Task-Classifying Retrieval Augmented Generation

Jingyuan Ju
School of Information Science and Technology
Beijing University of Chemical Technology
Beijing, China
15098126414@163.com

Peng Zang
Zouping Huimao New Material Technology Co., LTD.
Zouping, China
zangpeng@wqmail.cn

Hao Zhong
Binzhou Institute of Technology Weiqiao-UCAS Science and Technology Park
Binzhou, China
zhonghao@wqucas.com

Lingfeng Wang*
School of Information Science and Technology
Beijing University of Chemical Technology
Beijing, China
lfwang@buct.edu.cn
*Corresponding author

2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 979-8-3315-1936-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICAIBD64986.2025.11082008

***Abstract*—Hallucination phenomena are inevitably present in large language models (LLMs) for generation tasks. Retrieval Augmented Generation (RAG) technology alleviates this issue to some extent by incorporating external retrieved documents. However, existing Retrieval-Augmented Generation (RAG) methods still face limitations in terms of task adaptability and the richness of retrieval inputs. To address these issues, this paper proposes an innovative RAG framework that enhances both task classification and high-quality document retrieval, aiming to improve the accuracy and robustness of generated results.Specifically, we design a task identification module capable of automatically categorizing user inputs into three types of tasks: information retrieval, text generation, and general question answering. This classification then guides the subsequent processing flow. At the retrieval stage, we introduce a HyDE-inspired hypothetical document generation strategy, which constructs task-relevant hypothetical answers to improve the alignment between retrieved documents and user intent. To improve the quality of generation, this article introduces a task adaptive reflection and generation mechanism, implementing differentiated reflection processes and generation strategies for different task types, thereby improving the accuracy and completeness of the model's response in multi task scenarios. Experimental results demonstrate that the proposed approach achieves state-of-the-art performance across multiple natural language processing tasks, particularly excelling in complex question answering and open-domain information retrieval scenarios.**

Keywords—Retrieval-Augmented Generation (RAG), HyDE-based Retrieval, Reflection, Task - classifying framework

I. INTRODUCTION

Large-language models (LLMs) have demonstrated powerful language understanding and generation capabilities in the field of natural language processing, being able to generate fluent and context-relevant text [1-3]. The excellent performance of LLMs in a wide range of tasks is mainly attributed to their training on large-scale text data, which enables them to capture rich contextual semantics. However, LLMs inevitably face the problem of hallucinations, especially when generating factual information. The generated content often contains fictional or incorrect factual information [4]. This is particularly prominent in scenarios that require handling the latest information or domain - specific knowledge [5][6][7]. The weak dependence of the model on external information, coupled with the incompleteness of its internal representation, makes solving the hallucination problem a major challenge in building reliable generation systems.

Retrieval Augmented Generation (RAG) effectively alleviates the above-mentioned problems by introducing external knowledge. Its core idea is to input the retrieved relevant documents as reference information to the model, thereby enhancing factual accuracy during the generation process and reducing the occurrence of hallucinations [8]. In the RAG framework, the model input is supplemented with documents retrieved from an external knowledge base, which provide contextual support for the generation results [9].

Although RAG technology has shown significant potential, its performance highly depends on the accuracy and relevance of the retrieved documents [10]. If the retrieved documents contain irrelevant or misleading information, it will not only affect the quality of the generation results but may also exacerbate the hallucination phenomenon [11]. In response to this problem in the RAG framework, some emerging studies have attempted to introduce reflection mechanisms to improve the evaluation and utilization of retrieved documents. For example, Self-RAG proposed a RAG framework with a reflection mechanism. By using self - reflection to adjust the quality of the retrieved content, it further reduced hallucinations and improved the flexibility and robustness of the model in multi-task scenarios [12]. CRAG (Corrective Retrieval Augmented Generation) proposed a self-correction retrieval mechanism. By introducing a lightweight evaluator, it conducted a comprehensive evaluation of the quality of the retrieved documents and triggered different retrieval actions based on the evaluation results [13].

Although recent advances have improved RAG systems to some extent, they generally lack task sensitivity and tend to apply uniform evaluation and generation procedures across different task types. This one-size-fits-all strategy limits the effective use of retrieved information, especially in the context of diverse task objectives, where it may lead to information redundancy or contextual misalignment, ultimately compromising generation quality. To address these limitations, this paper proposes an adaptive Retrieval-Augmented Generation method that enhances two key components—task identification and high-quality retrieval input—to improve flexibility and accuracy across multi-task scenarios. The main contributions are as follows:

- HyDE-based Hypothetical Document Generation: Before retrieval, the system constructs hypothetical answers related to the user query and concatenates them with the original input. This guides the retrieval system toward a more relevant semantic space, improving the

This work was supported by the National Natural Science Foundation of China under Grant NO.62376017, N0.62272009, and Fundamental Research Funds for the Central Universities (buctrc202221).

quality and alignment of candidate documents from the outset.

- Task Classification Module: We design a task classifier based on large language models to accurately categorize user inputs into three task types: information retrieval, text generation, and general question answering. Each task type corresponds to specific document utilization strategies and reflection mechanisms, enabling task-aware, dynamically adjusted processing pathways.
- Adaptive Reflection Mechanism: Document utilization and generation strategies are dynamically adapted based on task type. For text generation tasks, evaluation indicators such as knowledge coverage and generation diversity are incorporated during the generation phase to enhance output quality. For information retrieval tasks, document assessment and optimization are completed during the retrieval stage, followed by direct response generation. In contrast, general question answering tasks bypass retrieval entirely and proceed directly to generation using the language model.

Experimental results across multiple natural language processing tasks demonstrate the superiority of the proposed RAG framework. By introducing a task-aware classification mechanism and stage-specific adaptive strategies, this study provides both theoretical foundation and practical guidance for developing more intelligent and reliable next-generation RAG systems.

II. Relevant Technology

A. Large Language Models (LLMs)

In recent years, large-language models (LLMs) have achieved breakthrough progress in natural language processing (NLP) tasks. LLMs can learn the structure and semantics of language by pretraining on a vast amount of text data, thus demonstrating strong language generation and understanding capabilities in various tasks [1-2]. Recent research advances have extended their capabilities to multimodal tasks, such as text-to-image and text-to-video generation [14]. This enables the generation and editing of images and videos based on detailed prompts [15], significantly broadening the application scope of generative artificial intelligence.

B. Retrieval Augmented Generation (RAG)

To alleviate the hallucination problem in LLMs, researchers proposed the Retrieval Augmented Generation (RAG) framework [8-9]. The core idea of RAG is to retrieve relevant documents from an external knowledge base and input them as auxiliary information into the generation model, thereby enhancing factual accuracy during the generation process. In RAG, the generation model relies not only on its own parameters but also on the retrieved documents, which provide more contextual information for the generation task. This method has achieved remarkable results in knowledge - intensive tasks, especially in complex question - answering and long - text generation tasks [10]. However, the performance of the RAG framework is highly dependent on the quality and relevance of the retrieved documents. If the retrieved documents contain errors or irrelevant information, it may affect the quality of the generated content and even trigger hallucinations [11]. To address this issue, Self - RAG introduced a self - reflection mechanism that can adaptively adjust the generation process based on the quality of the retrieved documents, thereby reducing hallucinations [12]. CRAG (Corrective Retrieval - Augmented Generation) proposed a correction mechanism, designing a lightweight evaluator to assess the quality of retrieved documents and trigger different retrieval strategies based on the evaluation results [13]. Fig. 1 presents these two typical RAG frameworks.

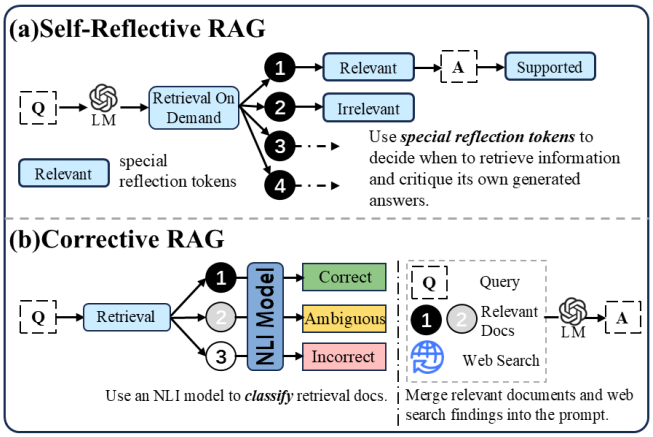

Fig. 1. Explanation of Two Typical RAG Frameworks with Reflection MechanismsGiven a knowledge - intensive query Q: (a) Self - Reflective RAG requires specialized instruction tuning of the general language model (LM) to generate specific tags for self - reflection; (b) Corrective RAG employs an external retrieval evaluator to enhance document qualityt

III. Building A Task-Classifying Retrieval Augmented Generation Framework

Our method is based on CRAG [13] as the foundation. To highlight the innovation of this article, we have made incremental improvements on the basic diagram of the CRAG framework to visually demonstrate the structural improvement of our method compared to CRAG.The task - classifying RAG framework proposed in this paper mainly comprises three parts, including the Hyde - based hypothesis document generation technique (Section A), the task classification module based on large models (Section B), and the adaptive reflection mechanism (Section C). Figure 2 provide an overview of the proposed method.

A. Task - Classification Module

Most existing Retrieval-Augmented Generation (RAG) methods adopt a uniform evaluation and generation process, overlooking the differences in generation goals and evaluation requirements across various task types. This one-size-fits-all approach often leads to inefficient use of retrieved documents in multi-task scenarios, thereby affecting the relevance and quality of the generated content. To address this limitation, this paper introduces a task classification module, which identifies the task type associated with user input before the generation process begins. This prior knowledge guides the selection of retrieval strategies, the adaptation of reflection mechanisms, and the customization of generation strategies, thereby enhancing the overall flexibility and task adaptability of the RAG system.

The core function of this module is to automatically identify the task type implied by the input and accordingly match it with a corresponding processing path. The task types are categorized as follows:

- Information Retrieval Task: The user queries for specific information, and the system needs to retrieve relevant data from databases, literature, or knowledge bases.
- Text Generation Task: The user requests the generation of a certain form of text, such as an article or report.
- General question Answering Task: Question Answering: When a user asks a specific question, the system needs to generate accurate and concise answers. Only the knowledge of the model itself is needed to answer, without searching.

This module adopts a few - shot learning method based on GPT - 4, achieving efficient classification through structured Prompt engineering. The Prompt template includes (1) a system instruction layer: clarifying the classification task objective; (2) an example demonstration layer: providing typical samples of each task category; (3) a format constraint layer: specifying the JSON output format.

In preliminary experiments, this approach demonstrated a classification accuracy of 92.3% (n=1000), an increase of 17.2% compared to traditional machine learning methods, and the inference latency was controlled within 300ms.

B. Hypothesis Document Generation Technique

In natural language processing, user queries are often unstructured and colloquial, whereas documents in the corpus tend to follow formal and domain-specific writing conventions. This stylistic discrepancy can hinder effective retrieval based on direct vector similarity. To mitigate this issue, we propose a hypothesis-enhanced retrieval strategy, inspired by the Hypothetical Document Embedding(HyDE) [16], which leverages large language models (LLMs) to bridge the semantic gap between user queries and document representations.

In our approach, instead of directly using the user query for retrieval, we first generate a hypothetical answer based on the query using an LLM. This hypothetical document (hypo_doc) is then split into multiple semantically coherent chunks, and each chunk is individually concatenated with the original user query to form a set of enriched retrieval prompts. These composite inputs are then embedded and used to perform vector-based retrieval against the document corpus.

To further refine the retrieval quality, we aggregate all retrieved results, followed by duplicate removal and semantic re-ranking based on relevance scores. This process ensures that the final set of candidate documents is both diverse and highly relevant to the user's intent.

For example, given a query such as "How to improve writing skills?", the system first generates a hypothesis like: "Read a wide variety of texts including novels, essays, poems, and academic articles to expose yourself to different writing styles and vocabulary." This hypothesis is split into several logical segments, each of which is paired with the original query and used as a retrieval input, helping the retriever to better target relevant documents in the knowledge base.

By integrating hypothesis generation, query fusion, and post-retrieval refinement, this method significantly improves the semantic alignment between user intent and retrieved content, thereby enhancing the overall effectiveness of the RAG pipeline.

C. Adaptive Reflection Mechanism

To address the differentiated quality requirements of retrieval results across various task types, we propose an adaptive reflection mechanism, which introduces task-aware evaluation logic into both the retrieval and generation stages. This mechanism selectively triggers reflection based on task classification outcomes, ensuring that knowledge-intensive tasks benefit from quality control while avoiding redundant processing for straightforward tasks.

1) Task-Aware Reflection Design

Upon completion of task classification, the reflection mechanism is conditionally activated based on task type:

- Question Answering: This task type typically requires concise, fact-based responses and often does not benefit from external retrieval. Therefore, no document retrieval or reflection is performed. The model directly generates answers based on the input query.
- Information Retrieval: For retrieval-focused tasks, the system performs document retrieval followed by retrieval-stage reflection only, using a retrieval quality assessor. The generation stage does not involve additional reflection or restructuring, as the output is primarily document-based.
- Text Generation: This task type benefits most from layered quality control. It involves both retrieval-stage and generation-stage reflection. In the generation stage, special attention is given to knowledge coverage (i.e., how comprehensively the response reflects the retrieved knowledge) and diversity (i.e., avoiding repetition or stylistic monotony).

2) Retrieval-Stage Reflection: CoT-Based Document Evaluator

For knowledge intensive tasks (information retrieval, text generation) other than general Q&A, this paper introduces the Chain of Thought (CoT) reasoning mechanism based on the big language model in the retrieval phase, and constructs a retrieval quality evaluator to identify whether the current retrieval results have sufficient information support. The evaluator simulates manual judgment of evidence relevance through multi-step reasoning process to evaluate whether the document has the key content required to answer questions.

Its core processes include: first, semantic analysis of user queries to clarify information requirements; Then analyze the content of the search results sentence by sentence to identify whether it contains evidence information that directly or indirectly supports the answer; Finally, the judgment of "sufficient/partially sufficient/insufficient" is made according to the type and coverage of evidence as the basis for subsequent reflection and supplementary inspection. This mechanism can

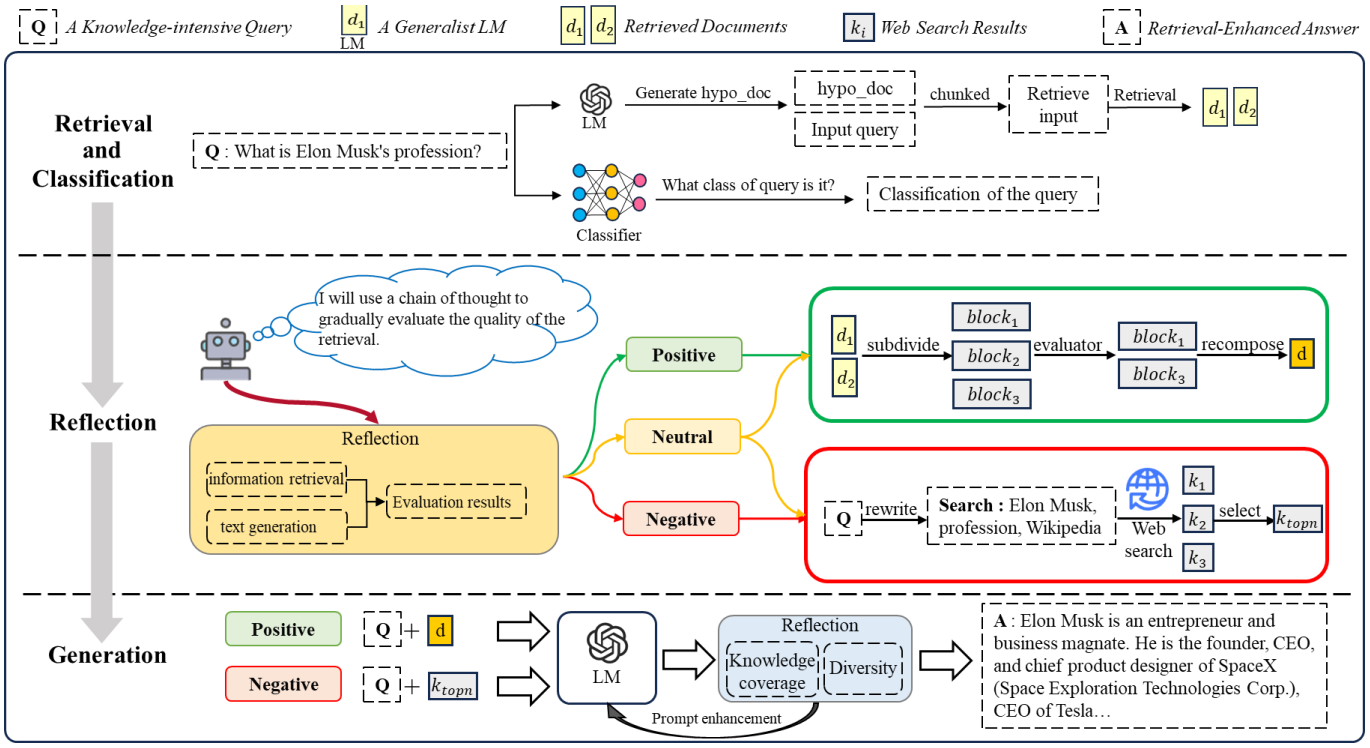

Fig. 2. Overview of the Task - Classifying RAG Framework :A Hyde - based hypothesis document generation technique is introduced to enhance the quality of retrieved documents. A task classifier is constructed to categorize user inputs into {information retrieval, text generation, question answering} tasks. An adaptive reflection mechanism is designed to dynamically evaluate the retrieved documents and choose different reflection strategies based on different tasks.

effectively filter information noise, fill knowledge gaps, and improve the accuracy and coverage of generated content.

Its core processes include: first, semantic analysis of user queries to clarify information requirements; Then analyze the content of the search results sentence by sentence to identify whether it contains evidence information that directly or indirectly supports the answer; Finally, the judgment of "sufficient/partially sufficient/insufficient" is made according to the type and coverage of evidence as the basis for subsequent reflection and supplementary inspection. This mechanism can effectively filter information noise, fill knowledge gaps, and improve the accuracy and coverage of generated content.

3) Generation-Stage Reflection: Task-Specific Strategies

In the generation stage, task-specific reflection strategies are employed to ensure output quality and task alignment:

For information retrieval tasks, which typically focus on extracting factual content, no additional generation-stage reflection is required once retrieval sufficiency is verified. The system proceeds directly with generation.

For text generation tasks, two additional reflection dimensions are introduced: knowledge coverage and diversity. Knowledge coverage evaluates whether the generated content sufficiently integrates key information points from the retrieved documents, while diversity ensures that the text avoids repetitive phrasing and maintains syntactic and stylistic variation. If either metric falls below a predefined threshold, the system triggers regeneration and targeted augmentation.

The calculation method of knowledge coverage is as follows: Let $K_D = \{k_1, k_2, \ldots, k_n\}$ be the knowledge points extracted from the retrieved documents D， $F_A = \{f_1, f_2, \ldots, f_m\}$ from the generated answer A。 Each knowledge point is encoded using BGE-m3, and semantic similarity is computed via cosine similarity. Define the match function as:

$$match(k_i, F_A) = \begin{cases} 1 & if\ \exists f_j \in F_A, sim(k_i, f_j) \geq \theta_{sim} \\ 0 & otherwise \end{cases} \quad (1)$$

Where $sim(\cdot)$ represents cosine similarityThen, the knowledge coverage score is:

$$Coverage(A) = \frac{1}{|K_D|} \sum_{k_i \in K_D} match(k_i, F_A) \quad (2)$$

We use distinct-2 to calculate diversity, It measures diversity by calculating the number of unique bigrams in the total bigrams. The formula is as follows:

$$Distinct\text{-}2 = \frac{unique_{bigrams}}{total_{bigrams}} \quad (3)$$

Where $unique_{bigrams}$ represents the number of unique bigrams in the generated text, and $total_{bigrams}$ represents the total number of bigrams in the generated text.

For general question answering tasks, which often do not require external knowledge, the system skips both retrieval and reflection steps and directly generates the answer based on the original user query.

These strategies ensure that the system adapts appropriately to task-specific requirements, enhancing both generation quality and efficiency.

IV. EXPERIMENTS

A. Experimental Setup

To verify the effectiveness of the proposed task classification module and adaptive reflection mechanism in enhancing the performance of RAG systems, we conducted experiments on four public natural language datasets. To ensure comparability between our method and CRAG and Self - RAG in the experiments, the large - model used for generation was the SelfRAG - LLaMA2 - 7b finetuned by Self - RAG, which was used to generate answers.

In the experiments, we used four datasets to evaluate the performance of the proposed method, as follows:

- PopQA [17]: Used for short - text generation tasks, where each question typically requires answering a factual knowledge about a single entity. Accuracy was used as the evaluation metric.
- Biography [18]: Used for long - text generation tasks, requiring the generation of a detailed biography about an entity. FactScore (Min et al., 2023) was used to evaluate the generated biographies.
- PubHealth [19]: Used for true/false question tasks, where the model needs to verify the truthfulness of health - related statements. Accuracy was used as the evaluation metric.
- Arc-Challenge [20]: Used for multiple - choice question tasks, where the model needs to select the correct description from 3 or 4 optional answers. Accuracy was used as the evaluation metric.

For the PopQA, PubHealth, and Arc - Challenge datasets, we followed the evaluation metrics used in previous studies, namely accuracy, to measure the model's performance on these tasks. Accuracy can intuitively reflect the model's ability to give correct answers in true/false and multiple - choice question tasks, as well as its ability to generate content consistent with the reference answers in short - text generation t asks.

For the Biography dataset, we used FactScore [21] as the evaluation metric. FactScore is mainly used to assess the factual accuracy of generated content in long - text generation tasks, calculating the score by comparing the generated text with real facts. This metric can more comprehensively assess the model's ability to grasp and present factual information in long - text generation tasks.

B. Comparative Experiments

We compared the method proposed in this paper with multiple baselines, including standard RAG [8], Self - RAG [12], and CRAG [13]. The experimental results are shown in TABLE I.

TABLE I. COMPARATIVE EXPERIMENTAL RESULTS

Method	PopQA	Bio	Pub	ARC
RAG	40.3	59.2	39.0	46.7
Self-RAG	54.9	81.2	72.4	**67.3**
CRAG	59.3	74.1	75.6	54.8
Ours	**62.5**	**82.3**	**75.9**	60.3

First, compared with the baseline methods, our proposed method achieved the best results on three datasets. Specifically, our method achieved an accuracy of 62.5% on the PopQA dataset, 82.3% on the Biography dataset, 75.9% on the PubHealth dataset, and 60.3% on the Arc-Challenge dataset. In contrast, the current state-of-the-art CRAG achieved an accuracy of 59.3%, 74.1%, 75.6%, and 54.8% on these four datasets, respectively. This indicates that our method demonstrates advanced performance across multiple datasets and shows significant improvement in various types of natural language processing tasks.

Second, our method also shows consistent improvements over the current state-of-the-art method CRAG across different task types. On the short-text generation task (PopQA), our method outperformed CRAG by 3.2 percentage points. For the long-text generation task (Biography), it achieved an improvement of 8.2 percentage points. On the closed-set tasks, our method gained 0.3 percentage points on PubHealth and 5.5 percentage points on ARC-Challenge. These results demonstrate the robustness and cross-task generalization capability of our method, especially in handling more complex or diverse input conditions.

C. Ablation Study

To further verify the effectiveness of each component in our method, we conducted ablation studies on the Biography dataset. Specifically, we removed the hypothesis document generation module and the adaptive reflection mechanism one by one, and observed their impacts on model performance. The results are shown in the table below:

TABLE II. ABLATION EXPERIMENTS

	Accuracy	Δ vs Full Model
Ours	82.3	-
w/o. Hyde	79.6	-2.7 (↓ 3.2%)
w/o. Reflection	77.8	-2.2 (↓ 5.4%)

These results confirm that both the hypothesis document generation and the reflection mechanism play key roles in improving performance. The hypothesis document enriches semantic context and filters noise to improve retrieval quality. The adaptive reflection mechanism enables dynamic evaluation and correction during generation, leading to more accurate and complete outputs. This is especially valuable for the Biography dataset, which involves long-form text generation and benefits significantly from enhanced content sufficiency and coherence.

V. CONCLUSION

This paper proposes a Task-Classfying RAG framework aimed at addressing the limitations of existing RAG systems in task-adaptive reflection and retrieval input enrichment. The framework consists of three key components: a task classification module based on large language models, a HyDE-based query understanding module, and an adaptive reflection mechanism. The task classification module precisely

categorizes user inputs into three task types—information retrieval, text generation, and general question answering, enabling task-specific reflection strategies. The HyDE-based query understanding module enriches retrieval inputs by integrating the original query with hypothesis documents as complementary information sources. The adaptive reflection mechanism performs differentiated pre-generation reflection tailored to the distinct generation objectives of each task.

Comparative experimental results demonstrate the superiority of our method across multiple natural language processing tasks. Specifically, in short-text generation, long-text generation, and fact verification, our method achieves consistent improvements in accuracy and FactScore metrics over baseline approaches, exhibiting strong generalization capability. Ablation studies further confirm the significant contributions of the HyDE-based query understanding and adaptive reflection modules to overall model performance. In conclusion, the proposed RAG framework enhances task-specific performance through task-sensitive reflection and generation mechanisms, providing theoretical and technical support for building smarter and more reliable generation systems.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal et al. "Language models are few-shot learners." Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.

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

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix et al. "Llama: Open and efficient foundation language models." CoRR, vol. abs/2302.13971, 2023.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu et al. "Survey of hallucination in natural language generation." ACM Comput. Surv., vol. 55, no. 12, pp. 248:1-248:38, 2023.

[5] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao et al. "Kilt: A benchmark for knowledge intensive language tasks." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2523-2544.

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." arXiv preprint arXiv:2311.05232, 2023.

[7] Z. Xu, S. Jain, M. Kankanhalli. "Hallucination is inevitable: An innate limitation of large language models." arXiv preprint arXiv:2401.11817, 2024.

[8] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks." In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 2020.

[9] K. Guu, K. Lee, Z. Tung, P. Pasupat, M.-W. Chang. "Retrieval augmented language model pre-training." In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 2020, pp. 3929-3938.

[10] C.-H. Tan, J.-C. Gu, C. Tao, Z.-H. Ling, C. Xu, H. Hu et al. "Tegtok: Augmenting text generation via task-specific and open-world knowledge." In Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1597-1609.

[11] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi et al. "Large language models can be easily distracted by irrelevant context." In Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 31210-31227.

[12] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection." arXiv preprint arXiv:2310.11511, 2023.

[13] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling. "Corrective Retrieval Augmented Generation." arXiv preprint arXiv:2401.15884, 2024.

[14] A. Singh. "A survey of ai text-to-image and ai text-to-video generators." In 2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC), 2023, pp. 32-36.

[15] A. Singh, A. Ehtesham, G. K. Gupta, N. K. Chatta, S. Kumar, and T. T. Khoei. "Exploring prompt engineering: A systematic review with SWOT analysis." 2024.

[16] M. Schmitt, J. Kieseler, J. G. Flekger, and I. Gurevych, "Precise Zero-Shot Dense Retrieval without Relevance Labels," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 1-10. doi: 10.18653/v1/2023.acl-long.123

[17] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 9802–9822, Toronto, Canada.

[18] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12076–12100, Singapore.

[19] T. Zhang, H. Luo, Y.-S. Chuang, W. Fang, L. Gaitskell, T. Hartvigsen et al. "Interpretable unified language checking," CoRR, vol. abs/2304.03728, 2023.

[20] S. Bhakthavatsalam, D. Khashabi, T. Khot, B. D. Mishra, K. Richardson, A. Sabharwal et al. "Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge," CoRR, vol. abs/2102.03315, 2021.

[21] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh et al. "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12076–12100, Singapore.