

# Rowen: Adaptive Retrieval-Augmented Generation for Hallucination Mitigation in LLMs

Hanxing Ding

[dinghanxing18s@ict.ac.cn](mailto:dinghanxing18s@ict.ac.cn)

State Key Laboratory of AI Safety,  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Beijing, China

Liang Pang\*

[pangliang@ict.ac.cn](mailto:pangliang@ict.ac.cn)

State Key Laboratory of AI Safety,  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Beijing, China

Zihao Wei

State Key Laboratory of AI Safety,  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Beijing, China

Huawei Shen

State Key Laboratory of AI Safety,  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Beijing, China

Xueqi Cheng

State Key Laboratory of AI Safety,  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Beijing, China

## 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<sup>1</sup>.

## CCS Concepts

- Information systems → Question answering.

## Keywords

Large Language Models, Hallucination Mitigation, Self-Consistency Check

\*Corresponding author

<sup>1</sup>Our Code: <https://github.com/dhx20150812/Rowen>.



This work is licensed under a Creative Commons Attribution 4.0 International License.  
*SIGIR-AP 2025, Xi'an, China*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2218-9/2025/12

<https://doi.org/10.1145/3767695.3769500>

## ACM Reference Format:

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Rowen: Adaptive Retrieval-Augmented Generation for Hallucination Mitigation in LLMs. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2025), December 7–10, 2025, Xi'an, China*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767695.3769500>

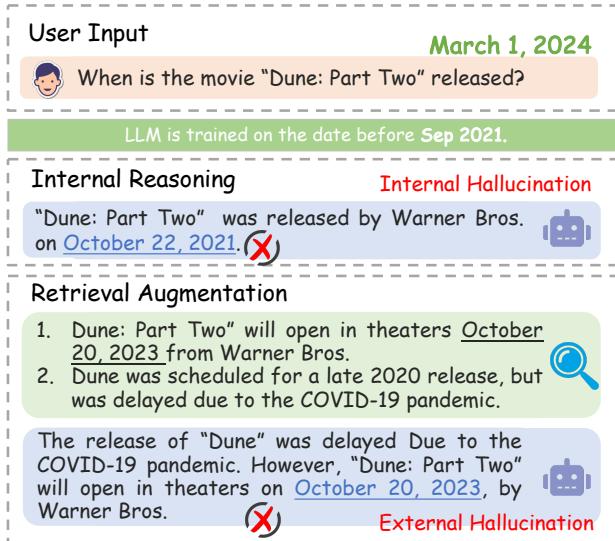
## 1 Introduction

In recent years, large language models (LLMs) have demonstrated impressive abilities in natural language understanding [9, 11], generation [31, 33], and reasoning [4, 36, 47]. 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* [14, 44, 46]. 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 [6, 15, 23, 35] or collaborative refinements involving interactions among multiple models [5, 7], aiming to enhance logical coherence in refined content. Despite their effectiveness, these self-improvement methods may be limited by LLMs' knowledge boundaries [20, 28] or may not fully exploit parametric knowledge [10], leading to what we term *internal hallucination*, illustrated in Figure 1.

Alongside these self-improvement strategies, Retrieval-Augmented Generation (RAG) [3] serves as a complementary method to overcome knowledge limitations. RAG employs a *retrieve-then-read* pipeline [18, 19], integrating relevant documents from external knowledge sources into the LLMs' generation process [3, 10, 21, 34, 41]. 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 [20, 29].

Drawing inspiration from the latest neuroscience research [27], which reveals how the human brain dynamically switches between internal thoughts and external sensations, we introduce a novel



**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*.**

method, termed Rowen (**R**etrieve **o**nly **w**hen 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 [22] and the StrategyQA dataset [8]. 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 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

In this section, we discuss recent works on hallucination detection and mitigation, focusing on uncertainty estimation methods for detection and post-hoc correction for mitigation.

### 2.1 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 [12, 25, 34]. 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 [1, 32, 38]. However, these methods tend to display a high degree of overconfidence when expressing their confidence [30, 38]. (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 [25, 35, 38, 48]. 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.

### 2.2 Post-hoc Correction for Hallucination Mitigation

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 [6, 15, 23, 35, 39] or through collaborative refinements or debates involving multiple models [5, 7]. Despite their strengths, LLMs are sometimes constrained by their knowledge boundaries and the complexity of the reasoning chain, resulting in occasional inaccuracies [20, 24, 28] 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 [3, 10, 21, 26, 34, 40, 41]. 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 [20, 29], 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 [2, 13, 16, 24].

### 3 Methodology

Our objective is to enhance the factuality of LLM responses by integrating parametric and external knowledge. We propose a framework called Rowen (Retrieve only when needed). Initially, we leverage LLMs' Chain-of-Thought (CoT) reasoning to generate 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.

---

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 1:** The prompt used to generate CoT answer.

#### 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  $x_0$  and prioritizing accuracy and fact-checking for answer elaboration, detailed in Table 1. After generating the CoT thought  $t_0$ , we ask  $\mathcal{M}$  to provide a concise answer  $r_0$  for the input query  $x_0$ . Our aim is to ensure high-quality responses are generated from the outset. The answer  $r_0$  is finalized as the ultimate response to the input query  $x_0$  after 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 [46]. Unfortunately, current consistency-based methods fail when LLMs provide consistent yet incorrect answers across different perturbations [43]. 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 Cross-Language / Model Perturbations.** 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  $x_0$  and instruct the model  $\mathcal{M}$  to provide a set of  $k$  semantically equivalent questions  $\mathcal{X} = \{x^1, x^2, \dots, x^k\}$ . 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.

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

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

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

$$\mathbf{r}_{\mathcal{T}}^j = \mathcal{M}(x_{\mathcal{T}}^j), j = 1, \dots, k. \quad (2)$$

**Cross-Model 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 source question:

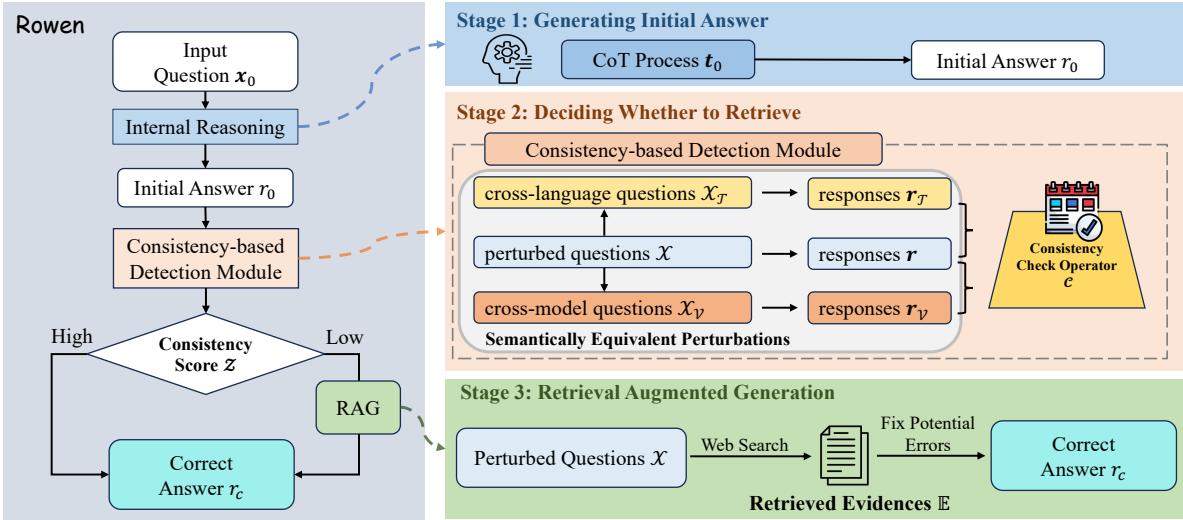
$$\mathbf{r}_{\mathcal{V}}^j = \mathcal{M}_{\mathcal{V}}(x^j), j = 1, \dots, k. \quad (3)$$

**3.2.2 Consistency Score Calculation.** In this step, we utilize the generated questions and answers from all previous stages to calculate a numerical consistency score that captures language-level and model-level cross-checking uncertainties.

---

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



**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.

**Cross-Language Consistency Score.** Let  $C(\cdot, \cdot)$  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" → 1.0, "False" → 0.0}. In our implementation, we leverage the model  $M$  and utilize the prompt provided in Table 2 to implement the cross-language checking operator and calculate the cross-language consistency score  $\mathcal{Z}$  as:

$$\mathcal{Z}_{CL} = \frac{1}{k} \sum_{j=1}^k C(\mathcal{P}^j, \mathcal{P}_T^j), \quad (4)$$

where  $\mathcal{P} = (x, r)$  and  $\mathcal{P}_T = (x_T, r_T)$  denote the QA pairs in the source language and target language, respectively.

---

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

**Cross-Model Consistency Score.** Similar to the cross-language consistency score calculation, we use the prompt in Table 3 to implement the checking operator  $C$  to calculate cross-model consistency score:

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

where  $\mathcal{P}_V = (x_V, r_V)$  denote the QA pairs generated by verifier model  $M_V$ .

**Hybrid Consistency Score.** 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:

$$\mathcal{Z}_{Hybrid} = \mathcal{Z}_{CL} + \alpha * \mathcal{Z}_{CM}, \quad (6)$$

where  $\alpha$  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  $r_0$ . We then introduce a retrieval-augmented generation procedure.

**Searching Relevant Knowledge.** To help the LM  $M$  correct errors, we search for supporting evidence from external sources like online webpages. We first ask the model  $M$  to generate search queries for each paraphrased question in  $X$ . These queries are input into the online search engine to retrieve relevant knowledge, denoted as  $E$ , used for correcting factual errors in  $r_0$ .

**Repairing Hallucinated Contents.** With the retrieved evidence  $E$ , the model reviews the original thought process  $t_0$  and initial answer  $r_0$ . The aim is to identify and correct inaccuracies, producing the refined answer  $r_c$ :

$$r_c = M(x_0, t_0, r_0, E). \quad (7)$$

The corrected answer  $r_c$  serves as the final response to question  $x_0$ .

Models	TruthfulQA			StrategyQA
	GPT-Judge ↑	BLEU ↑	Rouge-L ↑	Accuracy ↑
<i>Vanilla LLMs</i>				
ChatGPT (gpt-3.5-turbo)	47.92	10.17	<b>31.31</b>	61.40
<i>Self-improvement Methods</i>				
CoVe [6]	48.01	12.81	26.52	61.40
Multi-agent Debate [7]	50.83	3.94	21.05	65.73
Self-Reflection [15]	42.99	3.86	18.18	62.40
<i>Retrieval-augmented Methods</i>				
Factool [3]	34.50	1.34	12.22	67.20
Detect-and-Mitigate [34]	49.98	3.17	18.59	56.94
<i>Adaptive Retrieval Methods</i>				
FLARE [16]	45.04	11.59	26.83	61.19
Adaptive-Retrieval [24]	45.55	8.87	26.75	62.50
Self-RAG [2]	40.36	4.36	21.28	58.40
Adaptive-RAG [13]	46.02	10.29	26.24	68.50
LUQ [42]	55.08	5.79	21.44	71.00
<i>Our Framework</i>				
Rowen-CL	57.39	7.60	24.16	74.00
Rowen-CM	56.29	6.85	22.36	72.40
Rowen-Hybrid	<b>59.34</b>	<b>15.27</b>	31.15	<b>75.60</b>

**Table 4: 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.**

## 4 Experimental Setup

In this section, we outline the experimental setup of Rowen, detailing the datasets utilized, the evaluation metrics employed, the baseline methods considered, and the implementation specifics.

### 4.1 Datasets and Evaluation Metrics

We evaluate the hallucination mitigation performance on the following datasets:

**TruthfulQA.** We use the TruthfulQA dataset [22] to evaluate the ability of LLMs in generating truthful responses [17, 45]. 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<sup>2</sup>. 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 [8] comprises crowdsourced yes / no questions that require multi-step reasoning for accurate answers. We follow previous work [16] to randomly sample 500 examples due to the cost consideration of running experiments. We also follow the settings of Wei et al. [37] 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.

<sup>2</sup><https://github.com/sylinrl/TruthfulQA>

### 4.2 Baseline Methods

We consider the following methods as our baselines: (1) Vanilla LLMs, such as ChatGPT. (2) Self-improvement methods: **CoVe** [6] generates verification questions to self-analyze potential errors, systematically addressing each question to refine the baseline response. **Self-Reflection** [15] presents an interactive self-reflection methodology that incorporates knowledge acquisition and answer generation. **Multi-agent Debate** [7] utilize multiple LM agents to debate their individual responses over multiple rounds to arrive at a common final answer. (3) Retrieval-augmented methods: **Factool** [3] leverages various tools to gather evidence about the factuality of the generated content. **Detect and Mitigate** [34] actively detects hallucinations during generation by identifying potential hallucination through the logit output values of LLMs. (4) Adaptive retrieval methods: **FLARE** [16] adopts an active retrieval strategy that only retrieves when LLMs generate low probability tokens. **Adaptive-Retrieval** [24] only retrieves when necessary based on a pre-defined threshold for entity popularity. **Self-RAG** [2] trains a single arbitrary LM that adaptively retrieves passages on-demand. **Adaptive-RAG** [13] trains a query-complexity classifier to decide when to retrieve based on question complexity. **LUQ** [42] is a sampling-based uncertainty quantification for long text.

### 4.3 Implementation Details

**LLMs.** In our experiments, we validate Rowen using the GPT-3.5 language model, specifically the gpt-3.5-turbo-0613 version<sup>3</sup>.

<sup>3</sup><https://api.openai.com/v1/chat/completions>

We re-implement all baselines, except Self-RAG, using GPT-3.5 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<sup>4</sup>. This decision is made considering the substantial cultural disparities between the two languages, which can enhance the model's capability to detect semantic inconsistencies when responding to identical questions. For the cross-model detection module, we adopt Qwen-Max-0428<sup>5</sup>, a large instruction-tuned model for chat service, as the verifier LM<sup>6</sup>. We execute all experiments on 8 NVIDIA A800 80G GPUs.

*Retrieval Module.* Following Chern et al. [3], we utilize the Google Search API offered by Serper<sup>7</sup> to search the top pages and extract the most pertinent search snippets from the API's response. Subsequently, we parse the response to acquire various types of snippets, including answer boxes, knowledge graphs, and organic search results.

## 5 Experimental Results

In this section, we evaluate the effectiveness of Rowen for hallucination mitigation. In particular, we aim to answer the following research questions:

- **RQ1:** How does Rowen perform compared to existing strong methods for hallucination mitigation?
- **RQ2:** Is the proposed multilingual detection module superior to other hallucination detection methods?
- **RQ3:** How well does Rowen generalize to new datasets and models?
- **RQ4:** How do different hyperparameters affect the performance of Rowen?
- **RQ5:** Does Rowen eliminate both types of hallucinations as expected?
- **RQ6:** Does Rowen reduce unnecessary retrieval calls?

### 5.1 Main Results (RQ1)

We evaluate the effectiveness of Rowen on the TruthfulQA and StrategyQA datasets. Table 4 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 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

<sup>4</sup>We study the impact of the choice of target language in § 5.4 Table 7.

<sup>5</sup><https://qwenlm.github.io/blog/qwen-max-0428>

<sup>6</sup>We also study the impact of the choice of verifier LM in § 5.4 Table 8.

<sup>7</sup><https://serper.dev/>

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 (RQ2)

To validate the effectiveness of our proposed 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 [25] that measures information consistency between the different responses to determine hallucinations. (3) Consistency-based method [43], SAC<sup>3</sup>, that evaluates semantic-aware cross-check consistency, building upon the foundation of self-consistency principles.

Based on the results in Table 5, it is evident that logits-based methods perform moderately well in detecting hallucinations. Specifically, the entropy-based estimation exhibits superior performance. Besides, the combination of different variants in SelfCheckGPT leads to a slight performance improvement. Additionally, it is worth noting that SAC<sup>3</sup> 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<sup>3</sup>. We also notice that Rowen achieves notable hallucination mitigation with a minimal number of retrieval calls. This underlines the superior efficiency of our adaptive retrieval module.

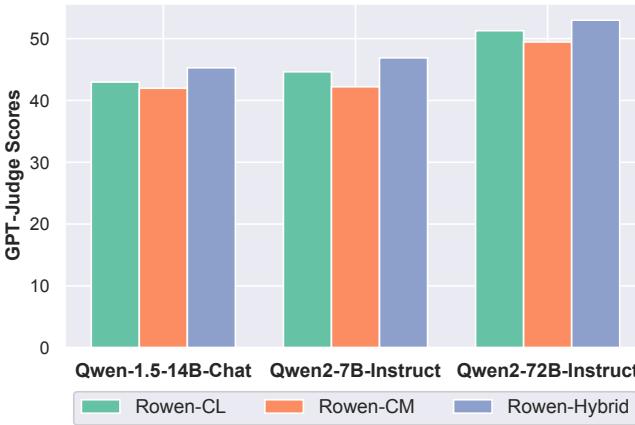
### 5.3 Scalability of Rowen (RQ3)

*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<sup>8</sup>. These models are chosen for the cross-language detection model  $\mathcal{M}$  due to their strong capabilities in

<sup>8</sup><https://huggingface.co/Qwen>

Models	TruthfulQA			StrategyQA		
	GPT-Judge $\uparrow$	BLEU $\uparrow$	Rouge-L $\uparrow$	Ratio(%)	Accuracy $\uparrow$	Ratio(%)
<i>LLaMa2-7B</i>						
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
<i>SelfCheckGPT</i>						
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
<i>Consistency</i>						
SAC <sup>3</sup> -Q	51.02	7.90	28.00	24.5	65.50	24.0
SAC <sup>3</sup> -all	52.22	9.37	29.56	20.8	67.00	24.5
Rowen-Hybrid	<b>59.34</b>	<b>15.27</b>	<b>31.15</b>	23.0	<b>75.60</b>	20.0

**Table 5: Performance comparison of applying other hallucination detection methods in adaptive retrieval scenarios. We also report the ratio of retrieval conducted by each method.**



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

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  $M_V$ , we choose to use Llama-3-8B-Instruct<sup>9</sup>.

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, nearly matching the baseline results on ChatGPT. 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

<sup>9</sup><https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

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	<b>39.98</b>	<b>57.31</b>	<b>69.04</b>	<b>78.46</b>

**Table 6: Hallucinations mitigation performance on the NQ dataset and TriviaQA dataset.**

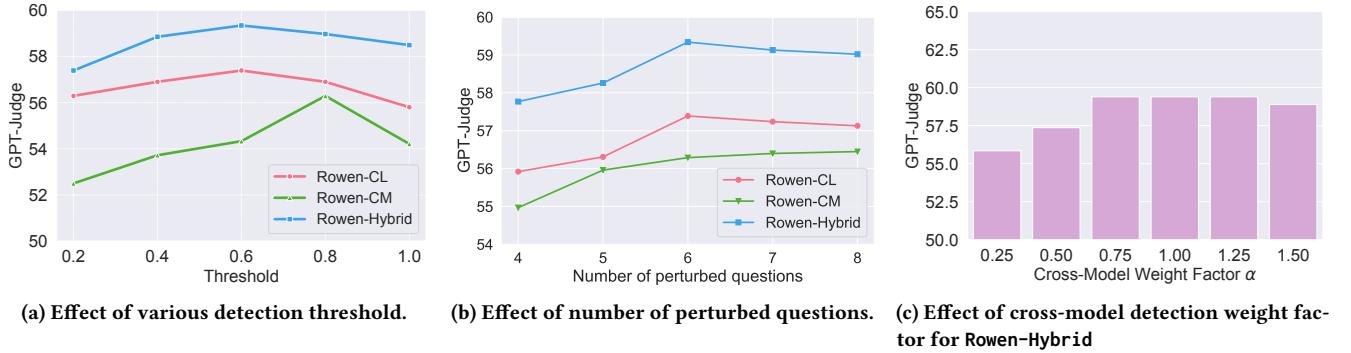
of intermediate length, namely TriviaQA and Natural Questions, given the diverse answer lengths in TruthfulQA (long answers) and StrategyQA (binary responses). We conduct comparisons against two strong adaptive retrieval baselines, reporting the metrics of EM (Exact Match) and F1 score on their dev sets. The experimental results are shown in Table 6.

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 enhancement over baseline methods.

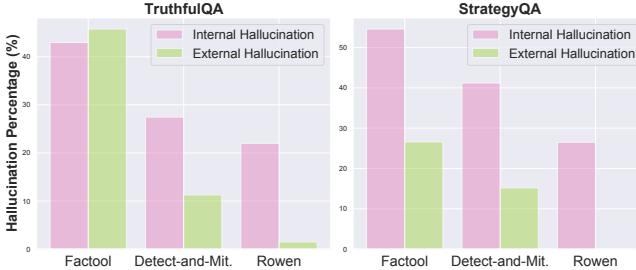
#### 5.4 Impact of Hyper-parameters in Rowen (RQ4)

We analyze the impact of key factors for hallucination mitigation on the TruthfulQA dataset, including detection thresholds, numbers of perturbed questions, and cross-model detection weights, as shown in Figure 4.

*Detection Threshold.* We experiment with various consistency thresholds to investigate their impact on hallucination mitigation effectiveness. As shown in Figure 4(a), we observe significant performance improvement when the detection threshold increases from 0.2 to 0.6, as this identifies more hallucinations and enhances



**Figure 4: Impact of key factors for hallucination mitigation on the TruthfulQA dataset.**



**Figure 5: The percentage of internal and external hallucinations across different methods on the both dataset.**

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.

*Number of Perturbed Questions.* We also study the effect of the number of perturbed question on hallucination mitigation and the results are shown in Figure 4(b). We observe that the performance of Rowen improves with an increasing number of question samples, yet the performance gain gradually diminishes after surpassing 6 question samples. This indicates that, in practice, using 6 question samples could achieve reasonably good performance at a relatively low computational cost.

*Cross-Model Detection Weight.* We also examine the impact of different cross-model detection weight for Rowen-Hybrid. Figure 4(c) 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 the accuracy of hallucination detection. However, when the weight factor exceeds 1.0, performance declines, with scores dropping to around 56 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.

*Impact of the Choice of Target Language.* To investigate how different language pairs affect the effectiveness of hallucination

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

**Table 7: Comparison of hallucination mitigation performance for Rowen-CL under different language choices on the TruthfulQA dataset.**

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

**Table 8: Comparison of hallucination mitigation performance for Rowen-CM under different verifier LM on the TruthfulQA dataset.**

mitigation for Rowen-CL, we conduct experiments on the TruthfulQA dataset with various target languages. The results of these experiments are presented in Table 7. 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 8 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

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	<b>59.34</b>	8	<b>75.60</b>	6

**Table 9: Analysis on LLM calls efficiency and hallucination mitigation performance across different methods.**

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	<b>0.9</b>	1.4
Rowen-Hybrid	1.5	<b>0.5</b>

**Table 10: Statistics on the average number of retrieval calls to answer each question.**

achieved 52.50% and 54.09%, respectively. These findings underscore the critical role of verifier LM selection in enhancing the effectiveness of hallucination mitigation tasks.

### 5.5 Quantitative Analysis (RQ5)

To verify whether Rowen effectively reduces internal and external hallucinations, we present a comparative analysis of the prevalence of both types 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 5 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.

### 5.6 Analysis of Computation Cost (RQ6)

*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 9, 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.

*Efficiency Analysis of Retrieval Calls.* To verify the retrieval efficiency of Rowen, in Table 10, 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. Factool generates the most retrieval API calls due to verifying each claim. Detect-and-Mitigate and FLARE identify low-confident concepts in LLM outputs and call search APIs with fewer API calls. Adaptive-RAG underestimates the difficulty of adversarial questions in the TruthfulQA dataset, resulting in poor truthfulness scores despite using the minimum number of retrievals. Rowen-Hybrid, which retrieves information only for uncertain responses, excels in both retrieval efficiency and factual accuracy, showcasing its superiority over other RAG methods. This demonstrates the superiority of Rowen-Hybrid over other RAG methods in terms of both efficiency and effectiveness.

## 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 two datasets demonstrate the effectiveness of our approach. We also conduct detailed analytical experiments to underscore the impact of each module and the choices of hyperparameters. 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.

## Acknowledgement

This work was supported by the Strategic Priority Research Program of the CAS under Grants No.XDB0680302, the National Natural Science Foundation of China (NSFC) under Grants No. 62276248, the Key Research and Development Program of Xinjiang Uyghur Autonomous Region Grant No. 2024B03026, the Beijing Nova Program under Grants No. 20250484765, and the Youth Innovation Promotion Association CAS under Grants No. 2023111.

## References

- [1] Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do Language Models Know When They’re Hallucinating References? *CoRR* abs/2305.18248 (2023).
- [2] Akari Asai, Zequi Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *CoRR* abs/2310.15111 (2023).
- [3] 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 (2023).
- [4] 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 (2023).
- [5] 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*. Association for Computational Linguistics, 12621–12640.
- [6] Shehzad Dhuliaiwala, 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 (2023).
- [7] 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 (2023).
- [8] 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 (2021).

- [9] 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.
- [10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *CoRR* abs/2311.05232 (2023).
- [11] 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. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *CoRR* abs/2305.08322 (2023).
- [12] Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. *CoRR* abs/2307.10236 (2023).
- [13] Soyeon 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 (2024).
- [14] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38.
- [15] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 1827–1843.
- [16] 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*. Association for Computational Linguistics, 7969–7992. <https://aclanthology.org/2023.emnlp-main.495>
- [17] 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 (2022).
- [18] 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. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Online.
- [19] 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*.
- [20] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large Language Models with Controllable Working Memory. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada.
- [21] Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models. *CoRR* abs/2305.14623 (2023).
- [22] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Dublin, Ireland.
- [23] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alom, 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 (2023).
- [24] 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. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada.
- [25] 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 (2023).
- [26] 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*. ACM, 751–760.
- [27] Craig Poskær and Mariam Aly. 2023. Switching between External and Internal Attention in Hippocampal Networks. *Journal of Neuroscience* 43, 38 (2023), 6538–6552. doi:10.1523/JNEUROSCI.0029-23.2023
- [28] 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 (2023).
- [29] 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 (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 31210–31227.
- [30] Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuxiang 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* (2024).
- [31] 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](https://github.com/tatsu-lab/stanford_alpaca).
- [32] 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 (2023).
- [33] 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 Lampe. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023).
- [34] 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 (2023).
- [35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. 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.
- [36] Yuhang Wang, Dongyuan Lu, Chao Kong, and Jitao Sang. 2023. Towards Alleviating the Object Bias in Prompt Tuning-based Factual Knowledge Extraction. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada.
- [37] 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*.
- [38] 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 (2023).
- [39] 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 (2023).
- [40] 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 (2024).
- [41] 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 (2023).
- [42] Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text Uncertainty Quantification for LLMs. *CoRR* abs/2403.20279 (2024).
- [43] Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Kumar Sriran. 2023. SAC<sup>3</sup>: 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*. Association for Computational Linguistics, 15445–15458.
- [44] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How Language Model Hallucinations Can Snowball. *CoRR* abs/2305.13534 (2023).
- [45] Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023. Alleviating Hallucinations of Large Language Models through Induced Hallucinations. *CoRR* abs/2312.15710 (2023).
- [46] 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. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR* abs/2309.01219 (2023).
- [47] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. 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.
- [48] Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada.