

KNOWHALU: HALLUCINATION DETECTION VIA MULTI-FORM KNOWLEDGE BASED FACTUAL CHECKING

Anonymous authors

Paper under double-blind review

ABSTRACT

As large language models (LLMs) become increasingly integral to a wide array of applications, ensuring the factual accuracy of their outputs and mitigating hallucinations is paramount. Current approaches, which primarily rely on *self-consistency checks* or *post-hoc fact-checking*, often fall short by disregarding the nuanced structure of queries and the diverse forms of contextual knowledge required for accurate response generation. To address these shortcomings, we introduce KnowHalu (pronounced “No Halu”), the first multi-form knowledge-based hallucination detection framework. We also introduce a new category of hallucinations, off-target hallucinations, which occur when responses are factually accurate but irrelevant or nonspecific to the query (e.g., answering “What’s the primary language in Barcelona?” with “European language”). In particular, KnowHalu employs a rigorous two-phase process to detect hallucinations. In the first phase, it isolates off-target hallucinations by analyzing the semantic alignment between the response and the query. In the second phase, it conducts a novel multi-form knowledge-based fact-checking through a comprehensive pipeline of reasoning and query decomposition, knowledge retrieval, knowledge form optimization, judgment generation, and judgment aggregation. Extensive evaluations demonstrate that KnowHalu significantly surpasses state-of-the-art (SOTA) baselines across diverse tasks, achieving over 15% improvement in question answering (QA) and 6% in summarization tasks when applied to the same underlying LLM. These results underscore the effectiveness and versatility of KnowHalu, setting a new benchmark for hallucination detection and paving the way for safer and more reliable LLM applications.

1 INTRODUCTION

Significant advancements have been achieved in the field of Natural Language Processing (NLP) with the advent of Large Language Models (LLMs). While these models excel in generating coherent and contextually relevant text, they are prone to ‘*hallucinations*’ — generating plausible but factually incorrect or unspecific information (Bang et al., 2023). This poses a considerable challenge, especially in applications demanding high factual accuracy, such as medical records analysis (Singhal et al., 2023), finance (Wu et al., 2023; Yang et al., 2023), and drug design (Vert, 2023; Savage, 2023).

To mitigate or detect hallucinations in LLMs, a series of approaches have been explored. For instance, self-consistency-based approaches detect hallucinations by identifying contradictions in responses that are stochastically sampled from the LLMs in response to the same query (Wang et al., 2022; Manakul et al., 2023; Mündler et al., 2023). Other approaches detect hallucinations by probing LLMs’ hidden states (Azaria & Mitchell, 2023) or output probability distributions (Manakul et al., 2023). These methods do not incorporate external knowledge and are thus limited by LLMs’ internal knowledge. Post-hoc fact-checking approaches have been recently shown to be effective even when LLMs’ internal knowledge proves inadequate and achieved SOTA hallucination detection (Peng et al., 2023; Semnani et al., 2023). However, due to the limitation of LLM reasoning capabilities, even if the extracted knowledge is correct, the models may still struggle to perform factual checking accurately, especially with complex queries or logic, such as multi-hop queries or those involving multiple factual assertions. Thus, how to fully leverage the inherent reasoning capabilities of the model is important.

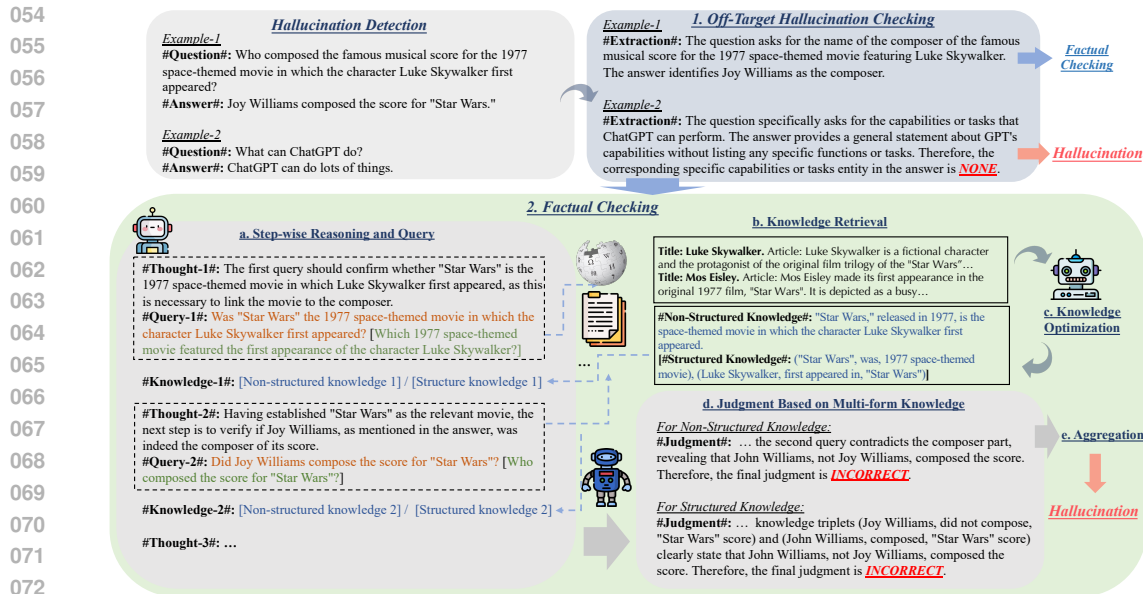


Figure 1: Overview of KnowHalu. The hallucination detection process starts with “Off-Target Hallucination Checking”, focusing on the early identification of off-target hallucinations by scrutinizing the specificity of the answers. For potential fabrication hallucinations, KnowHalu then provides a comprehensive “Factual Checking”, which consists of five steps: (a) “Step-wise Reasoning and Query” breaks down the original query into step-wise reasoning and sub-queries for detailed factual checking; (b) “Knowledge Retrieval” retrieves unstructured knowledge via RAG and structured knowledge in the form of triplets for each sub-query; (c) “Knowledge Optimization” leverages LLMs to summarize and refine the retrieved knowledge into different forms; (d) “Judgment Based on Multi-form Knowledge” employs LLMs to critically assesses the answer to sub-queries, based on each form of knowledge; (e) “Aggregation” provides a further refined judgment by aggregating predictions based on different forms of knowledge.

Recognizing this gap, our work proposes a novel multi-phase hallucination detection mechanism, KnowHalu (pronounced as “No halu”), the overall framework of which is presented in Figure 1. In particular, we first perform the *off-target hallucination checking*, where the answer indeed provides a fact but it is not a helpful response to the question, such as the answer “ChatGPT can do lots of things” for the question “What can ChatGPT do?” This type of hallucination has been extensively discussed in various works on hallucination in LLMs (Huang et al., 2023; Zhang et al., 2023b; Li et al., 2023). Despite this, the step of identifying off-target hallucinations remains critically underrepresented in current detection methodologies. Existing approaches often fail to discern answers that are factually correct but irrelevant to the posed questions. We then perform a *step-wise decomposition of queries*, which enables targeted retrieval of external knowledge pertinent to each logical step. For each decomposed logical step, we will perform the *multi-form knowledge based factual checking*, leveraging both the unstructured knowledge (e.g., normal semantic sentences) and structured knowledge (e.g., object-predicate-object triplets). This multi-form knowledge analysis captures a comprehensive spectrum of factual information, enhancing the reasoning capability of LLMs and ensuring a robust and thorough validation of each piece of retrieved knowledge. Finally, we perform the reasoning step by composing the step-wise factual checking results together and guide the LLMs to make the final judgment by providing related demonstrations. Our multi-step and multi-form knowledge based factual checking not only improves the verification accuracy but also enhances the model’s ability to handle intricate and layered queries.

Our extensive experimental evaluations reveal that KnowHalu significantly outperforms state-of-the-art (SOTA) baselines in detecting hallucinations. The experiments, conducted across diverse datasets and tasks, demonstrate not only the high accuracy of our method in factual verification but also its versatility in handling various types of queries.

In summary, we make the following key contributions:

- We introduce KnowHalu, a novel approach with two main phases (off-target hallucination detection and multi-step factual checking) for detecting hallucinations in texts generated by LLMs, leveraging multi-form knowledge for factual checking. In particular, we define the categories of *off-target hallucinations* for the first time.

Table 1: Different Types of Hallucinations for Question-Answer (QA) Task. (Row 1-5 showcase off-target hallucinations, and the last row shows fabrication hallucinations.)

Type of Hallucination	Category	Description	Example
Off-Target Hallucination	Vague or Broad Answers	Answers that are too general and do not address the specificities of the question.	#Question#: What is the primary language in Barcelona? #Answer#: European languages.
	Parroting or Reiteration	The response simply echoes part of the question without adding new or relevant information.	#Question#: What is the title of John Steinbeck’s novel about the Dust Bowl? #Answer#: Steinbeck wrote about the Dust Bowl.
	Misinterpretation of Question	Misunderstanding the question, leading to an off-topic or irrelevant response.	#Question#: What is the capital of France? #Answer#: France is in Europe.
	Negation or Incomplete Information	Pointing out what is not true without providing correct information.	#Question#: Who is the author of “Pride and Prejudice”? #Answer#: Not written by Charles Dickens.
	Overgeneralization or Simplification	Overgeneralizing or simplifying the answer.	#Question#: What types of movies has Christopher Nolan worked on? #Answer#: Biographical film.
Factual Hallucination	Fabrication	Introducing false details or assumptions not supported by the truth of facts	#Question#: When was “The Sound of Silence” released? #Answer#: 1966 (<i>Incorrect. The correct answer is 1964</i>)

- We are the first work to explore the influence of both the formulations of the queries and the forms of knowledge used for detecting hallucinations, highlighting our novel exploration into factors critical for improving hallucination detection accuracy.
- We further propose a verification mechanism where collections of facts are checked interdependently instead of in parallel, and an aggregation methodology based on the prediction results from different forms of knowledge to further reduce hallucinations in judgment itself. Our experiments show that our method achieves at least a **15.50%** improvement in hallucination detection on the question-answering task, and yields at least an additional **6.20%** improvement in the text summarization task when compared to the SOTA baselines using the same underlying LLMs.

2 RELATED WORK

Hallucination of LLMs. Hallucination in the LLM literature generally refers to LLMs generating nonfactual, irrelevant, or unspecific outputs. Such phenomena have been observed in a variety of tasks (Huang et al., 2023; Zhang et al., 2023b; Ji et al., 2023), such as translation (Lee et al., 2018), dialogue (Balakrishnan et al., 2019), summarization (Durmus et al., 2020), and question answering (Sellam et al., 2020). We summarize and describe the hallucination types for the QA task covered in this paper in Table 1. Benchmarks have been proposed to evaluate the extent to which LLMs hallucinate, such as Honovich et al. (2021), Huang et al. (2021), and Li et al. (2023).

Hallucination detection and mitigation. Methods that attempt to detect and mitigate hallucination without fact-checking include methods based on chain-of-thought (Wei et al., 2022; Huang et al., 2022; Dhuliawala et al., 2023), methods based on self-consistency (Wang et al., 2022; Huang et al., 2022; Mündler et al., 2023; Manakul et al., 2023), and methods that probe LLMs’ hidden states (Azaria & Mitchell, 2023) or output probability distributions (Manakul et al., 2023). Since these methods do not augment LLMs with external knowledge, they usually struggle when LLMs’ internal knowledge is inadequate. On the other hand, fact-checking-based methods (Roller et al., 2020; Komeili et al., 2021; Shuster et al., 2022a; 2021; 2022b; Izacard et al., 2022; Li et al., 2023; Semnani et al., 2023) rely on retrieved knowledge to prevent hallucinations in LLMs. However, these methods are often limited by the ways they retrieve knowledge and utilize the retrieved knowledge. For instance, Li et al. (2023) employ only a single query with the knowledge for detecting hallucinations in inputs that inherently necessitate multi-hop reasoning, which would benefit from a step-by-step query process. Semnani et al. (2023) instead introduces a robust framework initially designed to deliver fact-checked responses. This innovative method employs a comprehensive process that entails generating queries to fetch information from Wikipedia, summarizing and filtering the retrieved content, and then crafting a response informed by this vetted knowledge. However, when adapting WikiChat for hallucination detection, although it effectively uses retrieved knowledge for fact-checking in parallel, it occasionally neglects the coherence of the facts being verified, potentially leading to inaccuracies. Besides, during the fact-checking phase, WikiChat directly retrieves the evidence with the claim, which may not be effective when the claim itself is the result of hallucination as shown in Appendix B. Our approach,

in contrast, derives step-by-step queries with refined formulations and provides LLMs with either structured or unstructured knowledge for consecutive fact-checking, leading to better retrieval and higher knowledge utilization.

3 KNOWHALU

KnowHalu provides a systematic hallucination detection framework based on multi-step query and reasoning. It starts with “*Off-Target Hallucination Checking*” to pinpoint non-specific hallucinated answers, followed by “*Factual Checking*”, verifying the correctness of the answer through a multi-step process based on different forms of knowledge.

3.1 OFF-TARGET HALLUCINATION CHECKING.

Current approaches in hallucination detection can mainly identify fabricated hallucinations, i.e., answers with mismatching facts (Manakul et al., 2023; Peng et al., 2023; Semnani et al., 2023). Yet, hallucinations also emerge in other types, as outlined in Table 1, which extend beyond simple factual inaccuracies. A typical trait of these off-target hallucinations is their factual correctness while a lack of direct relevance and helpfulness to the original query. For instance, given a question, “*What is the primary language in Barcelona?*”, the hallucinated answer “*European languages*” is factually correct but fails in providing specific answers, thus being off-target, which is important for the quality of real-world LLMs. Note that we mainly focus on the potential hallucinations of LLMs in practice, so if the model rejects to answer a question, it will not be viewed as hallucinations. For example, if the answer to the question “*What is the capital of France?*” is “*France is in Europe,*” it is categorized as an off-target hallucination due to a misinterpretation of the question. Conversely, responses like “*Sorry, the question is hard*” or “*Sorry, I don’t know*” are not considered hallucinations in our work. These replies correctly interpret the question and reject to provide an answer, and thus do not fall under the “*Misinterpretation of Question*” category shown in Table 1. Instead, they acknowledge the difficulty in providing a specific answer, thereby offering a useful response. In fact, we should encourage models to refuse to answer questions that they do not know.

To bridge this gap, in KnowHalu, we first introduce a “*Off-Target Hallucination Checking*” phase, as depicted in the first row of Figure 1. This step aims to identify off-target hallucinations. A straightforward approach might involve prompting the language model to identify such hallucinations based on provided examples. However, this often results in high false positives, inaccurately flagging correct answers as hallucinations. To counter this, we address this challenge by solving an *extraction* task, which prompts the language model to extract specific entity or details requested by the original question from the answer. If the model fails to extract such specifics, it returns “*NONE*”. This extraction-based specificity check is designed to reduce false positives while effectively identifying off-target hallucinations. Responses yielding “*NONE*” are directly labeled as hallucinations, and the remaining generations will be sent to the next phase for further factual checking. Examples of instructions in the extraction task for each type of off-target hallucination are provided in Appendix A.1.

3.2 FACTUAL CHECKING

The *Factual Checking* phase consists of five key steps: (a) *Step-wise Reasoning and Query* breaks down the original query into sub-queries following the logical reasoning process and generates different forms of sub-queries; (b) *Knowledge Retrieval* retrieves knowledge for each sub-query based on existing knowledge database; (c) *Knowledge Optimization* summarizes and refines the retrieved knowledge, and maps them to different forms, such as unstructured knowledge (object-replicate-object triplet); (d) *Judgment Based on Multi-form Knowledge* assesses the answer for each sub-query based on multi-form knowledge; and (e) *Aggregation* combines insights of judgments based on different forms of knowledge and makes a further refined judgment.

a. Step-wise Reasoning and Query In this step, we aim to break down the original query into local sub-queries following the reasoning logic, and we will retrieve knowledge for each sub-query sequentially (details in steps b and c), which is similar with ReAct (Yao et al., 2023), to cumulatively perform factual checking along the reasoning process. One main challenge here is “*how do we craft precise and effective sub-queries, which can accurately retrieve the relevant knowledge at each*

216 *logical step?*” To address this challenge, we identify two key factors that significantly enhance the
 217 accuracy of knowledge retrieval for factual checking: (1) continuous and direct (one-hop) queries,
 218 and (2) the formulation of queries. We will analyze these two key factors, which lead to our design
 219 choice below.

220 First, we observe that multi-hop queries (e.g., Example-1 in Figure 1) often struggle to retrieve
 221 specific and related knowledge due to their inherently complex and ambiguous context. On the other
 222 hand, the one-hop queries are effective to retrieve the most relevant and useful knowledge. Thus, we
 223 decompose the original query into sequence of simpler and direct one-hop sub-queries following the
 224 logical reasoning process, which significantly enhance the retrieval accuracy for factual checking.
 225 Concretely, this iterative querying process starts by interpreting the original query as a series of
 226 logical steps to form sub-queries accordingly, and then perform factual checking for each sub-query.
 227 For instance, based on the example in Figure 1, the initial query first confirms whether “*Star Wars*” is
 228 indeed the 1977 space-themed movie featuring Luke Skywalker. This step is crucial for connecting
 229 the movie to its composer. Subsequent queries delve deeper, examining the accuracy of other specific
 230 details provided in the answer, such as the composer’s identity. The queries are intricately connected,
 231 each building upon the knowledge obtained from the previous one. This iterative process continuous
 232 to generate subsequent queries based on newly acquired knowledge, until the logical reasoning
 233 process is completed.

234 Second, we observe that the formulation of queries also plays a critical role for the final factual
 235 checking. In particular, queries with correct details will lead to high-quality knowledge retrieval;
 236 while queries with incorrect or unrelated entities may lead to poor and irrelevant knowledge retrieval.
 237 As a result, we propose two query formulations: *General Query* and *Specific Query*. The General
 238 Query avoids mentioning specific, potentially hallucinated details (e.g., “*Who composed the score for*
 239 *’Star Wars’?*”); the Specific Query is constructed based on the key entities mentioned in the answers
 240 (e.g., “*Did Joy Williams compose the score for ’Star Wars’?*”), as shown in Figure 1. More concrete
 241 examples illustrating the impact of these two query formulations on the retrieval outcomes for both
 242 correct and hallucinated details can be found in Appendix B.

243 In our experiments shown in Section 5.2, we examine how different query formulations —*general* and
 244 *specific* — affect knowledge retrieval and the accuracy of final hallucination detection, by leveraging
 245 only one or both query formulations. More detailed prompts are provided in Appendix A.2.

246 **b. Knowledge Retrieval** We perform knowledge retrieval for each sub-query generated from step
 247 a. In particular, for QA tasks, we adopt the Retrieval-Augmented Generation (RAG) framework
 248 developed based on Wikipedia knowledge base (Semnani et al., 2023), and the retrieval is based on
 249 ColBERT v2 (Santhanam et al., 2022b) and PLAID (Santhanam et al., 2022a). We retrieve Top-K
 250 relevant passages for each sub-query, each formatted as “Title: ..., Article: ...”. In addition, when
 251 we perform knowledge retrieval for summarization tasks, we treat the source document itself as the
 252 knowledge base for retrieval. In particular, we first segment the original documents into distinct text
 253 chunks. We then embed the sub-queries and text chunks into dense vectors using a text encoder.
 254 Similarly, we will retrieve the Top-K text chunks that exhibit the highest cosine similarity with the
 255 input sub-queries.

256 **c. Knowledge Optimization** The knowledge retrieved for each sub-query is usually a long and
 257 verbose passage with distracting irrelevant details. Thus, this step aims to leverage another LLM
 258 to distill useful information and optimize clear and concise knowledge, which could be in different
 259 forms. In particular, we propose two forms of knowledge, *unstructured* and *structured* knowledge.
 260 The unstructured knowledge represents the texts retrieved from given knowledge bases in a concise
 261 way, such as “*’Star Wars,’ released in 1977, is the space-themed movie in which the character Luke*
 262 *Skywalker first appeared.*” Since the unstructured text may not be precise for logical reasoning, we
 263 also retrieve structured knowledge as object-predicate-object triplets, such as (“*Star Wars*”, *was*,
 264 *1977 space-themed movie*) and (*Luke Skywalker, first appeared in, “Star Wars”*) (examples of our
 265 demonstrations are in Appendix A.3). Such multi-form knowledge will effectively assist LLMs to
 266 perform logical reasoning and final factual checking. In addition, if a query retrieves no relevant
 267 knowledge, the LLM is instructed to respond with “*No specific information is available*”.

268 **d. Judgment Based on Multi-form Knowledge** After obtaining the retrieved multi-form knowl-
 269 edge for sub-queries, we gather #*Query*# and #*Knowledge*# and present them to another LLM for
 hallucination judgment. The #*Judgment*# assesses the sub-query and its corresponding knowledge
 sequentially to ascertain if there is any contradiction to verification each detail in the answer. If there

Table 2: Performance of different methods for hallucination detection in QA task. Results of methods using external ground truth knowledge (i.e., knowledge provided by HaluEval) are shown inside the parentheses, and results generated based on Wiki knowledge are shown outside the parentheses.

Model	Method	TPR (%)	TNR (%)	Avg Inconclusive Rate (%)	Avg Acc (%)
GPT-4	Zero-Shot CoT	68.3	61.8	–	65.05
Starling-7B	SelfCheckGPT (Manakul et al., 2023)	89.7	30.3	–	60.00
	HaluEval (Vanilla) (Li et al., 2023)	33.2	80.3	–	56.75
	HaluEval (CoT) (Li et al., 2023)	68.7	26.0	–	47.35
	HaluEval (Knowledge) (Li et al., 2023)	33.0 (82.0)	60.3 (40.0)	–	46.65 (61.00)
	Self Consistency (Wang et al., 2022)	37.5 (80.1)	58.6 (41.8)	–	48.05 (60.95)
	KnowHalu (Structured)	68.1 (67.8)	65.6 (83.1)	10.50 (8.45)	66.85 (75.45)
	KnowHalu (Unstructured)	68.2 (72.4)	69.9 (85.9)	6.75 (4.85)	69.05 (79.15)
	KnowHalu (Aggregation)	68.7 (72.7)	75.9 (88.7)	6.75 (4.85)	72.30 (80.70)
Mistral-7B	SelfCheckGPT (Manakul et al., 2023)	67.9	45.9	–	56.90
	HaluEval (Vanilla) (Li et al., 2023)	40.9	41.3	–	41.10
	HaluEval (CoT) (Li et al., 2023)	48.0	41.1	–	44.55
	HaluEval (Knowledge) (Li et al., 2023)	50.9 (55.1)	47.9 (11.1)	–	49.40 (33.10)
	Self Consistency (Wang et al., 2022)	51.9 (55.0)	47.6 (10.9)	–	49.75 (32.95)
	KnowHalu (Structured)	54.0 (59.0)	67.3 (82.6)	12.35 (9.45)	60.65 (70.80)
	KnowHalu (Unstructured)	63.5 (62.5)	58.4 (82.2)	11.70 (6.90)	60.95 (72.35)
	KnowHalu (Aggregation)	60.7 (63.3)	65.4 (85.1)	10.80 (6.20)	63.05 (74.20)
GPT-3.5	WikiChat (Semnani et al., 2023)	16.0	82.2	–	49.10
	SelfCheckGPT (Manakul et al., 2023)	67.8	61.8	–	64.80
	HaluEval (Vanilla) (Li et al., 2023)	44.1	55.2	–	48.15
	HaluEval (CoT) (Li et al., 2023)	66.5	21.6	–	44.05
	HaluEval (Knowledge) (Li et al., 2023)	34.4 (38.1)	71.7 (75.7)	–	53.05 (56.90)
	Self Consistency (Wang et al., 2022)	37.1 (33.6)	64.7 (76.3)	–	50.90 (54.95)
	KnowHalu (Structured)	72.6 (75.7)	66.6 (80.0)	6.60 (7.10)	69.60 (77.85)
	KnowHalu (Unstructured)	77.3 (68.9)	53.2 (75.7)	11.90 (14.50)	65.25 (72.30)
	KnowHalu (Aggregation)	76.3 (77.5)	67.8 (83.1)	6.60 (7.05)	72.05 (80.30)

is any conflict between the answer the knowledge from a sub-query, the judgment is **INCORRECT**. On the other hand, if all the details of the answer are verified by the knowledge of sub-queries, the judgment is **CORRECT**. For some scenarios, where the knowledge is inadequate for a conclusive judgment, the output will be **INCONCLUSIVE**. Prompts used to guide this judgment process are shown in Appendix A.4.

e. Aggregation The judgment of hallucination above is based on each form of the retrieved knowledge (e.g., structured and unstructured). To mitigate the prediction uncertainty, here we aggregate these judgment based on the multi-form knowledge to make the final prediction.

The motivation for this aggregation mechanism is two-fold: 1) each knowledge form might uniquely identify cases that the other cannot, particularly when one yields an **INCONCLUSIVE** judgment and the other does not; 2) when the LLM makes a hallucinated judgment, it may lead to low confidence scores for the judgment of “**CORRECT**” or “**INCORRECT**.” Thus, we can rely on the alternative knowledge form if it demonstrates a higher confidence for the judgment.

Concretely, we treat the judgment based on one form of the knowledge—typically the form yielding better average accuracy—as the *base judgment* and that of the other forms of knowledge as *supplement judgment*. If the confidence score for the base judgment falls below the a specific threshold δ_1 and the supplement judgment maintains higher confidence above δ_2 , we will take the supplement judgment as the final prediction. In all other cases, the base judgment will perform as the final prediction. The corresponding pseudo-code and more details for this aggregation mechanism are provided in Appendix E.3.

4 EXPERIMENTS

We have evaluated KnowHalu on the standard HaluEval dataset (Li et al., 2023), comparing with SOTA hallucination detection baselines under different settings. We find that 1) KnowHalu consistently outperforms the baselines in terms of hallucination detection in different tasks, 2) different models benefit differently from knowledge forms; GPT-3.5 performs better with structured knowledge, while open-sourced models like Starling-7B or Mistral-7B are more effective with unstructured knowledge; 3) the aggregation of predictions from different knowledge forms can further improve detection accuracy. All experiments were conducted on a single NVIDIA A6000 GPU.

4.1 EXPERIMENTAL SETUP

Dataset. We conduct our experiments using the standard HaluEval dataset (Li et al., 2023), and focus on hallucination detection for two primary tasks: multi-hop QA and text summarization. For the

Table 3: Performance of different methods for hallucination detection in Text Summarization.

Model	Method	TPR (%)	TNR (%)	Avg Acc (%)
GPT-4	Zero-Shot CoT	43.0	83.0	63.0
	SelfCheckGPT (Manakul et al., 2023)	80.2	38.4	59.3
	HaluEval (Vanilla) (Li et al., 2023)	17.4	95.6	56.5
	HaluEval (CoT) (Li et al., 2023)	31.6	81.0	56.3
	Self Consistency (Wang et al., 2022)	24.0	89.0	56.5
Starling-7B	KnowHalu (Structured)	80.2	45.4	62.8
	KnowHalu (Unstructured)	65.0	67.2	66.1
	KnowHalu (Aggregation)	59.2	75.5	67.3
	SelfCheckGPT (Manakul et al., 2023)	29.0	91.8	60.4
	HaluEval (Vanilla) (Li et al., 2023)	79.0	9.0	44.0
Mistral-7B	HaluEval (CoT) (Li et al., 2023)	84.8	6.0	45.4
	Self Consistency (Wang et al., 2022)	83.4	6.6	45.0
	KnowHalu (Structured)	68.6	63.2	65.9
	KnowHalu (Unstructured)	67.0	67.4	67.2
	KnowHalu (Aggregation)	67.0	67.8	67.4
GPT-3.5	SelfCheckGPT (Manakul et al., 2023)	33.6	86.8	60.2
	HaluEval (Vanilla) (Li et al., 2023)	66.6	58.0	62.3
	HaluEval (CoT) (Li et al., 2023)	44.4	63.4	53.9
	Self Consistency (Wang et al., 2022)	66.6	59.0	62.8
	KnowHalu (Structured)	64.4	71.0	67.7
KnowHalu (Unstructured)	62.8	68.0	65.4	
KnowHalu (Aggregation)	69.0	68.0	68.5	

Table 4: Performance of KnowHalu using different query formulations, evaluated based on the Starling-7B model.

Query	Method	TPR	ARP	TNR	ARN	Avg Acc (%)
Specific Query	KnowHalu (Structured)	57.4	19.8	64.1	18.6	60.75
	KnowHalu (Unstructured)	66.0	11.5	64.9	10.7	65.45
General Query	KnowHalu (Structured)	65.6	15.5	58.7	18.9	62.15
	KnowHalu (Unstructured)	70.4	10.6	60.7	15.9	65.55
Combined Queries	KnowHalu (Structured)	68.1	9.0	65.6	12.0	66.85
	KnowHalu (Unstructured)	68.2	5.3	69.9	8.2	69.05

Table 5: Performance of KnowHalu using different number of retrieved Wiki passages K , evaluated based on the Starling-7B model.

Top- K Passages	Method	TPR	ARP	TNR	ARN	Avg Acc (%)
$K = 1$	KnowHalu (Structured)	61.1	16.1	64.3	16.6	62.70
	KnowHalu (Unstructured)	65.6	10.4	64.8	13.0	65.20
$K = 2$	KnowHalu (Structured)	68.1	9.0	65.6	12.0	66.85
	KnowHalu (Unstructured)	68.2	5.3	69.9	8.2	69.05
$K = 3$	KnowHalu (Structured)	67.2	7.8	66.2	9.9	66.70
	KnowHalu (Unstructured)	68.6	4.1	70.8	4.5	69.70
$K = 4$	KnowHalu (Structured)	67.1	8.0	65.8	11.1	66.45
	KnowHalu (Unstructured)	68.8	3.7	69.6	5.2	69.20
$K = 5$	KnowHalu (Structured)	64.2	9.9	67.3	9.3	65.75
	KnowHalu (Unstructured)	66.9	4.3	72.7	4.0	69.80

multi-hop QA task, the dataset comprises questions and correct answers from HotpotQA (Yang et al., 2018), with hallucinated answers generated by ChatGPT. In the text summarization task, the dataset includes documents and their non-hallucinated summaries from CNN/Daily Mail (See et al., 2017), along with hallucinated summaries generated by ChatGPT.

In our experiment, we randomly sample 1,000 pairs from the QA task as the test set. Each test pair comprises both a correct answer and a hallucinated answer to the same question. Additionally, we sampled 500 pairs from the summary task, with each pair containing both accurate and hallucinated counterparts for the same document. We use these balanced test sets for evaluation and comparison.

Baselines. For the QA hallucination detection task, our study evaluates seven state-of-the-art baselines, each chosen to demonstrate distinct aspects of model performance in detecting hallucinations. (1) The first three baselines from the HaluEval suite (Li et al., 2023)—*HaluEval (Vanilla)*, which makes judgments without external knowledge; *HaluEval (Knowledge)*, which utilizes external knowledge; and *HaluEval (CoT)*, employing Chain-of-Thought reasoning—are specifically tailored for the dataset we used, ensuring optimized performance on the HaluEval benchmark. (2) *GPT-4 (Zero-shot CoT)* leverages the intrinsic world knowledge of a model dated 2023-11-06, which is the same date as the wiki database we use for retrieval, testing whether a smaller model with systematic knowledge reasoning can outperform a larger model based on its extensive pre-trained knowledge. (3) *WikiChat (Semnani et al., 2023)* generates responses by retrieving and summarizing Wikipedia information, using the same data as our wiki database for retrieval, to ensure accuracy through fact-checking. (4) *SelfCheckGPT (Manakul et al., 2023)* uses the inherent knowledge of large language models to generate and evaluate multiple responses for consistency without external data. (5) *Self Consistency (Wang et al., 2022)* involves sampling hallucination judgments from the three HaluEval baselines multiple times and taking a majority vote; we apply it on all three HaluEval baselines and report the best outcome. For *SelfCheckGPT* and *Self Consistency*, we sample 20 times using a temperature of 1.0. The prompts employed to query *GPT-4 (Zero-shot CoT)*, *WikiChat*, and *SelfCheckGPT* for hallucination detection are in Appendix A.5.

Models. For our experiments, we use three models: Starling-7B (Starling-LM-7B-alpha) (Zhu et al., 2023) and Mistral-7B (Mistral-7B-Instruct-v0.2) (Jiang et al., 2023), which are two open-source models that have shown high performance on the LMSYS Chatbot Arena Leaderboard (LMSYS, 2023; Zheng et al., 2023); and GPT-3.5 (gpt-3.5-turbo-1106) (OpenAI, 2023), a closed-source model from OpenAI.

Metric. Our evaluation focuses on five key metrics: True Positive Rate (TPR), True Negative Rate (TNR), Average Accuracy (Avg Acc), and Average Inconclusive Rate. TPR quantifies the ratio of correctly identified hallucinations, TNR measures the ratio of correctly identified non-hallucinations, and Avg Acc denotes the overall accuracy. Average Inconclusive Rate represents the model capability of identifying *inconclusive* cases. Note that it is not always possible to successfully retrieve the corresponding knowledge to verify the answer. However, existing baselines based on external knowledge still require the model to provide a binary judgment (Yes/No); thus, the accuracy reported could be higher than their actual performance since some answers actually cannot be assessed with

the available knowledge. On the contrary, KnowHalu allows the INCONCLUSIVE option to provide more informative judgments based on our capable framework. In this way, the average accuracy metric is slightly unfair for KnowHalu since we aim to provide a more fine-grained detection; otherwise, its reported average accuracy should be even higher. Nevertheless, KnowHalu still beats all baselines significantly in terms of the average accuracy (Table 2).

4.2 HALLUCINATION DETECTION ON QA TASK

Setup. To detect hallucinations in the QA Task, we test two distinct knowledge sources. The first, which we refer to as “*off-the-shelf knowledge*,” is the knowledge provided in the HaluEval dataset. This consists of specific passages that are directly related to the question-answer pairs within the dataset, which serves as a natural upper bound for the quality of retrieved knowledge. The second knowledge source, which we refer to as “*Wiki retrieval knowledge*,” comes from the information retrieval system constructed over the Wikipedia Database as outlined in step (b) in Section 3.2. This system fetches the Top-K most relevant passages in response to a given query. We aim to evaluate the effectiveness of different hallucination detection approaches given these two knowledge sources.

When utilizing the “*Wiki retrieval knowledge*”, the number of fetched passages K is consistently set to 2 for all methods. For our method, we report the results considering different query formulations. A detailed analysis of the influence of different query formulations and K is presented in Section 5.2 and Section 5.3, respectively.

Results. The main results are shown in Table 2. We observe that leveraging our sequential reasoning and query approach, coupled with a well-formulated query for knowledge retrieval and the aggregation of two distinct forms of knowledge, KnowHalu consistently outperforms baselines by around 15% when using the same knowledge source with the same model.

Furthermore, our results reveal several intriguing observations: (1) relying solely on the pre-trained knowledge of LLMs and employing multiple samples to detect contradictions between different samples (as in SelfCheckGPT and Self Consistency) typically results in sub-optimal performance; (2) employing systematic, step-wise reasoning and querying enables a small 7B model (Starling-7B) within KnowHalu to achieve comparative performance with GPT-3.5; (3) using all the three smaller models in KnowHalu demonstrates superior detection performance when compared to the powerful GPT-4, which has implicit reasoning capabilities and knowledge; (4) the form of knowledge matters for different models—open-sourced models like Starling-7B and Mistral-7B appear to perform better with unstructured knowledge, while GPT-3.5 seems to benefit more from structured knowledge (i.e., triplets), enhancing the need for aggregation mechanisms. A comprehensive analysis of the individual contributions of each component within KnowHalu is in Section 5.

4.3 HALLUCINATION DETECTION ON SUMMARIZATION TASK

Setup. In the task of text summarization, the original document serves as the primary source of knowledge. During our experiments, we segment the document into passages with fewer than 40 words each. Both the input query and retrieved passages are encoded using BGE large model (BAAI, 2023) from FlagEmbedding (Xiao et al., 2023b; Zhang et al., 2023a; Xiao et al., 2023a). For each query, the Top-K relevant passages are retrieved for knowledge optimization, with the number of passages K set to 3. The impact of varying K is further analyzed in Section 5.3.

Unlike the QA task, to detect hallucinations in the summarization task, any detail in the summary that cannot be supported or identified in the original document will be considered as a hallucination, which means we have a complete knowledge source. As a result, the off-target hallucination checking phase is not required for this task, allowing us to move directly to factual checking. The judgment now only includes CORRECT and INCORRECT, as cases that would be classified as INCONCLUSIVE in the QA task are inherently INCORRECT in text summarization task. In addition, given that some summaries are quite lengthy, we segment the original summary into small parts, with each segment comprising no more than 30 words. Each segment is independently evaluated for hallucination, and the entire summary is labeled as a hallucination if any part receives an INCORRECT judgment.

Results. The main results are presented in Table 3. We observe that KnowHalu significantly outperforms the baselines, achieving performance increases of 8.0% with Starling-7B, 7.0% with Mistral-7B, and 6.2% with GPT-3.5. Notably, we can see that nearly all variations of KnowHalu surpass the powerful GPT-4, demonstrating a superior performance. In particular, GPT-3.5 model demonstrates a great advantage when utilizing structured knowledge, whereas both the Starling-7B and Mistral-7B models still benefit more from unstructured knowledge. Besides, as we can see, reliance on detecting contradictions between different sample summarizations (SelfCheckGPT) typ-

Table 6: Impact of Off-Target Hallucination Checking on Hallucination Detection Performance using the Starling-7B Model for QA Task. The table reports the True Positive Rate (TPR), True Negative Rate (TNR), and Average Accuracy (Avg Acc) for each method.

Method	TPR (%)	TNR (%)	Avg Acc (%)
HaluEval (Knowledge)	33.0 (82.0)	60.3 (40.0)	46.65 (61.00)
Pure Fact-Checking (Structured) + Off-Target Hallucination Checking	52.1 (48.2) 68.1 (67.8)	67.2 (86.6) 65.6 (83.1)	59.65 (67.40) 66.85 (75.45)
Pure Fact-Checking (Unstructured) + Off-Target Hallucination Checking	53.4 (56.0) 68.2 (72.4)	71.7 (88.1) 69.9 (85.9)	62.55 (72.05) 69.05 (79.15)

ically results in highly unbalanced TPR and TNR. For example, with GPT-3.5, it yields a TNR of 86.8% but a corresponding TPR of only 33.6%. At the same time, relying solely on the majority sample judgment of hallucinations (Self-Consistency) offers only marginal improvements. Additionally, we also observe some ‘lazy’ behaviors in GPT-3.5 during step-wise reasoning, as demonstrated in Appendix D. We also present the performance of various temporal versions of GPT-3.5 in Appendix C for comprehensive explorations.

5 ABLATION STUDIES

5.1 IMPACT OF OFF-TARGET HALLUCINATION CHECKING

Prompts provided by HaluEval (Li et al., 2023) not only cover cases of fabrication hallucination but also off-target hallucinations as shown in Table 1. Thus, our pipeline separates the process into two phases: (1) treating the detection of off-target hallucinations as an independent task; (2) if an off-target hallucination is detected, no further checking is required; otherwise, we proceed to a second-phase for factual checking. This approach raises two intriguing questions: (1) whether such decomposition improves hallucination detection performance, and (2) what is the performance when factual checking is conducted directly without any preliminary off-target hallucination demonstration. As shown in Table 6, incorporating off-target checking consistently enhances the detection of hallucinated cases, with an approximately 15% improvement in TPR and 2% in FPR. Furthermore, even without off-target checking, our standalone factual checking phase still surpasses the baseline using the same knowledge source, demonstrating the effectiveness of our multi-query and reasoning process.

5.2 FORMULATIONS OF QUERIES

We investigate the impact of different query formulations used for knowledge retrieval on hallucination detection. All experiments are conducted using Starling-7B with $K = 2$ for Wiki retrieval knowledge. We evaluate the following three approaches: (1) using only *specific queries*, (2) using only *general queries*, and (3) combining the Top-K results from both query types. The results are detailed in Table 4. As we can see, the formulation of the query is crucial in knowledge retrieval. In particular, *specific query* formulation enhances the accuracy for non-hallucinated cases but reduces that for hallucinated ones due to polluted context. Conversely, using *general queries* yields an inverse effect. Combining both query types improves the overall detection accuracy by at least 3.5% and reduces the abstention rate by over 5%, demonstrating that the combination of both query formulations indeed leads to more accurate and relevant knowledge retrieval. The results conducted with the off-the-shelf knowledge and the results for text summarization are presented in Appendix E.1.

5.3 NUMBER OF RETRIEVAL KNOWLEDGE

We explore how the number of retrieved Wiki passages, K , impacts the performance of hallucination detection in this section. Throughout this analysis, we consistently employ a combination of both specific and general queries for knowledge retrieval, focusing on assessing the influence of varying K . The results are presented in Table 5. We can observe that increasing the number of retrieved passages enhances detection accuracy and reduces the abstain rate. In addition, the performance converges when K is greater than 2, and additional knowledge will only provide marginal improvement, highlighting the potential efficiency of KnowHalu. We also provide the results for similar experiments conducted on text summarization in Appendix E.2.

5.4 AGGREGATION BASED ON MULTI-FORM KNOWLEDGE

We further explore mitigating hallucinations by implementing a confidence-based aggregation mechanism that utilizes various forms of knowledge, both structured and unstructured. The motivation of our approach is the observation that judgments susceptible to hallucinations typically have lower confidence levels compared to those that are accurate and free from hallucinations. Consequently,

Table 7: Latency per instance (in seconds) for various methods on the QA task using the Starling-7B model on one A6000 GPU.

Method	Knowledge Source	Latency (s/instance)
HaluEval (Knowledge)	Off-the-shelf	0.27
HaluEval (Knowledge)	Wiki	0.57
Self Consistency	Off-the-shelf	4.36
Self Consistency	Wiki	8.71
KnowHalu	Off-the-shelf	7.98
KnowHalu	Wiki	8.75
WikiChat	Wiki	10.62
SelfCheckGPT	-	23.89

we adopt a strategy where if a base judgment, derived from one form of knowledge, displays low confidence (below δ_1), and a supplementary judgment from a different form of knowledge shows significantly higher confidence (above δ_2), the latter is prioritized based on its reliability.

To select thresholds δ_1 and δ_2 , we employ a data-driven approach that utilizes the quantile of the confidence distribution associated with each form of knowledge. We achieve this by using a small validation set for both tasks, during which we collect confidence distributions for judgments obtained based on each knowledge type. This approach facilitates a more precise evaluation of δ_1 and δ_2 through an examination of various quantiles within these distributions. By adjusting δ_1 and δ_2 according to these quantiles on the validation set, we aim to identify the optimal configurations that yield the highest average accuracy for each task. In our experiments, we consistently utilize the judgments based on the form of knowledge that provides the best average accuracy as the base judgment. The specific values of δ_1 and δ_2 , along with a detailed description of the process for selecting them, are provided in Appendix E.3.

5.5 LATENCY COMPARISON

Our reasoning process, while sequential across multiple steps, benefits from shared key-value caching. We retain the cache from previous reasoning steps to accelerate subsequent ones, thereby reducing latency. Specifically, the latency for the methods using the Starling-7B model on one A6000 GPU per instance for the QA task is shown in Table 7. As we can see, while our method shows a slight increase in latency compared to the sample-based method Self Consistency (Wang et al., 2022) when using off-the-shelf knowledge, it incurs similar costs when using wiki-retrieved knowledge. Notably, our method consistently outperforms Self Consistency by at least 20%, regardless of the underlying LLM used, as shown in Table 2. Additionally, our method’s latency is lower than that of WikiChat (Semnani et al., 2023)—which extracts and verifies claims individually—and is significantly lower than that of another sample-based hallucination detection method, SelfCheckGPT (Manakul et al., 2023). Consequently, the latency introduced by our approach remains within acceptable limits, staying under 10 seconds per instance for detecting hallucinations in answers to multi-hop complex questions, such as those in the HaluEval dataset (based on HotpotQA) used in our experiments. Moreover, latency is expected to be significantly lower for simpler, one-hop questions, making our method a viable option in settings where higher accuracy is essential.

6 CONCLUSION

In this work, we have introduced KnowHalu, a novel framework for detecting hallucinations in text generated by Large Language Models (LLMs). Our approach stands out by employing a two-phase process: off-target Hallucination Checking and Factual Checking, which includes multi-form knowledge retrieval and optimization, along with an aggregation method for the final judgment. Through extensive experimentation on standard datasets, KnowHalu has demonstrated significant improvements over state-of-the-art baselines in detecting hallucinations in both QA and text summarization tasks. While the current implementation has limitations in handling extended dialogues and much longer responses, specifically the lack of correlation consideration between sentences within a single response, future extensions will focus on adapting the framework for dialogue systems and optimizing it for longer interactions.

REFERENCES

- 540
541
542 Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying. *arXiv preprint*
543 *arXiv:2304.13734*, 2023.
- 544 BAAI. bge-large-en-v1.5. <https://huggingface.co/BAAI/bge-large-en-v1.5>,
545 2023.
- 546
547 Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. Constrained
548 decoding for neural nlg from compositional representations in task-oriented dialogue. *arXiv*
549 *preprint arXiv:1906.07220*, 2019.
- 550 Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia,
551 Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of
552 chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.
- 553
554 Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt’s behavior changing over time? *arXiv*
555 *preprint arXiv:2307.09009*, 2023.
- 556 Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and
557 Jason Weston. Chain-of-verification reduces hallucination in large language models, 2023.
- 558
559 Esin Durmus, He He, and Mona Diab. Feqa: A question answering evaluation framework for
560 faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*, 2020.
- 561 Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2 :
562 Evaluating factual consistency in knowledge-grounded dialogues via question generation and
563 question answering. *arXiv preprint arXiv:2104.08202*, 2021.
- 564
565 Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han.
566 Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.
- 567 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
568 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language
569 models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*,
570 2023.
- 571
572 Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem
573 in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.
- 574 Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane
575 Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with
576 retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.
- 577
578 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,
579 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM*
580 *Computing Surveys*, 55(12):1–38, 2023.
- 581 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
582 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al.
583 Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- 584
585 Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv*
586 *preprint arXiv:2107.07566*, 2021.
- 587
588 Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in
589 neural machine translation. 2018.
- 589 Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-
590 scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023*
591 *Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.
- 592
593 LMSYS. chatbot-arena-leaderboard. [https://huggingface.co/spaces/lmsys/](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard)
[chatbot-arena-leaderboard](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard), 2023.

- 594 Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box
595 hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*,
596 2023.
- 597
598 Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations
599 of large language models: Evaluation, detection and mitigation. In *The Twelfth International
600 Conference on Learning Representations*, 2023.
- 601
602 OpenAI. GPT-3.5 Turbo API, Model Version 1106. [https://platform.openai.com/docs/
603 guides/gpt](https://platform.openai.com/docs/guides/gpt), 2023.
- 604
605 Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars
606 Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language
607 models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.
- 608
609 Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle
610 Ott, Kurt Shuster, Eric M Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint
arXiv:2004.13637*, 2020.
- 611
612 Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. Plaid: an efficient engine for
613 late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information
614 & Knowledge Management*, pp. 1747–1756, 2022a.
- 615
616 Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2:
617 Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022
618 Conference of the North American Chapter of the Association for Computational Linguistics:
Human Language Technologies*, pp. 3715–3734, 2022b.
- 619
620 Neil Savage. Drug discovery companies are customizing chatgpt: here’s how. *Nature Biotechnology*,
621 2023.
- 622
623 Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with
624 pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for
Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, 2017.
- 625
626 Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text
627 generation. *arXiv preprint arXiv:2004.04696*, 2020.
- 628
629 Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. WikiChat: Stopping the hallucination of
630 large language model chatbots by few-shot grounding on Wikipedia. In Houda Bouamor, Juan
631 Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP
2023*, pp. 2387–2413, Singapore, December 2023. Association for Computational Linguistics.
632 URL <https://aclanthology.org/2023.findings-emnlp.157>.
- 633
634 Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation
635 reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- 636
637 Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston.
638 Language models that seek for knowledge: Modular search & generation for dialogue and prompt
639 completion. *arXiv preprint arXiv:2203.13224*, 2022a.
- 640
641 Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung,
642 Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that
continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022b.
- 643
644 Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan
645 Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode
646 clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- 647
Jean-Philippe Vert. How will generative ai disrupt data science in drug discovery? *Nature Biotech-*
nology, pp. 1–2, 2023.

- 648 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-
649 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.
650 *arXiv preprint arXiv:2203.11171*, 2022.
- 651 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
652 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
653 *Neural Information Processing Systems*, 35:24824–24837, 2022.
- 654 Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhan-
655 jan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for
656 finance. *arXiv preprint arXiv:2303.17564*, 2023.
- 657 Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to
658 advance general chinese embedding, 2023a.
- 659 Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. Lm-cocktail: Resilient tuning of language
660 models via model merging, 2023b.
- 661 Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large
662 language models. *arXiv preprint arXiv:2306.06031*, 2023.
- 663 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov,
664 and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question
665 answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*
666 *Processing*, pp. 2369–2380, 2018.
- 667 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.
668 ReAct: Synergizing reasoning and acting in language models. In *International Conference on*
669 *Learning Representations (ICLR)*, 2023.
- 670 Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to
671 augment large language models, 2023a.
- 672 Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lema Liu, Tingchen Fu, Xinting Huang, Enbo Zhao,
673 Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: a survey on hallucination in large
674 language models. *arXiv preprint arXiv:2309.01219*, 2023b.
- 675 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
676 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
677 chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.
- 678 Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm
679 helpfulness & harmlessness with rlaif, November 2023.
- 680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701

702 A PROMPTS

703 A.1 PROMPT FOR OFF-TARGET HALLUCINATION CHECKING

704 For each type of hallucination listed in Table 1, we include one to two illustrative examples in the
 705 prompts for instructing the LLMs to implement the extraction task as shown in Table 8. This results
 706 in a compilation of 10 examples where the extraction outcome is “NONE,” indicative of off-target
 707 hallucinations. Additionally, we include 11 examples where the extraction successfully identifies
 708 the entity pertinent to the question. It’s important to note that these 11 examples feature a mix of
 709 correct and incorrect (fabricated) answers. This variety is intentional, as the primary objective in this
 710 phase is to filter out off-target hallucinations, thus allowing for the possibility of incorrect (fabricated)
 711 responses in the examples.

712 A.2 PROMPT FOR STEP-WISE REASONING AND QUERY

713 We present examples of prompts for Step-wise Reasoning and Query, illustrating our approach with
 714 both structured and unstructured knowledge forms for the QA task. These examples are detailed
 715 in Table 9 for structured knowledge and in Table 10 for unstructured knowledge, respectively.
 716 Specifically, we provide three examples of correct answers, three examples of hallucinated answers,
 717 and two examples where the knowledge is not available during the reasoning and query process.
 718 Besides, note that the prompt examples presented here are designed for combined queries. Therefore,
 719 in the demonstration of the line for *#Query#*, it starts with a specific query followed by a general
 720 query enclosed in brackets. This indicates that the LLMs are instructed to produce queries in this
 721 composite format, effectively integrating both specific and general inquiry approaches within a single
 722 output. During the knowledge retrieval step, both types of queries are employed for retrieval, and
 723 the retrieved passages for each are merged. In scenarios where only a specific query is adopted,
 724 the prompt will be adjusted to retain only the initial specific query, omitting the bracketed portion.
 725 Conversely, for general queries, the specific query is excluded, maintaining only the general query.
 726 The prompt structure for the text summarization task follows a similar format, also featuring three
 727 examples each for correct and hallucinated answers. Therefore, we only display the QA task prompts
 728 as representative examples to avoid redundancy.

729 A.3 PROMPT FOR KNOWLEDGE OPTIMIZATION

730 In our experiments, we extract related knowledge from either *Wiki retrieval knowledge* or *off-the-shelf*
 731 *knowledge* in an optimized format, corresponding to either structured knowledge (in triplets) or
 732 unstructured knowledge (in normal semantic text). For the QA task, examples of prompts based
 733 on Wiki retrieval knowledge, tailored for these two forms, are illustrated in Table 11 and Table 12,
 734 respectively. Given the similarity in format for off-the-shelf knowledge on QA task and the approach
 735 used for the Text Summarization task, we omit those details to prevent redundancy.

736 A.4 PROMPT FOR JUDGMENT

737 Upon completing the step-wise reasoning and query process, we accumulate sufficient information to
 738 proceed to the final judgment for each form of knowledge. At this stage, the *#Thought#* line, initially
 739 instrumental in guiding the query, becomes redundant and is therefore removed. We retain only the
 740 *#Query#* and *#Knowledge#* inputs for the LLM to facilitate the final judgment. The prompts used
 741 for this purpose on QA task, tailored to both structured and unstructured knowledge, are displayed
 742 in Table 13 for structured knowledge and in Table 14 for unstructured knowledge. Given the prompts
 743 for the text summarization task follow a similar format, we also omit their detailed presentation here
 744 to avoid repetition.

745 A.5 SETUP DETAILS FOR BASELINES

746 A.5.1 GPT-4 WITH ZERO-SHOT CoT

747 **QA.** For GPT-4 with zero-shot reasoning, we utilize the system prompt from HaluEval:

756 You are a hallucination detector. You MUST determine if the provided
 757 answer contains hallucination or not for the question based on the world
 758 knowledge. The answer you provided MUST be "Yes" or "No". You should
 759 first provide your judgment and then provide your reasoning steps.

760
 761 **And the corresponding user prompt is:**

762 You should try your best to determine if the answer contains non-factual
 763 or hallucinated information. The judgement you give MUST be "Yes" or "No".
 764 You should first provide your judgment and then provide your reasoning
 765 steps.

```
766 #Question#: {question}
767 #Answer#: {answer}
768 #Your Judgement#:
```

769
 770 **Text Summarization.** We utilize the system prompt from HaluEval:

771 You are a summary judge. You MUST determine if the provided summary
 772 contains non-factual or hallucinated information. The answer you give
 773 MUST be "Yes" or "No". You should first provide your judgment and then
 774 provide your reasoning steps.

775
 776 **And the corresponding user prompt is:**

777 You should try your best to determine if the summary contains non-factual
 778 or hallucinated information for the document. The judgement you give MUST
 779 be "Yes" or "No". You should first provide your judgment and then provide
 780 your reasoning steps.

```
781 #Document#: {document}
782 #Summary#: {summary}
783 #Your Judgement#:
```

784 785 A.5.2 SELF-CHECKGPT

786 **QA.** We start by prompting the underlying large language model (LLM) to generate an answer
 787 to the given question, and similarly with the SelfCheckGPT (Manakul et al., 2023), we sampled
 788 the answer 20 times at a temperature setting of 1.0. Next, we employ a prompt adapted from
 789 SelfCheckGPT (Manakul et al., 2023), which compares these 20 newly generated answers against
 790 the original reference answer provided in HaluEval:

```
791 Question: {question}
792 Answer-1: {original_answer}
793 Answer-2: {newly_sampled_answer}
794
```

795 Is the Answer-1 supported by the Answer-2 for the same question mentioned
 796 above? Provide the judgment with Yes or No.

797
 798 Judgment:

799
 800 The final judgment of hallucination is based on the majority vote from these comparisons.

801 **Text Summarization.** Similarly, we begin by prompting the underlying large language model (LLM)
 802 to generate a summary of the original document, and sample this summary 20 times at a temperature
 803 setting of 1.0. We then utilize a prompt adapted from SelfCheckGPT (Manakul et al., 2023), which
 804 assesses the hallucinations in the original summary provided in HaluEval against these 20 newly
 805 generated summaries. Given that summaries are typically lengthy, we follow the same protocol used
 806 in both KnowHalu and SelfCheckGPT by segmenting the original summary into smaller parts, each
 807 comprising no more than 30 words. Each segment is evaluated against the newly sampled summaries
 808 (used as the context) to determine the presence of hallucinations:

```
809 Context: {context}
810 Sentence: {segment}
```

810 Is the sentence supported by the context above? Answer Yes or No.

811
812 Judgment :

813
814 The hallucination judgment for each segment is based on the majority vote among the 20 new
815 generated summaries. If any segment of the original summary is identified as a hallucination, the
816 entire summary is labeled as such, consistent with the setting for KnowHalu.

817 A.5.3 WIKICHAT

818
819 For WikiChat (Semnani et al., 2023), originally crafted for generating factual responses rather than
820 conducting hallucination checks, we adapt its fact-checking capability. We prompt the LLM to
821 articulate claims regarding the hallucination status of an answer, followed by detailed reasoning.
822 This approach enables WikiChat’s fact-checking mechanism to verify these claims, facilitating the
823 generation of accurate judgments.

824 The user prompt employed to elicit reasoned judgments and claims is outlined as follows:

825 #Question#: {question} #Answer#: {answer} Is this answer to the question
826 a result of hallucination? Please indicate your judgment as either "Yes"
827 or "No," accompanied by a step-by-step explanation."

828 In leveraging WikiChat for hallucination detection, we utilize its comprehensive workflow, initially
829 designed for factual checking. This includes generating a query to fetch information from Wikipedia,
830 summarizing and filtering retrieved passages, and utilizing an LLM for initial response generation. By
831 prompting the LLM to articulate judgment alongside claim statements, we engage WikiChat’s fact-
832 checking mechanism against the retrieved evidence, thereby refining the response through iterations.
833 This process guarantees that the final determination of hallucination status is both evidence-based and
834 meticulously refined for precision. Additionally, we illustrate the detailed process by which WikiChat
835 generates the final fact-checked judgment in Table 15, providing insight into the iterative refinement
836 and fact-checking stages integral of WikiChat.

837 B ILLUSTRATIVE CASES OF RETRIEVAL OUTCOMES BASED ON QUERY 838 FORMULATION VARIATIONS

839 In this section, we present a range of illustrative cases that showcase the outcomes of retrieval based
840 on the specific nuances in query formulation. This includes both specific and general queries aimed
841 at verifying correct and hallucinated details. These examples utilize the same retrieval system built
842 upon the Wikipedia database, as constructed in WikiChat (Semnani et al., 2023). The objective
843 is to subjectively demonstrate the differences in retrieval effectiveness between these two types of
844 queries. We display the top two results of each retrieval instance. The complete set of example cases
845 is detailed in Table 16.

846 The findings indicate that specific queries, by leveraging additional information from the knowledge
847 base, can yield highly accurate results when verifying correct details. However, their effectiveness
848 diminishes when applied to hallucinated details, often leading to the retrieval of irrelevant or nonex-
849 istent information. In contrast, general queries excel in retrieving better results for hallucinated
850 details, as they avoid the pitfalls of specific, possibly inaccurate information. Yet, this approach
851 does not capitalize on the extra information provided in the answer, which can result in less effective
852 verification of correct details. Therefore, each query formulation has its distinct advantages and
853 limitations, necessitating a strategic choice based on the nature of the detail being verified.

854 C PERFORMANCE COMPARISON FOR DIFFERENT GPT-3.5 VERSIONS IN QA 855 TASK

856
857 To examine the impact of different GPT-3.5 model versions, we tested `gpt-3.5-turbo-0613`
858 and `gpt-3.5-turbo-1106` on the QA Task, with results presented in Table 17. Consistent
859 with observations made by Chen et al. (2023) in HotpotQA (Yang et al., 2018), newer versions of
860 GPT-3.5 exhibit a decline in performance. For instance, when simply prompting GPT-3.5 to answer
861 questions without Chain of Thought (CoT) or knowledge support, akin to HaluEval (Vanilla), the
862 performance drops by approximately 12%. In addition, with the integration of CoT (HaluEval(CoT)),
863 the performance with the newer version drops 6.90%, and with knowledge-enhanced prompts
(HaluEval(Knowledge)), the decrease is 10.75%.

864 However, interestingly, within our framework KnowHalu, the performance instead slightly improves
 865 by 1% to 2%, suggesting that our approach can leverage the inherent capabilities of the GPT-3.5
 866 model more efficiently. Moreover, we observe that GPT-3.5 models across both versions particularly
 867 benefit more from structured knowledge forms, i.e., using triplets for reasoning, indicating their
 868 proficiency in managing structured knowledge within the QA task explored here, which is also one of
 869 the main findings in our KnowHalu framework.

870 D FAILURE CASES FOR GENERATING QUERY WITH GPT-3.5 IN 871 SUMMARIZATION TASK

872 We find the performance of GPT-3.5 used in KnowHalu could have high uncertainty. This is
 873 primarily due to two key observations during our experiments: first, GPT-3.5 occasionally rejects
 874 certain summaries, which are recognized with inappropriate or potentially invasive content, an issue
 875 not analyzed in baselines; second, GPT-3.5 tends to demonstrate ‘lazy’ behavior during the step-wise
 876 querying process. For instance, it may prematurely conclude that a summary does not necessitate
 877 further verification, or it might simply conclude the correctness of a summary even without generating
 878 any query.

879 In specific, we first encountered instances where the model exhibited reluctance in generating queries,
 880 often citing concerns related to information privacy or involving sensitivity:

```
881 #Summary#: Under the ownership of miner Pat Burke, Ena was also the
882 infamous playground for Rene Rivkin and his mates in the 1980s.
883 #Thought-1#: The summary contains inaccurate and potentially defamatory
884 information about individuals. Therefore, it is not appropriate to create
885 queries to verify the details provided.
886 ...
```

```
887 #Summary#: Elderly man bludgeons wife to death before taking his own life
888 with knife.
889 #Thought-1#: This summary contains sensitive and potentially distressing
890 content. For ethical and respectful reasons, I will not be providing
891 queries or engaging in the verification process for this particular
892 summary.
893 ...
```

894 Besides, the model seemed reluctant to generate queries, either by confirming the summary’s correct-
 895 ness without further inquiry or declaring that no queries could be formulated due to the summary’s
 896 lack of specific details or claims needing verification:

```
897 #Summary#: Among the many famous military leaders who passed through its
898 gates was Field Marshall Montgomery himself.
899 #Thought-1#: This summary is incomplete and lacks specific details for
900 verification. Therefore, no queries can be formulated.
901 ...
```

```
902 #Summary#: Shaving cream is the recommended treatment for jellyfish
903 stings, but also helps remove any nematocysts that may be stuck to the
904 skin.
905 #Thought-1#: The summary does not contain any specific claims or details
906 that need verification. Therefore, no queries are needed.
907 ...
```

908 While the former case represents the ethical considerations of the GPT model, the latter reflects its
 909 limitations in engaging with content requiring nuanced understanding. Such behaviors, encompassing
 910 approximately 10% of the overall test set, may introduce a bias in accurately assessing the model’s
 911 actual performance. Noticeably, the Starling-7B model instead successfully generated the queries for
 912 both cases with the same prompt.

913 E ADDITIONAL EXPERIMENT RESULTS

914 E.1 FORMULATION OF THE QUERY

915 We explore the impact of query formulation when using the off-the-shelf knowledge for the QA task
 916 in Table 18. In this context, the specificity of the query formulation becomes less critical since there
 917

Algorithm 1 Aggregation for Final Judgment

Input: Base Judgment, Supplement Judgment, Confidence Thresholds: δ_1, δ_2
Output: Aggregated Judgment

```

1: if Base Judgment == 'INCONCLUSIVE' then
2:   return Supplement Judgment
3: else if  $\mathbb{P}(\text{Base Judgment}) < \delta_1$  and  $\mathbb{P}(\text{Supplement Judgment}) > \delta_2$  then
4:   return Supplement Judgment
5: else
6:   return Base Judgment
7: end if

```

is no actual retrieval process involved; the system consistently returns the same knowledge provided by HaluEval for any query. Consequently, as observed, the performance across different formulations remains relatively similar.

In the text summarization task, as detailed in Table 19, which requires an actual retrieval process for passages, we observe a pattern akin to the QA task utilizing Wiki retrieval knowledge. Here the number of retrieved passages is set to 3 consistently. In this context, specific queries are particularly beneficial for validating correct details, as indicated by a higher True Negative rate (TN). Conversely, general queries are more adept at identifying incorrect details, leading to a higher True Positive rate (TP). This distinction further underscores the importance of query formulation in enhancing the accuracy of hallucination detection.

E.2 NUMBER OF RETRIEVAL KNOWLEDGE

The impact of varying the number of retrieval passages for the text summarization task is provided in Table 20, where we consistently employ combined queries for knowledge retrieval. Notably, fewer retrieval passages results in a higher True Positive (TP) rate, as limited information tends to offer less support for the claims in the queries, thereby increasing the likelihood of identifying hallucinated content. Moreover, the results indicate that the average performance stabilizes beyond $K = 3$, suggesting an optimal balance in the amount of information required for effective hallucination detection.

E.3 AGGREGATION

In this section, we detail the selection of thresholds δ_1 and δ_2 , which is crucial for improving the performance of aggregating judgments based on different forms of knowledge. Notably, the confidence distribution for judgments, e.g., "INCORRECT" or "CORRECT," varies due to differences in both the prompts and the reasoning processes used for different forms of knowledge. Therefore, to better accurately gauge the confidence levels, instead of testing different absolute values for exploring the best setting, we leverage the quantile of the confidence distribution for each form of knowledge to identify these two values. Once they are identified, we will use the Algorithm 1 to make the aggregated judgment.

Therefore, we sample a small validation subset of 100 cases with both non-hallucinated and hallucinated answers for both the QA and text summarization tasks from HaluEval, separate from our test dataset, to collect confidence distributions for the judgments based on each form. This enables the determination of δ_1 and δ_2 by exploring various quantiles of these distributions; for instance, if the base judgment relies on "Structured Knowledge" and the supplementary judgment on "Unstructured Knowledge", we will adjust δ_1 across quantiles q_1 from 0.05 to 0.95 with a step size of 0.05 of the confidence distribution collected for structured knowledge; and δ_2 is set to a corresponding higher quantile q_2 , starting with the value q_1 (the relative confidence level for the supplementary judgment must be higher) and also incrementing by steps of 0.05, up to 0.95. The final pair of quantiles q_1 and q_2 that leads to the best average accuracy will determine the corresponding δ_1 and δ_2 to be used on our test set. The optimal δ_1 and δ_2 , with their corresponding quantiles q_1 and q_2 for each task leading to the best performance, are summarized in Table 21. Notice that the judgments may undergo different tokenizations. Specifically, 'INCORRECT' might be tokenized as 'INC', 'OR', 'RECT', whereas 'CORRECT' can be tokenized as 'COR', 'RECT' when using the GPT-3.5. Thus, in our experiments, to maintain consistency and mitigate the impact of varying tokenization patterns on

972 judgment, we consistently rely on the confidence score associated with the first token of the judgment
973 label as the representative probability for the entire judgment.
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

1026 You are now a very truthful and objective extractor in this task. Your role is to meticulously analyze each
1027 pair of a question and its corresponding answer. Focus on identifying the primary entity or information that
1028 the question seeks, and then extract the specific information or entity from the answer that directly
1029 corresponds to this aspect. Note that your task here is not to judge the correctness or relevance of the
1030 answer, but solely to identify and extract the corresponding entity or information. In your #Extraction#,
1031 please clearly state what the main aspect of the question asks, and then specify the entity or information
1032 from the answer that matches this aspect. If the answer does not contain the specific entity or information
1033 sought by the question, indicate that the corresponding specific entity in the answer is NONE.

1031 #Question#: What is the primary language spoken in both Barcelona and Valencia?
1032 #Answer#: In both Barcelona and Valencia, they speak European languages.
1033 #Extraction#: The question specifically asks for the primary language spoken in Barcelona and Valencia. The
1034 answer, however, fails to mention any specific language and only broadly refers to "European languages."
1035 Therefore, the corresponding specific language entity in this answer is NONE.

1035 #Question#: In the midst of 17th-century historical milestones like the rise of Baroque art, groundbreaking
1036 scientific discoveries by Galileo and Newton, and the expansion of global exploration and colonization, which
1037 locations served as the formal signatories for the momentous Peace of Westphalia, marking the end of the
1038 Thirty Years' War?
1039 #Answer#: Munster and Osnabruck, Germany, and it was signed in 1648.
1040 #Extraction#: The question primarily asks for the specific locations where the Peace of Westphalia treaties
1041 were signed, the answer provides these locations as Munster and Osnabruck in Germany.

1039 #Question#: What year did the scientist who developed the polio vaccine first publish his findings?
1040 #Answer#: The scientist who developed the polio vaccine published his findings in the 20th century.
1041 #Extraction#: The question asks for the specific year when the scientist who developed the polio vaccine first
1042 published his findings. The answer provides a general timeframe, the 20th century, but does not specify the
1043 exact year required by the question. Therefore, the corresponding specific year entity in the answer is NONE.

1043 #Question#: "A Brief History of Time" is a book by a renowned physicist who later won the Presidential Medal
1044 of Freedom in which year?
1045 #Answer#: "A Brief History of Time" was published before its author received the Presidential Medal of
1046 Freedom.
1047 #Extraction#: The question asks for the specific year when the author of "A Brief History of Time" received
1048 the Presidential Medal of Freedom. The answer only states that the book was published before the author
1049 received the award but does not specify the year required by the question. Therefore, the corresponding
1050 specific year entity in the answer is NONE.

1048 #Question#: In what year was the novel "To Kill a Mockingbird" by Harper Lee first published?
1049 #Answer#: 1960
1050 #Extraction#: The question asks for the specific year of the first publication of "To Kill a Mockingbird." The
1051 answer correctly provides this information as 1960, directly matching the specific detail requested in the
1052 question.

1052 #Question#: What films has the cinematographer Emmanuel Lubezki been involved in?
1053 #Answer#: Emmanuel Lubezki worked on some visually stunning films.
1054 #Extraction#: The question asks for specific films involving cinematographer Emmanuel Lubezki. The answer
1055 mentions that Lubezki worked on visually stunning films but fails to name any specific films. Therefore, the
1056 corresponding specific film titles in the answer are NONE.

1055 #Question#: Which movies did director Christopher Nolan and cinematographer Wally Pfister collaborate on?
1056 #Answer#: "Inception," "The Dark Knight," and "Memento."
1057 #Extraction#: The question seeks specific movies that director Christopher Nolan and cinematographer Wally
1058 Pfister collaborated on. The answer directly provides the names of these movies: "Inception," "The Dark
1059 Knight," and "Memento," thus fully addressing the specific aspect of the question.

1059 #Question#: In "Whiplash," who played the ambitious jazz drummer at a prestigious music conservatory?
1060 #Answer#: The actor in "Whiplash" played the ambitious jazz drummer at a prestigious music conservatory.
1061 #Extraction#: The question asks for the name of the actor who played the jazz drummer in "Whiplash." The
1062 answer does not provide the actor's name, only reiterating the role played in the movie. Therefore, the
1063 corresponding specific actor's name in the answer is NONE.

1062 #Question#: Which fantasy novel, known for its depiction of a realm heavily inspired by medieval architecture
1063 and folklore, draws its inspiration from the author's experiences in a city renowned for these elements?
1064 #Answer#: Jonathan Strange & Mr Norrell
1065 #Extraction#: The question seeks the title of a fantasy novel inspired by medieval architecture and folklore.
1066 The answer provides this title as "Jonathan Strange & Mr Norrell," directly addressing the specific aspect of
1067 the question.

1066 ...

1067 #Question#: Who has won more Grand Slam singles titles in tennis, Roger Federer or Rafael Nadal?
1068 #Answer#: Rafael Nadal
1069 #Extraction#: The question asks for the tennis player who has won more Grand Slam singles titles between Roger
1070 Federer and Rafael Nadal. The answer identifies Rafael Nadal as the player who has won more Grand Slam singles
1071 titles, directly addressing the specific comparison made in the question.

1071 #Question#: Is "The Starry Night" a painting by Vincent van Gogh that depicts a scene from his view in an
1072 asylum in Saint-Remy-de-Provence?
1073 #Answer#: no
1074 #Extraction#: The question asks if "The Starry Night" is a painting by Vincent van Gogh that depicts a scene
1075 from his view in an asylum in Saint-Remy-de-Provence. The answer provided is "no." This response directly
1076 addresses the specific query about the painting and its origin.

1075 In your #Extraction#, clearly identify the main focus of the question, and then pinpoint the specific entity
1076 or information in the answer that corresponds to this focus. If the answer lacks the particular entity or
1077 information requested by the question, state that the relevant specific entity in the answer is NONE. Notice,
1078 you DO NOT need to judge the correctness of the answer.
1079 #Question#: {question}
#Answer#: {answer}
#Extraction#:

Table 8: The prompt for Non-Fabrication Hallucination Checking.

1080 As a truthful and objective query specialist, your role is to craft precise queries for verifying the accuracy
1081 of provided answers. In the #Thought-k# section, start by identifying indirect reference not indicated in both
1082 the question and the answer, guiding the focus of your initial queries. Then, scrutinize each detail in the
1083 answer to determine what needs verification and propose the corresponding #Query-k#. For information not
1084 indicated in both, initiate with a direct query and a rephrased broader context version in brackets. For
1085 details given in the answer, include the claim in your query, such as "Did (entity from the answer) do
(action/question's focus)?" and append a more general query without specifying the key entity for a wider
context in brackets. Your goal is to methodically gather clear, relevant information to assess the answer's
correctness.

#Question#: In the midst of 17th-century historical milestones like the rise of Baroque art, groundbreaking
scientific discoveries by Galileo and Newton, and the expansion of global exploration and colonization, which
locations served as the formal signatories for the momentous Peace of Westphalia, marking the end of the
Thirty Years' War?
#Answer#: Munster and Osnabruck, Germany, and it was signed in 1648.
#Thought-1#: The first query should confirm whether the Peace of Westphalia was indeed signed in Munster and
Osnabruck, Germany, as provided by the answer.
#Query-1#: Was the Peace of Westphalia signed in Munster and Osnabruck, Germany? [Where was the Peace of
Westphalia signed?]
#Knowledge-1#: (Peace of Westphalia, signed in, Munster and Osnabruck, Germany)
#Thought-2#: Having confirmed the locations, the next step is to validate the year '1648' of the signing, as
mentioned in the answer.
#Query-2#: Was the Peace of Westphalia signed in the year 1648? [When was the Peace of Westphalia signed?]
#Knowledge-2#: (Peace of Westphalia, signed in, October 1648)
#Thought-3#: All the necessary information to judge the correctness of the answer has been obtained, so the
query process can now be concluded.

#Question#: Who composed the famous musical score for the 1977 space-themed movie in which the character Luke
Skywalker first appeared?
#Answer#: Joy Williams composed the score for "Star Wars."
#Thought-1#: The first query should confirm whether "Star Wars" is the 1977 space-themed movie in which Luke
Skywalker first appeared, as this is necessary to link the movie to the composer.
#Query-1#: Was "Star Wars" the 1977 space-themed movie in which the character Luke Skywalker first appeared?
[Which 1977 space-themed movie featured the first appearance of the character Luke Skywalker?]
#Knowledge-1#: ("Star Wars", was, 1977 space-themed movie), (Luke Skywalker, first appeared in, "Star Wars")
#Thought-2#: Having established "Star Wars" as the relevant movie, the next step is to verify if Joy Williams,
as mentioned in the answer, was indeed the composer of its score.
#Query-2#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
#Knowledge-2#: (Joy Williams, did not compose, "Star Wars" score), (John Williams, composed, "Star Wars"
score)
#Thought-3#: Found one wrong detail, we do not need further query.

#Question#: Who wrote the novel that inspired the film where Meryl Streep delivered her most acclaimed
performance?
#Answer#: William Styron
#Thought-1#: The first query should identify the film featuring Meryl Streep's most acclaimed performance, as
it's not specified in the question or answer.
#Query-1#: In which film did Meryl Streep deliver her most acclaimed performance? [What is Meryl Streep's most
critically acclaimed film role?]
#Knowledge-1#: (Meryl Streep's most acclaimed performance, cited as, "Sophie's Choice")
#Thought-2#: Knowing the film is "Sophie's Choice," next determine the novel that inspired it.
#Query-2#: Which novel inspired the film "Sophie's Choice"? [What novel was the film "Sophie's Choice" based
on?]
#Knowledge-2#: (film "Sophie's Choice", inspired by, novel of the same name)
#Thought-3#: With "Sophie's Choice" identified as the film and novel, the final step is to verify the author
"William Styron", as indicated by the answer.
#Query-3#: Did William Styron write the novel "Sophie's Choice"? [Who wrote the novel "Sophie's Choice"?]
#Knowledge-3#: No specific information is available.
#Thought-4#: All relevant queries have been exhausted, and no further related inquiries can be made at this
point.

...

#Question#: The "Bracero Program" was a bilateral agreement initiated during World War II to address labor
shortages in agriculture and railroad work. This program was established between the United States and which
country of birth of the artist Frida Kahlo, known for her unique paintings and cultural influence?
#Answer#: Mexico, which is bordered to the north by the United States.
#Thought-1#: The first query should confirm Frida Kahlo's country of birth, as it will directly link to the
country involved in the Bracero Program.
#Query-1#: What is the country of birth of artist Frida Kahlo? [Where was artist Frida Kahlo born?]
#Knowledge-1#: No specific information is available.
#Thought-2#: Since specific information about Frida Kahlo's country of birth isn't available, let's move
forward to verify if Mexico was the country involved in the "Bracero Program" with the United States, as
indicated in the answer.
#Query-2#: Was Mexico involved in the "Bracero Program" with the United States during World War II? [Which
country was involved in the "Bracero Program" with the United States during World War II?]
#Knowledge-2#: (Mexico, involved in, "Bracero Program" with United States), (Bracero Program, occurred during,
World War II)
#Thought-3#: Finally, confirm that Mexico is indeed bordered to the north by the United States, as this is
part of the answer provided.
#Query-3#: Is Mexico bordered to the north by the United States? [Which country is bordered to the north by
the United States?]
#Knowledge-3#: (Mexico, bordered to the north by, United States)
#Thought-4#: All the necessary information to judge the correctness of the answer has been obtained, so the
query process can now be concluded.

Please ensure that all queries are direct, clear, and explicitly relate to the specific context provided in
the question and answer. Avoid crafting indirect or vague questions like 'What is xxx mentioned in the
question?' Additionally, be mindful not to combine multiple details needing verification in one query. Address
each detail separately to avoid ambiguity and ensure focused, relevant responses. Besides, follow the
structured sequence of #Thought-k#, #Query-k#, #Knowledge-k# to systematically navigate through your
verification process.

#Question#: {question}
#Answer#: {answer}

1134 As a truthful and objective query specialist, your role is to craft precise queries for verifying the accuracy
1135 of provided answers. In the #Thought-k# section, start by identifying indirect reference not indicated in both
1136 the question and the answer, guiding the focus of your initial queries. Then, scrutinize each detail in the
1137 answer to determine what needs verification and propose the corresponding #Query-k#. For information not
1138 indicated in both, initiate with a direct query and a rephrased broader context version in brackets. For
1139 details given in the answer, include the claim in your query, such as "Did (entity from the answer) do
1140 (action/question's focus)?" and append a more general query without specifying the key entity for a wider
1141 context in brackets. Your goal is to methodically gather clear, relevant information to assess the answer's
1142 correctness.

#Question#: In the midst of 17th-century historical milestones like the rise of Baroque art, groundbreaking
1140 scientific discoveries by Galileo and Newton, and the expansion of global exploration and colonization, which
1141 locations served as the formal signatories for the momentous Peace of Westphalia, marking the end of the
1142 Thirty Years' War?
1143 #Answer#: Munster and Osnabruck, Germany, and it was signed in 1648.
1144 #Thought-1#: The first query should confirm whether the Peace of Westphalia was indeed signed in Munster and
1145 Osnabruck, Germany, as provided by the answer.
1146 #Query-1#: Was the Peace of Westphalia signed in Munster and Osnabruck, Germany? [Where was the Peace of
1147 Westphalia signed?]
1148 #Knowledge-1#: Yes, the Peace of Westphalia was signed in Munster and Osnabruck, Germany.
1149 #Thought-2#: Having confirmed the locations, the next step is to validate the year '1648' of the signing, as
1150 mentioned in the answer.
1151 #Query-2#: Was the Peace of Westphalia signed in the year 1648? [When was the Peace of Westphalia signed?]
1152 #Knowledge-2#: Yes, the Peace of Westphalia was signed in October 1648.
1153 #Thought-3#: All the necessary information to judge the correctness of the answer has been obtained, so the
1154 query process can now be concluded.

#Question#: Who composed the famous musical score for the 1977 space-themed movie in which the character Luke
1155 Skywalker first appeared?
1156 #Answer#: Joy Williams composed the score for "Star Wars."
1157 #Thought-1#: The first query should confirm whether "Star Wars" is the 1977 space-themed movie in which Luke
1158 Skywalker first appeared, as this is necessary to link the movie to the composer.
1159 #Query-1#: Was "Star Wars" the 1977 space-themed movie in which the character Luke Skywalker first appeared?
1160 [Which 1977 space-themed movie featured the first appearance of the character Luke Skywalker?]
1161 #Knowledge-1#: "Star Wars," released in 1977, is the space-themed movie in which the character Luke Skywalker
1162 first appeared.
1163 #Thought-2#: Having established "Star Wars" as the relevant movie, the next step is to verify if Joy Williams,
1164 as mentioned in the answer, was indeed the composer of its score.
1165 #Query-2#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
1166 #Knowledge-2#: No, the score for "Star Wars" was composed by John Williams, instead of Joy Williams.
1167 #Thought-3#: Found one wrong detail, we do not need further query.

#Question#: Who wrote the novel that inspired the film where Meryl Streep delivered her most acclaimed
1168 performance?
1169 #Answer#: William Styron
1170 #Thought-1#: The first query should identify the film featuring Meryl Streep's most acclaimed performance, as
1171 it's not specified in the question or answer.
1172 #Query-1#: In which film did Meryl Streep deliver her most acclaimed performance? [What is Meryl Streep's most
1173 critically acclaimed film role?]
1174 #Knowledge-1#: Meryl Streep's most acclaimed performance is often cited as her role in "Sophie's Choice."
1175 #Thought-2#: Knowing the film is "Sophie's Choice," next determine the novel that inspired it.
1176 #Query-2#: Which novel inspired the film "Sophie's Choice"? [What novel was the film "Sophie's Choice" based
1177 on?]
1178 #Knowledge-2#: The film "Sophie's Choice" is inspired by the novel with the same name.
1179 #Thought-3#: With "Sophie's Choice" identified as the film and novel, the final step is to verify the author
1180 "William Styron", as indicated by the answer.
1181 #Query-3#: Did William Styron write the novel "Sophie's Choice"? [Who wrote the novel "Sophie's Choice"?]
1182 #Knowledge-3#: No specific information is available.
1183 #Thought-4#: All relevant queries have been exhausted, and no further related inquiries can be made at this
1184 point.

1169 ...

1170 ...

#Question#: The "Bracero Program" was a bilateral agreement initiated during World War II to address labor
1171 shortages in agriculture and railroad work. This program was established between the United States and which
1172 country of birth of the artist Frida Kahlo, known for her unique paintings and cultural influence?
1173 #Answer#: Mexico, which is bordered to the north by the United States.
1174 #Thought-1#: The first query should confirm Frida Kahlo's country of birth, as it will directly link to the
1175 country involved in the Bracero Program.
1176 #Query-1#: What is the country of birth of artist Frida Kahlo? [Where was artist Frida Kahlo born?]
1177 #Knowledge-1#: No specific information is available.
1178 #Thought-2#: Since specific information about Frida Kahlo's country of birth isn't available, let's move
1179 forward to verify if Mexico was the country involved in the "Bracero Program" with the United States, as
1180 indicated in the answer.
1181 #Query-2#: Was Mexico involved in the "Bracero Program" with the United States during World War II? [Which
1182 country was involved in the "Bracero Program" with the United States during World War II?]
1183 #Knowledge-2#: Yes, Mexico was involved in the "Bracero Program" with the United States during World War II.
1184 #Thought-3#: Finally, confirm that Mexico is indeed bordered to the north by the United States, as this is
1185 part of the answer provided.
1186 #Query-3#: Is Mexico bordered to the north by the United States? [Which country is bordered to the north by
1187 the United States?]
1188 #Knowledge-3#: Yes, Mexico is bordered to the north by the United States.
1189 #Thought-4#: All the necessary information to judge the correctness of the answer has been obtained, so the
1190 query process can now be concluded.

Please ensure that all queries are direct, clear, and explicitly relate to the specific context provided in
1184 the question and answer. Avoid crafting indirect or vague questions like 'What is xxx mentioned in the
1185 question?' Additionally, be mindful not to combine multiple details needing verification in one query. Address
1186 each detail separately to avoid ambiguity and ensure focused, relevant responses. Besides, follow the
1187 structured sequence of #Thought-k#, #Query-k#, #Knowledge-k# to systematically navigate through your
1188 verification process.

#Question#: {question}
1189 #Answer#: {answer}

Table 10: The prompt for Step-wise Reasoning and Query based on Unstructured Knowledge.

1188
1189
1190
1191
1192
1193 As an objective responder, your primary role is to provide accurate answers in triplets form by extracting
1194 relevant information from available knowledge sources, which are presented as article titles and summaries.
1195 Your task involves carefully reviewing these articles to find information directly pertinent to the questions
1196 asked. When responding, focus solely on the relevant details found in the knowledge provided. If the provided
1197 knowledge does not contain the necessary details to answer a question, respond with "No specific information
1198 is available."

1197 #Query#: Was the Peace of Westphalia signed in Munster and Osnabruck, Germany? [Where was the Peace of
1198 Westphalia signed?]
1199 #Knowledge#: Title: Peace of Westphalia. Article: The Peace of Westphalia (,) is the collective name for two
1200 peace treaties signed in October 1648 in the Westphalian cities of Osnabruck and Munster. They ended the
1201 Thirty Years' War (1618-1648) and brought peace to the Holy Roman Empire, closing a calamitous period of
1202 European history that killed approximately eight million people. Holy Roman Emperor Ferdinand III, the
1203 kingdoms of France and Sweden, and their respective allies among the princes of the Holy Roman Empire,
1204 participated in the treaties. The negotiation process was lengthy and complex.
1205 Title: Peace of Westphalia. Article: Talks took place in two cities, because each side wanted to meet on
1206 territory under its own control. A total of 109 delegations arrived to represent the belligerent states, but
1207 not all delegations were present at the same time. Two treaties were signed to end the war in the Empire: the
1208 Treaty of Munster and the Treaty of Osnabruck.
1209 #Answer#: (Peace of Westphalia, signed in, Munster and Osnabruck, Germany)

1205 #Query#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
1206 #Knowledge#: Title: Star Wars (soundtrack). Article: Star Wars (Original Motion Picture Soundtrack) is the
1207 soundtrack album to the 1977 film "Star Wars", composed and conducted by John Williams and performed by the
1208 London Symphony Orchestra. Williams' score for "Star Wars" was recorded over eight sessions at Anvil Studios
1209 in Denham, England on March 5, 812, 15 and 16, 1977. The score was orchestrated by Williams, Herbert W.
1210 Spencer, Alexander Courage, Angela Morley, Arthur Morton and Albert Woodbury. Spencer orchestrated the scores
1211 for "The Empire Strikes Back" and "Return of the Jedi".
1212 Title: Music of Star Wars. Article: For the Disney+ series "The Book of Boba Fett", Ludwig Goransson composes
1213 the main theme, while Joseph Shirley composes the score. "Obi-Wan Kenobi". For the Disney+ series "Obi-Wan
1214 Kenobi", John Williams returned to write the main theme. Natalie Holt composed the rest of the score, making
1215 her the first woman to score a live-action "Star Wars" project. "Andor". For the Disney+ series "Andor",
1216 Nicholas Britell composes the score. "Ahsoka".
1217 #Answer#: (Joy Williams, did not compose, "Star Wars" score), (John Williams, composed, "Star Wars" score)

1213 #Query#: Did William Styron write the novel "Sophie's Choice"? [Who wrote the novel "Sophie's Choice"?]
1214 #Knowledge#: Title: Sophie's Choice (novel). Article: "Sophie's Choice" generated significant controversy at
1215 time of its publication. Sylvie Mathe notes that "Sophie's Choice", which she refers to as a "highly
1216 controversial novel", appeared in press in the year following the broadcast of the NBC miniseries "Holocaust"
1217 (1978), engendering a period in American culture where "a newly-raised consciousness of the Holocaust was
1218 becoming a forefront public issue."
1219 Title: Sophie's Choice (novel). Article: Sylvie Mathe notes that Styron's "position" in the writing of this
1220 novel was made clear in his contemporary interviews and essays, in the latter case, in particular "Auschwitz",
1221 "Hell Reconsidered", and "A Wheel of Evil Come Full Circle", and quotes Alvin Rosenfeld's summary of Styron's
1222 position, where Rosenfeld states that: Rosenfeld, summarizing, states, "The drift of these revisionist views,
1223 all of which culminate in Sophie's Choice, is to take the Holocaust out of Jewish and Christian history and
1224 place it within a generalized history of evil."
1225 #Answer#: No specific information is available.

1221 ...

1222 #Query#: Was Mexico involved in the "Bracero Program" with the United States during World War II? [Which
1223 country was involved in the "Bracero Program" with the United States during World War II?]
1224 #Knowledge#: Title: Latin America during World War II. Article: In addition to those in the armed forces, tens
1225 of thousands of Mexican men were hired as farm workers in the United States during the war years through the
1226 "Bracero" program, which continued and expanded in the decades after the war. World War II helped spark an era
1227 of rapid industrialization known as the Mexican Miracle. Mexico supplied the United States with more strategic
1228 raw materials than any other country, and American aid spurred the growth of industry. President Avila was
1229 able to use the increased revenue to improve the country's credit, invest in infrastructure, subsidize food,
1230 and raise wages.
1231 Title: Military history of Mexico. Article: Although most countries in the Western Hemisphere eventually
1232 entered the war on the Allies' side, Mexico and Brazil were the only Latin American nations that sent troops
1233 to fight overseas. The cooperation of Mexico and the United States in World War II helped bring about
1234 reconciliation between the two countries at the leadership level. In the civil arena, the Bracero Program gave
1235 thousands of Mexicans the opportunity to work in the US in support of the Allied war effort. This also granted
1236 them an opportunity to gain US citizenship by enlisting in the military.
1237 #Answer#: (Mexico, involved in, "Bracero Program" with United States), (Bracero Program, occurred during,
1238 World War II)

1233 #Query#: {question}
1234 #Knowledge#: {knowledge}
1235 #Answer#:

Table 11: The prompt of the knowledge optimization for Structured Knowledge.

1242
1243
1244
1245
1246
1247
1248 As an objective responder, your primary role is to provide accurate answers by extracting relevant information
1249 from available knowledge sources, which are presented as article titles and summaries. Your task involves
1250 carefully reviewing these articles to find information directly pertinent to the questions asked. When
1251 responding, focus solely on the relevant details found in the knowledge provided. If the provided knowledge
1252 does not contain the necessary details to answer a question, respond with "No specific information is
1253 available."
1254 #Query#: Was the Peace of Westphalia signed in Munster and Osnabruck, Germany? [Where was the Peace of
1255 Westphalia signed?]
1256 #Knowledge#: Title: Peace of Westphalia. Article: The Peace of Westphalia (,) is the collective name for two
1257 peace treaties signed in October 1648 in the Westphalian cities of Osnabruck and Munster. They ended the
1258 Thirty Years' War (1618-1648) and brought peace to the Holy Roman Empire, closing a calamitous period of
1259 European history that killed approximately eight million people. Holy Roman Emperor Ferdinand III, the
1260 kingdoms of France and Sweden, and their respective allies among the princes of the Holy Roman Empire,
1261 participated in the treaties. The negotiation process was lengthy and complex.
1262 Title: Peace of Westphalia. Article: Talks took place in two cities, because each side wanted to meet on
1263 territory under its own control. A total of 109 delegations arrived to represent the belligerent states, but
1264 not all delegations were present at the same time. Two treaties were signed to end the war in the Empire: the
1265 Treaty of Munster and the Treaty of Osnabruck.
1266 #Answer#: Yes, the Peace of Westphalia was signed in Munster and Osnabruck, Germany.
1267 #Query#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
1268 #Knowledge#: Title: Star Wars (soundtrack). Article: Star Wars (Original Motion Picture Soundtrack) is the
1269 soundtrack album to the 1977 film "Star Wars", composed and conducted by John Williams and performed by the
1270 London Symphony Orchestra. Williams' score for "Star Wars" was recorded over eight sessions at Anvil Studios
1271 in Denham, England on March 5, 812, 15 and 16, 1977. The score was orchestrated by Williams, Herbert W.
1272 Spencer, Alexander Courage, Angela Morley, Arthur Morton and Albert Woodbury. Spencer orchestrated the scores
1273 for "The Empire Strikes Back" and "Return of the Jedi".
1274 Title: Music of Star Wars. Article: For the Disney+ series "The Book of Boba Fett", Ludwig Goransson composes
1275 the main theme, while Joseph Shirley composes the score. "Obi-Wan Kenobi". For the Disney+ series "Obi-Wan
1276 Kenobi", John Williams returned to write the main theme. Natalie Holt composed the rest of the score, making
1277 her the first woman to score a live-action "Star Wars" project. "Andor". For the Disney+ series "Andor",
1278 Nicholas Britell composes the score. "Ahsoka".
1279 #Answer#: No, the score for "Star Wars" was composed by John Williams, instead of Joy Williams.
1280 #Query#: Did William Styron write the novel "Sophie's Choice"? [Who wrote the novel "Sophie's Choice"?]
1281 #Knowledge#: Title: Sophie's Choice (novel). Article: "Sophie's Choice" generated significant controversy at
1282 time of its publication. Sylvie Mathe notes that "Sophie's Choice", which she refers to as a "highly
1283 controversial novel", appeared in press in the year following the broadcast of the NBC miniseries "Holocaust"
1284 (1978), engendering a period in American culture where "a newly-raised consciousness of the Holocaust was
1285 becoming a forefront public issue."
1286 Title: Sophie's Choice (novel). Article: Stingo, a novelist who is recalling the summer when he began his
1287 first novel, has been fired from his low-level reader's job at the publisher McGraw-Hill and has moved into a
1288 cheap boarding house in Brooklyn, where he hopes to devote some months to his writing. While he is working on
1289 his novel, he is drawn into the lives of the lovers Nathan Landau and Sophie Zawistowska, fellow boarders at
1290 the house, who are involved in an intense and difficult relationship.
1291 #Answer#: No specific information is available.
1292 ...
1293 #Query#: Was Mexico involved in the "Bracero Program" with the United States during World War II? [Which
1294 country was involved in the "Bracero Program" with the United States during World War II?]
1295 #Knowledge#: Title: Latin America during World War II. Article: In addition to those in the armed forces, tens
1296 of thousands of Mexican men were hired as farm workers in the United States during the war years through the
1297 "Bracero" program, which continued and expanded in the decades after the war. World War II helped spark an era
1298 of rapid industrialization known as the Mexican Miracle. Mexico supplied the United States with more strategic
1299 raw materials than any other country, and American aid spurred the growth of industry. President Avila was
1300 able to use the increased revenue to improve the country's credit, invest in infrastructure, subsidize food,
1301 and raise wages.
1302 Title: Military history of Mexico. Article: Although most countries in the Western Hemisphere eventually
1303 entered the war on the Allies' side, Mexico and Brazil were the only Latin American nations that sent troops
1304 to fight overseas. The cooperation of Mexico and the United States in World War II helped bring about
1305 reconciliation between the two countries at the leadership level. In the civil arena, the Bracero Program gave
1306 thousands of Mexicans the opportunity to work in the US in support of the Allied war effort. This also granted
1307 them an opportunity to gain US citizenship by enlisting in the military.
1308 #Answer#: Yes, Mexico was involved in the "Bracero Program" with the United States during World War II.
1309 #Query#: {question}
1310 #Knowledge#: {knowledge}
1311 #Answer#:

Table 12: The prompt of the knowledge optimization for Unstructured Knowledge.

1296 You are responsible for forming the final judgment on answers based on query results. Review the #Question#
1297 and #Answer#, then examine the #Query-k# and the corresponding #Knowledge-k# in triplets form. In the
1298 #Judgment# section, determine if the knowledge supports or contradicts the answer, stating whether it is
1299 CORRECT, INCORRECT, or INCONCLUSIVE. Analyze if the knowledge confirms or refutes each aspect of the answer.
1300 If all aspects are supported, the answer is CORRECT. If any aspect is contradicted, the answer is INCORRECT.
1301 If information is insufficient for a definitive judgment, classify it as INCONCLUSIVE.

1302 #Question#: In the midst of 17th-century historical milestones like the rise of Baroque art, groundbreaking
1303 scientific discoveries by Galileo and Newton, and the expansion of global exploration and colonization, which
1304 locations served as the formal signatories for the momentous Peace of Westphalia, marking the end of the
1305 Thirty Years' War?
1306 #Answer#: Mnster and Osnabrck, Germany, and it was signed in 1648.
1307 #Query-1#: Was the Peace of Westphalia signed in Mnster and Osnabrck, Germany? [Where was the Peace of
1308 Westphalia signed?]
1309 #Knowledge-1#: (Peace of Westphalia, signed in, Mnster and Osnabrck, Germany)
1310 #Query-2#: Was the Peace of Westphalia signed in the year 1648? [When was the Peace of Westphalia signed?]
1311 #Knowledge-2#: (Peace of Westphalia, signed in, October 1648)
1312 #Judgment#: The answer claims that the Peace of Westphalia was signed in Mnster and Osnabrck, Germany, in the
1313 year 1648. The answer's mention of "Mnster and Osnabrck, Germany," is confirmed by the knowledge triplet
1314 (Peace of Westphalia, signed, Mnster and Osnabrck, Germany). Concerning the year, the answer indicates "1648,"
1315 and this is validated by the knowledge triplet (Peace of Westphalia, signed, October 1648). Therefore, both
1316 the locations and the year provided in the answer are corroborated by the respective knowledge triplets,
1317 leading to the judgment of CORRECT.

1318 #Question#: Who composed the famous musical score for the 1977 space-themed movie in which the character Luke
1319 Skywalker first appeared?
1320 #Answer#: Joy Williams composed the score for "Star Wars."
1321 #Query-1#: Was "Star Wars" the 1977 space-themed movie in which the character Luke Skywalker first appeared?
1322 [Which 1977 space-themed movie featured the first appearance of the character Luke Skywalker?]
1323 #Knowledge-1#: ("Star Wars", was, 1977 space-themed movie), (Luke Skywalker, first appeared in, "Star Wars")
1324 #Query-2#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
1325 #Knowledge-2#: (Joy Williams, did not compose, "Star Wars" score), (John Williams, composed, "Star Wars"
1326 score)
1327 #Judgment#: The answer claims that "Star Wars" is the 1977 space-themed movie in which the character Luke
1328 Skywalker first appeared and Joy Williams is the composer. The first query's knowledge, with the triplets
1329 ("Star Wars", was, 1977 space-themed movie) and (Luke Skywalker, first appeared, "Star Wars"), confirms "Star
1330 Wars" as the correct movie. However, the second query contradicts the composer part of the answer. The
1331 knowledge triplets (Joy Williams, did not compose, "Star Wars" score) and (John Williams, composed, "Star
1332 Wars" score) clearly state that John Williams, not Joy Williams, composed the score. Therefore, the final
1333 judgment is INCORRECT.

1334 #Question#: Who wrote the novel that inspired the film where Meryl Streep delivered her most acclaimed
1335 performance?
1336 #Answer#: William Styron
1337 #Query-1#: In which film did Meryl Streep deliver her most acclaimed performance? [What is Meryl Streep's most
1338 critically acclaimed film role?]
1339 #Knowledge-1#: (Meryl Streep's most acclaimed performance, cited as, "Sophie's Choice")
1340 #Query-2#: Which novel inspired the film "Sophie's Choice"? [What novel was the film "Sophie's Choice" based
1341 on?]
1342 #Knowledge-2#: (film "Sophie's Choice", inspired by, novel of the same name)
1343 #Query-3#: Did William Styron write the novel "Sophie's Choice"? [Who wrote the novel "Sophie's Choice"?]
1344 #Knowledge-3#: No specific information is available.
1345 #Judgment#: The answer claims that William Styron wrote the novel which inspired the film "Sophie's Choice,"
1346 featuring Meryl Streep's most acclaimed performance. The first query's knowledge (Meryl Streep's most
1347 acclaimed performance, cited as, "Sophie's Choice") confirms "Sophie's Choice" as the film. The second query's
1348 knowledge (film "Sophie's Choice", inspired by, novel of the same name) verifies that the film is based on a
1349 novel with the same name. However, the third query lacks specific information about the author of "Sophie's
1350 Choice." Since there is no direct confirmation that William Styron wrote the novel, the final judgment for the
1351 answer is INCONCLUSIVE.

1352 ...

1353 #Question#: The "Bracero Program" was a bilateral agreement initiated during World War II to address labor
1354 shortages in agriculture and railroad work. This program was established between the United States and which
1355 country of birth of the artist Frida Kahlo, known for her unique paintings and cultural influence?
1356 #Answer#: Mexico, which is bordered to the north by the United States.
1357 #Query-1#: What is the country of birth of artist Frida Kahlo? [Where was artist Frida Kahlo born?]
1358 #Knowledge-1#: No specific information is available.
1359 #Query-2#: Was Mexico involved in the "Bracero Program" with the United States during World War II? [Which
1360 country was involved in the "Bracero Program" with the United States during World War II?]
1361 #Knowledge-2#: (Mexico, involved in, "Bracero Program" with United States), (Bracero Program, occurred during,
1362 World War II)
1363 #Query-3#: Is Mexico bordered to the north by the United States? [Which country is bordered to the north by
1364 the United States?]
1365 #Knowledge-3#: (Mexico, bordered to the north by, United States)
1366 #Judgment#: The answer claims that Mexico, known as the birthplace of artist Frida Kahlo and bordered to the
1367 north by the United States, participated in the "Bracero Program." The first query lacks information about
1368 Frida Kahlo's birthplace, making it impossible to verify this specific part of the answer. The knowledge from
1369 the second query (Mexico, involved in, "Bracero Program" with United States), (Bracero Program, occurred
1370 during, World War II) directly confirms Mexico's involvement in the program, during World War II, and the
1371 third query's knowledge (Mexico, bordered to the north by, United States) affirms the geographical detail.
1372 Since both Frida Kahlo's birthplace and the country participating in the "Bracero Program" with the United
1373 States refer to the same country, the absence of direct information about Kahlo's birthplace does not affect
1374 the overall correctness of the answer. Therefore, the final judgment is CORRECT.

1375 #Question#: {question}
1376 #Answer#: {answer}
1377 {query_knowledge}
1378 #Judgment#:

Table 13: The prompt of the judgment based on Structured Knowledge.

1350
1351
1352 You are responsible for forming the final judgment on answers based on query results. Review the #Question#
and #Answer#, then examine the #Query-k# and the corresponding #Knowledge-k#. In the #Judgment# section,
1353 determine if the knowledge supports or contradicts the answer, stating whether it is CORRECT, INCORRECT, or
INCONCLUSIVE. Analyze if the knowledge confirms or refutes each aspect of the answer. If all aspects are
1354 supported, the answer is CORRECT. If any aspect is contradicted, the answer is INCORRECT. If information is
insufficient for a definitive judgment, classify it as INCONCLUSIVE.
1355
#Question#: In the midst of 17th-century historical milestones like the rise of Baroque art, groundbreaking
1356 scientific discoveries by Galileo and Newton, and the expansion of global exploration and colonization, which
locations served as the formal signatories for the momentous Peace of Westphalia, marking the end of the
1357 Thirty Years' War?
#Answer#: Mnster and Osnabrck, Germany, and it was signed in 1648.
#Query-1#: Was the Peace of Westphalia signed in Mnster and Osnabrck, Germany? [Where was the Peace of
1359 Westphalia signed?]
#Knowledge-1#: Yes, the Peace of Westphalia was signed in Mnster and Osnabrck, Germany.
1360 #Query-2#: Was the Peace of Westphalia signed in the year 1648? [When was the Peace of Westphalia signed?]
#Knowledge-2#: Yes, the Peace of Westphalia was signed in October 1648.
1361 #Judgment#: The answer claims that the Peace of Westphalia was signed in Mnster and Osnabrck, Germany, in the
year 1648. The answer's mention of "Mnster and Osnabrck, Germany," is supported by the first query, confirming
1362 that the treaties were indeed signed in these cities. Regarding the year, the answer specifies "1648," which
is supported by the second query, verifying the signing year as 1648. Both the location and year mentioned in
1363 the answer are thus validated by the respective queries, leading to a final judgment of CORRECT.
1364
#Question#: Who composed the famous musical score for the 1977 space-themed movie in which the character Luke
1365 Skywalker first appeared?
#Answer#: Joy Williams composed the score for "Star Wars."
1366 #Query-1#: Was "Star Wars" the 1977 space-themed movie in which the character Luke Skywalker first appeared?
[Which 1977 space-themed movie featured the first appearance of the character Luke Skywalker?]
1367 #Knowledge-1#: "Star Wars," released in 1977, is the space-themed movie in which the character Luke Skywalker
first appeared.
1368 #Query-2#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
#Knowledge-2#: No, the score for "Star Wars" was composed by John Williams, instead of Joy Williams.
1369 #Judgment#: The answer states that "Star Wars" is the 1977 space-themed movie in which the character Luke
Skywalker first appeared and Joy Williams is the composer. The first query supports the movie part of the
1370 answer, confirming "Star Wars" as the 1977 film featuring Luke Skywalker. However, the second query
contradicts the composer part, revealing that John Williams, not Joy Williams, composed the score. Therefore,
1371 the final judgment is INCORRECT.
1372
#Question#: Who wrote the novel that inspired the film where Meryl Streep delivered her most acclaimed
1373 performance?
#Answer#: William Styron
1374 #Query-1#: In which film did Meryl Streep deliver her most acclaimed performance? [What is Meryl Streep's most
critically acclaimed film role?]
1375 #Knowledge-1#: Meryl Streep's most acclaimed performance is often cited as her role in "Sophie's Choice."
#Query-2#: Which novel inspired the film "Sophie's Choice"? [What novel was the film "Sophie's Choice" based
1376 on?]
#Knowledge-2#: The film "Sophie's Choice" is inspired by the novel with the same name.
1377 #Query-3#: Did William Styron write the novel "Sophie's Choice"? [Who wrote the novel "Sophie's Choice"?]
#Knowledge-3#: No specific information is available.
1378 #Judgment#: The answer claims that William Styron wrote the novel that inspired the film "Sophie's Choice," in
which Meryl Streep delivered her most acclaimed performance. The first query confirms that Meryl Streep's most
1379 acclaimed performance is often cited as her role in "Sophie's Choice." The second query establishes that the
film "Sophie's Choice" was indeed inspired by a novel of the same name. However, the third query fails to
1380 provide specific information about the author of the novel "Sophie's Choice." Due to this lack of direct
confirmation about the author, the claim that William Styron wrote the novel cannot be conclusively verified.
1381 Consequently, the final judgment for the answer is INCONCLUSIVE.
1382
1383 ...
1384
#Question#: The "Bracero Program" was a bilateral agreement initiated during World War II to address labor
1385 shortages in agriculture and railroad work. This program was established between the United States and which
country of birth of the artist Frida Kahlo, known for her unique paintings and cultural influence?
1386 #Answer#: Mexico, which is bordered to the north by the United States.
#Query-1#: What is the country of birth of artist Frida Kahlo? [Where was artist Frida Kahlo born?]
1387 #Knowledge-1#: No specific information is available.
#Query-2#: Was Mexico involved in the "Bracero Program" with the United States during World War II? [Which
1388 country was involved in the "Bracero Program" with the United States during World War II?]
#Knowledge-2#: Yes, Mexico was involved in the "Bracero Program" with the United States during World War II.
1389 #Query-3#: Is Mexico bordered to the north by the United States? [Which country is bordered to the north by
the United States?]
1390 #Knowledge-3#: Yes, Mexico is bordered to the north by the United States.
#Judgment#: The answer claims that Mexico, known as the birthplace of artist Frida Kahlo and bordered to the
1391 north by the United States, participated in the "Bracero Program." The first query does not provide specific
information about Frida Kahlo's birthplace. However, the second query confirms that Mexico was indeed involved
1392 in the "Bracero Program" with the United States during World War II. The third query verifies that Mexico is
bordered to the north by the United States. Since both Frida Kahlo's birthplace and the country participating
1393 in the "Bracero Program" with the United States refer to the same country, the absence of direct information
about Kahlo's birthplace does not affect the overall correctness of the answer. Therefore, the final judgment
1394 is CORRECT.
1395
#Question#: {question}
1396 #Answer#: {answer}
1397 {query_knowledge}
1398 #Judgment#:

Table 14: The prompt of the judgment based on Unstructured Knowledge.

```

1404 {
1405   "retrieve_and_generate": {
1406     "initial_search_query": "director of the 2007 film Jin Ji-hee",
1407     "initial_search_query_time": 2007,
1408     "initial_search_bullets": [],
1409     "initial_search_results": []
1410   },
1411   "llm_utterance": "Yes. This answer is a result of hallucination. As of my last knowledge update in 2021,
1412   there is no record of a film in 2007 featuring Jin Ji-hee directed by Kim Pil-sung. Therefore, the
1413   information provided does not align with known facts.",
1414   "generate_and_correct": {
1415     "The answer provided is a result of hallucination.": {
1416       "label": "NOT ENOUGH INFO",
1417       "fixed_claim": "",
1418       "retrieval_results": [
1419         {
1420           "title": "Chronic hallucinatory psychosis",
1421           "paragraph": "As none is forthcoming he/she tries to account for their presence and the result is a
1422           delusion, and, most frequently, a delusion of persecution. Also, it needs to be noted that the
1423           delusion is a comparatively late arrival and is the logical result of the hallucinations. Sources. A
1424           paper read at the Quarterly Meeting of the Medico-Psychological Association on February 24, 1920,
1425           written by Robert Hunter Steen, King's College Hospital, London",
1426           "score": 20.4
1427         },
1428         {
1429           "title": "Tactile hallucination",
1430           "paragraph": "Tactile hallucinations are the result of a dysfunctional somatosensory and a
1431           dysfunctional awareness regions of the brain. Tactile sensory input is produced and conducted
1432           through the spinal cord and thalamus and it is received at the primary somatosensory cortex. Once it
1433           has reached the primary somatosensory cortex, it is distributed across the brain and it will not be
1434           processed unless it is important and one pays close attention to the information based on a specific
1435           context.",
1436           "score": 20.2
1437         }
1438       ]
1439     },
1440     "There is no record of a film in 2007 featuring Jin Ji-hee directed by Kim Pil-sung.": {
1441       "label": "REFUTES",
1442       "fixed_claim": "",
1443       "retrieval_results": [
1444         {
1445           "title": "Yim Pil-sung",
1446           "paragraph": "Inspired by the titular fairy tale, in \"Hansel and Gretel\" (2007) a young man (Chun
1447           Jung-myung) gets lost in a forest and stumbles into a house inhabited by three strange children
1448           (Shim Eun-kyung, Eun Ji-won and Jin Ji-hee) who refuse to let him leave. It received a Special
1449           Mention at the 12th Puchon International Fantastic Film Festival in 2008 and won two awards at the
1450           29th Fantasporto in 2009, the Special Jury Prize in the Fantasy competition and Best Film in the
1451           Orient Express sidebar.",
1452           "score": 21.9
1453         },
1454         {
1455           "title": "Venus and Mars (2007 film)",
1456           "paragraph": "Venus and Mars (; lit. \"Fight\") is a 2007 South Korean romantic comedy film. It was
1457           directed by Han Ji-seung and stars Sol Kyung-gu and Kim Tae-hee. Plot. Despite having little in
1458           common, Sang-min and Jin-ah fall in love and get married. Before long they come to realise that they
1459           actually can't stand each other, but the resulting split is far from amicable. Things come to a head
1460           when Jin-ah refuses to return the pendulum from Sang-min's beloved clock, and their petty squabbles
1461           degenerate into life-threatening violence. Release.",
1462           "score": 21.8
1463         }
1464       ]
1465     }
1466   ],
1467   "combined_evidences": [],
1468   "combined_utterance": "I'm sorry, but I couldn't find any relevant information to confirm or deny the
1469   director of the 2007 film Jin Ji-hee was in.",
1470   "feedback": [
1471     "- Relevant: The response is on-topic and directly addresses the user's question about the director of the
1472     2007 film Jin Ji-hee was in. 100/100",
1473     "- Temporally Correct: The response correctly uses the past tense to describe the search for the
1474     information. 100/100",
1475     "- Natural: The response sounds natural and maintains a polite tone. 100/100",
1476     "- Non-Repetitive: The response does not repeat any previous statement. 100/100"
1477   ],
1478   "feedback_scores": [
1479     100,
1480     100,
1481     100,
1482     100
1483   ],
1484   "refined_utterance": "I'm sorry, but I couldn't find any relevant information to confirm or deny the
1485   director of the 2007 film Jin Ji-hee was in.",
1486   "user_utterance": "#Question#: Who directed the 2007 film Jin Ji-hee was in? #Answer#: The director's name
1487   was Kim Pil-sung. Is this answer to the question a result of hallucination? Please indicate your judgment as
1488   either \"Yes\" or \"No,\" accompanied by a step-by-step explanation.",
1489   "agent_utterance": "I'm sorry, but I couldn't find any relevant information to confirm or deny the director
1490   of the 2007 film Jin Ji-hee was in.",
1491   "engine": "gpt-35-turbo",
1492   "generate_engine": "gpt-35-turbo",
1493   "draft_engine": "gpt-35-turbo",
1494   "pipeline": "early_combine",
1495   "wall_time_seconds": 20.4
1496 }

```

Table 15: An illustrative example of the process in WikiChat for retrieving, generating, and refining responses, resulting in a final fact-checked judgment.

Table 16: Retrieval outcomes for specific and general queries for both correct and incorrect details, the relevant knowledge is highlighted in purple.

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

Details for Verification	Top2 Retrieval Results for Specific Query	Top2 Retrieval Results for General Query
<p>Correct Detail: Peace of Westphalia was signed in Mnster and Osnabrck, Germany.</p>	<p>Query: Was the Peace of Westphalia signed in Mnster and Osnabrck, Germany?</p> <p>Title: Peace of Westphalia. Article: The Peace of Westphalia (,) is the collective name for two peace treaties signed in October 1648 <i>in the Westphalian cities of Osnabrck and Mnster</i>. They ended the Thirty Years' War (16181648) and brought peace to the Holy Roman Empire, ...</p> <p>Title: Peace of Westphalia. Article: In Mnster, negotiations took place between the Holy Roman Empire and France, as well as between the Dutch Republic and Spain who on 30 January 1648 signed a peace treaty ending the Eighty Years' War that was not part of the Peace of Westphalia ...</p>	<p>Query: Where was the Peace of Westphalia signed?</p> <p>Title: Peace of Westphalia. Article: The Peace of Westphalia (,) is the collective name for two peace treaties signed in October 1648 <i>in the Westphalian cities of Osnabrck and Mnster</i>. They ended the Thirty Years' War (16181648) and brought peace to the Holy Roman Empire, ...</p>
<p>Wrong Detail: Peace of Westphalia was signed in Peking, China.</p>	<p>Query: Was the Peace of Westphalia signed in Peking, China?</p> <p>Title: Sino-Portuguese Treaty of Peking. Article: Portugal followed up on this agreement by sending an envoy to Beijing, where a treaty of amity and commerce based on the protocol was drawn up. On 1 December 1887, the Treaty of Peking was signed by Chinese representatives Yikuang (Prince Qing) ...</p> <p>Title: Sino-Portuguese Treaty of Peking. Article: On 26 March 1887, Campbell and Portuguese Foreign Minister Henrique de Barros Gomes signed the four-point Lisbon Protocol: Art. 1st.A Treaty of friendship and commerce with the most favoured nation ...</p>	<p>Title: Breitenbach, Rhineland-Palatinate. Article: The Peace of Westphalia, a series of peace treaties signed between May and October 1648 in Osnabrck and Mnster, which brought the Thirty Years' War to an end, guaranteed the free practice of religion, and in the late 17th century, this law was also operative in Breitenbach, where it is likely that ...</p>
<p>Correct Detail: John Williams composed the score for "Star Wars."</p>	<p>Query: Did John Williams composed the score for "Star Wars"?</p> <p>Title: Star Wars: The Rise of Skywalker (soundtrack). Article: On January 10, 2018, it was confirmed that John Williams would return to compose and conduct "The Rise of Skywalker". The next month, Williams announced that it would be the last "Star Wars" film for which he would compose the score. In August 2019, ...</p> <p>Title: Star Wars (soundtrack). Article: Star Wars (Original Motion Picture Soundtrack) is the soundtrack album to <i>the 1977 film "Star Wars"</i>, composed and conducted by John Williams and performed by the London Symphony Orchestra. Williams' score ...</p>	<p>Query: Who composed the score for "Star Wars"?</p> <p>Title: Star Wars (soundtrack). Article: Star Wars (Original Motion Picture Soundtrack) is the soundtrack album to <i>the 1977 film "Star Wars"</i>, composed and conducted by John Williams and performed by the London Symphony Orchestra. Williams' score ...</p>
<p>Wrong Detail: Christopher Nolan composed the score for "Star Wars."</p>	<p>Query: Did Christopher Nolan composed the score for "Star Wars"?</p> <p>Title: List of awards and nominations received by Ludwig Gransson. Article: He received a second Academy Award nomination for Best Original Song thanks to "Lift Me Up", performed by Rihanna and written for the soundtrack of "Wakanda Forever". In 2020, he worked with Christopher Nolan in the film "Tenet", for which ...</p> <p>Title: Tales of the Jedi (TV series). Article: Additional music for the series is composed by Sean Kiner, Deana Kiner, David Glen Russell, Nolan Markey and Peter Lam. Walt Disney Records released the soundtrack for the first season of "Tales of the Jedi" digitally on October 26, 2022, alongside the ...</p>	<p>Title: Music of Star Wars. Article: James L. Venable and Paul Dinletir composed the music of (20032005) 2D animated series, Ryan Shore serves as the composer for "Star Wars: Forces of Destiny" (20172018) and "Star Wars Galaxy of Adventures" (20182020), and Michael Tavera composes the score to ...</p>

Table 17: Performance comparison of different methods for hallucination detection in QA task, evaluated based on different version of GPT-3.5. Results of methods using external ground truth knowledge (i.e., knowledge provided by HaluEval) are shown inside the parentheses, and results generated based on Wiki knowledge are shown outside the parentheses.

Model	Method	TPR (%)	TNR (%)	Avg Acc (%)
gpt-3.5-turbo-0613	HaluEval (Vanilla)	37.9	82.8	60.35
	HaluEval (CoT)	66.2	35.7	50.95
	HaluEval (Knowledge)	49.8 (42.6)	65.3 (92.7)	57.55 (67.65)
	KnowHalu (Structured)	65.3 (70.4)	56.9 (83.8)	61.10 (77.10)
	KnowHalu (Unstructured)	68.2 (67.5)	46.7 (70.7)	57.45 (69.10)
	KnowHalu (Aggregation)	68.4 (71.7)	61.0 (85.7)	64.70 (78.70)
gpt-3.5-turbo-1106	HaluEval (Vanilla)	44.1	55.2	48.15
	HaluEval (CoT)	66.5	21.6	44.05
	HaluEval (Knowledge)	34.4 (38.1)	71.7 (75.7)	53.05 (56.90)
	KnowHalu (Structured)	72.6 (75.7)	66.6 (80.0)	69.60 (77.85)
	KnowHalu (Unstructured)	77.3 (68.9)	53.2 (75.7)	65.25 (72.30)
	KnowHalu (Aggregation)	76.3 (77.5)	67.8 (83.1)	72.05 (80.30)

Table 18: Performance comparison of KnowHalu using different query formulations, evaluated using the Starling-7B model and off-the-shelf knowledge for the QA task. The table reports the True Positive Rate (TPR), Abstain Rate for Positive cases (ARP), True Negative Rate (TNR), Abstain Rate for Negative cases (ARN), and Average Accuracy (Avg Acc) for each method. The optimal results for structured knowledge and unstructured knowledge are respectively highlighted in bold.

Formulation	Method	TPR	ARP	TNR	ARN	Avg Acc (%)
Specific Query	KnowHalu (Structured)	60.7	12.8	86.6	6.1	73.65
	KnowHalu (Unstructured)	70.3	3.5	86.5	4.5	78.40
General Query	KnowHalu (Structured)	67.8	8.6	83.1	8.3	75.45
	KnowHalu (Unstructured)	72.4	4.2	85.9	5.5	79.15
Combined Queries	KnowHalu (Structured)	65.8	7.9	84.5	6.9	75.15
	KnowHalu (Unstructured)	68.9	5.1	84.2	5.3	76.55

Table 19: Performance comparison of KnowHalu using different query formulations, evaluated using the Starling-7B model for the Text Summarization task, the K is set to 3. The table reports the True Positive Rate (TPR), True Negative Rate (TNR), and Average Accuracy (Avg Acc) for each method. The optimal results for structured knowledge and unstructured knowledge are respectively highlighted in bold.

Formulation	Method	TPR	TNR	Avg Acc (%)
Specific Query	KnowHalu (Structured)	80.0	45.2	62.6
	KnowHalu (Unstructured)	67.6	61.6	64.6
General Query	KnowHalu (Structured)	87.8	34.2	61.0
	KnowHalu (Unstructured)	84.0	41.8	62.9
Combined Queries	KnowHalu (Structured)	80.2	45.4	62.8
	KnowHalu (Unstructured)	65.0	67.2	66.1

Table 20: Performance comparison of KnowHalu using different number of retrieved passages K , evaluated based on the Starling-7B model for Text Summarization Task. The table reports the True Positive Rate (TPR), True Negative Rate (TNR), and Average Accuracy (Avg Acc) for each method. The optimal results for structured knowledge and unstructured knowledge are respectively highlighted in bold.

Top- K Passages	Method	TPR	TNR	Avg Acc (%)
$K = 1$	KnowHalu (Structured)	84.8	39.8	62.3
	KnowHalu (Unstructured)	76.2	49.6	62.9
$K = 2$	KnowHalu (Structured)	81.0	43.8	62.4
	KnowHalu (Unstructured)	68.0	62.0	65.0
$K = 3$	KnowHalu (Structured)	80.2	45.4	62.8
	KnowHalu (Unstructured)	65.0	67.2	66.1
$K = 4$	KnowHalu (Structured)	79.6	45.2	62.4
	KnowHalu (Unstructured)	62.0	68.6	65.3
$K = 5$	KnowHalu (Structured)	81.0	43.8	62.4
	KnowHalu (Unstructured)	58.8	70.4	64.6

1566
 1567
 1568
 1569
 1570
 1571
 1572
 1573
 1574
 1575
 1576
 1577
 1578
 1579
 1580
 1581
 1582
 1583
 1584
 1585
 1586
 1587
 1588
 1589
 1590
 1591
 1592
 1593
 1594
 1595
 1596
 1597
 1598
 1599
 1600
 1601
 1602
 1603
 1604
 1605
 1606
 1607
 1608
 1609
 1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619

Table 21: Detailed selection of thresholds δ_1 and δ_2 , along with the corresponding quantiles, determined on the validation set for the best setting of each task. The query formulation presented in the table is used for the base judgment. Typically, the supplementary judgment adopts the same formulation to achieve the optimal performance, unless otherwise specified.

Task	Model	Knowledge Source	Formulation of Query	Knowledge Form for Base Judgment	$\delta_1 (q_1)$	$\delta_2 (q_2)$
QA	Starling-7B	off-the-shelf knowledge	General Query	Unstructured Knowledge	0.986236 (0.10)	0.995475 (0.25)
		Wiki retrieval knowledge	Combined Query	Unstructured Knowledge	0.999455 (0.70)	0.999241 (0.70)
	Mistral	off-the-shelf knowledge	General Query ¹	Unstructured Knowledge	0.995717 (0.40)	0.958954 (0.40)
		Wiki retrieval knowledge	Combined Query	Unstructured Knowledge	0.997418 (0.55)	0.999343 (0.65)
	GPT-3.5	off-the-shelf knowledge	Combined Query	Structured Knowledge	0.999978 (0.70)	0.999994 (0.75)
		Wiki retrieval knowledge	Combined Query	Structured Knowledge	0.999884 (0.30)	0.999934 (0.30)
Text Summarization	Starling-7B	Original Document	Combined Query	Unstructured Knowledge	0.999103 (0.25)	0.999436 (0.30)
	GPT-3.5	Original Document	Combined Query	Structured Knowledge	0.999881 (0.20)	0.999915 (0.45)

¹ Here, as a special case, the supplementary judgment instead uses a combined query formulation.