a large extent, a "numbers game."

5. Capacity Planning and Management in Practice: Capacity Planning: [...] Dynamic Balance: [...]

6. Global Cooperation and Market Recognition: International Cooperation: [...] Market Recognition: [...]

7. Ministry of Commerce Rebuttal: Ministry of Commerce Rebuttal: [...]

Through these multifaceted reports and analyses, China in 2024 has effectively rebutted the accusations of "overcapacity" from the U.S. and Western countries, showcasing the real situation and developmental achievements of its NEV industry.

LLM Decision

Relevance: All three answers are highly relevant to the question, directly addressing how China has refuted Western claims of "overcapacity" in 2024. Answer 1 and Answer 2 primarily focus on the economic and industrial aspects, while Answer 3 provides a broader perspective, including political motivations and government statements.

Accuracy: Answer 1 and Answer 2 correctly present factual arguments, such as China's high capacity utilization rates and international cooperation. Answer 3 expands on this by incorporating official responses from China's Ministry of Commerce and expert opinions, making it the most well-rounded and accurate. None of the answers appear to contain major inaccuracies.

Hallucination-Free: All three answers rely on logical arguments and data-driven reasoning, avoiding fabricated information. Answer 3, while more detailed, does not introduce any unverified claims, ensuring it remains fact-based.

Completeness: Answer 1 and Answer 2 cover core arguments such as government subsidies, market demand, and global cooperation. However, Answer 3 surpasses them in completeness by also discussing political motivations, international trade policies, expert insights, and direct rebuttals from Chinese government agencies. Thus, Answer 3 provides the most comprehensive response.

Figure 12: The figure displays the answers provided by different methods to the same question in the NowNewsQA dataset, along with the LLM's analysis of these answers, visually demonstrating the differences in answer quality across the various methods.