🍄 UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation

Anonymous ACL submission

Abstract

Large language models (LLMs) produce hallucinated text, compromising their practical utility in professional contexts. To assess the reliability of LLMs, numerous initiatives have developed benchmark evaluations for hallucination phenomena. However, they often employ constrained generation techniques to produce the evaluation dataset due to cost and time limitations. For instance, this may involve employing directed hallucination induction or deliberately modifying authentic text to generate hallucinations. These are not congruent with the unrestricted text generation demanded by real-world applications. Furthermore, a well-established Chinese-language dataset dedicated to the evaluation of hallucinations is presently lacking. Consequently, we have developed an Unconstrained Hallucination Generation Evaluation (UHGEval) benchmark, containing hallucinations generated by LLMs with minimal restrictions. Concurrently, we have established a comprehensive benchmark evaluation framework to aid subsequent researchers in undertaking scalable and reproducible experiments. We have also evaluated prominent Chinese LLMs and the GPT series models to derive insights regarding hallucination.

1 Introduction

Large language models (LLMs) have unparalleled proficiency in language generation, knowledge application, and intricate reasoning (Zhao et al., 2023). However, these models invariably manifest hallucination (Rawte et al., 2023), as they often generate content that is incongruent with user input, the model's output context, or factual information. Real-world hallucination examples from our UHGEval dataset can be observed in Fig. 1.

The fabricated news content depicted in Fig. 1 offers NO utility to journalists; on the contrary, the

Figure 1: Hallucinations from UHGEval. Using the IDs, you can locate the original news articles. *Note*: MOTIE denotes Ministry of Trade, Industry, and Energy. (In Chinese: Fig. 10)

verification and rectification of such content exacts a toll on the valuable time of journalists. To this concern, it is crucial to first formulate a comprehensive, stringent, and demanding benchmark for the assessment of hallucination in language generation (Zhang et al., 2023; Wang et al., 2023b).

While there have been a bunch of efforts to develop benchmarks for hallucination assessment, they always employ restricted techniques to produce particular kinds of hallucinated utterances. This approach is at odds with real-world scenarios where hallucinations arise in unrestricted, spontaneously generated content. For example, HaluEval specifies the type of hallucination in the prompt when generating hallucinated text: "You are trying to answer a question but misunderstand the question context and intention" (Li et al., 2023). Additionally, benchmarks such as HADES annotate hallucinations at a finer granularity by generating token-level hallucinations based on text perturbations (Liu et al., 2022), but the text perturbation method is still constrained. Besides, many benchmarks are centered on the evaluation of hallucinations in English, neglecting the assessment of such phenomena in Chinese. The extensive lexicon of Chinese characters, combined with the complexities introduced by Chinese word segmentation,
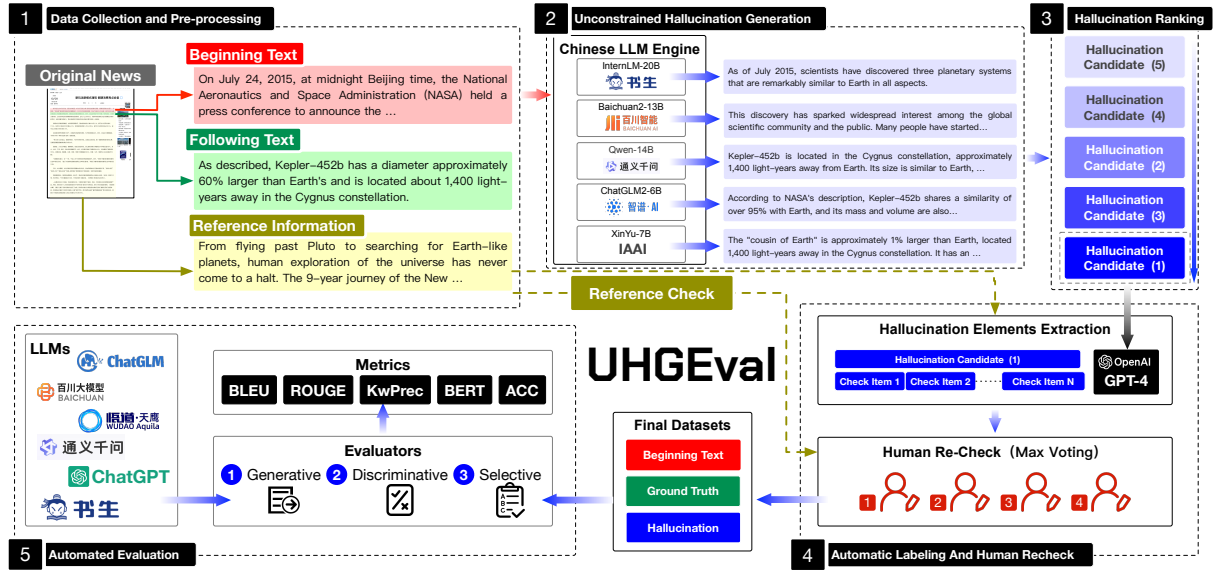
Note: We follow double-blind peer review requirements and refer to our author institutions as Org1, Org2, and Org3.

1

Figure 2: The process of creating 𝕌ℍ𝔾Eval. Steps 1 to 4 regarding the creation of the benchmark dataset are explained in Section 2; Step 5, concerning the evaluation framework, is detailed in Section 3. (In Chinese: Fig. 11)

renders the Chinese hallucination evaluation particularly arduous and deserving of focused scrutiny.

To address the aforementioned challenges, we introduce a novel benchmark for hallucination assessment, as depicted in Fig. 2. The benchmark dataset is composed of raw Chinese news articles and continuations of those articles freely generated by LLMs but annotated with hallucinations.

Furthermore, selecting texts from the news domain is intentional, given that news requires utmost precision in conveying factual information and exhibits minimal tolerance for hallucinations, presenting a considerable challenge for the majority of LLMs. Moreover, news data encompasses a wide range of topics, including medicine, technology, finance, sports, etc., incorporating features found in texts from other domains. Lastly, news articles are readily available and frequently employed as training corpora by a large number of LLMs, guaranteeing impartiality in the evaluation of many LLMs (Zhao et al., 2023).

Our contributions: (1) The development of an unconstrained hallucination evaluation dataset, comprising over 5000 items. Existing methods for constructing datasets often yield biases towards predefined directions, thereby hindering the full simulation of real-world hallucinations. (2) The establishment of a unified and diverse evaluation framework, 𝕌ℍ𝔾Eval, that encompasses discriminative, selective, and generative evaluations. Current benchmark methods for hallucination evaluation often exhibit a singular approach and lack task speci-ficity. (3) A comprehensive empirical analysis. We evaluated eight prominent Chinese LLMs and three classic GPT series models to explore the credibility of various LLMs. The aforementioned dataset, evaluation framework, and empirical results collectively constitute the 𝕌ℍ𝔾Eval benchmark, which is openly available on GitHub[1].

2 The UHGEval Dataset

2.1 Data Collection and Pre-processing

We amassed tens of thousands of historical news articles from leading Chinese news websites, covering the period from January 2015 to January 2017, to serve as the foundation for constructing the dataset. It is worth noting that the decision to eschew more recent news articles (e.g., from 2024) was made to better assess the model's understanding of existing knowledge. Indeed, the knowledge embedded within the training data of existing Chinese LLMs typically encompasses information about significant news between 2015 and 2017 (Zhao et al., 2023).

The collected news spans various topics, such as sports, education, science, society, finance, and more. This diversity underscores the advantage of choosing news texts for our dataset, as it enables the incorporation of a wide array of text genres. We hypothesize that the occurrence of hallucinations will vary as LLMs generate news across different

---

[1]Main: `https://anonymous.4open.science/r/UHGEval`
Dataset creation: `https://anonymous.4open.science/r/UHGEval-dataset`

Type	Share	Categories
DOC	27.52%	Politics, Law, Military, Education
NUM	43.34%	Sports, Economy, Market
KNO	6.55%	Science, Technology, Healthcare
GEN	22.59%	Society, Culture, Arts, Entertainment, Weather, Environmental Protection, Disasters, Accidents

Table 1: Statistics of collected news. DOC, NUM, KNO, and GEN denote document-intensive, number-intensive, knowledge-intensive, and general news, respectively.

categories. As a result, we have classified these diverse categories into four main types: document-intensive, number-intensive, knowledge-intensive, and general news, with details provided in Table 1.

In the data pre-processing stage, we divide a complete news article into three parts: the beginning text, the following text, and the reference information. The beginning text serves to guide the model in generating the continuation and is typically the opening portion of the news. During evaluation, the LLM is required to generate content following the beginning text. The following text comprises the subsequent sentences in the news article and serves as the ground truth for the continuation task. Finally, all the remaining text, after the beginning text is excluded, serves as a source of reference information. This section provides reference information for labeling and also acts as the reference text for the reference-based evaluation.

2.2 Unconstrained Hallucination Generation

Unlike directed hallucination generation (Li et al., 2023) or perturbation-based generation (Liu et al., 2022), we have adopted an unconstrained generation methodology for the continuation of natural language content, though it poses difficulties for subsequent annotations. This generation's fashion entails directly inputting the text to be continued into the model without any restrictive prompt instructions, thereby obtaining organic results.

Furthermore, current benchmarks for evaluating hallucination have predominantly relied on a single LLM to produce a hallucinated dataset. Notable examples include HaluEval (Li et al., 2023) and PHD (Yang et al., 2023b), which exclusively utilize ChatGPT, and FActScore (Min et al., 2023) and FACTOR (Muhlgay et al., 2023), which solely employ InstructGPT (Ouyang et al., 2022). In contrast, our methodology incorporates a suite of five distinct Chinese LLMs to generate hallucinated

content. These models include ChatGLM2-6B (Du et al., 2022), Baichuan2-13B (Yang et al., 2023a), Qwen-14B (Bai et al., 2023), InternLM-20B (InternLM, 2023), and Xinyu-7B. For additional information about the Xinyu series models, please refer to the Appendix B.1.

For each input news article, we concurrently generate five candidate continuations using 5 different LLMs without constraint. Overall, our approach engenders a more unconstrained and heterogeneous generation of hallucinations, mitigating the bias that may arise from the use of a single model or constrained prompting.

2.3 Hallucination Ranking

Given the unconstrained nature of our paradigm, the task of discerning whether the generated content is indeed hallucinated presents a significant challenge. Upon generating the continuations, an exclusive dependence on human annotation would incur substantial costs, whereas a purely machine-based approach, such as utilizing GPT4, could potentially yield less accurate results.

To navigate these complexities, we have adopted a two-stage annotation. This approach begins with an initial stage of hallucination ranking (Section 2.3), designed to sort the generated content based on the likelihood of hallucination. The ranking is then followed by the second stage of automatic labeling and human rechecking (Section 2.4).

Algorithm 1 Hallucination ranking

Require: $candidate : list[str]$
Ensure: $final : str$
  $candidate.sort(descend, by = fluency)$
  $picked \leftarrow candidate[: 3]$ ▷ More fluency

  $picked.sort(descend, by = likelihood)$
  $final \leftarrow picked[0]$ ▷ More Hallucination

Hallucination ranking is a crucial step in selecting the most appropriate continuation from a set of candidate continuations generated by LLMs. We employ a ranking process detailed in Algorithm 1. This process relies on two critical metrics: $fluency$, ensuring that the continuation does not become too nonsensical, and $likelihood$, which stands for the likelihood of hallucination occurrence, ensuring that the continuation includes a detectable level of hallucination. By employing such a ranking, it is guaranteed that, in the worst-case scenario, the $final$ candidate ranks at least

third in fluency and third in the likelihood of hallucination occurrence, achieving a balanced level. The two metrics are computed as follows.

Fluency This refers to the coherence and readability of the text. A fluent text should read smoothly, be grammatically correct, and make logical sense in the context of the continuation. To assess fluency, a reward model developed by the Org2 is employed, trained to score text quality based on fluency.

Likelihood of Hallucination Occurrence This dimension evaluates the extent to which the continuation may contain hallucinated content. To estimate the probability, we evaluate the lexical correlation between the generated continuation and the reference information. The lower the correlation, the more likely hallucinations are to occur. Despite existing metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), we believe that these rule-based methods may not effectively discover hallucinations. Therefore, we propose the keyword precision (kwPrec) metric.

This metric uses an LLM (e.g., GPT3.5-Turbo) to extract keywords from the continuation and determine whether these keywords have a match in the reference information. The ratio of all matches to the total keywords is then calculated. Since LLMs often extract appropriate keywords more effectively, kwPrec focuses more on factual relevance rather than expressional relevance. Fig. 3 illustrates the tokens segmented by our method compared to those obtained by BLEU-4 and ROUGE-L.

Figure 3: Tokenization results for BLEU-4, ROUGE-L, and kwPrec, using news$_{id=num\_000432}$ as an example. The meaning of the above sentence: Jiangsu is one of the most developed provinces in China for green food production.

2.4 Automatic Labeling and Human Rechecking

Through hallucination ranking, we can identify continuations that are both articulately expressed and likely to contain hallucinations. To detect continuations with confirmed hallucinations, we propose an annotation scheme that utilizes keywords, which

includes automatic labeling and subsequent human verification, as shown in Fig. 4.
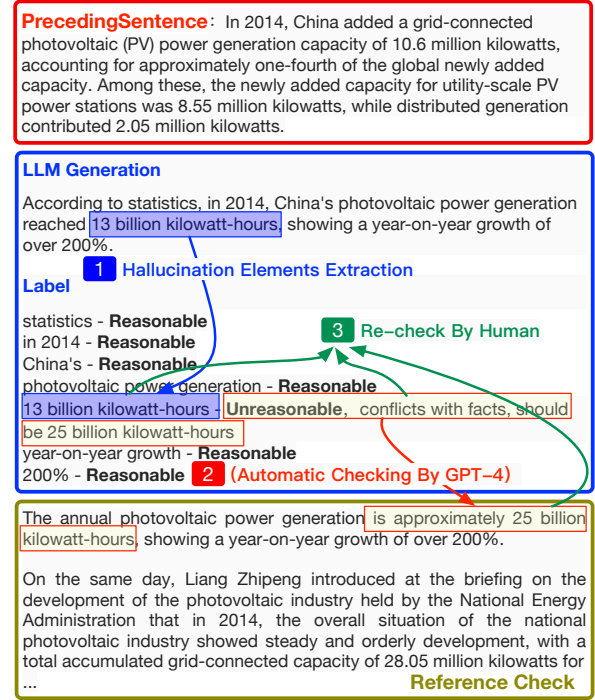

Figure 4: Labeling and rechecking. (In Chinese: Fig. 12)

Automatic labeling We utilize the keywords identified by GPT3.5-Turbo from the candidate continuations, similarly to the process used in the computation of kwPrec previously. These keywords act as the focal points for subsequent verification. Thereafter, we employ GPT4-0613 (OpenAI, 2023) to perform annotation on these keywords. It evaluates the validity of the keywords in the continuations by conducting a cross-reference with the provided original news and provides explanations for any detected unreasonable keywords.

Human rechecking We undertake a manual, one-to-one verification process by analyzing the annotated results and explanations provided by GPT4-0613 against the original news. This step ensures the accuracy of the machine-generated annotations. In the end, instances verified as accurate by annotators comprise the final UHGEval dataset. For details on manual annotation, please refer to Appendix A.1.

2.5 Dataset Statistics

Starting with 17,714 candidate hallucinated continuations, we curated a dataset of 5,141 hallucinated continuations, as detailed in the basic statistics in Table 2. For further analysis of statistics

and an example of the dataset, please refer to the Appendix A.2 and the Appendix A.3, respectively.

	DOC	KNO	NUM	GEN
#news	1242	320	2431	1148
avg. #hallu. kw.	2.15	1.99	2.54	2.12
avg. #kw.	8.43	8.09	8.07	8.17
#hallu. kw. / #kw.	25.47%	24.61%	31.44%	26.00%
avg. len. contn.	46.77	48.36	44.47	45.97
avg. len. begin.	102.15	102.66	103.20	102.86
avg. len. refer.	634.17	618.90	624.47	632.47

Table 2: Dataset basic statistics. # denotes quantity, avg. denotes average, len. denotes length, contn. denotes hallucinated continuations, begin. denotes news beginnings, and refer. denotes reference information.

3 Experiments

3.1 Models

Given that our dataset is tailored for the Chinese language generation domain, we selected eight widely used Chinese LLMs and three models from OpenAI. These LLMs are from eight base models: Aquila2 Base (BAAI, 2023), Baichuan2 Base (Yang et al., 2023a), GLM Base (Du et al., 2022), GPT Base[2], InternLM Base (InternLM, 2023), Qwen Base (Bai et al., 2023), BLOOMZ Base (Muennighoff et al., 2023), and LLaMA2 Base (Touvron et al., 2023). Refer to the Appendix B.1 for a detailed overview of the LLMs used in the experiments.

3.2 Evaluation Forms

In this study, we conducted a detailed analysis of evaluation methods across three dimensions: form, metric, and granularity. A more comprehensive report can be found in the Appendix B.2. Here, we introduce the three forms of evaluation.

Firstly, there is the discriminative evaluation, which involves having the model determine whether a continuation contains hallucinations. Secondly, similar to discriminative evaluation, selective evaluation allows LLMs to choose the continuation without hallucinations from options with and without such content. Lastly, we have generative evaluation. Specifically, the LLM under evaluation is provided with a beginning text and is then tasked with generating a continuation. Subsequently, various reference-based techniques are employed to assess whether the generated continuation includes hallucinations.

---

[2] https://openai.com

3.3 Evaluation Framework

To accommodate different forms of evaluation methods, we have developed a data-secure, easy-to-extend, and easy-to-use evaluation framework, as illustrated in Fig. 5. Refer to Appendix B.3 for a more detailed understanding of the various layers of the framework.
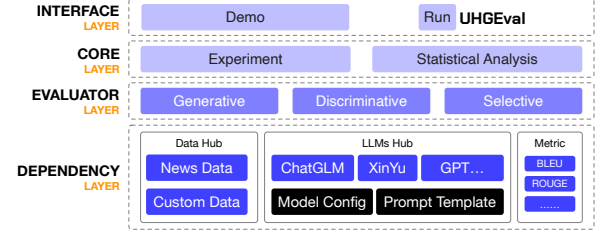

Figure 5: Evaluation framework

UHGEval is both intuitive and secure for users, offering efficient usage while concurrently ensuring the integrity of experimental results through robust resistance to exceptions and support for resuming evaluations post unexpected interruptions. For developers and researchers, the modules within the Dependency and Evaluator layers are fully interchangeable, thereby affording considerable flexibility for expansion.

3.4 Experimental Setup

To establish a robust experimental framework, we have set up some configurations as follows.

Prompt engineering We apply the technique of "intent + instruction + 3-shot (explainable) prompting." Intent delineates the role, instruction outlines the task, and the prompt incorporates three examples to aid the few-shot learning (Zhao et al., 2023). Furthermore, political content in examples is prohibited to adhere to content policies from model service providers. Explainable prompting entails not merely acquiring results, but also eliciting the model's rationale behind its responses. Refer to Appendix E to view the complete prompt templates.

Example Balancing To guarantee the reliability of experimental outcomes for all LLMs, we meticulously balance examples in discriminative and also in selective evaluations. Specifically, the LLM under evaluation will encounter an equal number of examples with and without hallucinations.

Hyperparameter settings Managing parameters for heterogeneous LLMs is a multifaceted endeavor, as different LLMs feature unique interface designs,

5

and the same parameters can have varying implications across LLMs. Despite these challenges, we commit to the principle of "guaranteeing overall output determinism while allowing for slight randomness, and aiming for consistent parameter settings across models." Consequently, we set the temperature to 0.1, the top_p to 0.9, the top_k to 5, and the random seed to 22.

Metrics For discriminative and selective evaluation, accuracy serves as the metric. For generative evaluation, metrics consist of 4-gram BLEU (BLEU-4), the longest common subsequence-based ROUGE (ROUGE-L), kwPrec, and BERTScore.

3.5 Results and Analysis

Results are presented in Table 3 and Table 4.

Discriminative evaluation Initially, the GPT series models' performance is notably superior in discriminative evaluation, showcasing their formidable foundational capabilities in knowledge recall, utilization, and judgment. Moreover, a comparison of experimental outcomes at the keyword and sentence levels reveals that accuracy is generally superior at the keyword level. This could stem from the fact that the hallucinated continuations in our dataset exhibit sufficient fluency, aligning with the fluency distribution of LLM outputs. This can potentially confuse the evaluated LLM, complicating the judgment of the continuation's authenticity. Conversely, keywords bypass fluency concerns, rendering keyword-level evaluation more amenable to LLMs. This observation implies that detecting hallucinations could be more dependable at the keyword level compared to the sentence level.

Selective evaluation Firstly, GPT4-1106 clinches the top spot, reaffirming the formidable foundational capabilities of the GPT series models. Concurrently, Xinyu2-70B attains second place, excelling as a model trained on the Chinese news corpus. This achievement, to a degree, confirms the merit of domain-specific LLMs. Secondly, when comparing the outcomes of the selective evaluation with those of the discriminative evaluation at the sentence level, most LLMs exhibit improved accuracy. We think, furnishing LLMs with more contrasting information alleviates the demand for the model's fact recall, thus diminishing the challenge of selective evaluation. Therefore, we posit that selective evaluation is comparatively simpler for LLMs.

Generative evaluation Overall, InternLM-20B, Xinyu2-70B, and Aquila-34B have achieved commendable results, but the performance of Aquila-34B could be attributed to its comparatively shorter average generation length. Additionally, the GPT series exhibits subpar performance, possibly due to the insubstantial amount of Chinese data in its training corpus. After all, the Chinese data incorporated into GPT's training from the Common Crawl corpus comprises less than 5%[3].

Evaluations by Type We focus on selective evaluation results and perform a comprehensive breakdown analysis of these across the four types, as illustrated in Table 4. Initially, most LLMs demonstrate enhanced accuracy for knowledge-intensive and document-intensive news. This may be because the training datasets for LLMs typically include substantial human knowledge and official documentation of major historical events. Furthermore, the majority of LLMs show reduced accuracy in general and number-intensive news. General news often contains societal minutiae, which are not the focus of LLM training. Regarding number-intensive news, it poses a considerable challenge for LLMs, given that encoding identical numbers with varied historical meanings is complex. However, GPT4-1106 attains especially high scores in the demanding number-intensive news.

3.6 Further Discussion

Each of the three evaluation forms possesses distinct advantages and drawbacks. Discriminative evaluation is often the method of choice for a range of standard benchmarks (Li et al., 2023; Cheng et al., 2023). This approach is intuitive, and the construction of evaluation prompts is straightforward. Selective evaluation resembles discriminative evaluation but is marginally less demanding because it includes a reference option for contrast. In both discriminative and selective evaluations, certain models might be suspected of conjecturing answers from a few shots due to inadequate reasoning skills, which can undermine the reliability of the outcomes. Consequently, the use of explainable prompting becomes essential. Generative evaluation most closely mirrors real-world applications. However, the generated content is unrestricted, which poses challenges for even the most dependable reference-based evaluation techniques.

---

[3] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

	Discriminative-Keyword			Discriminative-Sentence		Selective	
	avg. acc.	avg. #kws	#valid	avg. acc.	#valid	acc.	#valid
Aquila-34B	53.62%	3.00	3719	49.86%	5009	54.29%	4319
Baichuan2-13B	51.63%	**3.128**	4478	46.88%	5047	50.23%	5130
Baichuan2-53B	52.13%	2.98	1656	50.81%	1478	54.67%	4443
ChatGLM2-6B	50.80%	3.10	4289	43.87%	5130	43.59%	5130
GPT3.5-Turbo	53.72%	3.08	4183	50.02%	5039	49.03%	5103
GPT4-0613	**70.04%**	3.07	4100	**57.42%**	5024	55.20%	5047
GPT4-1106	<u>69.48%</u>	3.10	4189	<u>57.38%</u>	4903	**60.35%**	4752
InternLM-20B	50.92%	3.10	4388	51.01%	5130	49.43%	5130
Qwen-14B	52.86%	<u>3.125</u>	4478	50.58%	5130	54.74%	5130
Xinyu-7B	49.58%	3.12	4451	48.66%	5014	50.58%	5130
Xinyu2-70B	52.94%	3.12	4482	55.04%	5128	<u>57.93%</u>	5129

	Generative					
	avg. bleu	avg. rouge	avg. kwPrec	avg. bert	avg. len.	#valid
Aquila-34B	11.80%	6.04%	**34.36%**	67.51%	43.76	5130
Baichuan2-13B	8.84%	6.96%	25.51%	65.69%	46.04	5113
Baichuan2-53B	10.06%	<u>7.55%</u>	26.45%	67.65%	49.40	3837
ChatGLM2-6B	9.17%	7.17%	24.53%	64.89%	46.27	5094
GPT3.5-Turbo	9.02%	6.30%	27.74%	66.39%	39.04	5084
GPT4-0613	10.74%	7.19%	28.47%	67.36%	44.41	5109
GPT4-1106	8.62%	6.86%	30.94%	67.38%	44.83	5121
InternLM-20B	**14.89%**	**7.96%**	31.10%	<u>67.92%</u>	**51.55**	5125
Qwen-14B	12.72%	6.54%	32.95%	66.96%	45.85	5125
Xinyu-7B	10.30%	6.52%	28.64%	67.32%	49.84	4978
Xinyu2-70B	<u>13.41%</u>	7.05%	<u>33.93%</u>	**68.97%**	<u>51.10</u>	5130

Table 3: Discriminative, selective, and generative evaluation results. #kws denotes the number of keywords and #valid denotes the number of valid evaluations. In the same column, optimal values are bolded, and suboptimal values are underlined.

	KNO	DOC	GEN	NUM
Aquila-34B	**59.55%**	<u>54.97%</u>	53.74%	53.52%
Baichuan2-13B	**53.75%**	<u>52.10%</u>	48.43%	49.67%
Baichuan2-53B	**57.70%**	<u>57.46%</u>	56.26%	52.58%
ChatGLM2-6B	40.94%	**45.56%**	<u>44.23%</u>	42.63%
GPT3.5-Turbo	**55.21%**	51.06%	47.63%	47.85%
GPT4-0613	**59.87%**	<u>55.99%</u>	51.93%	55.73%
GPT4-1106	**68.73%**	<u>60.19%</u>	54.77%	62.04%
InternLM-20B	**51.88%**	<u>50.65%</u>	49.56%	48.43%
Qwen-14B	**62.81%**	<u>57.35%</u>	53.15%	53.09%
Xinyu-7B	48.44%	**52.02%**	<u>50.87%</u>	50.00%
Xinyu2-70B	**63.13%**	<u>61.47%</u>	54.46%	57.07%

Table 4: Evaluation by different types. In the same row, optimal values are bolded, and suboptimal values are underlined.

Therefore, employing a combination of metrics simultaneously, including lexical evaluation based on token coverage and semantic evaluation based on textual similarity, is imperative.

The foundational capabilities required of LLMs can be arrayed on a spectrum from simple to complex: generative, selective, and discriminative evaluation. Generative evaluation entails the direct invocation of parameters for continuation, bypassing the need for an extensive grasp of instructions. Selective evaluation necessitates a degree of inferential reasoning but offers comparative choices, rendering the level of difficulty moderate. Conversely, discriminative evaluation demands the precise retrieval of facts, thereby increasing the challenge.

4 Related Work

This section outlines benchmark datasets, their characteristics, and evaluation methodologies. These benchmarks are summarized in Table 5. For more related works, you may refer to Appendix C.

4.1 Benchmark dataset construction

Dataset construction usually involves three steps. Firstly, real-world texts for hallucination generation are collected, and most benchmarks directly use existing datasets, such as Wiki (Muhlgay et al., 2023), Alpaca (Li et al., 2023), PubMed (Pal et al., 2023), etc. Secondly, hallucinations are generated usually by LLMs such as GPT3.5-Turbo, and most works use a constrained hallucination generation (CHG) paradigm. STSN (Varshney et al., 2023) and XSum Hallu (Maynez et al., 2020) are the only two benchmarks that use UHG as we do. Thirdly, it is not certain that the content generated by the LLMs actually contains hallucinations, and often

Benchmark	Generation Method: Base Dataset	Annotation	Metric	Granularity	Lang.
ChineseFactEval (Wang et al., 2023a)	Manual	Manual	Acc	Sentence	CN
CSK-PN (Chen et al., 2023)	Direct: Common KGs	No Need	Acc	Word	EN
FACTOR (Muhlgay et al., 2023)	CHG: Wiki, News	Auto	FACTOR Acc	Sentence	EN
FActScore (Min et al., 2023)	CHG: Wiki	No Need	FActScore by Human	Short Sentence	EN
FactualityPrompts (Lee et al., 2022)	Direct: Wiki	Auto	NE Error, Entailment	Document, Sentence	EN
HADES (Liu et al., 2022)	CHG: Wiki	Manual	Acc, G-Mean, BSS, AUC, etc.	Word	EN
HalluQA (Cheng et al., 2023)	CHG, Manual: TruthfulQA, Wiki	Manual, Auto	Non-hallucination Rate	Sentence	CN
HaLoCheck (Elaraby et al., 2023)	CHG	No Need	HaLoCheck, selfcheckGPT	Sentence	EN
HaluEval (Li et al., 2023)	CHG: Alpaca, HotpotQA, etc.	Manual, Auto	Acc	Document	EN
HILT (Rawte et al., 2023)	CHG: NYT, Politifact	Manual	HVI	Word	EN
KoLA-KC (Yu et al., 2024)	Direct: Wiki, evolving dataset	Auto	BLEU, ROUGE	Document	EN
Med-HALT (Pal et al., 2023)	Direct: MedMCQA, PubMed, etc.	No Need	Acc, Pointwise Score	All	EN
PHD (Yang et al., 2023b)	CHG: Wiki	Manual	F1, Acc, Prec, Reca	Document	EN
SelfAware (Yin et al., 2023)	CHG: Quora, HowStuffWorks	Manual	F1, Acc	Sentence	EN
STSN (Varshney et al., 2023)	UHG	Manual	Acc, Prec, Reca	Sentence, Concept	EN
TruthfulQA (Lin et al., 2022)	Manual	Manual	Acc by Human or GPT-judge	Sentence	EN
UHGEval (Ours)	UHG: News	Auto, Manual	Acc, kwPrec, BERTScore, etc.	Sentence, Keyword	CN
XSum Hallu (Maynez et al., 2020)	UHG: XSum	Manual	ROUGE, BERTScore, Acc, etc.	Word, Document	EN

Table 5: Hallucination evaluation benchmarks sorted by name. In the Generation Method column, CHG refers to constrained hallucination generation, UHG refers to unconstrained hallucination generation, Manual indicates manually constructed, and Direct implies utilizing the base dataset without the need for generation. In the Annotation column, Auto denotes automatic machine annotation. In the Metric column, Acc, Prec, and Reca respectively indicate accuracy, precision, and recall. In the Lang. column, CN and EN respectively stand for Chinese and English.

requires annotation, which is mostly done by human involvement. There are also works using automatic machine labeling (Muhlgay et al., 2023; Lee et al., 2022; Cheng et al., 2023). These are the basic methods for constructing datasets, but there are also some other paradigms, such as constructing the dataset purely using manual labor, e.g. ChineseFactEval (Wang et al., 2023a), HADES (Liu et al., 2022), TruthfulQA (Lin et al., 2022), etc.

4.2 Benchmark dataset characteristics

Regarding the granularity of hallucinations labeled in the datasets, most studies assess hallucinations at the sentence and document levels, while a few examine them at the word (or keyword, concept) level. Concerning language, most evaluation datasets are in English. To our knowledge, the only two Chinese benchmarks, ChineseFactEval (Wang et al., 2023a) and HalluQA (Cheng et al., 2023) contain only 125 and 450 questions, respectively. Given the notably limited size of these datasets, our work significantly enhances the pool of data available for Chinese hallucination evaluation.

4.3 Evaluation schemes

Currently, building automatic metrics for evaluation is still dominant, and a small proportion of works use human evaluation (Min et al., 2023; Lin et al., 2022; Maynez et al., 2020). In terms of specific evaluation metrics, most works adopt common classification metrics, e.g., F1, accuracy, precision, and recall. some other works construct

their calculation methods, e.g., FACTOR (Muhlgay et al., 2023), FActScore (Min et al., 2023), HaLoCheck (Elaraby et al., 2023), etc. However, the above metrics are rule-based and can only evaluate the ability of LLMs to classify hallucinations, but not the ability of LLMs to generate content without hallucinations. Thus, some benchmarks explore even further in generative evaluation. For example, KoLA (Yu et al., 2024) evaluates knowledge creation (KC) using BLEU and ROUGE, and TruthfulQA (Lin et al., 2022) evaluates hallucinations using a specially trained classifier, GPT-judge.

5 Conclusion

LLMs are rapidly evolving, heralding a new era of potential applications within the realm of professional content generation. The progression of LLMs in this domain necessitates the establishment of robust benchmarks to steer their development effectively. In this work, we introduce a novel hallucination benchmark dataset using an unconstrained fashion, encompassing more than 5,000 instances annotated at the keyword level. Additionally, we propose a secure, scalable, and user-friendly evaluation framework to facilitate comprehensive assessments. Through meticulous experimentation on eleven prominent LLMs, our study has unearthed a series of enlightening findings. Looking ahead, our research endeavors will persist in exploring the intricacies of hallucination phenomena within professional content generation, aiming to further understand and enhance LLM capabilities.

Limitations

Dataset   Firstly, although we have utilized hallucination ranking, automatic labeling, human rechecking, and various other techniques mentioned in Appendix A to ensure the quality of data annotation, with over 5,000 data entries, there is still a possibility of labeling errors. We have mobilized the power of the open-source community to collectively improve our dataset. Additionally, the dataset creation process is flexible, allowing for dataset expansion into English and broader domains, such as mathematical reasoning and programming codes.

Framework   Although our framework simplifies the integration of LLMs through APIs or vLLM[4], users seeking to utilize custom or diverse Hugging-Face models may face initial hurdles. We need to further enhance the usability of our framework.

Constrained v.s. Unconstrained   We have determined that constrained generation cannot fully reflect real-world applications, but empirical analysis is required to prove this point. This may involve constructing a text classifier to determine the type of hallucination, followed by comparing the distribution of hallucinations in our dataset with those in other benchmark datasets to observe any significant deviations. We leave this for future work.

References

BAAI. 2023. Aquila2. https://github.com/FlagAI-Open/Aquila2.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yupeng Chang, Xu Wang, Jindong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Jiangjie Chen, Wei Shi, Ziquan Fu, et al. 2023. Say what you mean! large language models speak too positively about negative commonsense knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9890–9908, Toronto, Canada. Association for Computational Linguistics.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Marie-Catherine de Marneffe and Joakim Nivre. 2019. Dependency grammar. *Annual Review of Linguistics*, 5(1):197–218.

Zhengxiao Du, Yujie Qian, Xiao Liu, et al. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, et al. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

InternLM. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Nayeon Lee, Wei Ping, Peng Xu, et al. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, et al. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Tianyu Liu, Yizhe Zhang, Chris Brockett, et al. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, et al. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, et al. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

---

[4]https://github.com/vllm-project/vllm

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, et al. 2023. Crosslingual generalization through multitask finetuning. In *Annual Meeting of the Association for Computational Linguistics*.

Dor Muhlgay, Ori Ram, Inbal Magar, et al. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, et al. 2023. 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*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, et al. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Binjie Wang, Ethan Chern, and Pengfei Liu. 2023a. Chinesefacteval: A factuality benchmark for chinese llms. https://GAIR-NLP.github.io/ChineseFactEval.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, et al. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, et al. 2024. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Aiyuan Yang, Bin Xiao, Bingning Wang, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023b. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, et al. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, et al. 2024. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, et al. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Yafu Li, Leyang Cui, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

10

A The UHGEval Dataset

A.1 Dive into Human Rechecking Process

Least Hallucination Principle The keyword-based labeling scheme has inherent limitations. Languages exhibit a dependency structure (de Marneffe and Nivre, 2019). For instance, in the phrase "The rainbow is black," the words "rainbow" and "black" exhibit interdependence. One could contend that "black" is incorrect, while another could maintain that "rainbow" is erroneous, given that "night" is typically described as black. To address the challenges stemming from language dependency structures, we have adopted the *Least Hallucination Principle*. If a set of words can be selected, and their replacement with contextually appropriate words yields a reasonable sentence, then such a set of words is designated as a hallucinated word group. The words selected for annotation must meet the condition of comprising the minimal number of words in the group, as illustrated in Equation 1. In the equation, $\mathbf{W}$ is the set of keywords in a sentence, $\mathbf{w}$ is the hallucinated word group, $\text{correct}(\cdot)$ is the correction function that modifies hallucinated words to non-hallucinated words, and $\text{hallucinated}(\cdot)$ assesses whether a sentence composed of keywords hallucinated.

$$
\begin{aligned}
\min \quad & |\mathbf{w}| \\
\text{s.t.} \quad & \mathbf{w} \subset \mathbf{W} \\
& \mathbf{w}' = \text{correct}(\mathbf{w}) \\
& \text{false} = \text{hallucinated}(\mathbf{W} - \mathbf{w} + \mathbf{w}')
\end{aligned} \tag{1}
$$

By this principle, within the phrase "Journey to the West is an American novel and one of the Four Great Classics," the word "American" would be marked for annotation, as altering this single keyword to "Chinese" dispels the hallucination throughout the sentence.

Engagement of Annotators Additionally, we acknowledge that hallucination annotation may become somewhat tedious. Consequently, annotators are integrated throughout the entire process, participating in discussions instead of solely evaluating the accuracy of machine annotations. This approach also yields benefits for our work. For example, an annotator with a journalism background offered valuable professional insights into pinpointing news-related hallucinations, emphasizing that fact increment is a critical aspect of news writing.

A.2 Analysis of Final Dataset

We developed a conversion rate chart to depict the transition from candidate hallucinations to the final dataset, as depicted in Fig. 6. The conversion rate can be interpreted as the likelihood of hallucinations occurring across various categories. Our observations indicate a higher likelihood of hallucinations in number-intensive and general news, whereas this likelihood is reduced in knowledge-intensive and document-intensive news.
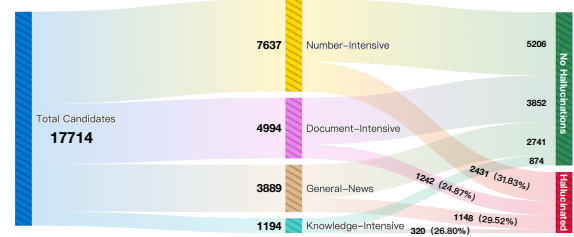

Figure 6: Conversion rates from candidates to hallucinations.

By analyzing the hallucinated word cloud depicted in Fig. 7 for each news category, we can draw the following conclusions: Number-intensive news often includes numeric values that are challenging to remember, like 0.09% and 6:3, which pose difficulties for both LLMs and humans. General news encompasses a diverse vocabulary, featuring terms such as "social media" and "friendship," which are often deemed less critical and thus challenging to incorporate into the training corpora of many LLMs. Knowledge-intensive news frequently features terms such as "according to incomplete statistics" and "key technology," which are prevalent in technical literature. However, LLMs may not always use these terms appropriately. Document-intensive news often contains terms associated with official statements, such as "representation," "president," and "spokesperson." This suggests that LLMs are susceptible to introducing unauthorized alterations to the content of documents.
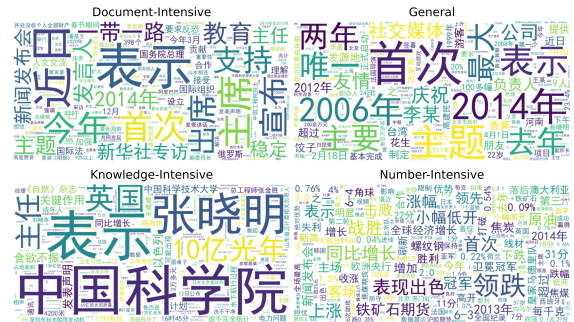

Figure 7: Hallucinated keywords in different types of news

A.3 An example from the UHGEval dataset

```
{
  "id": "num_000432",
  "headLine": "(Society) Jiangsu's First Selection of the Top 100 Green Foods Most Loved by Consumers",
  "broadcastDate": "2015-02-11 19:46:49",
  "type": "num",
  "newsBeginning": "Xinhua News Agency, Nanjing, February 11 (Reporter Li Xiang) 'Food is the paramount necessity of the
people, and safety is the top priority of food.' On February 11, Jiangsu announced the results of the 'First Consumers'
Favorite Green Foods' selection, with Lao Shan honey and 100 other foods receiving the title of 'Consumers' Favorite Green
Food'.",
  "hallucinatedContinuation": "Jiangsu is one of the most developed provinces in the country in terms of green food
production.",
  "generatedBy": "InternLM_20B_Chat",
  "appearedKeywords": ["Jiangsu", "national", "green food production"],
  "allKeywords": {
    "Jiangsu": "reasonable",
    "national": "reasonable",
    "green food production": "reasonable",
    "developed": "unreasonable, there is no factual evidence to prove that Jiangsu is one of the provinces with developed
green food production in the country, but what can be confirmed is that Jiangsu has active practices and promotions in green
food production",
    "province": "reasonable",
    "one of": "unreasonable, there is no specific factual evidence to show that Jiangsu is one of the developed provinces in
terms of green food production in the country"
  },
  "realContinuation": "61 award-winning production enterprises jointly signed a integrity pact, jointly building a green food
integrity alliance.",
  "newsRemainder": "61 award-winning production enterprises jointly signed an integrity pact, jointly building a green food
integrity alliance. This is an important measure for Jiangsu to ensure food safety and promote green food production.\n..."
}
```

Figure 8: An example from the UHGEval dataset. (In English)

```
{
  "id": "num_000432",
  "headLine": "（社会）江苏首次评选消费者最喜爱的百种绿色食品",
  "broadcastDate": "2015-02-11 19:46:49",
  "type": "num",
  "newsBeginning": "　新华社南京2月11日电（记者李响）"民以食为天，食以安为先"。江苏11日发布"首届消费者最喜爱的绿色食品"评选结果,
老山蜂蜜等100种食品获得消费者"最喜爱的绿色食品"称号。",
  "hallucinatedContinuation": "江苏是全国绿色食品生产最发达的省份之一。",
  "generatedBy": "InternLM_20B_Chat",
  "appearedKeywords": ["江苏", "全国", "绿色食品生产"],
  "allKeywords": {
    "江苏": "合理",
    "全国": "合理",
    "绿色食品生产": "合理",
    "发达": "不合理，没有事实证明江苏是全国绿色食品生产发达的省份，但可以确定的是，江苏在绿色食品生产上有积极的实践和推动",
    "省份": "合理",
    "之一": "不合理，没有具体的事实证据表明江苏是全国绿色食品生产发达的省份之一"
  },
  "realContinuation": "61家获奖生产企业共同签署诚信公约，共建绿色食品诚信联盟。",
  "newsRemainder": "61家获奖生产企业共同签署诚信公约，共建绿色食品诚信联盟。这是江苏保障食品安全、推动绿色食品生产的重要举措。
\n　此次评选由江苏省绿色食品协会等部门主办，并得到江苏省农委、省委农工办、省工商局、省地税局、省信用办、省消协等单位大力支持。评
选历时4个多月，经企业报名、组委会初筛、消费者投票等层层选拔，最终出炉的百强食品榜单由消费者亲自票选得出，网络、短信、报纸及现场投
票共310多万份票数，充分说明了评选结果的含金量。\n　食品安全一直是社会关注的热点。此次评选过程中，组委会工作人员走街头、进超市，
邀请媒体、消费者、专家深入产地开展绿色食品基地行，除了超市选购外，还搭建"诚信购微信商城""中国移动MO生活绿色有机馆"等线上销售平台，
开创江苏绿色食品"评展销"结合新局面……"
}
```

Figure 9: An example from the UHGEval dataset. (In Chinese)

B Experiments

B.1 LLMs Employed in This Research

All LLMs used in this study are detailed in Table 6.

Model	#Para.	Publisher	Date
GPT3.5-Turbo	175B*	OpenAI	2023.03*
GPT4-0613	NaN	OpenAI	2023.06
ChatGLM2	6B	Tsinghua	2023.06
Xinyu	7B	Org2&Org3	2023.06
InternLM	20B	ShLab	2023.07
Baichuan2	13B	Baichuan Inc.	2023.09
Baichuan2	53B	Baichuan Inc.	2023.09
Qwen	14B	Alibaba	2023.09
Aquila2	34B	BAAI	2023.10
Xinyu2	70B	Org2&Org3	2023.10
GPT4-1106	NaN	OpenAI	2023.11

Table 6: LLMs sorted by release date. All LLMs are chat models. Asterisk (*) denotes estimated value, NaN denotes no public data available, and 175B denotes 175billion.

GPT represents a series of LLMs developed by OpenAI (OpenAI, 2023). In this study, GPT3.5-Turbo, GPT4-0613, and GPT4-1106 are utilized. GLM constitutes a pre-training framework proposed by Tsinghua University (Du et al., 2022), and the ChatGLM2-6B chat model is employed. InternLM serves as an open-source, lightweight training framework, with its development team releasing a spectrum of models utilizing this framework (InternLM, 2023); the InternLM-20B open-source chat model is utilized in the present work. Baichuan2 comprises a series of expansive, multilingual base language models (Yang et al., 2023a), with both the open-source Baichuan2-7B chat model and the closed-source Baichuan2-53B chat model being employed in this investigation. Qwen encompasses a language model series characterized by distinct models with varying parameter counts (Bai et al., 2023), and the Qwen-14B open-source chat model is utilized in the current study. Aquila2 represents a language model series devised by BAAI, noted for surpassing comparable models in terms of performance (BAAI, 2023), and the Aquila2-34B chat model is employed in this research.

Besides, the Xinyu series models are the results of a collaborative research and development effort between the Org2 and the Org3. Xinyu-7B is an augmented large-scale language model derived from the foundational model, BloomZ-7B (Muennighoff et al., 2023) through continued pre-training, news-specific fine-tuning, and alignment optimiza-

tion. And, Xinyu2-70B is developed based on the open-source LLaMA2-70B (Touvron et al., 2023) framework, incorporating expansions to the Chinese lexicon, ongoing pre-training, and news-specific fine-tuning, thereby endowing it with a robust foundational capability in the news domain.

B.2 Evaluation Method

The evaluation of hallucinations can be decomposed into three principal dimensions: form, metric, and granularity. Form concerns how the model interacts with the evaluation dataset; metric refers to the precise computational approach utilized for performance assessment; and granularity signifies the depth of detail considered in the evaluation of hallucinations.

Form This encompasses human evaluation, discriminative evaluation, selective evaluation, and generative evaluation, among others. Human evaluation entails the direct application of human judgment to determine if the model's output contains hallucinations, representing a critical evaluation form (Chang et al., 2024). However, the drawbacks of this approach are evident: evaluating too many data points is tantamount to annotating a new dataset, with the associated time and financial expenditures proving prohibitive.

Discriminative evaluation enables LLMs to respond with binary answers of "yes" or "no" (Li et al., 2023; Cheng et al., 2023). Specifically, this evaluation modality involves presenting the LLM under scrutiny with an initial text followed by a continuation that may or may not include hallucinations. The LLM is tasked with producing a verdict as to the presence of hallucinations. Owing to the efficacy of few-shot prompting, this evaluation paradigm is relatively uncomplicated for LLMs to administer, as it facilitates the elicitation of the requisite responses. However, this method depends solely on the LLM's ability to draw upon the knowledge encoded within its parameters, necessitating the concurrent application of knowledge and reasoning, and thus requiring a robust foundational model capacity.

Selective evaluation allows LLMs to tackle multiple-choice questions by choosing between option A or B, as exemplified by PandaLM (Wang et al., 2024). Specifically, in selective evaluation, the LLM under evaluation is presented with an initial text followed by two continuations: one that includes hallucinations and another that does not.

The LLM's objective is to identify which of the two is hallucinated. This assessment method offers the LLM more contextual information than discriminative evaluation, thereby alleviating the burden of fact-checking and lessening the dependence on retrieving facts from its parameters. Consequently, this reduces the level of difficulty for the LLM.

However, both discriminative and selective evaluations encounter a substantial challenge. They are predicated on the assumption that "LLMs's capacity to produce reliable text is contingent upon their discernment between hallucinated and non-hallucinated content." These methods do not simulate the evaluation of the model's output for hallucinations. Consequently, generative evaluation is crucial as it directly evaluates the presence of hallucinations in the text generated by the LLM under evaluation. However, the challenge arises from the fact that it is not feasible to automatically and accurately ascertain if the newly generated text is hallucinated; if it were, annotated datasets would be redundant. In scenarios of unrestrained text generation, this issue becomes increasingly complex. This complexity stems from the fact that text generated without constraints may introduce a multitude of entities and facts absent in the reference material, complicating the verification of their accuracy. Despite these hurdles, generative evaluation continues to be a predominant strategy in Natural Language Generation (NLG) tasks (Novikova et al., 2017).

Metric Metrics include classification metrics such as accuracy, precision, recall, and others, which are applicable to human evaluation, discriminative evaluation, and selective evaluation. Generative evaluation, on the other hand, encompasses both lexical and semantic metrics. Lexical metrics evaluate the extent of token overlap between the generated text and the reference information, including metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and the newly proposed metric by us, kwPrec. Semantic metrics gauge the similarity in meaning between sentences, with examples including BERTScore (Zhang et al., 2020), GPT-judge (Lin et al., 2022), and GPTScore (Fu et al., 2023), among others.

Granularity Evaluations can be conducted at both the sentence and keyword levels. Owing to our annotation methodology, our dataset is marked at the keyword level to signify instances of hallucinations. This approach affords a broader spectrum of possibilities for configuring the evaluation task, enabling the evaluated model to address the presence of hallucinations at either the keyword level, the sentence level, or even the document level.

B.3 UHGEval Framework in Detail

The framework comprises four ascending layers: the dependency layer, the evaluator layer, the core layer, and the interface layer.

The dependency layer defines the essential foundational components needed for the evaluation framework, including datasets, LLM hubs, and various metrics. Importantly, each component is designed for extensibility: datasets can be replaced with custom ones, LLMs can be integrated via APIs or platforms like Hugging Face[5], and metrics can be customized to fit specific needs.

The evaluator layer, constituting the second layer, centers on an abstract class, Evaluator, and its various implementations. Within this layer, three distinct types are implemented: GenerativeEvaluator, DiscriminativeEvaluator, and SelectiveEvaluator. Users may also engineer custom evaluators, contingent upon adherence to the interface specifications of the abstract class, necessitating merely three function overloads.

The core layer, representing the third stratum, comprises two principal modules: experiment.py and analyst.py. The former facilitates experiments involving multiple LLMs, evaluators, and processes, whereas the latter is tasked with the statistical analysis of experimental outcomes.

The interface layer, serving as the final layer, orchestrates the user's interaction with $\mathbb{UHGE}$val. To streamline the initiation process, a succinct 20-line demonstration is offered, alongside a run.py script for launching experiments through the command line.

---

[5] https://huggingface.co/models

C More Related Works

C.1 Large Language Models

Language models are pivotal in computer science, evolving from statistical language models to neural language models, to pre-trained language models (PLMs), and now to the current generation of LLMs. The advent of models such as ChatGPT has seen contemporary LLMs exhibit new capabilities in handling complex tasks. These models can manage few-shot tasks via in-context learning and tackle mixed tasks by following instructions (Zhao et al., 2023).

LLMs can be classified according to two dimensions. The first dimension concerns the openness of the model weights. For example, open-source models include Meta's LLaMA (Touvron et al., 2023), Tsinghua University's GLM (Du et al., 2022), and Alibaba's Qwen (Bai et al., 2023), while closed-source models feature OpenAI's GPT (OpenAI, 2023), Baidu's ERNIE Bot (Sun et al., 2021), and Anthropic's Claude[6], among others. The second dimension differentiates between the use of a PLM or a supervised fine-tuned (SFT) model for specific inferences. A PLM is a language model trained on extensive unlabeled textual data to discern underlying patterns, structures, and semantic knowledge within the corpus. Conversely, an SFT model involves further training a PLM with labeled datasets tailored to a specific task, to improve performance in that area. Many open-source models, including LLaMA, GLM, and Qwen, have made their PLM weights publicly available. For SFT models, users can access the chat variants of open-source models or the API services provided by closed-source models. In our research, we focus primarily on evaluating closed-source GPT series models and open-source Chinese chat models.

C.2 Hallucinations in LLM

Despite remarkable advancements in LLMs, they continue to encounter challenges, with hallucination being one of the most notable. Hallucination in language models refers to generating content that strays from factual accuracy, leading to unreliable outputs. Hallucinations occur when the generated content is not aligned with user input, deviates from the model's previous outputs, or is at odds with established real-world knowledge (Zhang et al., 2023).

Specific examples include inaccuracies in age, currency, scores, and other numerical values; citing fictional statements; inventing non-existent characters; and muddling timelines by merging events from different periods (Rawte et al., 2023).

Regarding the causes of hallucinations, several factors can be responsible (Zhang et al., 2023). One contributing factor is the use of inaccurate or incomplete training data. During training, LLMs fine-tune their parameters with vast quantities of text data. However, this data may be flawed, harboring errors, inaccuracies, or gaps in information. Another factor involves inconsistencies in contextual information. While LLMs typically consider previously generated context when producing content, challenges in managing long-term dependencies or understanding complex contexts can result in inconsistencies. Additionally, hallucinations can arise from lacking or erroneous world knowledge. Although LLMs gain considerable world knowledge via training data, they may be deficient in specific domain knowledge or misinterpret certain facts, leading to hallucinations. Furthermore, model limitations, including generation strategies and alignment methods, can also play a role in hallucinations during content creation.

---

[6]https://www.anthropic.com/index/introducing-claude

D Figures in Chinese

Organization hallucinated
id=doc_003726

韩国产业通商资源部 韩国航天工业公司 表示，韩国政府仍将继续推进这一出口计划。

Statistics hallucinated
id=num_000691

节日期间，全国公路客运量达到 2.5 3.1 亿人次，同比增长 8.9% 3.2%。

Knowledge hallucinated
id=kno_000410

镰状细胞病是一种严重的遗传性血液疾病，易引起 疼痛性动脉硬化 贫血，栓塞等。

Timeline hallucinated
id=gen_005626

国家艺术基金于 2012 2013 年正式成立，其宗旨是为了支持全国范围内的艺术创作和艺术人才培养。

Figure 10: Hallucinations from 𝕌ℍ 𝔾Eval. Using the IDs, you can locate the original news articles. (In English: Fig. 1)
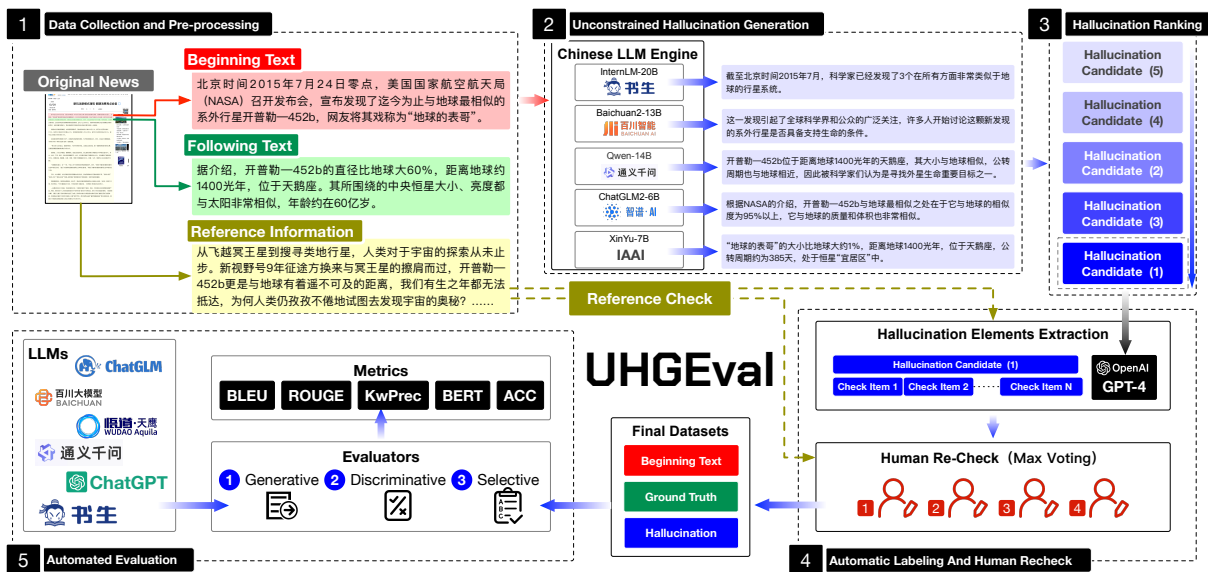

Figure 11: The process of creating 𝕌ℍ 𝔾Eval. Steps 1 to 4 regarding the creation of the benchmark dataset are explained in Section 2; Step 5, concerning the evaluation framework, is detailed in Section 3. (In English: Fig. 2)

Figure 12: Labeling and rechecking. (In English: Fig. 4)

E Prompt Templates

In these templates, the orange text represents intent and instruction, the green text represents demonstrations, and the black text represents specific questions.

You are a news worker for Xinhua News Agency. You need to determine whether a key term in the news continuation is realistic. Please directly state whether it is realistic or not, and provide the reason.

The beginning of the news: """《European Basketball League's Second Phase Group Stage: Alba Berlin Faces Grim Qualification Prospects》\n2015-03-05 06:01:47\n　　Xinhua, Berlin, March 4 (Reporter Wang Dong) — The ninth round of the 2014-2015 season European Basketball League's second phase group stage kicks off tonight. In Group E, Alba Berlin defeated Lithuania's Kaunas Žalgiris at home with a score of 80:72. After nine rounds, with four wins and five losses, they are firmly in fifth place in the group, yet their qualification prospects still look grim."""
Continuation: """Alba Berlin's head coach, John Cummings, stated that although their current ranking is not ideal, they will still give their all and strive for victory in every match."""
Realistic or not: The continuation's """John Cummings""" is not realistic because such a person does not exist.

Demonstration 2; Demonstration 3

The beginning of the news: """《{headLine}》\n{broadcastDate}\n{newsBeginning}"""
Continuation: """{continuation}"""
Realistic or not: The continuation's """{keyword}"""

你是一名新华社新闻工作者。你需要判断新闻续写中的某个关键词是否符合现实。请直接输出符合现实或不符合现实，同时给出原因。

新闻的开头："""《（体育）欧洲篮球联赛第二阶段小组赛阿尔巴出线形势不乐观》\n2015-03-05 06:01:47\n　　新华社柏林3月4日电（记者王东）2014-2015赛季欧洲篮球联赛第二阶段小组赛第九轮今晚开始。在E组的比赛中，柏林阿尔巴队主场以80：72力挫立陶宛考纳斯萨拉基利斯队，9轮过后4胜5负，稳居小组第五名，但是出线形势仍不乐观。"""
续写："""柏林阿尔巴队主教练约翰·卡明斯表示，虽然目前的排名并不理想，但他们仍然会全力以赴，争取每个比赛的胜利。"""
是否符合现实：续写中的"""约翰·卡明斯"""不符合现实，因为不存在此人。

示例2；示例3

新闻的开头："""《{headLine}》\n{broadcastDate}\n{newsBeginning}"""
续写："""{continuation}"""
是否符合现实：续写中的"""{keyword}"""

Figure 13: Prompt template for discriminative evaluation (keyword level)

You are a news worker for Xinhua News Agency. You need to assess whether the news continuation is realistic.

The beginning of the news: """《European Basketball League's Second Phase Group Stage: Alba Berlin Faces Grim Qualification Prospects》\n2015-03-05 06:01:47\n   Xinhua, Berlin, March 4 (Reporter Wang Dong) — The ninth round of the 2014-2015 season European Basketball League's second phase group stage kicks off tonight. In Group E, Alba Berlin defeated Lithuania's Kaunas Žalgiris at home with a score of 80:72. After nine rounds, with four wins and five losses, they are firmly in fifth place in the group, yet their qualification prospects still look grim."""
Continuation: """Alba Berlin's head coach, John Cummings, stated that although their current ranking is not ideal, they will still give their all and strive for victory in every match."""
Judgment: The continuation is not realistic because there is no person named "John Cummings".

Demonstration 2; Demonstration 3

The beginning of the news: """《{headLine}》\n{broadcastDate}\n{newsBeginning}"""
Continuation: """{continuation}"""
Judgment:

你是一名新华社新闻工作者。你需要判断新闻续写是否符合现实。

新闻的开头："""《（体育）欧洲篮球联赛第二阶段小组赛阿尔巴出线形势不乐观》\n2015-03-05 06:01:47\n   新华社柏林 3 月 4 日电（记者王东）2014-2015 赛季欧洲篮球联赛第二阶段小组赛第九轮今晚开始。在 E 组的比赛中，柏林阿尔巴队主场以 80：72 力挫立陶宛考纳斯萨拉基利斯队，9 轮过后 4 胜 5 负，稳居小组第五名，但是出线形势仍不乐观。"""
续写："""柏林阿尔巴队主教练约翰·卡明斯表示，虽然目前的排名并不理想，但他们仍然会全力以赴，争取每个比赛的胜利。"""
判断：续写不符合现实，因为并无"约翰·卡明斯"这个人。

示例 2；示例 3

新闻的开头："""《{headLine}》\n{broadcastDate}\n{newsBeginning}"""
续写："""{continuation}"""
判断：

Figure 14: Prompt template for discriminative evaluation (sentence level)

Now, I hope you can act as a sequel evaluator, where you need to choose the continuation that is more realistic and accurate.

The beginning of the news: """《European Basketball League's Second Phase Group Stage: Alba Berlin Faces Grim Qualification Prospects》\n2015-03-05 06:01:47\n   Xinhua, Berlin, March 4 (Reporter Wang Dong) — The ninth round of the 2014-2015 season European Basketball League's second phase group stage kicks off tonight. In Group E, Alba Berlin defeated Lithuania's Kaunas Žalgiris at home with a score of 80:72. After nine rounds, with four wins and five losses, they are firmly in fifth place in the group, yet their qualification prospects still look grim."""
A: """Alba Berlin's head coach, John Cummings, stated that although their current ranking is not ideal, they will still give their all and strive for victory in every match."""
B: """\n   The European Basketball League games are usually scheduled for Thursdays and Fridays, but the group stage match between Alba and Kaunas was moved forward a day."""
Which continuation is more realistic and accurate: B is more realistic and accurate

Demonstration 2; Demonstration 3

The beginning of the news: """《{headLine}》 \n{broadcastDate}\n{newsBeginning}"""
A: """{contn1}"""
B: """{contn2}"""
Which continuation is more realistic and accurate:

现在，你我希望你作为一个续写评测器，你需要选择出更符合现实，更准确的续写。

新闻的开头："""《（体育）欧洲篮球联赛第二阶段小组赛阿尔巴出线形势不乐观》\n2015-03-05 06:01:47\n   新华社柏林 3 月 4 日电（记者王东）2014-2015 赛季欧洲篮球联赛第二阶段小组赛第九轮今晚开始。在 E 组的比赛中，柏林阿尔巴队主场以 80：72 力挫立陶宛考纳斯萨拉基斯队，9 轮过后 4 胜 5 负，稳居小组第五名，但是出线形势仍不乐观。"""
A："""柏林阿尔巴队主教练约翰·卡明斯表示，虽然目前的排名并不理想，但他们仍然会全力以赴，争取每个比赛的胜利。"""
B："""\n   欧洲篮球联赛一般安排在每周四和周五进行，但是阿尔巴和考纳斯的这场小组赛提前一天进行。"""
哪个续写更符合现实，更准确：B 更符合现实，更准确

示例 2；示例 3

新闻的开头："""《{headLine}》 \n{broadcastDate}\n{newsBeginning}"""
A："""{contn1}"""
B："""{contn2}"""
哪个续写更符合现实，更准确：

Figure 15: Prompt template for selective evaluation

You are a news worker for Xinhua News Agency. I hope you can assist me in completing a news article. Please write a continuation based on the text I have already prepared. Here's an example:

The text already written:
《(Cultural Relics and Archaeology) The First Discovery of Tang Dynasty Pear Garden Disciples' Tomb Inscriptions in Luoyang》
2016-10-27 15:14:41
　　Xinhua, Zhengzhou, October 27 - Two Tang Dynasty Pear Garden disciples' tomb inscriptions recently appeared at Luoyang Normal University, with experts preliminarily speculating that the tomb owners were the couple of the Sogdian musician Cao Qianlin from the Tang Dynasty. This is the first discovery of Tang Dynasty Pear Garden disciples' tomb inscriptions in Luoyang, adding valuable data to the study of ancient Silk Road cultural exchanges.
Text for continuation:
<response>
\n These two tomb inscriptions, currently on display at the Heluo Culture International Research Center Relics Exhibition Hall at Luoyang Normal University, were unearthed in the Zhanggou Community of the Longmen Garden District in Luoyang city. The tomb inscription of Cao Qianlin is 47 cm in length and width, with the cover engraved in seal script "Tomb Inscription of the Late Mr. Cao of the Great Tang", and the text of the inscription is in regular script, with clear and visible handwriting.
</response>

The text I have already written is: 《{headLine}》 \n{broadcastDate}\n{newsBeginning}
Please complete the text for continuation (write the continuation text between <response></response>):

你是一名新华社新闻工作者。我希望你能辅助我完成一篇新闻的撰写。请你根据我已经写好的文本为我续写一段话。下面是一个例子：

已经写好的文本：
《（文物考古）洛阳首现唐代梨园弟子墓志》
2016-10-27 15:14:41
　　新华社郑州 10 月 27 日专电（记者桂娟）两方唐代梨园弟子墓志日前现身洛阳师范学院，专家初步推测墓主人为唐代粟特乐人曹乾琳夫妇。这是洛阳首次发现唐代梨园弟子墓志，为古代丝路文化交流研究再添宝贵资料。
续写的文本：
<response>
\n　　正在洛阳师范学院河洛文化国际研究中心文物陈列馆展出的这两方墓志，出土于洛阳市龙门园区张沟社区。其中，曹乾琳墓志长宽各 47 厘米，盖文篆书"大唐故曹府君墓志铭"，墓志文字为楷书，字迹清晰可见。
</response>

现在我已经写好的文本是：《{headLine}》 \n{broadcastDate}\n{newsBeginning}
请你完成要续写的文本（续写的文本写在<response></response>之间）：

Figure 16: Prompt template for generative evaluation