Retrieve Only When It Needs: Adaptive Retrieval Augmentation for Hallucination Mitigation in Large Language Models

Hanxing Ding[1,2] **Liang Pang**[1*] **Zihao Wei**[1,2] **Huawei Shen**[1,2] **Xueqi Cheng**[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
{dinghanxing18s, pangliang, weizihao22z, shenhuawei, cxq}@ict.ac.cn

Abstract

Hallucinations present a significant challenge for large language models (LLMs). The utilization of parametric knowledge in generating factual content is constrained by the limited knowledge of LLMs, potentially resulting in internal hallucinations. While incorporating external information can help fill knowledge gaps, it also introduces the risk of irrelevant information, thereby increasing the likelihood of external hallucinations. To balance the use of parametric knowledge within LLMs and external information, in this study, we present Rowen, a novel framework that enhances LLMs with an adaptive retrieval augmentation process tailored to address hallucinated outputs. Rowen introduces a consistency-based hallucination detection module, which assesses the model's uncertainty regarding the input query by evaluating the semantic inconsistencies in various responses generated across different languages or models. When high uncertainties in the responses are detected, Rowen activates the retrieval of external information to rectify the model outputs. Through comprehensive empirical experiments, we demonstrate that Rowen surpasses the current state-of-the-art in both detecting and mitigating hallucinated content within the outputs of LLMs[1].

1 Introduction

In recent years, large language models (LLMs) have demonstrated impressive abilities in natural language understanding (Hendrycks et al., 2021; Huang et al., 2023c), generation (Touvron et al., 2023; Taori et al., 2023), and reasoning (Zhang et al., 2023e; Wang et al., 2023a; Chu et al., 2023). Despite their successes, it has been widely observed that even state-of-the-art LLMs often generate factually incorrect or nonsensical outputs, referred to as *hallucinations* (Ji et al., 2023a; Zhang
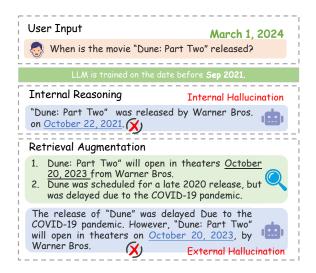

Figure 1: The limited knowledge of LLMs poses a challenge for generating accurate answers, referred to as *Internal Hallucination*, when faced with the latest or domain-specific questions. Additionally, retrieval-augmented generation occasionally faces the risk of error accumulation, where irrelevant evidence may infiltrate the generation phase and lead to nonfactual responses, known as *External Hallucination*.

et al., 2023d,b). These unreliable outputs pose significant risks in practical deployments of LLMs.

Efforts to enhance the factual accuracy of LLM outputs have been substantial. These studies often utilize LLMs' extensive parametric knowledge and advanced logical reasoning capabilities. They employ approaches like self-reflection (Wang et al., 2023b; Ji et al., 2023b; Madaan et al., 2023; Dhuliawala et al., 2023) or collaborative refinements involving interactions among multiple models (Cohen et al., 2023; Du et al., 2023), aiming to enhance logical coherence in refined content. Despite their effectiveness, these self-improvement methods may be limited by LLMs' knowledge boundaries (Ren et al., 2023; Li et al., 2023a) or may not fully exploit parametric knowledge (Huang et al., 2023a), leading to what we term *internal hallucination*, illustrated in Figure 1.

Alongside these self-improvement strategies,

Retrieval-Augmented Generation (RAG) (Chern et al., 2023) serves as a complementary method to overcome knowledge limitations. RAG employs a *retrieve-then-read* pipeline (Karpukhin et al., 2020; Lewis et al., 2020), integrating relevant documents from external knowledge sources into the LLMs' generation process (Huang et al., 2023a; Ye et al., 2023; Varshney et al., 2023; Chern et al., 2023; Li et al., 2023b). However, as depicted in Figure 1, RAG methods are susceptible to *external hallucination* when irrelevant evidence is incorporated, potentially leading to cumulative errors and compromising output accuracy (Li et al., 2023a; Shi et al., 2023).

Drawing inspiration from the latest neuroscience research (Poskanzer and Aly, 2023), which reveals how the human brain dynamically switches between internal thoughts and external sensations, we introduce a novel method, termed Rowen (**R**etrieve **o**nly **when** it needs). Rowen involves an innovative consistency-based uncertainty estimation module, which perturbs semantically equivalent questions and then evaluates the semantic inconsistency of various responses across diverse languages and models when subjected to these perturbed queries. To mitigate *internal hallucinations*, we trigger a retrieval process to fetch relevant information when Rowen detects inconsistencies in LLMs' responses, indicating internal reasoning failures. This helps LLMs refine their reasoning chains and rectify potential hallucinations. To reduce *external hallucinations*, Rowen minimizes the risk of incorporating erroneous information by optimizing retrieval phases. If the perturbed answers convey consistent content, suggesting that LLMs are capable of generating the correct answer themselves, we directly adopt the original answer produced by internal reasoning. This method integrates parametric knowledge within LLMs with retrieved sources, ensuring a balanced integration of internal reasoning and external evidence to effectively mitigate hallucinations.

We evaluate the effectiveness of Rowen on the TruthfulQA dataset (Lin et al., 2022) and the StrategyQA dataset (Geva et al., 2021). Remarkably, our approach excels on the TruthfulQA dataset, yielding an impressive GPT-Judge score of 59.34%, marking a substantial improvement over the state-of-the-art (SOTA) baseline by a significant margin (+16.74%). Similarly, on the StrategyQA dataset, our approach achieves an accuracy of 75.60%, surpassing existing self-improvement and RAG-based baselines with notable superiority. These results comprehensively underscore the powerful capability of Rowen in mitigating hallucinated outputs of LLMs. Furthermore, our adaptive retrieval strategy significantly reduces unnecessary retrievals, thereby enhancing the efficiency of RAG systems.

2 Related Works

Mitigating hallucinations in the inference time could be a cost-effective and controllable way. A line of research harnesses the extensive parametric knowledge and robust logical reasoning capabilities of LLMs to ensure logical consistency either through self-reflection within a single model (Wang et al., 2023b; Ji et al., 2023b; Xu et al., 2023; Madaan et al., 2023; Dhuliawala et al., 2023) or through collaborative refinements or debates involving multiple models (Cohen et al., 2023; Du et al., 2023). Despite their strengths, LLMs are sometimes constrained by their knowledge boundaries and the complexity of the reasoning chain, resulting in occasional inaccuracies (Ren et al., 2023; Li et al., 2023a; Mallen et al., 2023) termed *internal hallucination*. To address this knowledge gap, retrieval-augmented generation methods leverage external knowledge as supplementary evidence to aid LLMs in providing accurate responses (Huang et al., 2023a; Ye et al., 2023; Niu et al., 2012; Xu et al., 2024; Varshney et al., 2023; Chern et al., 2023; Li et al., 2023b). However, these approaches, while effective, occasionally encounter the challenge of error accumulation, where irrelevant evidence may seep into the generation process, leading to incorrect responses (Li et al., 2023a; Shi et al., 2023), a phenomenon referred to as *external hallucination*. Our work only performs retrieval augmentation when hallucinations are detected, thereby maximizing the utilization of both the parametric knowledge and externally retrieved information.

There are also some adaptive retrieval methods that assess the difficulty of questions or the confidence in responses to decide whether to retrieve documents (Jiang et al., 2023; Mallen et al., 2023; Asai et al., 2023; Jeong et al., 2024). Due to space limitations, we discuss these works in Appendix C.

3 Methodology

Our objective is to enhance the factuality of LLM responses by integrating parametric and external knowledge. We propose a framework called Rowen (**R**etrieve **o**nly **when** needed). Initially, we leverage LLMs' Chain-of-Thought (CoT) reasoning to gen-
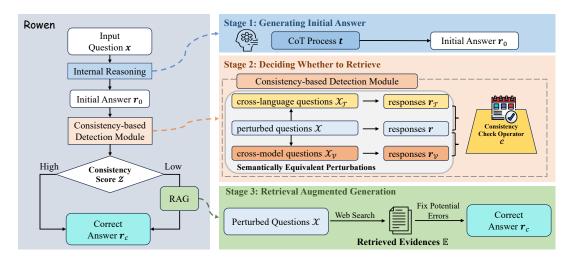
Figure 2: Overview of Rowen framework. We start by producing an initial response using CoT thinking. Then, a consistency-based detection module evaluates the semantic consistency of responses to the same question in different languages or models. If inconsistencies arise, a retrieval-augmented mechanism is engaged to fetch external information, helping to rectify the reasoning and correct any inaccuracies. Otherwise, the initial answer is retained.

erate an initial response (§ 3.1). To mitigate internal hallucinations, Rowen employs a consistency-based hallucination detection module that assesses the reliability of the initial response (§ 3.2). If high uncertainties are found, the initial answer is refined using external information via retrieval augmentation (§ 3.3), resulting in the final response. Otherwise, the initial response is considered the final output. For external hallucinations, Rowen resorts to external knowledge only when high uncertainties are found, ensuring that the final answer is both accurate and reliable.

3.1 Stage 1: Generating Initial Answer

To maximize the exploitation of the parametric knowledge in LLMs, we initially employ their Chain-of-Thought (CoT) reasoning to generate a preliminary response. This process involves: critically assessing the validity of the information in the input query x and prioritizing accuracy and fact-checking for answer elaboration, detailed in Appendix F Table 9. After generating the CoT thought t, we ask the model $\mathcal{M}$ to provide a concise answer r_0 for the input query x. The answer r_0 will serve as the final response to the input query x if our detection module ensures it is free from hallucinations.

3.2 Stage 2: Deciding Whether to Retrieve

To decide when to retrieve, we leverage model uncertainty, which refers to the confidence level of model outputs and serves as a crucial indicator for deciding when to trust LLMs (Zhang et al., 2023d). Unfortunately, current consistency-based methods

fail when LLMs provide consistent yet incorrect answers across different perturbations (Zhang et al., 2023a). This issue may arise because these methods focus exclusively on semantic coherence within a single language or model.

To tackle this issue, we propose novel cross-language and cross-model detection modules that assess semantic consistency among responses for the same question across different languages or models. If inconsistencies are detected in these responses, we flag them as potentially inaccurate and invoke a retrieval process.

3.2.1 Diversifying Question Verbalizations

To facilitate subsequent consistency-based hallucination detection, we begin by leveraging advancements in LLM prompting to generate semantically equivalent perturbations. Initially, we start with an input x and instruct the model $\mathcal{M}$ to provide a set of k semantically equivalent questions $\mathcal{X} = \{x^1, x^2, \ldots, x^k\}$. We use the prompt: "For the question [ORIGINAL QUESTION], please provide k semantically equivalent questions" with a high decoding temperature to generate diverse perturbed expressions.

After obtaining the diverse verbalized questions, we prompt the LM $\mathcal{M}$ to generate its candidate responses according to the questions. We employ a greedy decoding strategy to avoid unpredictable randomness of the LM $\mathcal{M}$ as much as possible.

$$r^j = \mathcal{M}(x^j), j = 1, \ldots, k , \quad (1)$$

where k is the length of the generated semantically equivalent questions $\mathcal{X}$.

3.2.2 Cross-Language Consistency-based Detection

Recent finding reveals that multilingual models do not always learn knowledge that is fully aligned across different languages (Gao et al., 2024). Building on this insight, we propose that this inconsistency can be strategically used as an indicator for the necessity of information retrieval. When language models provide inconsistent answers to the same query in different languages, it signals a lack of certainty about the relevant knowledge. In such cases, retrieving external knowledge is crucial for accurate responses.

To achieve cross-language detection, we first incorporate a cross-language consistency check (Rowen-CL) to evaluate the semantic consistency of responses to the same question across different languages. We introduce language-level perturbations by asking the model $\mathcal{M}$ to translate the source-language questions $\mathcal{X}$ into corresponding paraphrased questions $\mathcal{X}_\mathcal{T} = \{\boldsymbol{x}_\mathcal{T}^1, \boldsymbol{x}_\mathcal{T}^2, \ldots, \boldsymbol{x}_\mathcal{T}^k\}$ in the target language. The model $\mathcal{M}$ is then instructed to generate corresponding answers to each question in the target language.

$$\boldsymbol{r}_\mathcal{T}^j = \mathcal{M}(\boldsymbol{x}_\mathcal{T}^j) = \mathcal{M}(\text{translate}(\boldsymbol{x}^j)), \ j = 1, \ldots, k. \tag{2}$$

To captures language-level cross-checking uncertainties, we utilize the generated questions and answers from all previous stages to calculate a numerical consistency score to calculate a numerical consistency score. Let $\mathcal{C}(\cdot, \cdot)$ denote a semantic equivalence checking operator that takes two QA pairs as inputs and returns "True" if they are semantically equivalent, and "False" otherwise. We then map the response to a numerical semantic equivalent score: {"True" $\rightarrow$ 1.0, "False" $\rightarrow$ 0.0}. We leverage the model $\mathcal{M}$ and utilize the prompt provided in Appendix F Table 10 to implement the cross-language checking operator and calculate the cross-language consistency score $\mathcal{Z}$ as:

$$\mathcal{Z}_{\text{CL}} = \frac{1}{k} \sum_{j=1}^k \mathcal{C}(\mathcal{P}^j, \mathcal{P}_\mathcal{T}^j), \tag{3}$$

where $\mathcal{P}^j = (\boldsymbol{x}^j, \boldsymbol{r}^j)$ and $\mathcal{P}_\mathcal{T}^j = (\boldsymbol{x}_\mathcal{T}^j, \boldsymbol{r}_\mathcal{T}^j)$ denote the QA pairs in the source language and target language, respectively.

3.2.3 Cross-Model Consistency-based Detection

Besides language-level cross-checking, we also introduce a cross-model detection module (Rowen-CM) to evaluate the semantic consistency of responses to the same question across different models. We adopt an additional verifier LM $\mathcal{M}_\mathcal{V}$ for model-level cross-checking and instruct the verifier LM $\mathcal{M}_\mathcal{V}$ to provide answers for each semantically equivalent question:

$$\boldsymbol{r}_\mathcal{V}^j = \mathcal{M}_\mathcal{V}(\boldsymbol{x}^j), \ j = 1, \ldots, k. \tag{4}$$

Similar to the cross-language consistency score calculation, we use the prompt in Appendix F Table 11 to implement the checking operator $\mathcal{C}$ to calculate cross-model consistency score:

$$\mathcal{Z}_{\text{CM}} = \frac{1}{k} \sum_{j=1}^k \mathcal{C}(\mathcal{P}^j, \mathcal{P}_\mathcal{V}^j), \tag{5}$$

where $\mathcal{P}_\mathcal{V}^j = (\boldsymbol{x}_\mathcal{V}^j, \boldsymbol{r}_\mathcal{V}^j)$ denote the QA pairs generated by verifier model $\mathcal{M}_\mathcal{V}$.

3.2.4 Hybrid Consistency-based Detection

The different variants of Rowen capture various aspects of uncertainty in the original response, complementing each other effectively. We propose integrating the cross-language and cross-model consistency scores to create a unified hybrid consistency score:

$$\mathcal{Z}_{\text{Hybrid}} = \mathcal{Z}_{\text{CL}} + \alpha * \mathcal{Z}_{\text{CM}}, \tag{6}$$

where α is a weight factor for the cross-model consistency score.

3.3 Stage 3: Retrieval Augmented Generation

If the consistency score $\mathcal{Z}$ falls below a threshold, it indicates possible hallucinated content in the original response $\boldsymbol{r}_0$. We then introduce a retrieval-augmented generation procedure.

Searching Relevant Knowledge To help the LM $\mathcal{M}$ correct errors, we search for supporting evidence from external sources like online webpages. We first ask the model $\mathcal{M}$ to generate search queries for each paraphrased question in $\mathcal{X}$, following the prompt in Appendix F Table 12. These queries are input into the online search engine to retrieve relevant knowledge, denoted as $\mathbf{E}$, used for correcting factual errors in $\boldsymbol{r}_0$.

Repairing Hallucinated Contents With the retrieved evidence $\mathbf{E}$, the model reviews the original thought process $\boldsymbol{t}$ and initial answer $\boldsymbol{r}_0$. The aim is to identify and correct inaccuracies, producing the refined answer $\boldsymbol{r}_c$:

$$\boldsymbol{r}_c = \mathcal{M}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{r}_0, \mathbf{E}). \tag{7}$$

Appendix F Table 13 shows the prompt for repairing the original answer. The corrected answer r_c serves as the final response to question x.

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

TruthfulQA We use the TruthfulQA dataset (Lin et al., 2022) to evaluate the ability of LLMs in generating truthful responses (Zhang et al., 2023c; Kadavath et al., 2022). In our study, we focus on the generation task in TruthfulQA. Therefore, to evaluate the factuality of responses from LLMs, we calculate the GPT-Judge score, obtained by fine-tuning the babbage-002 model using the original fine-tuning data from their official repository[2]. We also report the BLEU and Rouge-L scores to evaluate the lexical overlap between generated responses and ground-truth references.

StrategyQA The StrategyQA dataset (Geva et al., 2021) comprises crowdsourced yes / no questions that require multi-step reasoning for accurate answers. We follow previous work (Jiang et al., 2023) to randomly sample 500 examples due to the cost consideration of running experiments. We also follow the settings of Wei et al. (2022) to generate both the reasoning process as well as the final answer. We present the exact-match accuracy of the generated yes / no answers compared to the gold-standard answers.

4.2 Baseline Methods

We consider the following methods as our baselines: (1) Vanilla LLMs, such as ChatGPT. (2) Self-improvement methods: **CoVe** (Dhuliawala et al., 2023) generates verification questions to self-analyze potential errors, systematically addressing each question to refine the baseline response. **Self-Reflection** (Ji et al., 2023b) presents an interactive self-reflection methodology that incorporates knowledge acquisition and answer generation. **Multi-agent Debate** (Du et al., 2023) utilize multiple LM agents to debate their individual responses over multiple rounds to arrive at a common final answer. (3) Retrieval-augmented methods: **Factool** (Chern et al., 2023) leverages various tools to gather evidence about the factuality of the generated content. **Detect and Mitigate** (Varshney et al., 2023) actively detects hallucinations during generation by identifying potential hallucination through the logit output values of LLMs. (4) Adaptive retrieval methods: **FLARE** (Jiang et al., 2023) adopts an active retrieval strategy that only retrieves when LLMs generate low probability tokens. **Adaptive-Retrieval** (Mallen et al., 2023) only retrieves when necessary based on a pre-defined threshold for entity popularity. **Self-RAG** (Asai et al., 2023) trains a single arbitrary LM that adaptively retrieves passages on-demand. **Adaptive-RAG** (Jeong et al., 2024) trains a query-complexity classifier to decide when to retrieve based on question complexity. **LUQ** (Zhang et al., 2024) is a sampling-based uncertainty quantification for long text.

4.3 Implementation Details

In our experiments, we validate Rowen using the ChatGPT language model. We re-implement all baselines, except Self-RAG, using ChatGPT to ensure a fair comparison. For semantic perturbations, we configure the temperature to 1.0 to generate diverse perturbed expressions. Otherwise, the temperature is set to 0.0 to obtain high-quality deterministic outputs. Considering the diversity of expressions and the latency in generating perturbations, we produce $k = 6$ semantically equivalent questions. For the cross-language detection module, English serves as the source language while Chinese is employed as the target language. For the cross-model detection module, we adopt Qwen-Max-0428[3], a large instruction-tuned model for chat service, as the verifier LM. Following Chern et al. (2023), we utilize the Google Search API offered by Serper[4] to search the top pages and extract the most pertinent search snippets from the API's response. We also conduct extensive experiments to study the impact of various hyperparameters in Appendix D.

5 Experimental Results

5.1 Main Results on Hallucination Mitigation

We evaluate the effectiveness of Rowen on the TruthfulQA and StrategyQA datasets. Table 1 presents the overall performance of Rowen compared to several strong baselines. Rowen demonstrates superior performance on both datasets, with higher GPT-Judge score and accuracy, indicating the effectiveness of our proposed method.

Vanilla ChatGPT shows a certain level of accuracy in answering factual questions, achieving

---

[2]https://github.com/sylinrl/TruthfulQA

[3]https://qwenlm.github.io/blog/qwen-max-0428
[4]https://serper.dev/

Models	TruthfulQA			StrategyQA
	GPT-Judge ↑	BLEU ↑	Rouge-L ↑	Accuracy ↑
Vanilla LLMs				
ChatGPT (gpt-3.5-turbo)	47.92	10.17	**31.31**	61.40
Self-improvement Methods				
CoVe (Dhuliawala et al., 2023)	48.01	12.81	26.52	61.40
Multi-agent Debate (Du et al., 2023)	50.83	3.94	21.05	65.73
Self-Reflection (Ji et al., 2023b)	42.99	3.86	18.18	62.40
Retrieval-augmented Methods				
Factool (Chern et al., 2023)	34.50	1.34	12.22	67.20
Detect-and-Mitigate (Varshney et al., 2023)	49.98	3.17	18.59	56.94
Adaptive Retrieval Methods				
FLARE (Jiang et al., 2023)	45.04	11.59	26.83	61.19
Adaptive-Retrieval (Mallen et al., 2023)	45.55	8.87	26.75	62.50
Self-RAG (Asai et al., 2023)	40.36	4.36	21.28	58.40
Adaptive-RAG (Jeong et al., 2024)	46.02	10.29	26.24	68.50
LUQ (Zhang et al., 2024)	55.08	5.79	21.44	71.00
Our Framework				
Rowen-CL	57.39	7.60	24.16	74.00
Rowen-CM	56.29	6.85	22.36	72.40
Rowen-Hybrid	**59.34**	**15.27**	31.15	**75.60**

Table 1: Experimental results of mitigating hallucinations on the TruthfulQA dataset and StrategyQA dataset. Rowen-Hybrid achieves a detection accuracy of 59.0% on the TruthfulQA dataset and 73.0% on the StrategyQA dataset following meticulous adjustments of hyper-parameters in § D.

scores of 47.92% and 61.40% on the two datasets, respectively. While self-improvement methods perform better than the vanilla LM on both datasets, they are still limited by their knowledge boundaries and suffer from *internal hallucinations*. RAG methods demonstrate relatively better performance compared to self-improvement methods, highlighting the benefits of integrating external knowledge. However, Factool falls short on the TruthfulQA dataset, and the Detect-and-Mitigate method underperforms on the StrategyQA dataset. This may be attributed to error accumulation caused by unnecessary retrieval (*external hallucinations*).

We also conduct additional experiments to compare with four adaptive retrieval methods. Notably, Adaptive-Retrieval faces challenges on the TruthfulQA dataset due to some questions lacking explicit entities, causing it to struggle in deciding when to retrieve based on entity popularity, leading to poor performance. Besides, Self-RAG's effectiveness is hindered by the capabilities of the LLaMa model, resulting in inferior performance.

Compared to the aforementioned baselines, Rowen demonstrates significant performance gains on both datasets. Both Rowen-CL and Rowen-CM exhibit excellent hallucination mitigation capabilities, even when compared to strong adaptive retrieval methods. Specifically, Rowen-Hybrid

achieves a GPT-Judge score of 59.34% on the TruthfulQA dataset, surpassing the strongest baseline by 16.74%. Additionally, Rowen-Hybrid attains an accuracy of 75.60% on the StrategyQA dataset, significantly outperforming existing baselines. These results underscore Rowen's ability to effectively leverage parametric knowledge and external information for maximum advantage.

5.2 Effect of Detection Module

To validate the superiority of our hallucination detection module, we compare its performance in adaptive retrieval scenarios for hallucination mitigation with strong detection methods: (1) average token-level probability / entropy that utilizes the probabilities / entropies of tokens generated by a proxy LM (e.g., LLaMa2-7B) as a metric to measure hallucination. (2) SelfCheckGPT (Manakul et al., 2023) that measures information consistency between the different responses to determine hallucinations. (3) Consistency-based method (Zhang et al., 2023a), SAC³, that evaluates semantic-aware cross-check consistency, building upon the foundation of self-consistency principles.

Based on the results in Table 2, it is evident that logits-based methods perform moderately well in detecting hallucinations, especially on the StrategyQA dataset. Specifically, the entropy-based es-

Models	TruthfulQA				StrategyQA	
	GPT-Judge ↑	BLEU ↑	Rouge-L ↑	Ratio(%)	Accuracy ↑	Ratio(%)
LLaMa2-7B						
Average Token Probability	50.19	12.51	31.11	46.5	71.50	23.0
Average Token Entropy	52.22	9.18	28.37	59.0	72.00	26.5
SelfCheckGPT						
w/ BERTScore	51.38	8.43	26.85	27.2	67.50	21.0
w/ MQAG	52.76	7.69	26.92	54.4	69.00	34.0
w/ Ngram	52.41	5.01	21.60	34.9	66.50	32.0
Combination	53.10	6.69	24.04	51.3	69.50	30.0
Consistency						
SAC^3-Q	51.02	7.90	28.00	24.5	65.50	24.0
SAC^3-all	52.22	9.37	29.56	20.8	67.00	24.5
Rowen-Hybrid	**59.34**	**15.27**	**31.15**	23.0	**75.6**	20.0

Table 2: Performance comparison of applying other hallucination detection methods in adaptive retrieval scenarios. We also report the ratio of retrieval conducted by each method.

timation exhibits superior performance. Besides, the combination of different variants in SelfCheck-GPT leads to a slight performance improvement. Additionally, it is worth noting that SAC^3 achieve competitive performance on detecting potential factual errors accurately.

Finally, we observe that Rowen significantly outperforms these strong hallucination detection baselines, especially the monolingual detection method SAC^3. We also notice that Rowen achieves notable hallucination mitigation with a minimal number of retrieval calls. This underlines the superior efficiency of our hallucination detection module.

5.3 Scalability of Rowen

Scalability to Open-Source LLMs In addition to ChatGPT, we also assess Rowen's effectiveness when employing open-source language models: Qwen1.5-14B-Chat, Qwen2-7B-Instruct, and Qwen2-72B-Instruct[5]. These models are chosen for the cross-language detection model $\mathcal{M}$ due to their strong capabilities in following Chinese instructions, a critical feature for effective cross-language detection. The Llama-series models are not considered due to their weaker performance in generating Chinese responses. Instead, the Qwen-series models have demonstrated state-of-the-art performance in Chinese, which aligns with our research goals. For verifier LM $\mathcal{M_V}$, we choose to use Llama-3-8B-Instruct[6].

Figure 3 shows the results of three variants of Rowen on three open-source models on TruthfulQA dataset. Our Rowen methods achieve strong performance when applied to open-source LLMs,

Figure 3: Experimental results of Rowen with open-source LLMs on TruthfulQA.

nearly matching the baseline results on Chat-GPT. Specifically, the Multi-agent Debate baseline scores 50.83, while the Detect-and-Mitigate baseline scores 49.98. These findings further prove the effectiveness and scalability of our proposed Rowen method within open-source LLMs.

Scalability to Other Datasets We assess the Rowen model's scalability by examining its performance on datasets with answers of intermediate length, namely TriviaQA and Natural Questions, given the diverse answer lengths in TruthfulQA (long answers) and StrategyQA (binary responses). We conduct comparisons against two strong adaptive retrieval baselines, reporting the metrics of EM (Exact Match) and F1 score on their dev sets. The experimental results are shown in Table 3.

In short, our Rowen model, with its different variants, consistently outperforms the baselines on both datasets, suggesting a notably more robust capability in handling intermediate-length answers. Particularly, Rowen-Hybrid emerges as the most effective, achieving the highest scores across both metrics and datasets. This denotes a significant

Methods	NQ		TriviaQA	
	EM	F1	EM	F1
FLARE	32.50	43.91	59.00	68.34
Adaptive-RAG	35.04	48.44	58.00	68.97
Rowen-CL	38.08	55.81	66.50	76.66
Rowen-CM	37.36	53.27	65.00	74.98
Rowen-Hybrid	**39.98**	**57.31**	**69.04**	**78.46**

Table 3: Hallucinations mitigation performance on the NQ dataset and TriviaQA dataset.

enhancement over baseline methods.

5.4 Analysis of Computation Cost for Rowen

Efficiency Analysis of LLM Calls To evaluate the efficiency of LLM calls in Rowen, we compare various methods—including Self-Reflection, Multi-agent Debate and three different Rowen variants—by analyzing the number of API calls required to answer a single question and their effectiveness in mitigating hallucinations. From the results shown in Table 4, we find that Rowen variants, particularly Rowen-CL and Rowen-CM, achieve significantly better hallucination mitigation compared to the baseline methods, while using a similar number of API call. To further improve efficiency, We design two useful speedup strategies to improve the efficiency of Rowen and describe them in Appendix E.

Methods	TruthfulQA		StrategyQA	
	GPT-Judge	# Calls	Accuracy	# Calls
Self-Reflection	42.99	6	62.40	5
Multi-agent Deb.	50.83	6	65.73	6
Rowen-CL	57.39	6	74.00	5
Rowen-CM	56.29	5	72.40	4
Rowen-Hybrid	**59.34**	8	**75.60**	6

Table 4: Analysis on LLM calls efficiency and hallucination mitigation performance across different methods.

Efficiency Analysis of Retrieval Calls To verify the retrieval efficiency of Rowen, in Table 5, we compare the average number of retrieval calls made by Factool, Detect-and-Mitigate, FLARE, and Rowen-Hybrid to answer a question across two datasets. Rowen-Hybrid, which retrieves information only for uncertain responses, excels in both retrieval efficiency and factual accuracy, showcasing its superiority over other RAG methods.

5.5 Quantitative Analysis of Hallucinations

To verify whether Rowen effectively reduces internal and external hallucinations, we present a comparative analysis of the prevalence of both types

Methods	# Num of Retrieval Calls	
	TruthfulQA	StrategyQA
Factool	12.5	11.6
Detect-and-Mit.	7.2	5.5
FLARE	2.1	3.9
Adaptive-RAG	**0.9**	1.4
Rowen-Hybrid	1.5	**0.5**

Table 5: Statistics on the average number of retrieval calls to answer each question.

of hallucinations across three methods—Factool, Detect-and-Mitigate, and Rowen-Hybrid. Internal hallucinations refer to the generation of wrong responses using only the parameterized knowledge of LLMs, while external hallucinations refer to the generation of incorrect responses using noisy documents introduced after retrieval. Figure 4 shows that Factool and Detect-and-Mitigate are significantly prone to both types of hallucinations. Both baseline methods have high levels of internal and external hallucinations, struggling with internal coherence and external fact alignment. In contrast, Rowen-Hybrid effectively reduces external hallucinations by timely integrating external knowledge, thereby avoiding unnecessary information retrieval and mitigating potential errors.
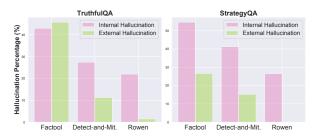

Figure 4: The percentage of internal and external hallucinations across different methods on the both dataset.

6 Conclusion and Future Work

In this paper, we introduce Rowen, a novel method designed to actively identify potential hallucinated responses in LLMs and correcting factual errors through an adaptive retrieval process. Our experiments on four datasets demonstrate the effectiveness of our approach. We also conduct detailed analytical experiments to underscore the impact of each module and the choices of hyper-parameters. Moving forward, we plan to expand our work by exploring how to effectively utilize retrieved evidence, even when irrelevant documents are integrated, and devise an effective method to minimize the influence of knowledge conflicts between internal knowledge and retrieved documents.

Limitations

Our method has several drawbacks. Firstly, it focuses on post-hoc correction, limiting its ability to detect hallucination only after generating a complete sentence rather than in real-time during the generation of contexts. Secondly, for ensuring accurate answers, our approach requires multiple requests to the API interface. This involves generating the CoT answer, perturbing and responding to the input, correcting the original answer, and so on, leading to increased latency. To address this, we can expedite the entire process by parallelizing API calls for both the OpenAI API and Google Search API. Additionally, reducing the number of perturbations can strike a better balance between detection precision and inference speed. Thirdly, our approach neglects to analyze and utilize the factual nature of the retrieved evidence, which we leave for future work.

Ethics Statement

This study delves into the detection and mitigation of hallucinations in language models (LMs), presenting broader implications for the realm of natural language processing (NLP) and addressing ethical concerns related to trust and reliability. Our method has the potential to enhance the accuracy and dependability of LMs, minimizing the risks associated with misinformation and biased outputs. Moreover, it foster accountability and trust in AI systems, contributing to the ongoing development of more reliable language models.

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *CoRR*, abs/2305.18248.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *CoRR*, abs/2309.15402.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12621–12640. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *CoRR*, abs/2309.11495.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pre-training and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *arXiv preprint arXiv:2404.04659*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *CoRR*, abs/2307.10236.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023c. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *CoRR*, abs/2403.14403.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea

Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1827–1843. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7969–7992. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. 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, December 6-12, 2020, virtual*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023b. Self-checker: Plug-and-play modules for fact-checking with large language models. *CoRR*, abs/2305.14623.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 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)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *CoRR*, abs/2303.08896.

Shuzi Niu, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2012. Top-k learning to rank: labeling, ranking and evaluation. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 751–760. ACM.

Craig Poskanzer and Mariam Aly. 2023. Switching between external and internal attention in hippocampal networks. *Journal of Neuroscience*, 43(38):6538–6552.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *CoRR*, abs/2307.11019.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. *arXiv preprint arXiv:2404.17287*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *CoRR*, abs/2305.14975.

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

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *CoRR*, abs/2307.03987.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *CoRR*, abs/2306.13063.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *CoRR*, abs/2304.14732.

Shicheng Xu, Liang Pang, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. List-aware reranking-truncation joint model for search and retrieval-augmented generation. *CoRR*, abs/2402.02764.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *CoRR*, abs/2309.06794.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: long-text uncertainty quantification for llms. *CoRR*, abs/2403.20279.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Kumar Sricharan. 2023a. Sac³: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15445–15458. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023b. How language model hallucinations can snowball. *CoRR*, abs/2305.13534.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023c. Alleviating hallucinations of large language models through induced hallucinations. *CoRR*, abs/2312.15710.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023d. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023e. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

A Kernel Density Estimation Visualization

In our work, we hypothesize that lower consistency levels correlate with increased uncertainty in the model, consequently leading to a higher probability of inaccurate predictions. To verify this hypothesis, we analyze the consistency distribution through kernel density estimation plots on the TruthfulQA dataset. As illustrated in Figure 5, in the lower range of consistency check scores (left-skewed), the proportion of incorrect predictions is higher. This indicates that low consistency check scores are often associated with incorrect model predictions. On the contrary, in the higher range of consistency check scores (right-skewed), the proportion of correct predictions is higher. This indicates that
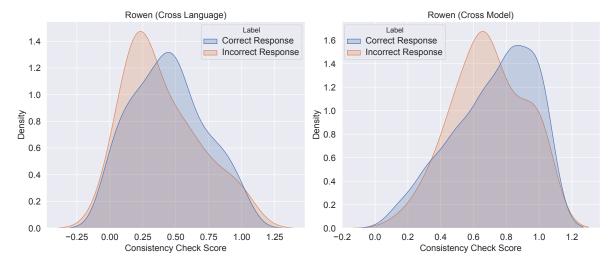
Figure 5: Kernal density estimation plots for consistency on the TruthfulQA dataset.

high consistency check scores are usually associated with correct model predictions. There is some overlap between the distributions of different labels, but the overall trend shows that the higher the consistency check score, the higher the likelihood of correct predictions. These findings suggest that consistency check scores can be an effective indicator of prediction accuracy: the higher the score, the more accurate the model prediction is likely to be. These findings provide empirical support for our hypothesis. Additionally, we observe that the suitable threshold points for distinguishing correct from incorrect answers differ between Rowen-CL and Rowen-CM, with the optimal threshold for Rowen-CM being higher. This observation aligns with the conclusion in § D.

B Related Work on Exploring Uncertainty for Hallucination Detection

Uncertainty refers to the confidence level of the model outputs, and it serves an important indicator for identifying and eliminating hallucinations, so it can assist users in determining when to trust LLMs. In general, methods for exploring uncertainty for hallucination detection can be categorized into three types: (1) Logit-based estimation relies on accessing the model's logits to calculate token-level probabilities or entropy, which are used to measure uncertainty (Huang et al., 2023b; Varshney et al., 2023). However, this approach can pose challenges for black-box closed-source models. (2) Verbalized-based estimation involves prompting language models to express their uncertainty using specific prompts (Xiong et al., 2023; Tian et al., 2023; Agrawal et al., 2023). However, these meth-

ods tend to display a high degree of overconfidence when expressing their confidence (Xiong et al., 2023; Tao et al., 2024). (3) To overcome these limitations, consistency-based estimations are adopted to measure the consistency score among multiple responses provided by the model for a given question (Manakul et al., 2023; Xiong et al., 2023; Wang et al., 2023b; Zhao et al., 2023). The underlying assumption suggests that when language models struggle with indecision and fabricate facts, they tend to provide logically inconsistent responses to identical questions. In this work, we propose that cross-language and cross-model consistency can offer highly sensitive signals for identifying hallucinations. Therefore, we utilize cross-language and cross-model detection modules that cross-check answers to the same question across different languages or models. This cross-checking paradigm serves as a powerful mechanism to identify hallucinations in LLMs.

C Discussions

There is a line of recent studies that concentrate on mitigating hallucinations with adaptive retrieval (Jiang et al., 2023; Mallen et al., 2023; Asai et al., 2023; Jeong et al., 2024). Our work diverges from existing methodologies by: (1) Traditional self-consistency method (Wang et al., 2023b) may falter when LLMs provide coherent yet incorrect responses to the same query, which is observed in SAC[3]. Our approach extends self-consistency uncertainty estimation to multilingual and multi-model settings, yielding superior hallucination mitigation performance compared to conventional methods (see Table 1). (2) FLARE (Jiang et al., 2023) relies on the token probability of model outputs, a lim-

itation for commercial closed-source models. In contrast, our method overcomes this constraint by assessing semantic similarity in generated content, which is suitable for white / black-box models. (3) Self-RAG (Asai et al., 2023) necessitates intricate retraining processes and significant computational resources (4 Nvidia A100 GPUs and 80GB memory). In contrast, our method effectively mitigates hallucination without the need for retraining, underscoring its suitability for resource-constrained scenarios. (4) Adaptive-Retrieval (Mallen et al., 2023) employs Wikipedia entity popularity for retrieval decisions, which proves ineffective for adversarial questions lacking clear entities in datasets like TruthfulQA. (5) Adaptive-RAG (Jeong et al., 2024) dynamically decide whether to retrieve based on a pre-trained classifier, which is a smaller LM trained to predict the complexity level of incoming queries. Adaptive-RAG could be influenced by the accuracy of the classifier, given the absence of training datasets for query-complexity classifier, potentially leading to incorrect labeling of queries.

Our approach, examining linguistic diversity (cross-language) and model response consistency (cross-model), effectively identifies and mitigates hallucinations. This enhances model robustness and reliability, providing a solid foundation for applying LLMs. Our method is novel, practical, flexible, training-free, and broadly applicable, offering new insights and optimizing LLM research and application.

D Impact of Hyper-parameters in Rowen

Impact of Detection Threshold. We experiment with various consistency thresholds to investigate their impact on hallucination mitigation effectiveness. As shown in Figure 6, we observe significant performance improvement when the detection threshold increases from 0.2 to 0.6, as this identifies more hallucinations and enhances factual accuracy by retrieving more external information. However, further increasing the threshold introduces more noise and irrelevant evidence, degrading performance. This suggests that an appropriate threshold is crucial for balancing detection accuracy and the retrieval process. Additionally, Rowen-CL and Rowen-CM have different optimal thresholds of 0.6 and 0.8, respectively, and the Rowen-Hybrid method demonstrates the most stable performance across varying thresholds.
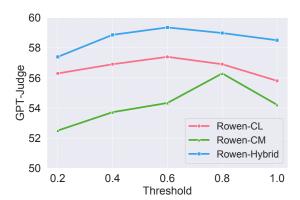

Figure 6: Effect of various detection threshold on the TruthfulQA dataset.

Impact of Number of Perturbed Questions. We investigate the impact of the number of perturbed questions k of Rowen on the performance of hallucination mitigation. The results in Figure 7 reveal that the mitigation performance improves with the increase in the number of perturbations, particularly notable in the transition from 4 to 6 perturbations. This enhancement is attributed to the greater number of perturbations aiding in higher detection accuracy. However, further increasing the perturbation count beyond 6 yields only marginal improvements in performance (even no improvements). Thus, after considering the results, we determine that the optimal number of perturbations for achieving the best hallucination mitigation performance is 6.
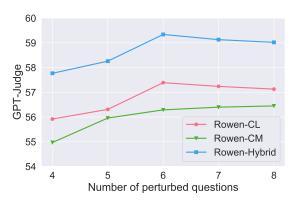

Figure 7: Effect of number of perturbed questions on the TruthfulQA dataset.

Impact of Cross-Model Detection Weight. We also examine the impact of different cross-model detection weight for Rowen-Hybrid. We ensure that for any given α, the detection threshold is recalibrated to maintain optimal performance. Figure 8 shows that GPT-Judge scores significantly increase as the weight rises from 0.25 to 1.0, indicating that introducing cross-model checking can improve

Language Pair	Discrepancy	GPT-Judge
English - German	Low	50.55
English - French	Medium	52.26
English - Chinese	High	57.39

Table 6: Comparison of hallucination mitigation performance for `Rowen-CL` under different language choices on the TruthfulQA dataset.

the accuracy of hallucination detection. However, when the weight factor exceeds 1.0, performance declines, with scores dropping to around 58 at a weight of 1.5. This suggests that while a moderate weight enhances factuality, excessively a high value allows the cross-model detection module to exert too much influence, reducing effectiveness.
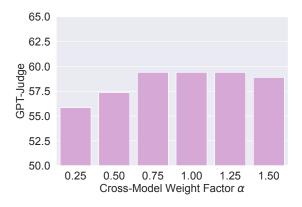

Figure 8: Effect of cross-model detection weight factor for `Rowen-Hybrid` on the TruthfulQA dataset.

Impact of the Choice of Target Language To investigate how different language pairs affect the effectiveness of hallucination mitigation for `Rowen-CL`, we conduct experiments on the TruthfulQA dataset with various target languages. The results of these experiments are presented in Table 6. It is observed that the combination of English as the source language and German as the target language yields the least favorable results. This may be attributed to their shared Germanic language family roots, which results in numerous linguistic similarities and overlapping cultural references. Conversely, as the cultural divergence between the source and target languages widens, we witness an enhancement in the performance of hallucination mitigation. This trend substantiates the hypothesis that cultural disparities between languages play a pivotal role in identifying hallucinations and bolstering the factuality of the generated responses.

Impact of the Choice of Verifier LM To investigate the impact of different verifier LMs on the

efficacy of hallucination mitigation for `Rowen-CM`, we conducted experiments using various verifier LMs on the TruthfulQA dataset. The experimental results presented in Table 7 demonstrate a significant improvement in hallucination mitigation as the capability of the verifier LM increases. Specifically, Qwen-Max exhibited the highest effectiveness in this task, achieving an efficacy of 56.29%, while Qwen-Turbo and Qwen-Plus achieved 52.50% and 54.09%, respectively. These findings underscore the critical role of verifier LM selection in enhancing the effectiveness of hallucination mitigation.

Verifier LM	Capability	GPT-Judge
Qwen-Turbo	Low	52.50
Qwen-Plus	Medium	54.09
Qwen-Max	High	56.29

Table 7: Comparison of hallucination mitigation performance for `Rowen-CM` under different verifier LM on the TruthfulQA dataset.

Impact of Perturbation Objects In evaluating cross-language and cross-model perturbations, as well as calculating subsequent consistency scores, applying CL and CM perturbations to the original question appears to be a viable approach. However, recent studies suggest that merely perturbing questions might lead to responses that are consistent but not necessarily accurate (Zhang et al., 2023a). This implies that question-level consistency alone does not suffice for assessing factual correctness, as these methods predominantly emphasize semantic coherence.

To validate this, we conduct preliminary experiments by applying CL and CM perturbations to the original question (question-level consistency checking) and compare the results with those of Rowen. The results, as shown in Table 8, demonstrate that Rowen consistently outperforms the approach of applying perturbations to the original question. This reinforces that Rowen remains the most effective method for ensuring accuracy.

E   Strategies to improve efficiency

We have implemented two helpful strategies to improve the efficiency without compromising the performance by:

(1) Our cross-language detection and cross-model detection modules can operate in parallel. This significantly reduces latency, as it allows for

Object	Variants	TruthfulQA	StrategyQA
		GPT-Judge	Accuracy
Input	CL	55.34	72.00
	CM	55.85	72.00
Perturbations	CL	57.39	74.00
	CM	56.29	72.40

Table 8: Comparisons of applying CL and CM perturbations to the original question or perturbated questions.

the simultaneous evaluation of semantic consistency across different languages and models.

(2) We have optimized the generation process for responding to multiple semantically equivalent perturbed questions by leveraging parallel API calls. This approach expedites the retrieval of diverse responses, which are crucial for our consistency checks.

By employing these parallel processing strategies, we ensure that the efficiency of our method is improved while preserving its effectiveness. We believe these optimizations adequately address the concerns about the method's efficiency and demonstrate our commitment to developing a scalable and high-performing framework.

F Prompt Instruction Examples

In this section, we provide these instructions used in our experiments.

```
When responding to my question, please first
evaluate the validity of the information or
the assumption underlying the question. Once
you've established its truth or existence, then
proceed to deliver a detailed explanation or
answer. Prioritize accuracy and fact-checking
before diving into elaboration or conjecture.
```

Table 9: The prompt used to generate CoT answer.

```
Given the question Q, and two potential answers:
answer A in English and answer B in Chinese.
Your task is to determine if the content and
meaning of A and B are equivalent in the
context of answering Q. Consider linguistic
nuances, cultural variations, and the overall
conveyance of information. Respond with a
binary classification. If A and B are
equivalent, output 'True.', otherwise output
'False'
```

Table 10: The instruction for determining whether two QA pairs in different languages are semantically equivalent.

```
Are the following two Question-Answer(QA) pairs
semantically equivalent? Provide your best
guess that it is correct (True or False). Given
ONLY the guess (True or False), no other words
or explanation.
```

Table 11: The instruction for determining whether two QA pairs generated by different models are semantically equivalent.

```
Transform the following question into a query
suitable for information retrieval and then use
the query to find relevant evidence. Follow
these steps for a short but comprehensive
approach:
Original Question:
1. Simplify the question into key terms and
concepts.
2. Remove any elements that are not essential
for a search query.
3. If necessary, rephrase the question to focus
on the most critical aspects for retrieval.
```

Table 12: The prompt used to reformulate search queries.

```
You are an advanced AI with the ability to
process and synthesize information efficiently.
When provided with a question, you should
consider the following aspects in your
response:
1. The specific question asked:
2. The reasoning process:
3. The short answer previously given related
to the question:
4. Evidences that have been retrieved:
Use this information to generate a concise,
accurate, and relevant answer that reflects
the latest understanding of the topic based on
the input provided.
Your response should be clear, direct,
and provide the most up-to-date information
available while maintaining coherence with the
previous discussion points.
```

Table 13: The instruction for repairing hallucinated contents in the outputs of LLMs.