FactCHD: Benchmarking Fact-Conflicting Hallucination Detection

Xiang Chen[1,4], **Duanzheng Song**[2], **Honghao Gui**[1,4], **Chenxi Wang**[2,4], **Ningyu Zhang**[2,4*], **Yong Jiang**[3], **Fei Huang**[3], **Chengfei Lyu**[3], **Dan Zhang**[2], **Huajun Chen**[1,4*]
[1]College of Computer Science and Technology, Zhejiang University
[2]School of Software Technology, Zhejiang University
[3]Alibaba Group
[4]Zhejiang University - Ant Group Joint Research Center for Knowledge Graphs
{xiang_chen, sdz, guihonghao, sunnywcx, zhangningyu, dan.zhang, huajunsir}@zju.edu.cn
{yongjiang.jy, f.huang, chengfei.lcf}@alibaba-inc.com

Abstract

Despite their impressive generative capabilities, LLMs are hindered by fact-conflicting hallucinations in real-world applications. The accurate identification of hallucinations in texts generated by LLMs, especially in complex inferential scenarios, is a relatively unexplored area. To address this gap, we present **FACTCHD**, a dedicated benchmark designed for the detection of fact-conflicting hallucinations from LLMs. **FACTCHD** features a diverse dataset that spans various factuality patterns, including vanilla, multi-hop, comparison, and set operation. A distinctive element of **FACTCHD** is its integration of fact-based evidence chains, significantly enhancing the depth of evaluating the detectors' explanations. Experiments on different LLMs expose the shortcomings of current approaches in detecting factual errors accurately. Furthermore, we introduce **TRUTH-TRIANGULATOR** that synthesizes reflective considerations by tool-enhanced ChatGPT and LoRA-tuning based on Llama2, aiming to yield more credible detection through the amalgamation of predictive results and evidence.

1 Introduction

Large Language Models (LLMs) [Zhao *et al.*, 2023] are susceptible to generating text that, while seemingly credible, can be factually inaccurate or vague, leading to the spread of misinformation online [Yin *et al.*, 2023; Ji *et al.*, 2023; Huang *et al.*, 2023]. This issue, referred to as *fact-conflicting hallucination* [Zhang *et al.*, 2023], arises from the incorporation of incorrect or obsolete knowledge into the models' parameters and from the models' inherent limitations in complex cognitive ability. These shortcomings constrain LLMs' deployment in critical domains like finance, healthcare, and law, and amplify the propagation of erroneous information. Therefore, it is crucial to effectively detect fact-conflicting hallucinations for mitigating or editing them [Yao *et al.*, 2023].

However, traditional fact verification tasks [Wadden *et al.*, 2020a; Wadden *et al.*, 2020b] are not suitable for LLM-based "QUERY-RESPONSE" data. Moreover, existing hallucination evaluation benchmarks [Li *et al.*, 2023; Muhlgay *et al.*, 2023], predominantly centering on vanilla facts and textual content, lack in-depth exploration of complex operations among facts, thus rendering their coverage against fact-conflicting hallucinations suboptimal. To bridge this gap, we introduce an inherently rigorous[Wang *et al.*, 2023b], yet authentic, task scenario: *fact-conflicting hallucination* detection, devoid of explicit claims or evidence. As shown in Figure 1, when confronted with a query and its generated response, detectors, are impelled to harness both their intrinsic knowledge and external resources, while rendering a factual judgment accompanied by an elucidative explanation.

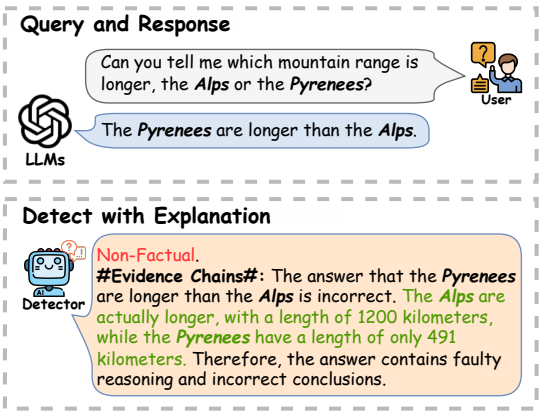

Figure 1: Illustration of fact-conflicting hallucination detection example from FACTCHD, where the green part represents factual explanation core (body part) in the chain of evidence.

In paving the way for future strides in hallucination evaluation, we introduce a new benchmark, **Fact**-**C**onflicting **H**allucination **D**etection (FACTCHD), tailored for LLMs and encompassing a variegated array of factuality patterns, including **Vanilla**, **Multi-hops**, **Comparison**, and **Set-Operation** patterns. For example in Figure 1, querying whether the "Alps" or "Pyrenees" are higher, exemplifies a comparison pattern, assessing the relative relationships among facts. Drawing inspiration from the adage *"to know it and to know the reason why of it"* by *Zhuzi*, FACTCHD extends beyond mere

* Corresponding author.

Datasets	# Source	Domains	Factuality Pattern	Evaluation	Expandability
FEVER [Thorne *et al.*, 2018]	Text	General	VAN.	CONSIST.	✗
CLIMATE-FEVER [Diggelmann *et al.*, 2020]	Text	Climate Change	VAN.	CONSIST.	✗
HEALTH-FEVER [Sarrouti *et al.*, 2021]	Text	Health	VAN.	CONSIST.	✗
SCI-FACT [Wadden *et al.*, 2020a]	Text	Scientific	VAN.	CONSIST.	✗
COVERT [Mohr *et al.*, 2022]	Text	COVID-19	VAN.	CONSIST.	✗
TABFACT [Chen *et al.*, 2020]	Table	General	VAN.	CONSIST.	✗
HOVER [Jiang *et al.*, 2020]	Text	General	MUL.	CONSIST.	✗
FEVEROUS [Aly *et al.*, 2021]	Text+Table	General	VAN.	CONSIST.	✗
HALUEVAL [Li *et al.*, 2023]	Text	General	VAN.	Hallucination	✗
FACTCHD (ours)	**KGs** () & Text	General&Vertical	VAN.&MUL.&COM.&SET.	FACT.+CHAIN.	✔

Table 1: Comparison with existing fact-checking datasets. Our FACTCHD include both general and vertical domains, such as health, COVID-19, climate, science, and medicine (genes, virus and disease). VAN., MUL., COM. and SET. are abbreviations for the distinct factuality patterns.

"QUERY-RESPONSE" labeling of hallucinations, incorporating golden chains of evidence to assess if detectors can provide coherent explanations for factualness judgment. Acknowledging the challenges of collecting comprehensive data through exhaustive human annotation, we propose a scalable data construction approach that harnesses existing knowledge graphs (KGs) and textual knowledge to create simulated hallucination instances with ChatGPT, verified with human annotation, for the efficient development of FACTCHD.

We evaluate the performance of various LLMs (such as Alpaca, Llama2-chat, and ChatGPT) using our FACTCHD benchmark across multiple settings: zero-shot, in-context learning, specialized detection tuning, and knowledge enhancement via retrieval/tools. The results indicate that specialized detection tuning and knowledge enhancement notably improve the detection of fact-conflicting hallucinations. Additionally, we present **TRUTH-TRIANGULATOR** framework grounded in "Triangulation" theory [Valenza, 2016]. This system comprises three roles: ChatGPT, enhanced with tools as the *Truth Seeker*; a detect-specific expert based on Llama2-7B-LoRA as the *Truth Guardian*; and the *Fact Verdict Manager*, which amasses evidence from another role to fortify the reliability and accuracy of the derived conclusions. TRUTH-TRIANGULATOR emphasizes the use of cross-referencing generators to astutely evaluate and adjudicate responses with potential factual discrepancies. Key insights are summarized as:

- We present FACTCHD[1], a large-scale, multi-domain evaluation benchmark with diverse factual patterns and interpretable evidence chains, setting a new standard for detecting fact-conflicting hallucinations from LLMs.
- We introduce a scalable data construction strategy leveraging KGs, etc. to efficiently develop a fact-conflicting hallucination dataset, offering stronger applicability due to the authentic, broad domain coverage of KGs.
- We devise a triangulation-based framework TRUTH-TRIANGULATOR that employs cross-referencing generators for verifying LLM responses.

2 Related Work

Hallucination in LLMs. Despite LLMs [Zhao *et al.*, 2023; Chen *et al.*, 2022; OpenAI, 2022] like ChatGPT demonstrating remarkable understanding and execution of user instructions, they are prone to confidently generating misleading "hallucinations" [Ji *et al.*, 2023; Wang *et al.*, 2023a]. These hallucinations, as categorized by [Zhang *et al.*, 2023], can be input-conflicting, context-conflicting, or fact-conflicting—with the latter being especially problematic due to the propagation of inaccurate factual information online. While previous studies have extensively examined hallucinations within natural language generation (NLG) for various NLP tasks [Shuster *et al.*, 2021; Creswell and Shanahan, 2022; Das *et al.*, 2023; Mallen *et al.*, 2022], HaluEval [Li *et al.*, 2023] has emerged as a recent benchmark for assessing LLMs' recognition of such errors. Different from HaLuEval which only analyzes ChatGPT's ability to evaluate whether hallucinatory, we specifically focus on the evaluation of fact-conflicting hallucination by constructing an interpretable benchmark that can serve as a public platform for checking the factual errors of context with explanation.

Factuality Detection. Our research is also related to prior works on fact verification in NLP tasks [Thorne *et al.*, 2018; Wadden *et al.*, 2020b; Gupta *et al.*, 2022; Dziri *et al.*, 2022; Liang *et al.*, 2023; Rashkin *et al.*, 2021; Kryściński *et al.*, 2020; Kang *et al.*, 2024], expanding the understanding of factuality beyond binary judgments. The FRANK framework [Pagnoni *et al.*, 2021] offers a detailed typology of factual errors, while [Dhingra *et al.*, 2019] explores lexical entailment in text generation. The FACTOR score by [Muhlgay *et al.*, 2023] evaluates LMs based on the likelihood of factual content. Existing benchmarks often miss the "QUERY-RESPONSE" context of LLMs, focusing solely on accuracy without providing explanatory rationales. Our contribution lies in: (1) presenting "QUERY-RESPONSE" formatted data for LLMs; (2) introducing metrics for interpretability in detecting fact-conflicting hallucinations; and (3) validating the effectiveness of TRUTH-TRIANGULATOR through cross-referential verification from multiple sources. The comparison with other datasets is detailed in Table 1.

3 Preliminaries

Task Formulation. Detecting fact-conflicting hallucinations in LLMs involves discerning factual errors in responses to human queries. A comprehensive detector must not only classify responses as factual or non-factual but also provide explanations for its judgments. We define the task as follows: **Input:** A question Q paired with an LLM-generated response R,

[1]Data is available at https://github.com/zjunlp/FactCHD.

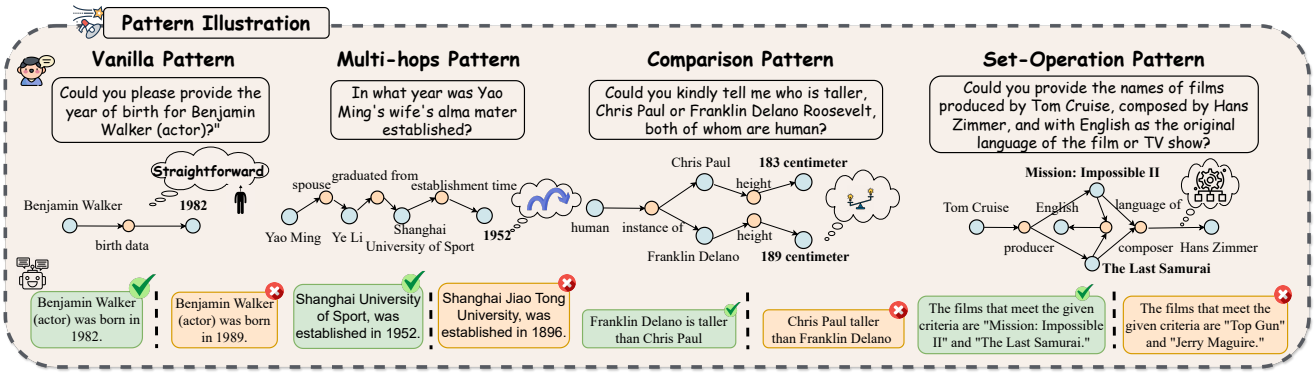

Figure 2: Overview of the factuality patterns invloved in our FACTCHD.

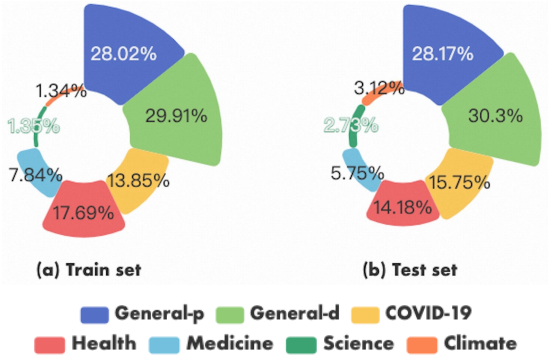

Figure 3: Domain distribution of FACTCHD, where "-p" and "-d" denote domains derived from Wikipedia and Wikidata, respectively.

which may contain various fact conflicts. **Output:** A combined label and explanation sequence $A = [l, e]$, where l is the binary factuality label (FACTUAL or NON-FACTUAL), and e articulates the rationale behind the assigned label. We evaluate the quality of e using the $ExpMatch$ metric through the golden evidence chains in FACTCHD.

Factuality Patterns. We aim to explore distinct patterns of factual errors in our FACTCHD, vividly illustrated in Figure 2. These include the *(1) vanilla pattern* dealing with factual statements that can be objectively verified using established sources, the *(2) multi-hops pattern* involving the process of concluding by connecting multiple pieces of facts, the *(3) comparison pattern* referring to the act of evaluating and comparing relative worth and relations between different pieces of facts, and the *(4) set-operation pattern* involving manipulating and combining sets of elements using operations to analyze relations between different facts. We generated corresponding "QUERY-RESPONSE" examples based on these patterns.

Knowledge-Driven Factual Foundation for Reliable "QUERY-RESPONSE" Generation. KGs [Liang *et al.*, 2024] serve as a rich trove of structured entities and relations, ideal for compositional reasoning and anchoring factual data. Alongside this, textual knowledge is critical for nuanced inference beyond basic facts. Our research focuses on collecting existing knowledge and integrating it into prompts as a factual foundation for "QUERY-RESPONS" and chain of evidence generation, as outlined below: (1) We use 438 widespread relations from Wikidata [Vrandecic and Krötzsch, 2014] and PrimeKG [Chandak *et al.*, 2023] to create varied subgraphs via K-hop walks, forming the knowledge base for generating "QUERY-RESPONS" examples with multi-hop reasoning, fact comparison, and set operation patterns. (2) We utilize text knowledge from datasets such as FEVER [Thorne *et al.*, 2018], Climate-Fever [Diggelmann *et al.*, 2020], Health-Fever [Sarrouti *et al.*, 2021], COVID-FACT [Saakyan *et al.*, 2021], and SCIFACT [Wadden *et al.*, 2020a] for generating examples with vanilla pattern.

Split	#Sample	VAN.	MULTI.	COMP.	SET-OP.
Train	51,383	31,986	8,209	5,691	5,497
Test	6,960	4,451	1,013	706	790

Table 2: Data Statistic of our FACTCHD

4 FACTCHD Benchmark Construction

Building on the aforementioned preliminaries, we develop FACTCHD, a dataset containing a wealth of training instances and an additional 6,960 carefully selected samples for evaluating fact-conflicting hallucinations from LLMs. Our dataset maintains a balanced representation of FACTUAL and NON-FACTUAL categories, offering a robust framework for assessment. The statistics and domain distribution of FACTCHD are depicted in Table 2 and Figure 3. Next, we outline the design principles of our benchmark as follows.

4.1 Collect Realistic Data as Demonstration

Subsequently, adhering to the defined factuality patterns, we manually craft corresponding queries and utilize open-source LLMs (such as ChatGLM, Alpaca, and Vicuna) to generate authentic responses. By manually annotating these responses for hallucinations, we acquire "QUERY-RESPONSE" examples annotated with hallucination presence. Our objective is to curate a robust dataset of hallucination cases that closely emulate real-world examples, providing a demonstrative foundation and establishing standards for prompt refinement to ensure alignment between generated hallucinations and these demonstrations.

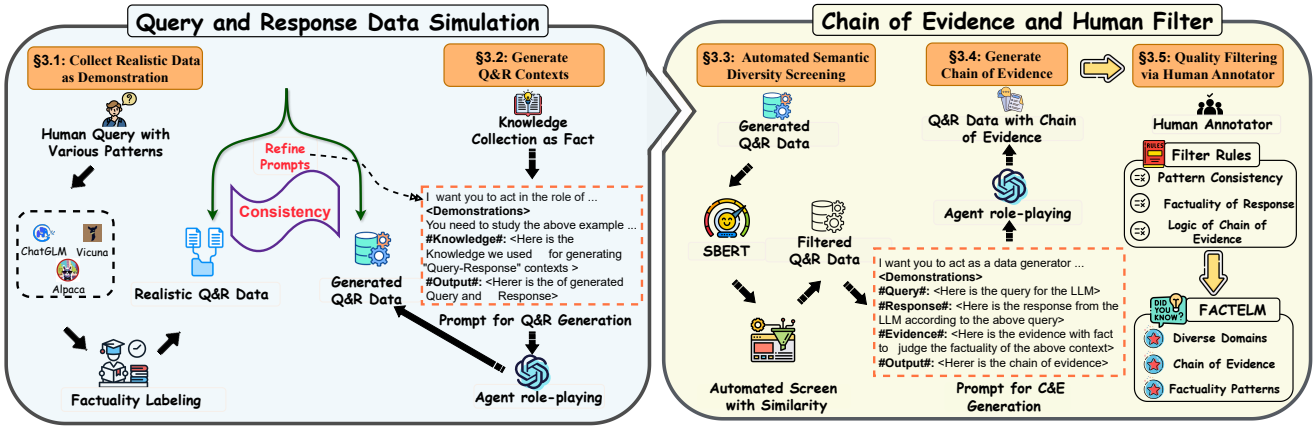

Figure 4: Overview of the construction process of FACTCHD.

4.2 Generate "QUERY-RESPONSE" Contexts

Generate "QUERY-RESPONSE" based on Knowledge with ChatGPT. We incorporate the knowledge collected on distinct factuality patterns into customized prompts, specifying whether to generate factual or non-factual responses. This guides ChatGPT in producing "QUERY-RESPONS" instances across various factuality categories with golden labels.

Refine Prompts with Consistency. In the initial phase of generating "QUERY-RESPONSE" instances, we assess the consistency of five samples per pattern against demonstrations, using majority consensus. We subjectively evaluate each context's adherence to the style and form of realistic demonstrations, employing iterative refinements of prompts to ensure that the generated data exhibits realistic patterns. We aim for an at least 95% consistency rate before scaling up the generation of "QUERY-RESPONSE" instances.

4.3 Automated Semantic Diversity Screening

To increase "QUERY-RESPONSE" context diversity, we employ Sentence-BERT (SBERT)[Reimers and Gurevych, 2019] to automatically compute semantic similarity matrices, filtering out near-duplicate samples to preserve dataset variety. During filtration, we removed 1,542 training and 832 test samples, guaranteeing a diverse final dataset. This careful pruning of semantically similar entries promotes a varied collection of queries and responses, bolstering the benchmark's effectiveness for assessing diverse instances.

4.4 Generate Chain of Evidence

Our benchmark evaluates the detectors' ability to not only identify hallucinations but also to provide effective explanations. It necessitates ChatGPT's generation of coherent golden evidence chains grounded in factual knowledge for substantiating judgments. Utilizing subgraph or textual facts outlined in §section 3, along with the previously generated "QUERY-RESPONSE" pairs, ChatGPT delivers thorough justifications for labels assigned to "QUERY-RESPONSE" contexts. These golden evidence chains are critical for assessing the explanatory validity of hallucination detectors.

4.5 Quality Filtering via Human Annotator

We craft filter rules to ensure pattern consistency, response factuality, and logical evidence chains for quality control in the annotation. Several educated annotators are uniformly trained and utilize both their expertise and search tools for rigorous sample vetting. To minimize subjectivity, we organized them into groups of three, incorporating a voting mechanism for the evaluation of data. Simultaneously judged mismatches by annotators led to sample discarding, resulting in final removal counts of 565 and 258 samples from the training and test sets, respectively. We involve Fleiss's Kappa (κ) as a measure of inter-annotator agreement to assess the reliability of our annotations. We calculate κ over the remaining annotated test set, resulting in a κ value of 0.858, indicating substantial agreement.

5 Experiments

5.1 Metric Definition

FACTCLS Metric. We employ the FACTCLS, denoted by the Micro F1 score, to evaluate binary factuality classification performance. This metric focuses on the distribution $p(l|Q\&R)$, classifying instances as either FACTUAL or NON-FACTUAL. With a specific emphasis on identifying non-factual examples, we designate NON-FACTUAL as the positive class and FACTUAL as the negative class.

EXPMATCH Metric. In FACTCHD, the golden evidence chain features introductory/expository statements (head-tail part) and a factual explanation core (body part). The former contextualizes with phrases such as 'Therefore, there is an incorrect conclusion in this query and response', and the latter delivers the in-depth, fact-based reasoning process. For hallucination detectors, prompts guide their outputs akin to the gold evidence chain for quality assessment. Given the paramount importance of aligning factual explanations over expository parts, we introduce EXPMATCH, a metric employing segmented matching with weighted averaging. It computes Score_{bd} via span-based Micro F1 for unigram overlap between generated and reference bodies, and Score_{ht} via ROUGE-L

	Evaluator	VANILLA		MULTI-HOPS		COMPARISON		SET-OPERATION		AVERAGE	
		CLS.	EXP.	CLS.	EXP.	CLS.	EXP.	CLS.	EXP.	CLS.	EXP.
Zero-Shot	GPT-3.5-turbo	55.12	22.79	59.54	29.84	16.66	18.89	55.46	28.23	52.82	24.03
	text-davinci-003	52.06	17.72	59.92	25.30	25.50	16.09	48.58	25.71	50.98	19.57
	Alpaca-7B	29.66	11.72	5.20	25.60	8.88	17.95	13.08	21.37	23.10	13.66
	Vicuna-7B	35.26	24.62	17.54	34.39	9.34	24.88	14.96	31.41	28.84	26.84
	Llama2-7B-chat	3.57	26.78	5.49	33.87	10.53	35.25	12.61	33.27	5.77	29.41
ICL (4-shot)	GPT-3.5-turbo	62.02	37.29	65.66	51.85	32.2	48.11	64.74	50.14	↑8.22 61.04	↑17.93 41.96
	text-davinci-003	56.52	39.36	55.02	58.22	8.50	48.53	50.34	51.82	↑1.9 52.88	↑24.88 44.45
	Alpaca-7B	35.82	31.01	18.12	40.16	8.86	29.28	6.70	31.52	↑5.24 28.34	↑16.76 32.23
	Vicuna-7B	41.36	42.51	29.24	58.35	19.36	41.55	13.46	53.60	↑6.3 35.14	↑19.14 45.98
	Llama2-7B-chat	31.00	39.08	39.13	54.38	10.50	41.83	27.96	51.73	↑24.48 30.25	↑13.61 43.02
Det.(tune)	Alpaca-7B-LoRA	73.14	49.00	63.34	70.83	69.92	59.88	68.18	63.75	↑42.32 70.66	↑22.73 54.96
	Vicuna-7B-LoRA	73.52	48.07	64.72	71.74	67.34	62.08	50.36	66.04	↑34.44 69.58	↑9.00 54.98
	Llama2-7B-chat-LoRA	77.41	47.91	67.70	67.30	62.27	57.03	78.68	65.94	↑44.48 74.73	↑10.69 53.71
Knowledge	Alpaca-7B-LoRA (wiki)	73.86	49.44	67.3	69.97	68.24	60.25	67.38	63.00	↑0.66 71.32	↑0.11 55.07
	Vicuna-7B-LoRA (wiki)	75.14	49.56	65.46	72.71	65.10	63.51	55.42	66.65	↑1.28 70.86	↑1.30 56.28
	Llama2-7B-chat-LoRA (wiki)	77.14	46.71	69.61	64.17	66.05	49.73	78.08	64.52	↑1.13 75.86	↑0.87 54.58
	GPT-3.5-turbo (tool)	69.71	38.60	69.92	48.43	44.08	47.26	74.21	45.65	↑7.59 68.63	↑0.41 42.37
	TRUTH-TRIANGULATOR	80.97	47.08	75.01	64.21	66.27	55.70	80.87	65.25	78.15	52.52

Table 3: Results on FACTCLS and EXPMATCH (abbreviated as **CLS.** and **EXP.**) along with FACTCHD estimated by each method. The shadow and shadow in each row represent the top-2 FACTCLS scores for the four factuality patterns. The ↑up and ↓down arrows respectively indicate positive/negative performance changes in the AVERAGE score compared to the corresponding upper-level method.

to assess similarity based on the longest common word subsequence for the head-to-tail part. EXPMATCH, combining Score_{bd} and Score_{ht}, evaluates the explanations generated from detectors as:

$$\text{EXPMATCH} = \alpha \times \text{Score}_{bd} + (1 - \alpha) \times \text{Score}_{ht}. \quad (1)$$

We initially set α to 0.7, emphasizing the body part[2].

5.2 Experimental Settings

Evaluation Models. We evaluate various leading LLMs on FACTCHD benchmark, focusing on OpenAI API models, including `text-davinci-003` (InstructGPT) and `GPT-3.5-turbo` (ChatGPT). Additionally, we explore the adoption of open-source models such as Llama2-chat, Alpaca[Taori *et al.*, 2023] and Vicuna [Chiang *et al.*, 2023], which are fine-tuned variants of the LLaMA [2023].

Implementation Details. Using Azure's OpenAI ChatGPT API, we generate samples with a temperature of 1.0 to control the diversity of generated samples, while limiting the maximum number of tokens to 2048 to ensure concise responses. We use a frequency penalty of zero and a Top-p of 1.0 to ensure unrestricted token selection during generation. For evaluations, we standardize the temperature at 0.2 to minimize randomness.

Baseline Strategy Settings This study investigates the effectiveness of various baseline strategies in detecting fact-conflicting hallucinations. These strategies include: (1) ZERO-SHOT LEARNING, which evaluates model performance without prior training in hallucination detection; (2) IN-CONTEXT LEARNING, using 4-shot samples as demonstrations to prompt the model and evaluate its ability to handle hallucination detection; (3) DETECT-SPECIFIC EXPERT MODEL, fine-tuning the LoRA [Hu *et al.*, 2022] parameters using our specialized training set to leverage domain expertise; (4) KNOWLEDGE ENHANCEMENT, enhancing the model's capabilities by integrating external knowledge through retrieval techniques or tool-based enhancements. These experiments aim to explore both the potential and limitations of these baseline methods in hallucination detection.

[2]Preliminary evaluations reveal that scores with α at 0.75 and 0.8 align with our established conclusions. If a generated result lacks factual or non-factual information, indicating a prediction failure, we attribute an EXPMATCH value of 0 to that particular example during metric calculation.

5.3 Empirical Experiment Results

Zero-shot Learning Performance

The empirical results presented in Table 3 show that LLMs with zero-shot learning struggle to identify implicit factuality in "QUERY-RESPONSE" contexts. The performance of the open-source 7B LLMs in zero-shot learning is notably poor, indicating deficiencies in both instruction comprehension and internal knowledge representation. Even the ChatGPT shows limited proficiency in distinguishing between factual and non-factual "QUERY-RESPONSE" samples, achieving only a 52.82% FactCls score in zero-shot. Llama2-chat is susceptible to false negatives, resulting in reduced sensitivity in hallucination detection.

In-context Learning Performance

The incorporation of few-shot information significantly improves fact-conflicting hallucination detection in the GPT-3.5-turbo, Alpaca-7B, Vicuna-7B, and Llama2-7B-chat models. Compared to models operating without this additional context, these enhancements result in an average increase of approximately ↑6% in the FACTCLS and ↑18% in EXPMATCH scores. However, the impact of integrating few-shot information on *text-davinci-003* is relatively modest, indicating its limited proficiency in managing few-shot learning. The

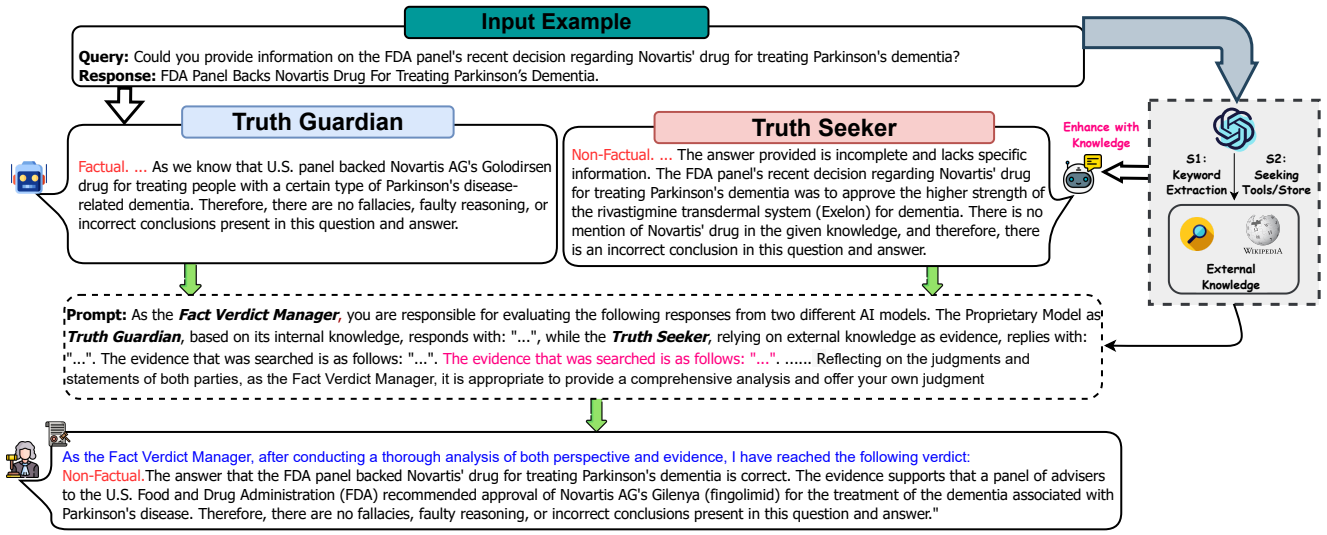

Figure 5: Overview TRUTH-TRIANGULATOR. Here we designate the as the "Truth Guardian" based on Llama2-7B-chat-LoRA while as the "Truth Seeker" based on GPT-3.5-turbo (tool) in our experiments. We want the "Fact Verdict Manager" to collect evidence from different viewpoints to enhance the reliability and accuracy of the obtained conclusion.

improvement from incorporating few-shot information into Llama2-chat is significant, possibly because of the model's superior contextual learning abilities compared to Alpaca and Vicuna, leading to a better understanding of demonstrations.

Detect-Specific Expert Performance

We investigate the effectiveness of tuning LLMs with LoRA for domain-specific expertise in detecting fact-conflicting hallucinations. By training these models on our trainset, we enhance their detection task proficiency. Our findings show that all tested open-source 7B models improve, with Llama2-chat outperforming its counterparts. After fine-tuning on hallucination data, Llama2-chat-7B achieves a FACTCLS score of 74.73% and an EXPMATCH score of 53.71% on our benchmark. This success with a 7B model underscores the viability of using such models as hallucination detectors and encourages further exploration in this area.

Knowledge Enhancemant

Retrieval Enhancement. This study enhances LLM-based detection with Wikipedia-sourced facts, using BM25 for initial document retrieval to select the top five most relevant paragraphs. Our experiments show that given the 7B model's limited ability to process long texts, providing knowledge during fine-tuning modestly improves performance. However, compared to the detect-specific expert without knowledge augmentation, the improvement achieved by combining Wikipedia retrieval with fine-tuning the LoRA parameters is relatively modest. We attribute these observations to two primary factors: our dataset spans multiple domains while Wikipedia encompasses merely a subset, and our retrieval method being elementary, occasionally yields lower-quality evidence.

Tool Enhancement. Considering the labor-intensive nature of tapping into external knowledge bases, we investigate leveraging ChatGPT's advanced contextual understanding for tool-enhanced hallucination detection, drawing from prior research

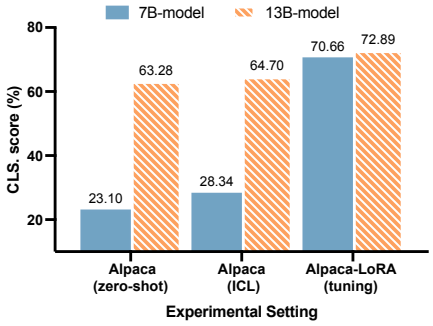

Figure 6: Analysis of model capacity impact.

such as [Chern *et al.*, 2023; Qin *et al.*, 2023]. By using a search engine to inform ChatGPT's analysis of "QUERY-RESPONSE" samples, we enable the model to identify hallucinations with supporting evidence from search results. Results in Table 3 demonstrate significant enhancements over ChatGPT's few-shot performance, highlighting the efficacy of tool learning.

5.4 Exploring Triangulation for Truth

We categorize tool-enhanced ChatGPT as the **Truth Seeker**, which aims to make informed judgments by seeking external knowledge. However, the information returned by external knowledge sources may inevitably be incomplete, erroneous, or redundant, thus potentially misleading the large-scale model. On the other hand, the detect-specific expert as the **Truth Guardian** relies on its knowledge and expertise in the task, tending towards more conservative predictions. To address these challenges, we propose the TRUTH-TRIANGULATOR framework inspired by the "Triangulation for Truth" theory, involving verifying and confirming information by cross-referencing multiple independent perspectives.

Query and Returned Evidence	Detection Opinion from Both Sides	Prediction with Triangulation
QUERY: Who was the mother of Afonso II, the third king of Portugal? **RESPONSE:**The mother of Afonso II, the third king of Portugal, was Queen Urraca of Castile. **Correct Answer: Dulce Berenguer of Barcelona. Returned by the tool:** 1.—died March 25, 1223, Coimbra), the third king of Portugal (1211–23), under whom the reconquest of the south from the Muslims was continued. born: 1185?, Coimbra, Port. Afonso II was the son of King Sancho I and Queen Dulcia, daughter of Ramón Berenguer IV of Barcelona. 2...	**TRUTH GUARDIAN:** FACTUAL. The answer that Queen Urraca of Castile was the mother of Afonso II, the third king of Portugal, is correct. This is supported by the evidence that Afonso II was the third king of Portugal and that Queen Urraca of Castile was his mother. Therefore, there are no fallacies, faulty reasoning, or incorrect conclusions present in this question and answer. **TRUTH SEEKER:** NON-FACTUAL. The answer stating that the mother of Afonso II, the third king of Portugal, was Queen Urraca of Castile is incorrect. According to the provided knowledge, the mother of Afonso II was Queen Dulcia, daughter of Ramón Berenguer IV of Barcelona. Therefore, there is an incorrect conclusion ...	**TRUTH-TRIANGULATOR:** As the Fact Verdict Manager, after conducting a thorough analysis of both perspectives and evidence, I have reached the following verdict: NON-FACTUAL. The answer that Queen Urraca of Castile was the mother of Afonso II, the third king of Portugal, is incorrect. The evidence provided shows that Afonso II was the son of King Sancho I and Queen Dulcia, daughter of Ramón Berenguer IV of Barcelona. Therefore, the answer contains false assumptions and incorrect conclusions. The Truth Seeker's response is supported by the provided evidence, while the Proprietary Model's response is not accurate as it did not consider the external evidence.
QUERY: Who is the Jeff Probst Show-nominated television producer's first wife? **RESPONSE:**His first wife was psychotherapist Shelley Wright. **Correct Answer: Shelley Wright.** **Returned by the tool:**1. Probst was married to his first wife, psychotherapist Shelley Wright, from 1996 to 2001. In 2004, after filming on Survivor: Vanuatu wrapped, he began dating one of the contestants, a 23 year old Julie Berry - he was about 20 years her senior. They broke up in early 2008. 2. ...	**TRUTH GUARDIAN:** FACTUAL. The answer that Jeff Probst's first wife was psychotherapist Shelley Wright is correct. According to the evidence provided, Jeff Probst was previously married to Shelley Wright, who is a *psychotherapist* in addition to being a television producer. Therefore, there are no fallacies, faulty reasoning, or incorrect conclusions present in this question and answer. **TRUTH SEEKER:** FACTUAL. According to the provided knowledge, Jeff Probst's first wife was psychotherapist Shelley Wright. Therefore, the answer is correct and there are no fallacies, faulty reasoning, or incorrect conclusions present in this question and answer.	**TRUTH-TRIANGULATOR:** As the Fact Verdict Manager, after evaluating the responses from both AI models and examining the provided evidence, I have reached the following verdict: FACTUAL. The answer that Jeff Probst's first wife was psychotherapist Shelley Wright is correct. Both AI models correctly identified that there are no fallacies, faulty reasoning, or incorrect conclusions present in the question and answer. The evidence provided confirms that Jeff Probst was married to Shelley Wright from 1996 to 2001, and that she is a psychotherapist in addition to being a television producer. Therefore, the answer is factual and accurate.

Table 4: Case analysis of out-of-distribution examples from ChatGPT using TRUTH-TRIANGULATOR.

Figure 5 illustrates our approach of fine-tuning ChatGPT as a **Fact Verdict Manager**, leveraging the output and knowledge from **Truth Seeker** and **Truth Guardian** to boost conclusion reliability. Table 3 reveals our model's superior performance to Lamma2-7b-chat-LoRA and GPT-3.5-turbo (tool), highlighting the benefits of triangulation in reducing single-source error and enhancing truth verification.

5.5 Experimental Analysis

Examining the Influences of Model Capacity. Figure 6 illustrates that transitioning from 7B to 13B models notably improves the detection of fact-conflicting hallucinations, particularly in zero-shot and in-context learning scenarios. Alpaca-13B outperforms ChatGPT, which can beascribed to the consistently adopted command prompt may be more friendly to Alpaca-13B [3]. Interestingly, when models are fine-tuned with training data, the impact of model capacity on performance improvement appears minimal. This implies that further training larger LLMs as hallucination detectors may yield limited benefits, and it is necessary to explore alternatives with higher upper limits for enhancing detection performance.

Enhancement lies in Accurate Evidence. We also explore the direct use of intrinsic facts from the dataset as golden evidence in the LoRA fine-tuning process. As shown in Table 5, integrating factual information (**w/ gold evidence**) leads to a significant improvement in the FACTCLS score, indicating that the modest improvement from retrieval may stem from the lower quality of the obtained evidence. This underscores the potential for substantial improvement by seeking the most accurate facts for evaluation.

Complete "QUERY-RESPONSE" Context. Our dataset differs from traditional fact-checking datasets in its inclusion of the "QUERY-RESPONSE" context. To examine its impact, we conduct experiments by excluding the "query" during the Alpaca-7B-LoRA fine-tuning process. The results in Table 5 reveal that the "**w/o query**" scenario leads to a significant 40% decrease in the FACTCLS score, highlighting the essential role of a comprehensive "QUERY-RESPONSE" context in extracting valuable information for informed decision-making.

[3]We keep consistent prompts for each LLM throughout all experiments. Given that LLMs are acknowledged for their pronounced sensitivity to prompts, we do not assert that our prompts are universally optimal.

Variants	CONV.	MULTI.	COMP.	SET.	AVG.
Alpaca-7B-LoRa	73.16	63.34	69.92	68.18	70.66
w/ golden evidence	82.68	96.10	70.50	87.90	83.56
w/o query	39.0	10.04	9.24	9.52	30.48

Table 5: Ablation analysis on input context.

Real-World Application Case Study. We extended the testing of TRUTH-TRIANGULATOR to real-world instances of ChatGPT-generated hallucinations beyond the FACTCHD benchmark to demonstrate its wide-ranging utility. The out-of-distribution case analysis, detailed in Table 4, delineates our model's strengths and limitations. These real-world tests confirm TRUTH-TRIANGULATOR's adeptness in making accurate assessments, particularly where expert and augmented ChatGPT evaluations diverge, thus bolstering the credibility of detecting fact-conflicting hallucinations in authentic, uncontrolled settings.

6 Conclusion and Future Work

We introduce FACTCHD, a meticulously designed benchmark tailored for the evaluation of fact-conflicting hallucinations from LLMs, which is notably enriched with a kaleidoscope of patterns and substantiated evidence chains to fortify the robust elucidation of factuality assessments. Moreover, we delineate TRUTH-TRIANGULATOR that employs the principle of triangulation to discern the veracity of information, deploying cross-referencing generators to arbitrate responses. Moving forward, we will broaden our evaluative scope to encompass various modalities and granularity in hallucination detection.

Acknowledgments

We would like to express gratitude to the anonymous reviewers for kind comments. This work was supported by the National Natural Science Foundation of China (No. 62206246), Response-driven intelligent enhanced control technology for AC/DC hybrid power grid with high proportion of new energy (5100-202155426A-0-0-00), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), CCF-Tencent Rhino-Bird Open Research Fund, and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[Aly *et al.*, 2021] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: fact extraction and verification over unstructured and structured information. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[Chandak *et al.*, 2023] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[Chen *et al.*, 2020] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *ICLR 2020*. OpenReview.net, 2020.

[Chen *et al.*, 2022] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW 2022*, pages 2778–2788. ACM, 2022.

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

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[Creswell and Shanahan, 2022] Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*, 2022.

[Das *et al.*, 2023] Souvik Das, Sougata Saha, and Rohini K Srihari. Diving deep into modes of fact hallucinations in dialogue systems. *arXiv preprint arXiv:2301.04449*, 2023.

[Dhingra *et al.*, 2019] Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. Handling divergent reference texts when evaluating table-to-text generation. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL 2019*, pages 4884–4895. Association for Computational Linguistics, 2019.

[Diggelmann *et al.*, 2020] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

[Dziri *et al.*, 2022] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar R. Zaïane, Mo Yu, Edoardo Maria Ponti, and Siva Reddy. Faithdial: A faithful benchmark for information-seeking dialogue. *Trans. Assoc. Comput. Linguistics*, 10:1473–1490, 2022.

[Gupta *et al.*, 2022] Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Dialfact: A benchmark for fact-checking in dialogue. In *ACL 2022*, pages 3785–3801, 2022.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[Huang *et al.*, 2023] Yue Huang, Qihui Zhang, Philip S. Yu, and Lichao Sun. Trustgpt: A benchmark for trustworthy and responsible large language models. *CoRR*, abs/2304.10513, 2023.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[Jiang *et al.*, 2020] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of EMNLP 2020*, 2020.

[Kang *et al.*, 2024] Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: certified generation risks for retrieval-augmented language models. *CoRR*, abs/2402.03181, 2024.

[Kryściński *et al.*, 2020] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *EMNLP*, pages 9332–9346, 2020.

[Li *et al.*, 2023] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *CoRR*, abs/2305.11747, 2023.

[Liang *et al.*, 2023] Ke Liang, Sihang Zhou, Yue Liu, Lingyuan Meng, Meng Liu, and Xinwang Liu. Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. *CoRR*, abs/2307.03591, 2023.

[Liang *et al.*, 2024] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Trans. Knowl. Data Eng.*, 36(1):226–238, 2024.

[Mallen *et al.*, 2022] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.

[Mohr *et al.*, 2022] Isabelle Mohr, Amelie Wührl, and Roman Klinger. Covert: A corpus of fact-checked biomedical COVID-19 tweets. In *LREC 2022*, pages 244–257, 2022.

[Muhlgay *et al.*, 2023] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. *CoRR*, abs/2307.06908, 2023.

[OpenAI, 2022] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt, 2022.

[Pagnoni *et al.*, 2021] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *NAACL*, pages 4812–4829, 2021.

[Qin *et al.*, 2023] Yujia Qin, Shengding Hu, and et al. Tool learning with foundation models. *CoRR*, abs/2304.08354, 2023.

[Rashkin *et al.*, 2021] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *CoRR*, abs/2112.12870, 2021.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP 2019*, pages 3980–3990, 2019.

[Saakyan *et al.*, 2021] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. Covid-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *ACL/IJCNLP 2021*, pages 2116–2129, 2021.

[Sarrouti *et al.*, 2021] Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. Evidence-based fact-checking of health-related claims. In *Findings EMNLP 2021*, pages 3499–3512, 2021.

[Shuster *et al.*, 2021] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of EMNLP 2021*, 2021.

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

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018.

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

[Valenza, 2016] Joyce Valenza. Truth, truthiness, triangulation: A news literacy toolkit for a "post-truth" world. *School Library journal*, 2016.

[Vrandecic and Krötzsch, 2014] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[Wadden *et al.*, 2020a] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *EMNLP 2020*, pages 7534–7550, 2020.

[Wadden *et al.*, 2020b] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, pages 7534–7550, Online, November 2020.

[Wang *et al.*, 2023a] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, and et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521, 2023.

[Wang *et al.*, 2023b] Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *CoRR*, abs/2310.00935, 2023.

[Yao *et al.*, 2023] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *EMNLP 2023*, abs/2305.13172, 2023.

[Yin *et al.*, 2023] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *CoRR*, abs/2305.18153, 2023.

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

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, and et al. A survey of large language models. *CoRR*, abs/2303.18223, 2023.