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

Jie Zhang[1*], Bo Tang[2,3*], Wanzi Shao[1], Wenqiang Wei[3], Jihao Zhao[3,4], Jianqing Zhu[5], Zhiyu Li[3], Wen Xi[5], Zehao Lin[3], Feiyu Xiong[3], Yanchao Tan[1†]

[1]College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China
[2]AIDS and SIAR, University of Science and Technology of China, Suzhou 215123, China
[3]MemTensor (Shanghai) Technology Co., Ltd., Shanghai 201306, China
[4]School of Information, Renmin University of China, Beijing 100086, China
[5]China 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) (Achiam et al. 2023; Guo et al. 2025; Yang et al. 2025) have achieved significant breakthroughs in natural language processing, particularly in tasks such as text generation and question-answering systems (Yang et al. 2024b; Laban et al. 2024). However, relying solely on internal parametric knowledge often leads LLMs to 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 this issue and enhance reliability, 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.

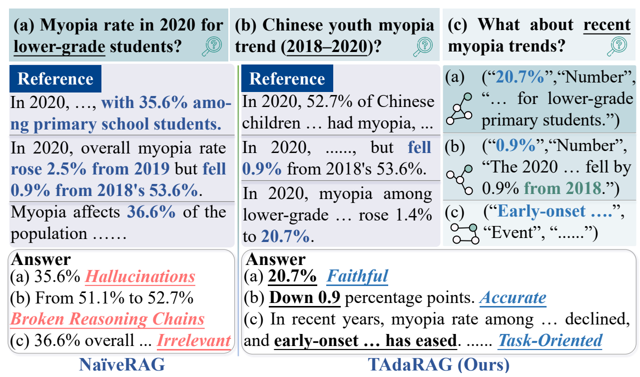

Figure 1: An illustrative example where the RAG system fails to generate correct responses due to truncation, leading to hallucinations, broken reasoning and irrelevant details. Our proposed TAdaRAG addresses these issues by integrating the task-adaptive knowledge graph dynamically.

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 enhances language models by retrieving information from external knowledge bases to improve text accuracy and credibility (Borgeaud et al. 2022; Guu et al. 2020). 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. 2024b; Zhang, Fang, and Chen 2024). Recent advancements also explore adaptive retrieval strategies to align retrieved content with query intent (Mo et al. 2024). However, existing methods process input text coarsely, introducing irrelevant information and omitting key details, which limits 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 (Achiam et al. 2023; 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 capabilities, laying a crucial foundation for on-the-fly construction of high-quality KGs.

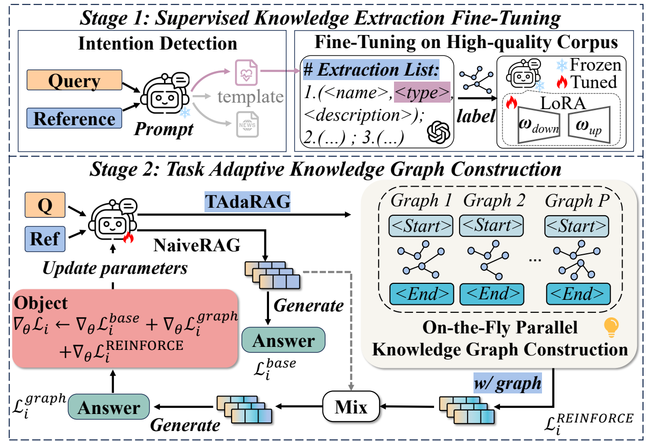

Figure 2: An illustration of our proposed TAdaRAG framework and its two-stage training: (1) Supervised Knowledge Extraction Fine-tuning and (2) Task-Adaptive Knowledge Graph Construction.

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 $g_i = \{g_i^1, g_i^2, \ldots, g_i^p\}$. Learnable tokens <|startextraction|> and <|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 concat$(\cdot)$ denotes the concatenation.

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

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

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

where $f(\cdot) = \text{softmax}(\text{lmhead}(\cdot))$, $\text{lmhead}(\cdot)$ 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 $\pi_\theta(y_i \mid x_i, \tilde{g}^{(i)})$.

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:

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

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:

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

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:

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

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 (Yang et al. 2024a) 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**	<u>42.90</u>	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***	<u>49.88</u>	**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	20.10
Advanced RAG	RQ-RAG	31.50	31.90	38.80	37.40	34.10	21.00
	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***	<u>46.83</u>	**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 → 39.31) and multi-hop question answering (HotpotQA: 42.90 → 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 → 36.41), indicating improved scalability and precision in integrating task-oriented knowledge for long-text summarization.

Long-Context Task Analysis (RQ2) We further compared `TAdaRAG` with three long-context mechanisms: 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. On 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. Establishing a gold-standard answer for RAG tasks is difficult, which becomes especially challenging in real-world industry scenarios where responses can vary widely and exhibit subjectivity. Thus, 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

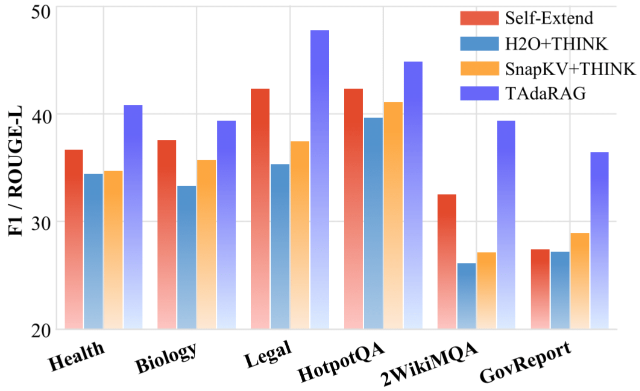

Figure 3: Long-context experiments on Mistral-7B.

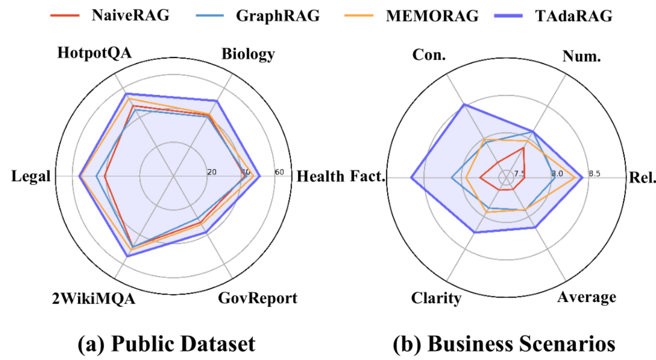

Figure 4: Experiments on Qwen2.5-14B.

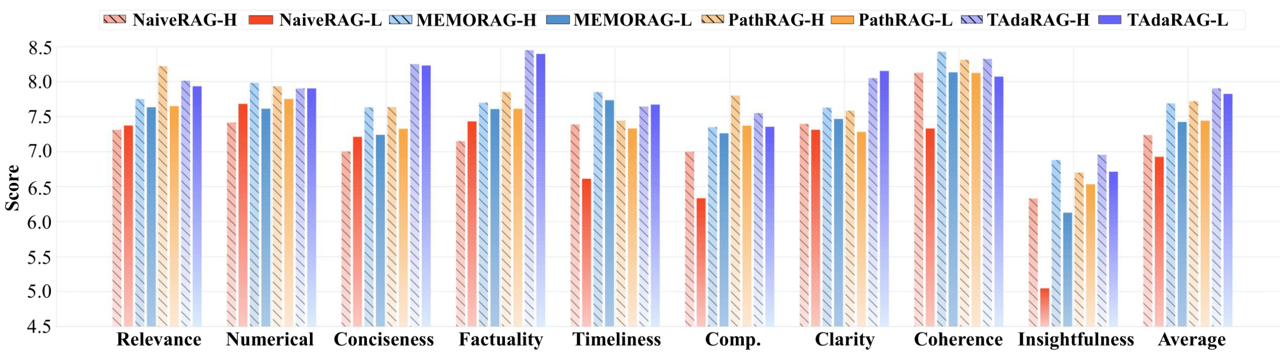

Figure 5: Multi-faceted comparison of different methods based on human (H) and GPT-4o (L). Higher values indicate better performance, with a maximum value of 10.

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) We present the comprehensive evaluation and analysis as follows.

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

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	0.680	0.694

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

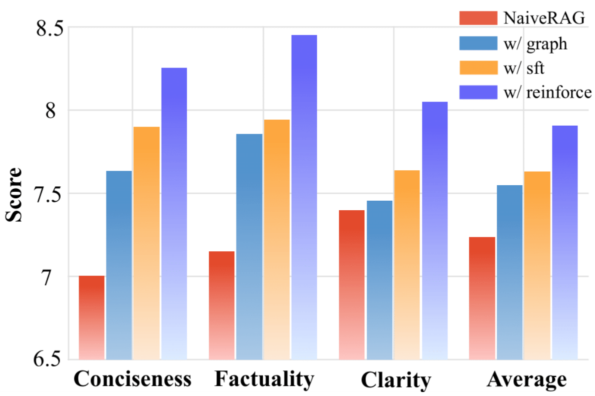

Figure 6: Ablation study on NowNewsQA.

further suggest that human ratings are stable and reliable, 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.

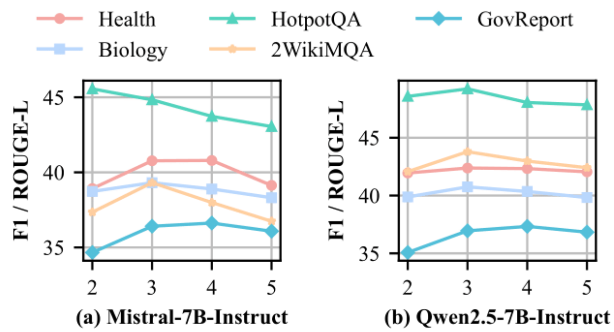

Figure 7: Parallel Subgraphs vs. Performance

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

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report.

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 Srikumar, 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.

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.; et al. 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.; et al. 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. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

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, A.; et al. 2024a. Qwen2.5 technical report. arXiv:2412.15115.

Yang, A.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, X.; Sun, K.; Xin, H.; Sun, Y.; Bhalla, N.; Chen, X.; Choudhary, S.; Gui, R.; Jiang, Z.; Jiang, Z.; et al. 2024b. 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*.