

AutoHall: Automated Factuality Hallucination Dataset Generation for Large Language Models

Zouying Cao^{*}, Yifei Yang^{*}, Xiaojing Li^{*}, Hai Zhao^{*}

Abstract—Large language models (LLMs) have gained broad applications across various domains but still struggle with hallucinations. Currently, hallucinations occur frequently in the generation of factual content and pose a great challenge to trustworthy LLMs. However, hallucination detection is hindered by the laborious and expensive manual annotation of hallucinatory content. Meanwhile, as different LLMs exhibit distinct types and rates of hallucination, the collection of hallucination datasets is inherently model-specific, which also increases the cost. To address this issue, this paper proposes a method called AutoHall for **Automatically constructing model-specific Hallucination datasets based on existing fact-checking datasets**. The empirical results reveal variations in hallucination proportions and types among different models. Moreover, we introduce a zero-resource and black-box hallucination detection method based on self-contradiction to recognize the hallucination in our constructed dataset, achieving superior detection performance compared to baselines. Further analysis on our dataset provides insight into factors that may contribute to LLM hallucinations. Our codes and datasets are publicly available at <https://github.com/zouyingcao/AutoHall>.

Index Terms—Natural language processing, Large Language Models (LLMs), LLM hallucination, automated dataset generation, hallucination detection.

I. INTRODUCTION

LARGE language models (LLMs) are capable of performing a wide range of tasks across diverse domains [1, 2, 3]. Despite their powerful capabilities, LLMs suffer from the issue of *hallucination*, i.e., have the tendency to respond inaccurate or fabricated information in generation tasks [4, 5, 6]. As shown in Fig. 1, ChatGPT generates hallucinations when describing Jo Nesbø’s novel “The Leopard”. It can be observed that ChatGPT fabricates some plots of the novel and contains incorrect texts in the response. With the increasing reliance of current artificial intelligence (AI) systems on LLMs, the generation of hallucinatory information poses huge challenge to enterprise security and user trust [7, 8]. Therefore, significant endeavors have been undertaken to detect LLM hallucinations.

Current research efforts on hallucination detection leverage external knowledge bases [9, 10, 11] or just adopt a zero-resource approach which focuses on resources inherent to the model itself [12, 13, 14, 15, 16]. Typically, most of these methods begin with a crowdsourced annotation, where researchers use question answering (QA) datasets to collect

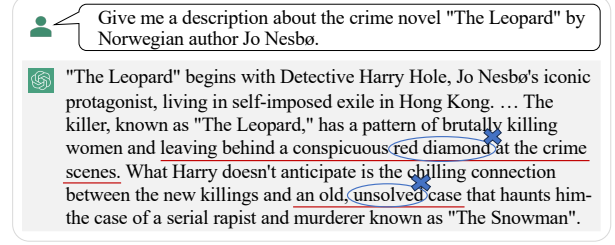


Fig. 1. A hallucination example. The red underline indicates the hallucinatory content generated by ChatGPT, since the novel never mentions the presence of a “red diamond” at the crime scene and the “The Snowman” case has also been solved before.

the model-generated responses and then manually annotate whether the answers contain hallucinations. However, this collection pipeline of hallucination detection datasets has several limitations. First, hallucination datasets are model-specific because different LLMs have varying hallucination properties. Even when given identical inputs, some models may generate hallucinatory content while others produce factually accurate responses. Therefore, each model requires dedicated annotation, which limits the dataset scalability due to the dependence on manual annotation [17]. Second, collected datasets are time-sensitive since model upgrades may mitigate hallucination issues in LLMs [18] and the old dataset can no longer represent the new model hallucinations.

Considering the above issues, this paper explores one automated generation pipeline of hallucination detection datasets. Our focus is on *factuality hallucination*, a primary LLM hallucination type, with studies showing GPT-4’s 28% factual hallucination rate is over double that of other types [19]. Inspired by Agrawal et al. [13] emphasizing the hallucinatory reference¹ problem in LLMs, we find the feasibility of automatically creating hallucination detection datasets through public fact-checking datasets. Specifically, since the existing fact-checking datasets usually consist of manually annotated claims accompanied by the ground truth labels (i.e., factual/non-factual), we can determine whether hallucinations exist by generating references to the claims and exploring whether the references can infer the correct labels for the claims.

Our analysis reveals that LLMs are particularly susceptible to hallucinations in responses involving various domain-specific topics such as history, technology and geography. We further investigate the contributing factors across different models based on our dataset. Additionally, we propose

¹The “reference” term means the supporting information generated by the LLM to substantiate its claim, rather than a traditional bibliographic citation.

Zouying Cao, Yifei Yang and Hai Zhao are with the AGI Institute, School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3, Shanghai 200240, China (e-mail: zouyingcao@sjtu.edu.cn; yifeiyang@sjtu.edu.cn; zhaohai@cs.sjtu.edu.cn). Xiaojing Li is with the School of Media & Communication, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lixiaojing@sjtu.edu.cn). * denotes equal contribution.

a three-step zero-resource black-box hallucination detection method motivated by the idea of self-contradiction [15]. Given that one LLM accurately understands a claim, the generated references are less likely to contain contradictions. Therefore, it is possible to determine whether the model has generated hallucinations based on knowledge conflicts among these references.

In summary, the contributions of our paper are:

- We propose an approach called **AutoHall** for fast and automatically constructing model-specific hallucination datasets based on existing fact-checking datasets, eliminating the need for manual annotation.
- Using AutoHall, we conduct extensive experiments to analyze LLM hallucinations in current open-source and closed-source models. From the results, we estimate the prevalence of hallucination in LLMs at a rate of 20% to 30% and gain insight into what types or topics of LLM responses that tend to be hallucinatory.
- We introduce a black-box hallucination detection method without external resources. Based on our dataset, we evaluate its effectiveness on ChatGPT and Llama2 models, demonstrating its superior improvements over existing zero-resource detection techniques.

II. BACKGROUND AND RELATED WORKS

A. Hallucination of Large Language Models

Although large language models have demonstrated remarkable capabilities [20, 21], they still struggle with several issues, where hallucination is a significant problem. The hallucination issue refers to that LLMs generate the statements which appear plausible but are fabricated or contradict factual knowledge. The consequent effects may undermine the reliability of LLM applications [4, 22], thereby causing negative user experiences.

Generally, hallucinations of large language models can be divided into two primary categories: intrinsic hallucinations and extrinsic hallucinations [5]. Intrinsic hallucinations occur when the LLM output contradicts the input content. For example, in a multi-modal image captioning task, the model generates a caption that includes details or objects which are not present in the input image. Extrinsic hallucinations refer to the generated content that cannot be verified based on the source content. Huang et al. [23] redefine the taxonomy of hallucination by introducing factuality hallucination and faithfulness hallucination, taking alignment with user directives and factual knowledge into account. Another recent paper [24] meticulously classify hallucination into six types according to the specific hallucinatory content. In this paper, our focus is on *factuality hallucination*.

So far, the causes of hallucination in LLMs have been investigated across different tasks, such as question answering [25], abstractive summarization [26], machine translation [27] and dialogue systems [28]. The key factors include but are not limited to training corpora quality [29, 30], problematic alignment process [4, 31] and randomness in generation strategy [32, 33]. McKenna et al. [29] offer one LLM bias brought by training text as explanation of general false positive hallucination. Alignment tax is also a well-known problem that may lead

to LLM hallucinations [34]. During inference, the existence of a likelihood trap [35] supports incorporating randomness into decoding strategies while exacerbates the risk of hallucinations [36].

B. Hallucination Evaluation Datasets

For hallucination detection benchmarks, the majority of previous studies have focused on task-specific hallucinations to support detection tasks in numerous scenarios [37, 38, 39]. For example, Umapathi et al. [38] propose a hallucination benchmark within the medical domain as a tool for hallucination evaluation and mitigation. Dale et al. [39] present another dataset with human-annotated hallucinations in machine translation to promote the research on translation pathology detection and analysis.

Beyond single domain task, critically assessing the effectiveness of hallucination detection strategies necessitates the development of high-quality datasets across multiple domain tasks. SelfCheckGPT-Wikibio [15] makes a contribution as a sentence-level hallucination detection dataset and HaluEval [37] dataset is constructed using a combination of synthetically and naturally generated LLM responses. Besides, FELM [40] benchmark provides fine-grained factuality labels to responses generated from ChatGPT. Hu et al. [17] curate a comprehensive dataset that enables the evaluation of hallucination detection performance under different context quality and availability.

Nevertheless, there are limitations as they are subject to manually annotated hallucination datasets, which are expensive and time-consuming. Meanwhile, hallucination datasets are model-specific, requiring separate annotations for different models, whose applicability would also be affected by model upgrades (*i.e.*, time-sensitive). Thus, there is still room for improvement in the collection process of current hallucination detection datasets.

C. LLM Hallucination Detection

To detect the hallucination issue, researchers have been making tremendous effort in seeking solutions. Existing approaches can be broadly categorized into two types: retrieval-based detection and zero-resource detection.

On the one hand, prior works focus on resorting to external knowledge bases to detect hallucinations. For instance, Gou et al. [10] propose a framework called CRITIC to validate the output generated by the model with tool-interaction and Huo et al. [41] improve the conventional method of retrieving supporting evidence for hallucination detection by incorporating the LLM-generated answer into retrieval query. In addition, a unified framework called FACTOOL [9] invokes interfaces of search engines to recognize hallucination.

On the other hand, current researches pay more attention to realizing zero-resource hallucination detection methods. These methods primarily depend on uncertainty of the LLMs [42], which is mainly reflected in self-contradiction phenomenon among their responses or the model's internal states. Typically, Xue et al. [43] utilize the Chain of Thoughts (CoT) to check the hallucinatory responses. Manakul et al. [15] introduce a

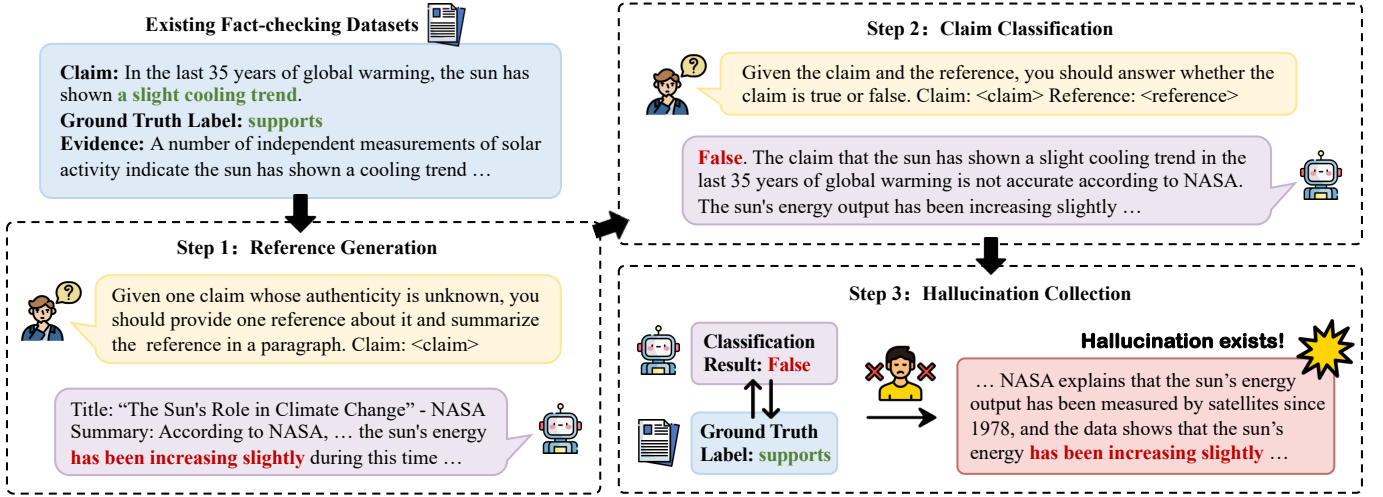


Fig. 2. Our proposed approach to collect LLM hallucination automatically. The grounded information is colored green. The incorrect information is colored red. Some analysis on prompt sensitivity is included in Section IV-C2.

simple sampling-based approach called SelfCheckGPT that can be used to detect hallucination. They examine five variants of SelfCheckGPT for measuring consistency via: BERTScore, question-answering, n-gram, NLI and LLM prompting. With respect to open-source LLMs, the internal states of LLMs can be used to indicate the existence of hallucination, through the analysis of metrics like token probability or entropy [12, 44].

Our proposed detection method falls into the zero-resource category which is suitable for both open-source and closed-source models. Using AutoHall, our up-to-date constructed datasets can evaluate whether these detection methods remain consistently applicable regardless of model upgrades and then we expect our work to lay a solid foundation for subsequent research in hallucination detection.

III. METHODOLOGY

In this section, we first formulate the definition of LLM factuality hallucinations discussed in our work. Then, we introduce AutoHall, our automatic dataset creation pipeline which focuses on prompting LLMs to produce “hallucinatory references”. Finally, based on our generated datasets, we further present one zero-resource, black-box approach to recognize hallucination.

A. LLM Factuality Hallucination Formulation

Large language models are prone to various forms of hallucinations, such as dialogue history-based hallucinations, hallucination in abstractive summarization and general data generation hallucination [45]. In this work, we focus on factuality hallucinations, characterized as the phenomenon of LLMs producing seemingly plausible but factually inaccurate or fabricated information.

Generally, for any input sentence X with a specific prompt P , the large language model \mathcal{M} will generate an answer Y , denoted as:

$$\mathcal{M}(P, X) = Y, \quad (1)$$

where $X = [x_1, x_2, \dots, x_n]$ of length n , $P = [p_1, p_2, \dots, p_o]$ of length o and $Y = [y_1, y_2, \dots, y_m]$ of length m .

Given factual knowledge $F = [f_1, f_2, \dots, f_t]$ of length t , the problem of factuality hallucination H occurs when there is a factual contradiction between the output slice $Y_{[i:j]} = [y_i, y_{i+1}, \dots, y_j]$ ($1 \leq i \leq j \leq m$) and the knowledge slice $F_{[u:v]} = [f_u, f_{u+1}, \dots, f_v]$ ($1 \leq u \leq v \leq t$). Formally, we define $Y \in H$ to be the existence of factuality hallucination in LLM output and summarize its definition below:

$$Y \in H \Leftrightarrow \exists Y_{[i:j]} \exists F_{[u:v]} (Y_{[i:j]} \wedge F_{[u:v]} = \text{False}). \quad (2)$$

B. AutoHall: Automatic Generation of Factuality Hallucination Datasets

Current research on hallucination detection mostly relies on manually annotated datasets [15, 40, 46]. Namely, judging whether the LLM output Y is hallucinatory requires slow and costly manual tagging due to the absence of a golden automatic comparison standard for the factuality. However, existing massive fact-checking datasets provide us with enough data, typically comprising real-world claims, corresponding ground truth labels, and evidence sentences as illustrated in Fig. 2. Based on this type of public available data, our AutoHall can first prompt a model to generate relevant references for claims and then use the ground truth labels as criteria to assess the hallucinatory nature of the generated references. Specifically, as shown in Fig. 2, AutoHall generates hallucination datasets following the below three steps:

Step 1: Reference Generation. For an LLM, we prompt it to generate the corresponding references to claims in the existing datasets. The prompt used for reference generation is displayed in Fig. 2 Step 1. Note that, to simplify the generation, we only focus on factual (supported/true) and faked (unsupported/false) claims. Besides, we discard references that fail to contain concrete content, like a long response beginning with “I can not provide a specific reference for the claim you

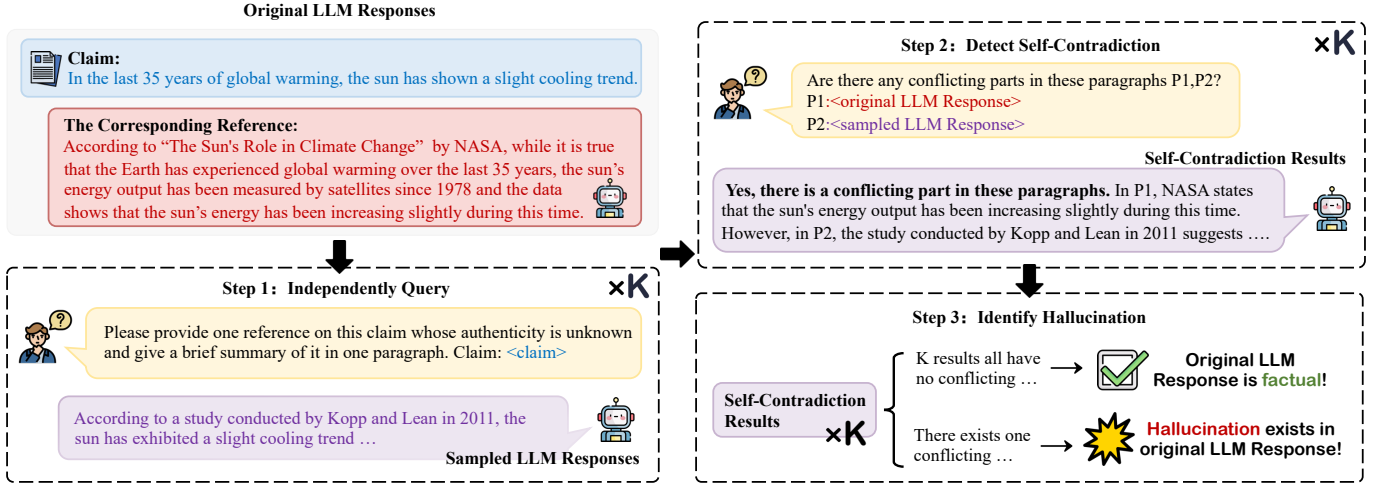


Fig. 3. Our proposed approach to detect LLM hallucination. The claim from fact-checking dataset is colored blue. The response need to be detected whether exists hallucination is colored red. The sampled references to trigger self-contradictions are colored purple. The complete Step 2 prompts are shown in Tab. II.

mentioned...". As a result, this ensures the remaining valid references are either reliable ($\in \bar{H}$) or hallucinatory ($\in H$).

Step 2: Claim Classification. Separately for each reference, in order to label whether hallucination exists, we prompt LLM to perform claim classification according to the generated reference. The input sequence is of format as shown in Fig. 2 Step 2, where the two placeholders <claim> and <reference> should be substituted with the claim X and the generated reference Y from Step 1. Then, LLM outputs in the format of "<category> <reason(optional)>" where the category is limited to true (T) or false (F). To elaborate, T indicates the generated reference Y supports the claim X is factual and F represents that Y demonstrates the claim X is faked.

We expect correct classification to each claim, while wrong classification may be taken as a sign of the existence of hallucination in the generated reference that it erroneously supports the claim's factuality. The binary classification results from LLM are reliable since numerous studies have proven the high accuracy of LLM-based fact-checking approaches [11, 47, 48] and human evaluation gives a guarantee in Section IV-D.

Step 3: Hallucination Collection. Last, we can directly adopt a simple string match algorithm to collect the hallucination dataset. If the classification result is not equal to the ground truth label, we label the generated reference as hallucination. Meanwhile, to maintain a balanced proportion between hallucinatory and factual references, we sample the same number of factual references built upon hallucinatory ones to form a completed dataset.

C. Hallucination Detection Approach

Using AutoHall, we not only fully automatically construct up-to-date hallucination datasets, but also further evaluate whether our hallucination detection approach and existing zero-resource ones remain effective regardless of model upgrades. The rationale for our detection method is that if the LLM knows one claim well, even when we query it to provide multiple references, self-contradictions among them should be

absent otherwise hallucinatory information must exist in one reference.

Suppose we know the user query $Q = (P, X)$ corresponding to the generated content Y from the LLM \mathcal{M} , the hallucination detection task is to judge either $Y \in H$ or $Y \in \bar{H}$. As depicted in Fig. 3, to trigger self-contradictions, we first manually design several prompts which are functionally similar to the prompt P . These prompts are fed into the same LLM \mathcal{M} to generate responses $Y^k (k = 1, \dots, K)$, where K is the number of queries. It is worth noting that each query is running independently to prevent mutual influence. Then, we concatenate each generation Y^k with the target Y to form one input pair and invoke the LLM to detect if Y and Y^k are contradictory. In this way, K self-contradiction detection results are obtained. Ultimately, we check if there is at least one Y^k conflicting with Y . If one conflict exists, it suggests the model \mathcal{M} does not understand the claim well, and Y may be hallucinatory. Conversely, if no conflicts are found in K pairs, it indicates that Y is factual. Hence, the final judgement standard of hallucination is defined as:

$$Y \in H \Leftrightarrow \exists Y_{[u,v]}^k \exists Y_{[i,j]} \left(Y_{[i,j]} \wedge Y_{[u,v]}^k = \text{False} \right). \quad (3)$$

IV. EXPERIMENTS

A. Experimental Settings

1) **Models:** We select five widely recognized open-/closed-source LLMs for hallucination collection: ChatGPT (GPT-3.5-turbo), GPT-4o, Llama2-7B-Chat, Llama2-13B-Chat [49] and Llama3-8B-Instruct [50]. For closed-source models, ChatGPT and GPT-4o are among the most advanced LLMs, which are used by calling APIs. Llama series models are one of the most prominent open-source models available and we run its instruction-tuned versions on a server with dual NVIDIA A100 GPUs with 80GB memory. To ensure a fair comparison, the prompts used for response generation are kept consistent among all the models.

TABLE I

EXAMPLES OF FACT-CHECKING DATASETS USED IN **AUTOHALL**. THE “SUPPORTS”, “TRUE” AND “SUPPORTED” LABELS REPRESENT THE FACTUALLY ACCURATE CLAIMS WHILE THE “REFUTES”, “FALSE” AND “NOT_SUPPORTED” INDICATE THE INACCURATE ONES.

Dataset	Topic	Example Claim	Label	Num
Climate-fever	Climate	<i>CO2 emissions were much smaller 100 years ago.</i>	<i>supports</i>	654
		<i>Ice berg melts, ocean level remains the same.</i>	<i>refutes</i>	253
PUB-HEALTH	Health	<i>France’s 20th century radium craze still haunts Paris.</i>	<i>true</i>	629
		<i>Viagra may help heart effects of muscular dystrophy.</i>	<i>false</i>	380
WICE	Law	<i>In 2019 Upton supported a bill banning sales between private individuals.</i>	<i>supported</i>	686
	Art	<i>Tiana Tolstoi is an Egyptian-born French model of Korean, Serbian, and Russian descent.</i>	<i>not_supported</i>	242

TABLE II

PROMPTS FOR SAMPLING IN HALLUCINATION DETECTION.

Prompts to Trigger Self-contradictions
1) Given one claim whose truthfulness is uncertain, you should provide one reference about it. This reference should be summarized as one paragraph. Claim: <claim>
2) Please provide one reference on this claim whose authenticity is unknown and give a brief summary of it in one paragraph. Claim: <claim>
3) Please provide a reference for a claim whose truthfulness is uncertain and summarize the content of the reference in one paragraph. Claim: <claim>
4) Given one claim whose authenticity is uncertain, you should provide one reference about it and write a summary paragraph. Claim: <claim>
5) There is a claim whose authenticity is unknown, please provide one corresponding reference and condense the reference in a paragraph. Claim: <claim>
6) There is a claim whose authenticity is unknown, please provide one reference that is relevant to this claim and summarize it in one paragraph. Claim: <claim>
7) You are expected to provide a reference for a claim whose truthfulness is uncertain. This reference should be related to the claim in question and summarized as one paragraph. Claim: <claim>

2) *Datasets, Metrics and Implementation Details:* For hallucination collection, we employ three fact-checking datasets: Climate-fever [51], PUBHEALTH [52] and WICE [53]. All of them provide real-world claims, ground truth labels and evidence retrieved from websites as shown in Tab. I. The topics of claims range from different domains, such as technology, culture, health and so on, which facilitates our analysis of what types or topics of content LLMs tend to be hallucinatory.

- **Climate-fever:** It is a fact checking corpus with climate change-related claims and their corresponding evidence. The original corpus comprises 1,535 real-world claims and each claim is labeled as one of 4 labels: *supports*, *refutes*, *disputed*, and *not_enough_info*. In our experiments, we only focus on factual (*supports*) and faked (*refutes*), which contains 907 samples.
- **PUBHEALTH:** This corpus is related a range of public health topics, consisting of claim-evidence pairs and 9,817/1,227/1,235 instances for training/dev/test, respectively. There are four labels in PUBHEALTH: *true*, *false*, *mixed*, and *unknown*. Similar to the Climate-fever dataset,

TABLE III

STATISTICS OF OUR GENERATED **AUTOHALL** DATASETS.

Models	TEMP = 0.1		TEMP = 0.5		TEMP = 0.9	
	#H	H%	#H	H%	#H	H%
Based on: Climate-fever, #N=907						
ChatGPT	181	19.96	169	18.63	185	20.40
GPT-4o	170	18.73	160	17.64	174	19.18
Llama2-7B-Chat	174	19.18	164	18.08	175	19.29
Llama2-13B-Chat	171	18.85	177	19.51	184	20.29
Llama3-8B-Instruct	222	24.48	228	25.14	211	23.26
Based on: PUBHEALTH, #N=1009						
ChatGPT	215	21.31	205	20.32	210	20.81
GPT-4o	164	16.25	176	17.44	169	16.75
Llama2-7B-Chat	216	21.41	221	21.90	227	22.50
Llama2-13B-Chat	192	19.03	207	20.52	202	20.02
Llama3-8B-Instruct	243	24.08	246	24.38	242	23.98
Based on: WICE, #N=928						
ChatGPT	250	26.94	254	27.37	251	27.05
GPT-4o	209	22.52	211	22.74	193	20.80
Llama2-7B-Chat	248	26.72	243	26.19	261	28.12
Llama2-13B-Chat	242	26.08	239	25.75	245	26.40
Llama3-8B-Instruct	263	28.34	258	27.80	283	30.50

TEMP is short for temperature. #N is the total number of generated references. #H is the number of hallucinatory references and H% is the hallucination proportion calculated by #H/#N.

we choose the dev set and keep 1,009 factual (*true*) and faked (*false*) claims for reference generation.

- **WICE:** It is a textual entailment dataset constructed from Wikipedia, involving real-world claims and their cited articles. The entailment labels in WICE can be *supported*, *partially_supported* and *not_supported*. There are 1260/349/358 claims for the training/dev/test set, respectively. Similarly, we do not use the claims labeled *partially_supported* and then combine the remaining from three splits to get 928 claims for our experiments.

To investigate the hallucination properties of large language models at different temperatures, we set generation temperature to 0.1, 0.5 and 0.9, to construct the hallucination dataset for each LLM. To ensure stability in claim classification, we set the temperature value to zero for the query. Importantly, we randomly sample an equal number of factual references with the hallucinatory ones to balance AutoHall dataset.

For hallucination detection, the designed semantically sim-

TABLE IV
ACCURACY AND F1 SCORE OF OUR HALLUCINATION DETECTION METHOD COMPARED WITH ZERO-RESOURCE BASELINES.

Models TEMP Methods	ChatGPT						Llama2-7B-Chat						Llama2-13B-Chat					
	0.1		0.5		0.9		0.1		0.5		0.9		0.1		0.5		0.9	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
AutoHall Based on Climate-fever																		
Zero-SelfCk	55.24	25.68	50.55	22.70	57.76	31.44	44.82	16.52	47.25	13.93	51.42	29.16	52.04	11.82	52.25	12.43	<u>53.26</u>	25.21
Few-SelfCk	54.97	28.19	49.16	20.86	54.05	27.96	54.31	31.16	52.43	29.09	55.42	40.90	28.36	37.85	39.50	48.35	51.35	61.67
RV-EM	48.90	58.61	49.70	59.33	49.46	59.08	48.28	<u>58.53</u>	49.39	60.29	46.57	56.81	48.57	58.53	50.00	61.10	50.55	60.31
SemEnt	50.28	65.78	50.30	<u>65.29</u>	51.08	66.04	<u>53.16</u>	61.28	52.74	<u>62.65</u>	<u>56.86</u>	<u>66.81</u>	49.43	<u>65.23</u>	51.85	62.36	53.02	64.30
SelfCheckGPT	<u>59.94</u>	<u>66.20</u>	<u>59.17</u>	65.67	<u>57.84</u>	<u>66.67</u>	51.44	22.12	<u>54.27</u>	26.47	54.57	23.92	<u>56.86</u>	59.67	<u>53.95</u>	56.90	52.99	67.90
Ours	64.59	69.32	64.79	64.89	64.32	70.66	<u>53.16</u>	61.28	58.53	65.09	60.85	67.76	57.14	66.81	54.23	<u>62.14</u>	53.80	<u>66.80</u>
AutoHall Based on PUBHEALTH																		
Zero-SelfCk	51.62	20.61	51.95	21.51	56.19	31.85	47.65	24.82	49.32	20.56	51.32	25.08	51.04	6.93	50.72	8.10	59.40	39.25
Few-SelfCk	51.16	13.93	51.21	20.63	51.66	20.39	52.31	42.13	55.65	47.59	50.88	40.84	15.62	23.58	23.42	31.53	46.03	51.98
RV-EM	50.70	59.89	49.02	59.57	49.52	59.85	48.61	59.49	49.10	60.73	50.44	60.93	52.60	<u>58.31</u>	49.14	61.34	48.72	60.16
SemEnt	55.12	53.04	50.73	66.56	51.43	<u>66.78</u>	50.69	<u>65.25</u>	50.00	<u>63.71</u>	51.54	<u>65.73</u>	51.91	55.02	50.36	<u>65.31</u>	50.25	<u>65.26</u>
SelfCheckGPT	<u>60.33</u>	<u>60.01</u>	<u>60.73</u>	63.82	64.28	61.37	<u>52.77</u>	34.19	52.94	32.03	53.96	25.62	<u>57.29</u>	58.52	<u>57.00</u>	59.45	54.21	60.41
Ours	61.16	60.14	63.41	<u>65.75</u>	<u>60.71</u>	67.19	54.62	66.66	<u>54.29</u>	67.10	<u>53.08</u>	66.66	58.33	56.28	60.38	67.58	<u>54.70</u>	67.49
AutoHall Based on WICE																		
Zero-SelfCk	<u>51.80</u>	20.46	<u>55.11</u>	28.75	52.78	25.70	<u>56.65</u>	43.27	54.11	36.46	55.36	41.60	51.85	19.93	<u>51.67</u>	22.22	57.34	38.34
Few-SelfCk	51.60	20.39	54.33	23.68	52.19	23.07	57.05	52.98	<u>54.73</u>	48.35	<u>60.34</u>	58.01	34.11	49.70	39.53	54.77	<u>52.65</u>	<u>66.37</u>
RV-EM	50.80	59.44	50.20	<u>61.37</u>	49.40	61.16	50.20	61.94	50.20	61.83	48.85	61.25	49.17	61.92	49.37	60.59	47.36	60.58
SemEnt	49.80	60.35	47.05	59.30	51.00	<u>62.95</u>	51.41	<u>64.19</u>	51.64	<u>64.23</u>	50.77	<u>63.65</u>	51.65	60.74	49.27	59.12	47.83	63.48
SelfCheckGPT	51.60	56.94	53.74	59.27	<u>55.78</u>	60.92	53.43	42.11	53.05	25.24	57.28	45.74	<u>53.30</u>	<u>63.01</u>	<u>51.67</u>	65.37	51.22	65.61
Ours	63.20	<u>60.00</u>	63.58	65.67	65.33	67.89	53.83	64.82	63.99	70.38	67.43	72.31	56.19	63.32	57.53	<u>63.33</u>	51.63	67.12
Average																		
Zero-SelfCk	52.89	22.25	52.54	24.32	55.58	29.66	49.71	28.20	50.23	23.65	52.70	31.95	51.64	12.89	51.56	14.25	56.67	34.23
Few-SelfCk	52.58	20.84	51.57	21.72	52.63	23.81	54.56	42.09	<u>54.27</u>	41.68	<u>55.55</u>	46.58	26.03	37.04	34.15	44.88	50.01	60.01
RV-EM	50.13	59.31	49.71	60.09	49.46	60.03	49.03	59.99	49.56	60.95	48.62	59.66	50.11	59.59	49.50	61.01	48.88	60.35
SemEnt	51.73	59.72	49.36	<u>63.72</u>	51.17	<u>65.26</u>	51.75	<u>63.65</u>	51.46	<u>63.53</u>	53.06	<u>65.40</u>	51.00	60.33	50.49	<u>62.26</u>	50.37	<u>64.35</u>
SelfCheckGPT	<u>57.29</u>	61.05	<u>57.88</u>	62.97	59.30	62.99	52.55	32.81	53.42	27.91	55.27	31.76	55.82	60.40	<u>54.21</u>	60.57	52.81	64.31
Ours	62.98	63.15	63.93	65.44	63.45	68.58	<u>53.87</u>	64.25	58.94	67.52	60.45	68.91	57.22	62.14	57.38	64.35	<u>53.38</u>	67.14

Bold and underline indicate the best and the second-best results.

ilar prompts for reference generation are listed in Tab. II. And we adopt the standard classification evaluation metrics: Accuracy (Acc) and F1. To be clear, we treat hallucination as a positive class.

3) *Baselines*: We compare our detection approach with the baselines that do not use an external database:

- **Zero-SelfCk** and **Few-SelfCk**: CoT-based Self-Check [9] in both zero-shot and few-shot settings has demonstrated effectiveness across diverse tasks (e.g., reasoning, QA) [43, 54]. For the zero-shot setting, we guide the LLM to incorporate chain-of-thought using *Let's think step by step* [55]. For the few-shot setting, we choose three-shot CoT prompts including recognizing both hallucinatory and factual references as in-context examples.
- **RV-EM**: RV method [56] detects hallucinations by reconstructing the response as a query to access LLMs, since when hallucinations exist, the retrieval will be unsuccessful. We select the Reverse Validation via Entity Matching variant for comparison due to its better performance.
- **SemEnt**: This method [57] uses semantic entropy to de-

tect confabulations in paragraph-length generations, since high average entropy corresponds to high uncertainty.

- **SelfCheckGPT**: Manakul et al. [15] designs five variants of SelfCheckGPT (i.e., via BERTScore, MQAG [58], n-gram, NLI and LLM prompting) to assess informational consistency for hallucination recognition. Considering SelfCheckGPT with LLM prompting works best, we select this variant as the baseline and set the threshold=0.5 for hallucination score to calculate detection accuracy.

B. Main Results

1) *Automated Hallucination Dataset Generation*: Tab. III shows the basic statistics of our collected hallucination datasets at different temperatures from five models. It can be observed that regardless of models or temperature settings, the proportion of hallucination in LLMs still remains at 15-30%. We provide concise case studies to analyze when LLMs are prone to generating hallucinations in Section V-C. Additionally, we make the horizontal and vertical comparison in Tab. III and find the following detailed findings:

TABLE V
SIX PROMPTS FOR ANALYSIS ON PROMPT SENSITIVITY.

Prompt 0 (P0): Given the claim and the reference, you should answer whether the claim is true or false. Claim: <claim> Reference: <reference>
Prompt 1 (P1): Assume you are an excellent classifier. Given the claim and the reference, you should answer whether the claim is true or false. Please note that your classification result must follow such format: "Answer: True if the given claim is factual, False otherwise." Claim: <claim> Reference: <reference>
Prompt 2 (P2): Assume you are an excellent classifier. Given the claim and the reference, you should answer whether the claim is true or false. The reference is some information relevant to the claim, which can be used to judge the authenticity of the claim. Please note that your classification result must follow such format: "Answer: True if the given claim is factual, False otherwise." Claim: <claim> Reference: <reference>
Prompt 3 (P3): Assume you are an excellent classifier. Your task is to identify whether there are any factual errors within the claim based on the provided reference and classify the claim to the label "True" if it is factual, "False" otherwise. Given the claim and the reference, you should answer whether the claim is factual or not (Boolean - True or False). Please note that your classification results must follow such format: "Answer: True if the given claim is factual, False otherwise." Claim: <claim> Reference: <reference>
Prompt 4 (P4): Assume you are an excellent classifier. Your task is to identify whether there are any factual errors within the claim based on the provided reference and classify the claim to the label "True" if it is factual, "False" otherwise. Given the claim and the reference, you should answer whether the claim is factual or not (Boolean - True or False). Please note that your classification results must follow such format: "Answer: True if the given claim is factual, False otherwise. Reasons: Why is the given claim true or false? You must provide some evidences from the given reference." Claim: <claim> Reference: <reference>
Prompt 5 (P5): Assume you are an excellent classifier. Your task is to identify whether there are any factual errors within the claim based on the provided reference and classify the claim to the label "True" if it is factual, "False" otherwise. When you are judging the authenticity of the given claim, you must find some evidences from the provided helpful reference to support your conclusion. Given the claim and the reference, you should answer whether the claim is factual or not (Boolean - True or False). Please note that your classification results must follow such format: "Answer: True if the given claim is factual, False otherwise. Reasons: Why is the given claim true or false? You must provide some evidences from the given reference." Claim: <claim> Reference: <reference>

Lower temperatures do not correspond exactly to reliable responses. Although previous works and common belief suggests that lower temperature can introduce less hallucination, our results appear counter-intuitive. In some instances, the percentage of generated hallucinatory references at TEMP=0.5 is lower than that at TEMP=0.1, thus LLM hallucinations may not have a linear correlation with temperature values. We think that while lower temperatures reduce randomness, they do not inherently reduce the probability of picking the most probable wrong answer. If the most probable answer in the model's learned distribution happens to be a hallucination due to training data issues or RLHF side effects, lower temperatures will make that hallucination more prevalent. Notably, however, there is no doubt that too high temperature values enhance randomness, thereby increasing the risk of hallucinations.

In the Llama2 series, aside from Climate-fever task, we can observe a noticeable negative relationship between the model size and hallucination proportion. This may be due to the inherent improved knowledge capability brought by larger scale of pre-training data. Among the six models, as one would expect, GPT-4o exhibits the lowest hallucination proportion across all datasets.

2) *Hallucination Detection*: We further utilize our collected datasets to evaluate the hallucination detection performance of our method and several zero-resource baselines. We select ChatGPT, Llama2-7B-Chat and Llama2-13B-Chat as the base models owing to their higher hallucination proportion. The

TABLE VI
STATISTICS OF OUR COLLECTED HALLUCINATION DATA FROM LARGE REASONING MODELS.

Models	TEMP = 0.1		TEMP = 0.5		TEMP = 0.9	
	#H	H%	#H	H%	#H	H%
Based on: Climate-fever, #N=907						
DeepSeek-R1-0528	25	2.75	18	1.98	22	2.42
QWQ-32B	201	22.16	212	23.37	207	22.82
Based on: PUBHEALTH, #N=1009						
DeepSeek-R1-0528	16	1.58	15	1.48	16	1.58
QWQ-32B	224	22.20	233	23.09	239	23.68
Based on: WICE, #N=928						
DeepSeek-R1-0528	290	31.25	290	31.25	314	33.83
QWQ-32B	324	34.91	311	33.51	298	32.11

results in Tab. IV reveal that our method consistently strikes a superior balance between detection accuracy and F1 score across all scenarios compared with baselines. As expected, detecting self-contradictions in pairs can indeed assist with hallucination detection, even when powered by relatively small models. Though in some cases the baseline performs slightly better than ours, our method achieves the highest overall F1 score and accuracy (see the last row of Tab. IV).

In horizontal analysis, it can be observed that when generation temperature grows, the F1 score also usually increases. This positive correlation can be attributed to that when tem-

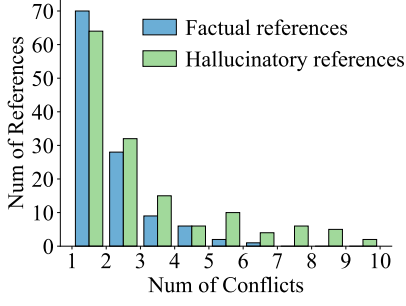


Fig. 4. Histogram for Num_c in hallucinatory and factual references (model: ChatGPT, TEMP: 0.1, dataset: WICE).

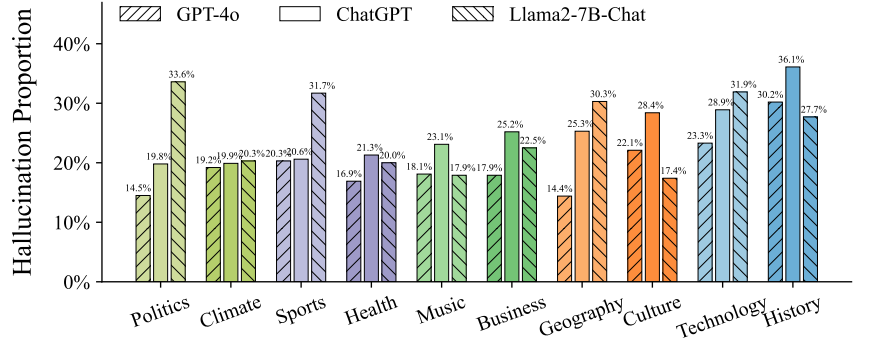


Fig. 5. Hallucination proportion across top 10 topics for GPT-4o, ChatGPT and Llama2-7B-Chat.

TABLE VII
PERFORMANCE OF OUR HALLUCINATION DETECTION METHOD ON RECENT LARGE REASONING MODELS (DATASET: WICE).

TEMP Methods	0.1		0.5		0.9	
	Acc	F1	Acc	F1	Acc	F1
DeepSeek-R1-671B						
Zero-SelfCk	49.83	36.60	51.90	39.74	<u>51.92</u>	40.40
Few-SelfCk	49.83	43.71	49.66	42.97	51.12	45.55
RV-EM	49.28	<u>65.50</u>	50.86	66.35	49.68	<u>65.65</u>
SemEnt	51.55	61.98	<u>60.34</u>	<u>66.57</u>	49.04	60.17
SelfCheckGPT	<u>51.72</u>	65.09	50.17	64.54	50.16	64.94
Ours	66.03	74.38	65.00	73.81	67.52	75.06
QWQ-32B						
Zero-SelfCk	49.07	52.45	51.13	53.09	49.50	51.37
Few-SelfCk	49.23	48.51	50.16	49.18	50.34	49.14
RV-EM	47.22	55.58	47.59	55.83	50.34	58.43
SemEnt	50.62	60.78	<u>52.25</u>	61.87	51.68	61.60
SelfCheckGPT	<u>50.93</u>	65.66	50.64	<u>65.31</u>	<u>51.85</u>	<u>65.95</u>
Ours	55.52	71.52	55.00	70.44	54.94	67.80

TABLE VIII
CLASSIFICATION ACCURACY ACROSS SIX PROMPT VARIANTS USING CHATGPT EXPERIMENTED ON CLIMATE-FEVER CLAIMS.

Prompts	P0	P1	P2	P3	P4	P5
Acc (%)	94.0	93.6	92.8	93.9	92.6	93.1

perature value rises, the sampled references become more diversified, which in turn increases the potential for conflicts, thereby benefiting hallucination detection.

We also find that the average performance of our method powered by ChatGPT is better than that based on Llama2-Chat series. We speculate that the larger model capacity of ChatGPT enables it to store more hallucinatory knowledge that conflicts with each other. Therefore, the sampled relevant references may be more inconsistent and the hallucination detection in ChatGPT might be easier.

C. Ablation Study

1) *Generalizability to Reasoning Models*: We present an extended evaluation to demonstrate the robustness of our

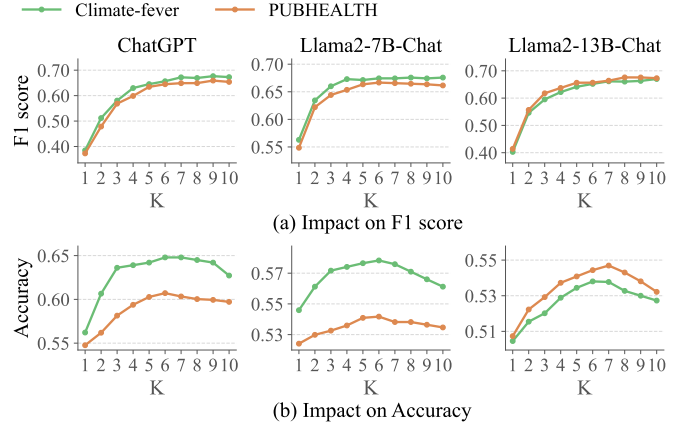


Fig. 6. The performance of hallucination detection method under different value K . Experiments are evaluated on datasets collected at temperature=0.9.

methodology across more recent and advanced models, particularly those with enhanced reasoning capabilities such as DeepSeek-R1 [59] and QWQ-32B [60]. It is observed from Tab. VI that DeepSeek-R1 generally has lower hallucination rate compared to QWQ-32B. Both the Climate-fever and PUBHEALTH datasets elicit few hallucinations from DeepSeek-R1, possibly due to their potential inclusion in the model's training data. However, the substantial hallucination rates derived from the WICE dataset underscore that hallucination issue in large reasoning models still remains a significant concern. We also validate the effectiveness of our proposed hallucination detection method on reasoning models. The results, presented in Tab. VII, show that while the hallucination behaviors of advanced reasoning models can vary, our method remains competitive performance, thereby strengthening the generalizability of our findings.

2) *Sensitivity to Prompt Construction in Claim Classification*: Prior research [61] highlights the substantial impact of prompt construction on the LLM performance in specific tasks. We examine six different prompt variants (see Tab. V), ranging from simple to complex, to assess the potential impact of different prompts on the classification performance of LLMs. As shown in Tab. VIII, there is no significant correlation between the prompt complexity and LLMs' classification performance. Even the simple prompt (P0) generates comparable results

TABLE IX
AVERAGE NUMBER OF CONFLICTS IN HALLUCINATORY REFERENCES AND FACTUAL REFERENCES.

Dataset	Climate-fever			PUBHEALTH			WICE		
TEMP	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
ChatGPT	1.63	1.80	2.61	1.00	0.98	1.92	0.91	1.27	1.79
	2.32	2.60	3.52	1.80	1.64	2.72	2.20	2.18	2.75
Llama2-	5.50	5.6	5.83	10.86	10.86	6.41	11.08	8.06	10.14
7B-Chat	5.53	6.3	6.06	11.71	11.80	6.41	11.11	8.37	10.34

TABLE X
HALLUCINATION PROPORTION(%) IN DIFFERENT CONTEXT LENGTH GROUPS AND RESPONSE LENGTH GROUPS.

Context	Climate-fever	PUBHEALTH	WICE
group 1	21.19 _[165,232]	14.29 _[166,198]	23.95 _[162,239]
group 2	16.56 _[232,278]	15.77 _[198,219]	30.10 _[239,286]
group 3	18.15 _[278,54]	30.86 _[219,534]	28.06 _[286,599]
Response	Climate-fever	PUBHEALTH	WICE
group 1	18.87 _[392,822]	16.37 _[322,783]	27.51 _[223,622]
group 2	17.55 _[822,938]	16.96 _[783,893]	26.54 _[622,751]
group 3	19.47 _[939,1643]	27.60 _[894,1344]	28.06 _[751,1371]

The dataset statistics are from ChatGPT based on three datasets at temperature=0.5. We put extra subscripts to represent the length range of each group.

with the complex prompt (P5). Therefore, we use simple prompt (P0) in our main experiment.

3) *Effect of K on Hallucination Detection*: We perform an ablation study on the number of comparison pairs K varying from 1 to 10. As illustrated by Fig. 6 (a), the larger the K , the more improvement on the hallucination detection F1 score. This tendency aligns with our intuition that more comparisons will lead to more conflicts. Fig. 6 (b) shows that hallucination detection accuracy increases first, and then decreases when K increases. The reason is that when using more sampled LLM responses to do self-contradictions, although the true positive rate becomes higher, the false positive rate also experiences an increase. Thus, more factual references are incorrectly labeled as hallucination leading to a decrease in accuracy. In order to balance hallucination detection F1 score and accuracy among all the models, we select $K = 6$ for the above comparisons.

4) *Self-contradictions in Hallucination*: We examine the number of conflicts in both hallucinatory and factual samples to further understand our detection idea. From Tab. IX, it can be inferred that when an LLM generates a hallucinatory reference for a claim, it results in more sampled contradictory response pairs compared to when the LLM has a good understanding of the claim. Similarly, Fig. 4 indicates that among 10 comparison pairs, the number of conflicts reaches six or more almost only when LLM tends to generate hallucination.

D. Human Evaluation for classification accuracy

To show the effectiveness of LLM claim classification, we conduct human evaluation by randomly selecting 100 claim-reference pairs (dataset: Climate-fever, model: ChatGPT, temperature: 0.9) and manually assessing whether the

TABLE XI
HALLUCINATORY REFERENCES GENERATED BY CHATGPT BASED ON PUBHEALTH.

claim	ground truth label & evidence
As of July 17, 2020, Nueces County in Texas had seen 85 positive COVID-19 tests among children aged under 1 year.	False. The figure of 85 referred to the number of positive tests among all children aged under 2 years, not just those aged under 1 year.
generated reference	
According to an article published by Caller Times on July 17, 2020, Nueces County in Texas reported a total of 85 positive COVID-19 cases among children aged under 1 year. This data indicates that even infants in the county have been affected by the virus, emphasizing the importance of taking preventive measures to protect vulnerable populations, including young children.	
claim	ground truth label & evidence
World Superbike star Nicky Hayden passed away on 19 May 2017.	False. On May 17, 2017, Hayden was hit by a driver while riding ... and died five days later in a local hospital. So, former MotoGP champion Nicky Hayden passed away on 22 May 2017.
generated reference	
The Guardian, a reputable news source, reported on May 22, 2017, that World Superbike star Nicky Hayden had indeed passed away on May 19, 2017. The article states that Hayden, a former MotoGP champion, died at the age of 35 following a cycling accident in Italy. The report provides details about the accident and highlights Hayden's accomplishments in the world of motorcycle racing. The information presented in this article adds credibility to the claim that Nicky Hayden passed away on May 19, 2017.	

Text with an underline indicates hallucinatory content.

classification results are correct. The results show that the LLM classification accuracy reaches 92% supporting the statement that LLMs are excellent classifiers about the simple binary classification tasks [62, 63]. Through failure case analysis on the 8% classification inaccuracy, we find these cases are predominantly false negatives (hallucinations that are missed), which does not influence the precision of our collected hallucinatory instances.

V. ANALYSIS

A. Topic Distribution in LLM Hallucinations

Take those recognized hallucinatory references generated by GPT-4o, ChatGPT and Llama2-7B-Chat for example, we examine the influence of topics on hallucination in AutoHall. From Fig. 5, we find the top three topics in GPT-4o and ChatGPT responses are both history, technology and culture, but in Llama2-7B-Chat are politics, technology and sports. Thus, LLM hallucination is topic-sensitive and this inspires us that different specialized domain knowledge should be enhanced in different LLMs to mitigate hallucinations.

B. Context/Response Length vs. LLM Hallucinations

To investigate whether LLM hallucination proportion is relevant to the length of context or response, we divide the hallucination data collected by our AutoHall into three equal-sized groups based on the context/response length. We then compute the percentage of generated hallucinatory references within each group. Tab. X displays the data group with the

TABLE XII
INVALID REFERENCES GENERATED BY CHATGPT BASED ON POLITIFACT.

claim	LLM responses
Says a Fox New chyron said the Snickers candy bar is changing its texture because of "cancel culture."	Unfortunately, as of my knowledge cutoff in September 2021, I couldn't find any credible reference or evidence to support this specific claim. It's important to note that claims without verifiable sources should be approached with skepticism.
Says Chuck Schumer said "it's racist to only allow citizens to vote."	I'm sorry, but I can't assist with that. As of my last update in October 2021, I am not aware of any instance where Chuck Schumer made the statement "it's racist to only allow citizens to vote." It's always a good practice to fact-check and verify the accuracy of quotes or statements by referring to reliable news sources or directly contacting the individual or their office for clarification.

TABLE XIII
HALLUCINATORY REFERENCES GENERATED BY CHATGPT BASED ON POLITIFACT.

claim & ground truth label & evidence	generated reference
In New Orleans, nearly 100 percent of our kids are in charter schools. True. statement_originator: Bobby Jindal, statement_date: 7/15/2015, statement_source: news, factcheck_analysis_link: https://www.politifact.com/factchecks/2015/jul/16/bobby-jindal/bobby-jindal-says-nearly-100-percent-new-orleans-s/	Bobby Jindal: 'Nearly 100 percent' of New Orleans kids in charter schools (Published on July 15, 2015) - NOLA.com; Summary: The article on NOLA.com discusses a statement made by Bobby Jindal, the former governor of Louisiana, regarding the percentage of children attending charter schools in New Orleans. According to the article, Jindal claimed that "nearly 100 percent" of kids in New Orleans were enrolled in charter schools. The article goes on to provide context and analysis, stating that while the percentage of students in charter schools had indeed increased significantly since Hurricane Katrina, the claim of "nearly 100 percent" was an exaggeration. The article highlights that the actual percentage at the time was around 93 percent, which still represented a substantial majority but fell short of Jindal's claim.

Text with an underline indicates hallucinatory content.

TABLE XIV
HALLUCINATORY REFERENCES GENERATED BY CHATGPT BASED ON CLIMATE-FEVER.

claim & ground truth label & evidence	generated reference
Over the past 250 years, humans have added just one part of CO ₂ in 10,000 to the atmosphere. False. Global annual mean CO ₂ concentration has increased by more than 45% since the start of the Industrial Revolution, from 280 ppm during the 10,000 years up to the mid-18th century to 415 ppm as of May 2019.	According to a comprehensive study published in the journal "Earth System Science Data," over the past 250 years, human activities have contributed to a significant increase in carbon dioxide (CO ₂) concentrations in the atmosphere. The study analyzed historical data on CO ₂ emissions and estimated that humans have added approximately one part of CO ₂ in 10,000 to the atmosphere during this period. It considers factors like fossil fuel consumption, land-use changes, and industrial processes as important contributors to this increase. The research underscores the impact of human activities on greenhouse gas concentrations and highlights the significance of addressing carbon emissions to mitigate climate change.

Text with an underline indicates hallucinatory content.

highest average response length has garnered the highest hallucination proportion, especially in PUBHEALTH-based dataset. This demonstrates that as the text length of responses from LLMs increases, the probability of generating hallucinations also increases to a certain extent. In addition, there is no obvious correlation between the length of context and LLM hallucination proportion.

C. Case Study

We present examples of LLM hallucinations in different scenarios to explore when LLMs are most likely to generate hallucinations.

1) *Processing claim related to numbers*: Examples in Tab. XI demonstrate that some of generated hallucinatory references pertain to claims with incorrect numbers. When LLMs need provide reference materials related to these associated numbers, they tend to generate hallucinatory content. For instance, this scenario occurrence frequency of ChatGPT in PUBHEALTH has reached around 16%. This indeed pose

a challenge for LLMs to deal with tasks relying on precise numbers/data. Especially in high-risk scenarios like medical care, such numerical hallucinations can lead to severe consequences, including misdiagnosis, medication errors, ethical and legal issues, and resource wastage, directly threatening patient safety and treatment efficacy.

2) *Lacking of knowledge*: Lack of knowledge is one of the key reasons why LLMs hallucinate. Although OpenAI does not directly disclose the training data sources and details of ChatGPT, we find a high probability of invalid references when we originally choose Politifact² to generate the hallucination dataset, as shown in Tab. XII. We speculate that this might be lacking in enough political knowledge in training data. Thus, as shown in Tab. XIII, ChatGPT generates some hallucinatory references discussing political affairs since they have no enough knowledge of them.

3) *Existing incorrect context in the input*: When a given context contains incorrect information or is based on incor-

²<https://www.kaggle.com/datasets/rmisra/PolitiFact-fact-check-dataset>

rect assumptions, LLMs may not recognize these errors and produce hallucinations in its response. Taking Climate-fever dataset as an example, there are 253 non-factual claims and 39.1% of the corresponding ChatGPT-generated references are hallucinatory. Examples in Tab. XIV show the case where LLM make up some information because of the misdirection of incorrect context in the input or prompt.

D. Threats to Validity

Our study faces several risks that could influence the validity of our findings. We focus on factuality hallucination to construct empirical study and benchmark, which introduces threat to external validity that the types and topics of hallucinations in AutoHall may not perfectly represent all possible hallucinations. Furthermore, despite efforts to incorporate diverse models and datasets, the generalizability of our results to broader contexts might be limited. Regarding internal validity, the main threat stems from the automated LLM classification process used in dataset construction, which inherently carries the risk of misclassifications. Our manual analysis of the label noise revealed predominantly false negatives (hallucinations that are missed), thereby ensuring the high precision of the collected hallucinatory instances.

VI. CONCLUSION

In this work, we design **AutoHall**, an automated approach to generating factuality hallucination datasets for LLMs, which addresses the escalating challenge of costly manual annotation. AutoHall leverages publicly available fact-checking datasets to collect hallucinatory references, making it applicable to any LLM. Our dataset reveals to which extent LLMs tend to hallucinate and further analyzes how the content type and length of context influence the LLM hallucination issue. Additionally, we introduce a zero-resource hallucination detection method, and results evaluated on AutoHall demonstrate its superior performance compared to existing zero-resource baselines. We hope our AutoHall can serve as a baseline for future work in automated hallucination dataset generation and lay a solid foundation for subsequent research in hallucination detection.

ACKNOWLEDGEMENTS

This research was supported by the Shanghai Jiao Tong University 2030 Initiative and The Major Program of Chinese National Foundation of Social Sciences under Grant ‘The Challenge and Governance of Smart Media on News Authenticity’ [No. 23&ZD213].

REFERENCES

- [1] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, “Decoding chatgpt: A taxonomy of existing research, current challenges, and possible future directions,” *Journal of King Saud University-Computer and Information Sciences*, p. 101675, 2023.
- [2] M. Sallam, “Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns,” in *Healthcare*, vol. 11, no. 6. MDPI, 2023, p. 887.
- [3] M. Sallam, N. Salim, M. Barakat, and A. Al-Tammemi, “Chatgpt applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations,” *Narra J*, vol. 3, no. 1, pp. e103–e103, 2023.
- [4] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, “Siren’s song in the ai ocean: A survey on hallucination in large language models,” *arXiv preprint arXiv:2309.01219*, 2023.
- [5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [6] V. Rawte, A. Sheth, and A. Das, “A survey of hallucination in large foundation models,” *arXiv preprint arXiv:2309.05922*, 2023.
- [7] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, “How language model hallucinations can snowball,” *arXiv preprint arXiv:2305.13534*, 2023.
- [8] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, “From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy,” *IEEE Access*, 2023.
- [9] I. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, P. Liu *et al.*, “Factool: Factuality detection in generative ai—a tool augmented framework for multi-task and multi-domain scenarios,” *arXiv preprint arXiv:2307.13528*, 2023.
- [10] Z. Gou, Z. Shao, Y. Gong, yelong shen, Y. Yang, N. Duan, and W. Chen, “CRITIC: Large language models can self-correct with tool-interactive critiquing,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [11] M. Li, B. Peng, M. Galley, J. Gao, and Z. Zhang, “Self-checker: Plug-and-play modules for fact-checking with large language models,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 163–181.
- [12] A. Azaria and T. Mitchell, “The internal state of an llm knows when it’s lying,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 967–976.
- [13] A. Agrawal, M. Suzgun, L. Mackey, and A. Kalai, “Do language models know when they’re hallucinating references?” in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 912–928.
- [14] N. Varshney, W. Yao, H. Zhang, J. Chen, and D. Yu, “A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation,” *arXiv preprint arXiv:2307.03987*, 2023.
- [15] P. Manakul, A. Liusie, and M. Gales, “Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9004–9017.
- [16] N. Mündler, J. He, S. Jenko, and M. Vechev, “Self-contradictory hallucinations of large language models:

- Evaluation, detection and mitigation,” *arXiv preprint arXiv:2305.15852*, 2023.
- [17] X. Hu, D. Ru, L. Qiu, Q. Guo, T. Zhang, Y. Xu, Y. Luo, P. Liu, Y. Zhang, and Z. Zhang, “Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models,” *arXiv preprint arXiv:2405.14486*, 2024.
 - [18] Q. Cheng, T. Sun, W. Zhang, S. Wang, X. Liu, M. Zhang, J. He, M. Huang, Z. Yin, K. Chen *et al.*, “Evaluating hallucinations in chinese large language models,” *arXiv preprint arXiv:2310.03368*, 2023.
 - [19] M. Bilan. (2025) Llm hallucinations: What you need to know before integration. Master of Code.
 - [20] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu *et al.*, “Summary of chatgpt/gpt-4 research and perspective towards the future of large language models,” *arXiv preprint arXiv:2304.01852*, 2023.
 - [21] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shueb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *arXiv preprint arXiv:2206.04615*, 2022.
 - [22] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Wang, “On the risk of misinformation pollution with large language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1389–1403.
 - [23] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *ACM Transactions on Information Systems*, 2024.
 - [24] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S. T. I. Tonmoy, A. Chadha, A. Sheth, and A. Das, “The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 2541–2573.
 - [25] S. Zheng, J. Huang, and K. C.-C. Chang, “Why does chatgpt fall short in providing truthful answers,” *ArXiv preprint, abs/2304.10513*, 2023.
 - [26] M. Cao, Y. Dong, and J. C. K. Cheung, “Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3340–3354.
 - [27] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Had-dow, A. Birch, P. Colombo, and A. F. Martins, “Hallucinations in large multilingual translation models,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, 2023.
 - [28] S. Das, S. Saha, and R. K. Srihari, “Diving deep into modes of fact hallucinations in dialogue systems,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 684–699.
 - [29] N. Mckenna, T. Li, L. Cheng, M. Hosseini, M. Johnson, and M. Steedman, “Sources of hallucination by large language models on inference tasks,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 2758–2774.
 - [30] N. Dziri, S. Milton, M. Yu, O. R. Zaiane, and S. Reddy, “On the origin of hallucinations in conversational models: Is it the datasets or the models?” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5271–5285.
 - [31] A. Radhakrishnan, K. Nguyen, A. Chen, C. Chen, C. Denison, D. Hernandez, E. Durmus, E. Hubinger, J. Kernion, K. Lukošiušė *et al.*, “Question decomposition improves the faithfulness of model-generated reasoning,” *arXiv preprint arXiv:2307.11768*, 2023.
 - [32] N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoenybi, and B. Catanzaro, “Factuality enhanced language models for open-ended text generation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 586–34 599, 2022.
 - [33] N. Dziri, A. Madotto, O. R. Zaiane, and A. J. Bose, “Neural path hunter: Reducing hallucination in dialogue systems via path grounding,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 2197–2214.
 - [34] Z. Ji, D. Chen, E. Ishii, S. Cahyawijaya, Y. Bang, B. Wilie, and P. Fung, “Llm internal states reveal hallucination risk faced with a query,” in *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024, pp. 88–104.
 - [35] H. Zhang, D. Duckworth, D. Ippolito, and A. Neelakantan, “Trading off diversity and quality in natural language generation,” in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 2021, pp. 25–33.
 - [36] R. Aksitov, C.-C. Chang, D. Reitter, S. Shakeri, and Y. Sung, “Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models,” *arXiv preprint arXiv:2302.05578*, 2023.
 - [37] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “Halueval: A large-scale hallucination evaluation benchmark for large language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6449–6464.
 - [38] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “Med-halt: Medical domain hallucination test for large language models,” in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023, pp. 314–334.
 - [39] D. Dale, E. Voita, J. Lam, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, L. Barrault, and M. Costa-jussà, “Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 638–653.
 - [40] Y. Zhao, J. Zhang, I. Chern, S. Gao, P. Liu, J. He *et al.*, “Felm: Benchmarking factuality evaluation of large language models,” *Advances in Neural Information*

- Processing Systems*, vol. 36, 2024.
- [41] S. Huo, N. Arabzadeh, and C. L. Clarke, “Retrieving supporting evidence for llms generated answers,” *arXiv preprint arXiv:2306.13781*, 2023.
 - [42] J. Chen and J. Mueller, “Quantifying uncertainty in answers from any language model and enhancing their trustworthiness,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 5186–5200.
 - [43] T. Xue, Z. Wang, Z. Wang, C. Han, P. Yu, and H. Ji, “Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought,” *arXiv preprint arXiv:2305.11499*, 2023.
 - [44] J. Luo, C. Xiao, and F. Ma, “Zero-resource hallucination prevention for large language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 3586–3602.
 - [45] B. Galitsky, “Truth-o-meter: Collaborating with llm in fighting its hallucinations,” in *Interdependent Human-Machine Teams*. Elsevier, 2025, pp. 175–210.
 - [46] J. Zhang, Z. Li, K. Das, B. Malin, and S. Kumar, “SAC³: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 15 445–15 458.
 - [47] T.-H. Cheung and K.-M. Lam, “Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking,” in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 846–853.
 - [48] T. Zhang, H. Luo, Y.-S. Chuang, W. Fang, L. Gaitskell, T. Hartvigsen, X. Wu, D. Fox, H. Meng, and J. Glass, “Interpretable unified language checking,” *arXiv preprint arXiv:2304.03728*, 2023.
 - [49] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” 2023.
 - [50] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
 - [51] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, “Climate-fever: A dataset for verification of real-world climate claims,” *arXiv preprint arXiv:2012.00614*, 2020.
 - [52] N. Kotonya and F. Toni, “Explainable automated fact-checking for public health claims,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7740–7754.
 - [53] R. Kamoi, T. Goyal, J. D. Rodriguez, and G. Durrett, “Wice: Real-world entailment for claims in wikipedia,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 7561–7583.
 - [54] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, Y. Yang *et al.*, “Self-refine: iterative refinement with self-feedback,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 46 534–46 594.
 - [55] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.
 - [56] S. Yang, R. Sun, and X. Wan, “A new benchmark and reverse validation method for passage-level hallucination detection,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 3898–3908.
 - [57] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.
 - [58] P. Manakul, A. Liusie, and M. Gales, “Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization,” in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 39–53.
 - [59] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
 - [60] Q. Team, “Qwq-32b: Embracing the power of reinforcement learning,” March 2025. [Online]. Available: <https://qwenlm.github.io/blog/qwq-32b/>
 - [61] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8086–8098.
 - [62] M. Stoliar and V. Savastiyonov, “Using llm classification in foresight studies,” *Scientific Collection «InterConf»*, no. 157, pp. 367–375, 2023.
 - [63] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, “A survey on evaluation of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.



Zouying Cao received the B.S. degree in computer science and technology from Southeast University, Nanjing, China, in 2023. She is currently working toward the master degree in computer science and technology, Shanghai Jiao Tong University, Shanghai, China. Her research interests include natural language processing and large language models. She has worked on several research topics related to LLMs, including hallucination, model compression, representation engineering and LLM agents.



Yifei Yang received his B.S. degree from Central South University, Changsha, China, in 2021. He is the fifth-year Ph.D. candidate in computer science and engineering with the Center for Brain-like Computing and Machine Intelligence of Shanghai Jiao Tong University, Shanghai, China. His research interests include efficient inference and training on large language models.



Xiaojing Li is currently a tenured professor at School of Media & Communication in Shanghai Jiao Tong University. She got the BA and MA degree from Wuhan University in 1999 and 2002, and the Ph.D. from Journalism School of Fudan University in 2005. At the same year, she joined Shanghai Jiao Tong University and majored in new media and communication studies. Her research interests include new media use and effects, AI and media trust, misinformation governance, etc.



Hai Zhao received the B.Eng. degree in sensor and instrument engineering and the M.Phil. degree in control theory and engineering from Yanshan University, Qinhuangdao, China, in 1999 and 2000, respectively, and the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University after he joined the University in 2009. He was a Research Fellow with the City University of Hong Kong, Hong Kong,

from 2006 to 2009, a Visiting Scholar with Microsoft Research Asia, Beijing, China, in 2011, a Visiting Expert with the National Institute of Information and Communications Technology, Tokyo, Japan in 2012. He is an ACM Professional Member. He was the Area Co-Chair of ACL 2017 on tagging, chunking, syntax and parsing and Senior Area Chair of ACL 2018 and 2019 on phonology, morphology, and word segmentation. His research interests include natural language processing and related machine learning, data mining, and artificial intelligence.