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

Xun Liang*, Shichao Song*, Simin Niu*, Zhiyu Li†✉, Feiyu Xiong†, Bo Tang†, Zhaohui Wy†, Dawei He‡, Peng Cheng‡, Zhonghao Wang‡, Haiying Deng‡

*School of Information, Renmin University of China, Beijing, China

†Institute for Advanced Algorithms Research, Shanghai, China

‡State Key Laboratory of Media Convergence Production Technology and Systems, Beijing, China

Email: {xliangs, songshichao, niusimin}@ruc.edu.cn, {lizy, xiongfy, tangb}@iaar.ac.cn

{hedawei, chengpeng, wangzhonghao, denghaiying}@xinhua.org

***Abstract*—Large language models (LLMs) have emerged as pivotal contributors in contemporary natural language processing and are increasingly being applied across a diverse range of industries. However, these large-scale probabilistic statistical models cannot currently ensure the requisite quality in professional content generation. These models often produce "hallucinated" text, compromising their practical utility in professional contexts. To assess the authentic reliability of LLMs in text generation, numerous initiatives have developed benchmark evaluations for hallucination phenomena. Nevertheless, these benchmarks frequently utilize constrained generation techniques due to cost and temporal constraints. These techniques encompass the use of directed hallucination induction and strategies that deliberately alter authentic text to produce hallucinations. These approaches 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 in text generation is presently lacking. Consequently, we have developed an Unconstrained Hallucination Generation Evaluation (UHGEval) benchmark, designed to compile outputs produced with minimal restrictions by LLMs. Concurrently, we have established a comprehensive benchmark evaluation framework to aid subsequent researchers in undertaking scalable and reproducible experiments. We have also executed extensive experiments, evaluating prominent Chinese language models and the GPT series models to derive professional performance insights regarding hallucination challenges.**

***Index Terms*—large language models, llms, hallucination, benchmark, unconstrained generation**

I. Introduction

With the proliferation of extensive textual corpora, the advent of high-performance GPUs, and the refinement of advanced deep learning paradigms, Large language models (LLMs) have exhibited unparalleled proficiency in a multitude of natural language processing (NLP) tasks, including language generation, knowledge application, and intricate reasoning. Concurrently, noteworthy advancements have been realized in the domains of human alignment, engagement with external environments, and the manipulation of tools [1].

Organization hallucinated id=doc_003726	~~The MOTIE in South Korea~~ Korea Aerospace Industries stated that the South Korean government will continue to advance this export plan.
Statistics hallucinated id=num_000691	During the holiday, the national highway passenger traffic reached ~~250~~ 310 million person-times, representing a year-on-year increase of ~~8.9%~~ 3.2%.
Knowledge hallucinated id=kno_000410	Sickle cell disease is a severe hereditary blood disorder that can lead to ~~atherosclerosis~~ anemia, infarction, and other complications.
Timeline hallucinated id=gen_005626	China National Arts Fund was officially established in ~~2012~~ 2013 with the aim of supporting artistic creation and the cultivation of artistic talent nationwide.

Fig. 1. Real-world hallucination examples from UHGEval. Using the IDs, you can locate the corresponding original Chinese news articles within our dataset. *Note*: MOTIE denotes Ministry of Trade, Industry, and Energy.

However, LLMs invariably manifest hallucinations [2]. Hallucination is characterized by generated content that is incongruent with user input, the model's own output context, or factual information. Real-world examples of hallucination from our UHGEval dataset can be observed in Fig. 1.

Owing to reliability concerns, these circumstances markedly hinder the practical deployment of LLMs. Furthermore, in specialized domains like medicine, law, finance, and journalism, hallucination presents a significant impediment to deployment [3], [4]. These fields require stringent standards of content timeliness, accuracy, and logical consistency, attributable to their dynamic and highly specialized characteristics. During the training data collection phase, LLMs may exhibit a deficiency in domain-specific knowledge, yielding outdated content. In the pre-training phase, constraints in model parameters or training methodologies may engender parameter inaccuracies, thwarting the retrieval of accurate content. During the supervised fine-tuning phase, incongruent datasets might yield excessively positive incorrect responses. In the inference phase, the absence of a rollback mechanism can precipitate a cumulative escalation of hallucinations, un-

* The authors contribute equally.

✉ Corresponding author.

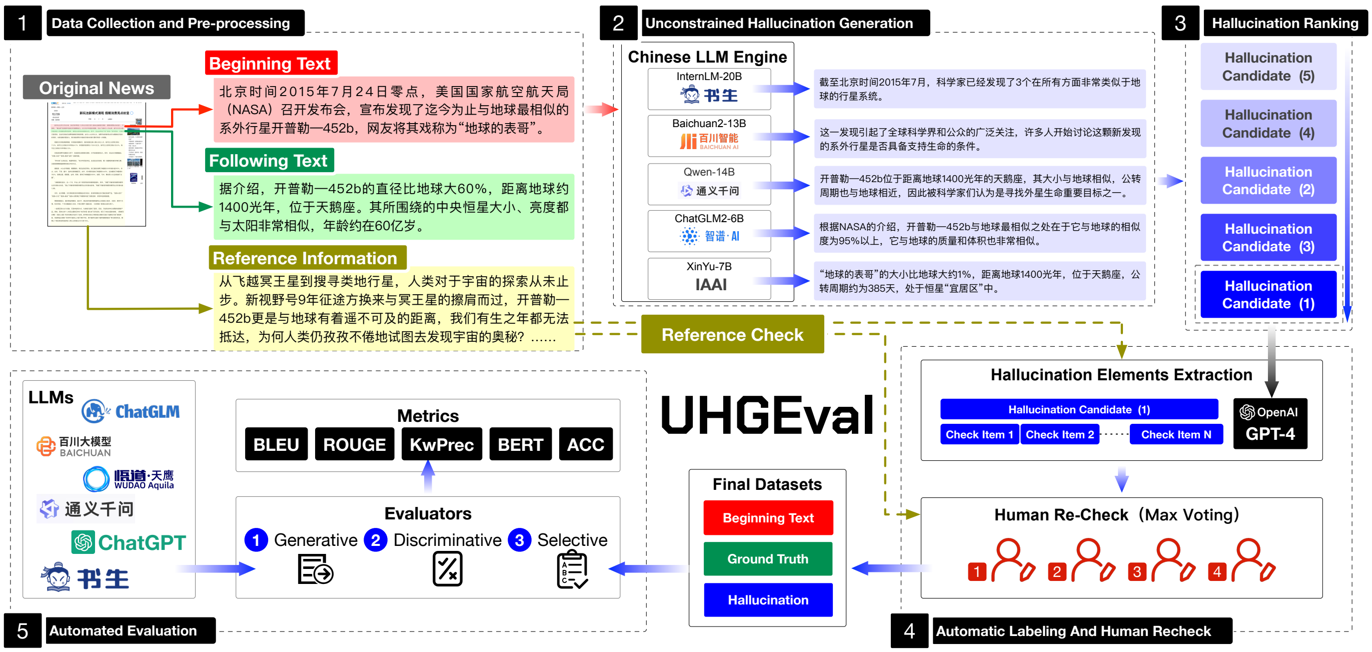

Fig. 2. The process of creating UHGEval. Steps 1 to 4 regarding the creation of the benchmark dataset are explained in Section II; Step 5, concerning the evaluation framework, is detailed in Section III.

dermining the logical integrity of responses [5]. For example, erroneous medical guidance, imprecise legal stipulations, and fabricated journalistic narratives substantially restrict the practical utility of LLMs in real-world contexts [3]. The fabricated news content depicted in Fig. 1 offers NO utility to journalists; on the contrary, the verification and rectification of such content exacts a toll on the valuable time of journalists.

Achieving professional-level generation necessitates confronting the significant challenge of devising novel training methodologies and model architectures. However, prior to these developments, it is crucial to formulate a comprehensive, stringent, and demanding benchmark for the assessment of hallucination in language generation [5], [3]. Without such a benchmark, conducting a comparative evaluation of efforts aimed at controlling hallucination would prove to be arduous.

While there have been initiatives to develop benchmarks for hallucination assessment, the majority of these methods employ restricted techniques to produce particular kinds of hallucinated utterances. This approach to generation is at odds with real-world scenarios where hallucinations may 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” [6]. Additionally, benchmarks such as HADES annotate hallucinations at a finer granularity by generating token-level hallucinations based on text perturbations [?], but the text perturbation method is still constrained. Ultimately, the majority of 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, 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 comprised of news articles. Selecting texts from this domain is intentional, given that news requires utmost precision in conveying factual information and exhibits minimal tolerance for hallucinations. Constructing an evaluation dataset within this sphere presents a considerable challenge for the majority of LLMs. Concurrently, news articles are of exceptional quality, readily available, and frequently employed as training corpora by a large number of LLMs, guaranteeing impartiality in the evaluation of many LLMs [1]. In light of these factors, we collected a considerable volume of raw news articles, established an efficient, professional-grade hallucination assessment dataset, and formulated an evaluation framework named UHGEval. It is significant to note that our dataset was produced in an entirely unconstrained fashion. We permit models to compose freely and subsequently sift through the content to identify hallucinations.

Our contributions are as follows: (1) The development of an unconstrained hallucination evaluation dataset. Existing methods for constructing datasets often yield biases towards predefined directions, thereby hindering the full simulation of real-world hallucinations. We have created a hallucination evaluation dataset comprising over 5000 items, generated without intervention, closely mirroring real-world scenarios. (2) The establishment of a unified and diverse evaluation framework. Current benchmark methods for hallucination evaluation often exhibit a singular approach and lack task specificity. We have developed 𝕌ℍ𝔾Eval, a unified, flexible, and robust evaluation framework that encompasses generative, discriminative, and selective modalities, along with sentence-level and keyword-level granularity. (3) A comprehensive empirical analysis. We conducted detailed experiments with the proposed benchmark on 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].

II. The UHGEval Benchmark Dataset

A. Data Collection and Pre-processing

To enhance the authenticity of the news continuation dataset, 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 the inclusion of more recent news articles (e.g., from 2023) was made to better assess the model's understanding of existing knowledge and past news events. Indeed, the knowledge embedded within the training data of existing Chinese LLMs typically encompasses information pertaining to significant news between 2015 and 2017 [1].

Considering the different categories of news, such as sports, education, science, and society, the generated hallucinations typically exhibit certain differences. Therefore, when curating the initial news collection for continuation, we endeavored to ensure that the distribution of the collection aligns with the original distribution by randomly sampling from the entire news dataset. Furthermore, we have categorized the collected news examples into four major types: document-intensive, number-intensive, knowledge-intensive, and general news, as shown in Table I. We hypothesize that the likelihood of generating hallucinations varies for different types of news. For example, number-intensive news frequently contains various numerical data, such as years, scores, and values, which may predispose the model to fabricating numbers or introducing minor deviations. Document-intensive news, on the other hand, primarily references official documents, such as factual policy documents, official statements, standard explanations, and legal clauses. In this case, the model may be inclined to fabricate specific policy or document names, or create detailed but fictional policy content. Knowledge-intensive news is characterized by an emphasis on enduring truths and analytical reasoning, which can render the model susceptible to flawed reasoning or the retrieval of incorrect facts. In addition to these three types, we also categorize culturally relevant general news as a separate category for experimental control.

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.

TABLE I
Statistics of Collected News

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

Note: In the table, DOC denotes document-intensive news; KNO demotes knowledge-intensive news; NUM denotes number-intensive news; GEN denotes general news. The same as below.

Filtering Settings. To ensure the overall quality of the final evaluation dataset, we have implemented the following filters: We consider only the categories listed in Table I, which correspond to the most frequently occurring categories in the original news collection. For news length, we set parameters such that the body length of the selected news falls between 630 and 870 characters, while the beginning text spans between 80 and 120 characters and consists of 2 to 5 sentences. These length parameters reflect the average values in the original news collection and were chosen to avoid overburdening the annotation process at a later stage.

B. Unconstrained Hallucination Generation

Historically, benchmarks for evaluating hallucination have predominantly relied on a single LLM to produce hallucinated dataset. Notable examples include HaluEval [6] and PHD [8], which exclusively utilize ChatGPT, and FActScore [9] and FACTOR [10], which solely employ InstructGPT [11]. In contrast, our methodology incorporates a suite of five distinct Chinese LLMs to generate hallucinated content. These models include ChatGLM2-6B [12], Baichuan2-13B [13], Qwen-14B [14], InternLM-20B [15], and the Xinyu series model, Xinyu-7B. Xinyu-7B is an augmented large-scale language model derived from the foundational BloomZ-7B [16] through continued pre-training, news-specific fine-tuning, and alignment optimization. Furthermore, Xinyu2-70B is developed based on the open-source LLaMA2-70B [17] 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. The Xinyu series models are the results of a collaborative research and development effort between the Institute for Advanced Algorithms Research, Shanghai (IAAR, SH), and the State Key Laboratory of Media Convergence Production Technology and Systems of the Xinhua News Agency. Xinyu-7B and Xinyu2-70B will also be utilized in the experiment phase.

[1]https://github.com/IAAR-Shanghai/UHGEval

Our approach engenders a more heterogeneous generation of hallucinations, mitigating the bias that may arise from the use of a single model and promoting equity within the dataset. This is due to the varying architectures and training corpora inherent to different LLMs. Furthermore, we have adopted an unconstrained generation methodology for the continuation of natural language content. This entails directly inputting the text to be continued into the model without any restrictive prompt instructions, thereby obtaining organic results. For each input example, we concurrently generate five candidate continuations. To maintain consistency across all models, we employ uniform parameter settings, with a temperature coefficient set at 1.0 and max_new_tokens limited to 1024.

C. Hallucination Ranking

Given the unconstrained nature of our generation paradigm, the task of discerning whether the generated content is indeed hallucinated presents a significant challenge. Upon generating the continuations, a straightforward reliance on human verification is infeasible. An exclusive dependence on human annotation would incur substantial costs and may not be sustainable at scale, 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 methodology. This approach begins with an initial phase of hallucination ranking, which is designed to preliminarily sort the generated content based on the likelihood of hallucination. The ranking is then followed by a combination of automatic labeling and human recheck. The integration of hallucination ranking and machine labeling serves a pivotal role in streamlining the subsequent human verification process. This hybrid approach aims to enhance the efficiency and accuracy of human checks, effectively bridging the gap between the scalability of automated processes and the critical discernment of human judgment.

Hallucination ranking is a crucial step in the process of evaluating and selecting the most appropriate continuation from a set of candidate continuations generated by LLMs. The objective of this step is to identify a continuation that not only demonstrates high quality in terms of coherence and readability but also includes an appropriate level of hallucination — misinformation or fabrications that are not supported by the input or real-world knowledge.

To strike this balance, the selection process takes into account two primary dimensions:

Fluency. This refers to the naturalness 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 Institute for Advanced Algorithms Research (IAAR) is employed. This model is trained to evaluate the quality of text and can assign scores to each continuation based on its fluency. By using this model, the top three continuations that exhibit the highest fluency are retained for further consideration.

Likelihood of Hallucination Occurrence. This dimension evaluates the extent to which the continuation may contain hallucinated content. For hallucination occurrence likelihood ranking, 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 lexical metrics based on n-gram coverage, such as BLEU [18] and ROUGE [19], we believe that these rule-based methods may not effectively discover hallucinated keywords. Therefore, we propose the keyword precision (kwPrec) metric.

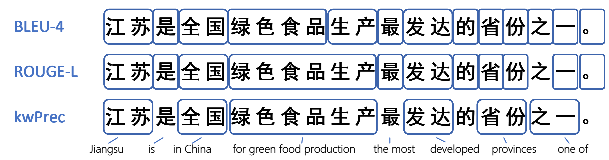

Fig. 3. Tokenization results for BLEU-4, ROUGE-L, and kwPrec, using news$_{id=num_000432}$ as an example. The meaning of the above sentence is: Jiangsu is one of the most developed provinces in China for green food production. *Note*: We ignore tokens that cause overlap.

This approach initially uses an LLM (here, we use 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.

After implementing this method, we calculate the kwPrec for each of the three candidate continuations, selecting the one with the lowest value as the final candidate. Through the screening in these two stages, we can ensure that, in the worst case scenario, the final candidate continuation ranks third in fluency and third in the likelihood of hallucination occurrence, achieving a balanced level.

By considering both fluency and the likelihood of hallucination, the process aims to filter out continuations that are either too nonsensical or too conservative (lacking any hallucinated content). The ideal candidate continuation would be one that is coherent and engaging but also contains a detectable level of hallucination, which can then be used for further analysis, such as studying the model's tendencies to hallucinate or for training systems to detect and mitigate such hallucinations.

The final candidate continuations will undergo further annotation to determine the presence and degree of hallucination, which can involve additional automated tools and human judgment. This multi-faceted approach helps ensure that the final selected continuation is both high-quality and relevant for the purposes of the hallucination evaluation benchmark.

D. Automatic Labeling And Human Recheck

Through the application of 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

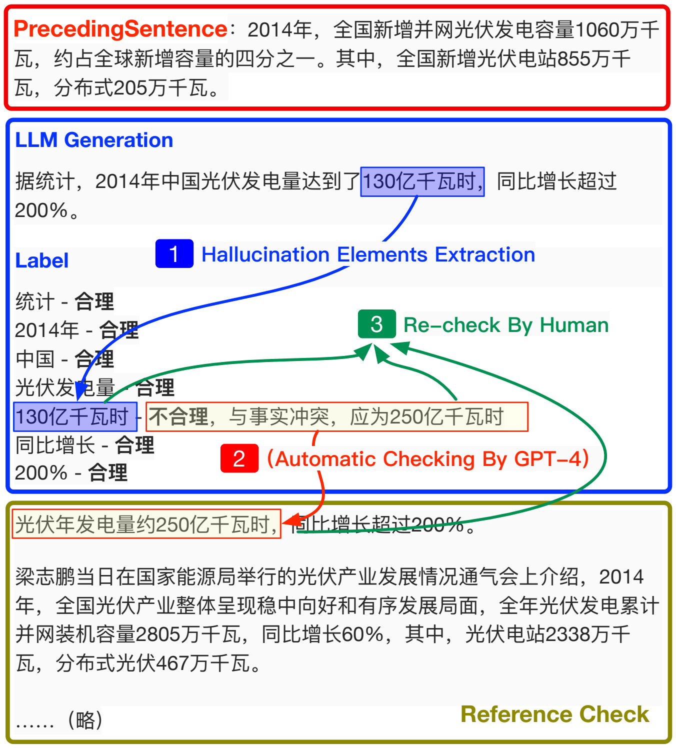

Fig. 4. The process of automatic labeling and human recheck.

that utilizes keywords, which includes automatic labeling and subsequent human verification, as shown in Fig. 4.

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 [20] to perform annotation on these keywords. GPT4-0613 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 recheck. 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 is implemented to ensure the accuracy of the machine-generated annotations. In the end, instances verified as accurate by annotators comprise the final $\mathbb{UHG}$Eval dataset.

However, the keyword-based annotation scheme exhibits inherent limitations. Languages exhibit a dependency structure among words [21]. 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 the erroneous term, given that "night" is typically described as black. To address the annotation 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 semantically coherent 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, correct$(\cdot)$ is the correction function that modifies hallucinated words to non-hallucinated words, and hallucinated$(\cdot)$ assesses whether a sentence composed of a set of keywords hallucinated.

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

In accordance with 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.

Additionally, we acknowledge that the task of 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.

E. Data Statics

Starting with 17,714 candidate hallucinated continuations, we curated a dataset of 5,141 hallucinated continuations, as detailed in the basic statistics in Table II. Additionally, we developed a conversion rate chart to depict the transition from candidate hallucinations to the final dataset, as depicted in Fig. 5. 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.

TABLE II
DATASET BASIC STATICS

	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

Note: In the table, # denotes quantity, avg. denotes average, len. denotes length, contn. denotes hallucinated continuations, begin. denotes news beginnings, and refer. denotes reference information. The same as below.

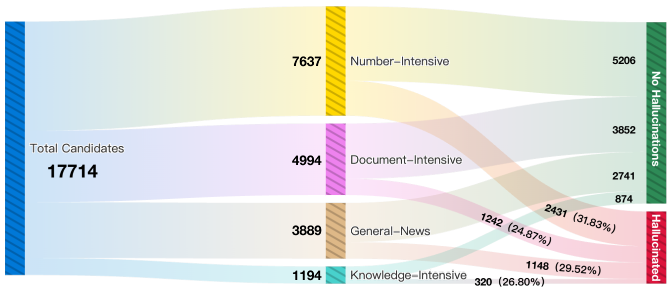

Fig. 5. Conversion rates from candidates to hallucinations.

By analyzing the hallucinated word cloud depicted in Fig. 6 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 documents.

Fig. 6. Word clouds of hallucinated keywords in different types of news

III. Experiments

A. Models

Given that our dataset is tailored for the Chinese language generation domain, we selected eight widely-used Chinese LLMs and three foundational models from OpenAI, as detailed in Table III. These include eight base models: GPT Base, GLM Base, BLOOMZ Base, InternLM Base, Baichuan2 Base, Qwen Base, Aquila2 Base, and LLaMA2 Base.

[2]https://openai.com/blog/new-models-and-developer-products-announced-at-devday

TABLE III
Models Sorted by Release Date

Model	Parm.	Type	Publisher	Release
GPT3.5-Turbo [1]	175B*	Chat	OpenAI	2023.03*
GPT4-0613 [20]	NaN	Chat	OpenAI	2023.06
ChatGLM2 [12]	6B	Chat	Tsinghua	2023.06
Xinyu	7B	Chat	IAAR&Xinhua	2023.06
InternLM [15]	20B	Chat	ShLab	2023.07
Baichuan2 [13]	13B	Chat	Baichuan Inc.	2023.09
Baichuan2 [13]	53B	Chat	Baichuan Inc.	2023.09
Qwen [14]	14B	Chat	Alibaba	2023.09
Aquila2 [22]	34B	Chat	BAAI	2023.10
Xinyu2	70B	Chat	IAAR&Xinhua	2023.10
GPT4-1106[2]	NaN	Chat	OpenAI	2023.11

Note: In the table, asterisk (*) denotes estimated value, NaN denotes no public data available, and 175B denotes 175 billion.

GPT represents a series of LLMs developed by OpenAI [20]. In this study, GPT3.5-Turbo, GPT4-0613, and GPT4-1106 are utilized. GLM constitutes a pre-training framework proposed by Tsinghua University [12], and the ChatGLM2-6B chat model is employed. BLOOMZ is a variant derived via multitask prompted fine-tuning (MTF) of the pre-trained BLOOM model [16], and following supplementary training, it is integrated into Xinyu-7B. InternLM serves as an open-source, lightweight training framework, with its development team releasing a spectrum of models utilizing this framework [15]; the InternLM-20B open-source chat model is utilized in the present work. Baichuan2 comprises a series of expansive, multilingual base language models [13], with both the open-source Baichuan2-7B chat model and the closed-source Baichuan2-53B model being employed in this investigation. Qwen encompasses a language model series characterized by distinct models with varying parameter counts [14], 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 [22], and the Aquila2-34B chat model is employed in this research. LLaMA2 constitutes a suite of pre-trained and fine-tuned LLMs, with scales ranging from 7 billion to 70 billion parameters [17]. Following additional training, LLaMA2-70B is incorporated into Xinyu2-70B.

B. Evaluation Method

For the evaluation of hallucinations in LLMs, the task is decomposed into three principal dimensions: form, metric, and granularity. Form concerns the manner in which 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.

In terms of 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 [23]. However, the drawbacks of this approach are evident: evaluating in excess of 5000 data points is tantamount to creating 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" [6], [24]. 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.

Similar to discriminative evaluation, selective evaluation allows LLMs to tackle multiple-choice questions by choosing between option A or B, as exemplified by PandaLM [25]. 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. Specifically, the LLM under evaluation is provided with an initial text and is then tasked with generating a continuation. Subsequently, various reference-based techniques are utilized to determine if the continuation includes hallucinations. However, the challenge arises from the fact that it is not feasible to automatically and accurately ascertain if 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 [26].

In terms of metrics, these 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 [18], ROUGE [19], and the newly proposed kwPrec. Semantic metrics gauge the similarity in meaning between sentences, with examples including BERTScore [27], GPT-judge [28], and GPTScore [29], among others.

In terms of 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 sentence level or keyword level.

C. Evaluation Framework

In order to accommodate different forms of evaluation methods, we have developed a of data-secure, easy-to-extend and easy-to-use evaluation framework, as illustrated in Fig. 7.

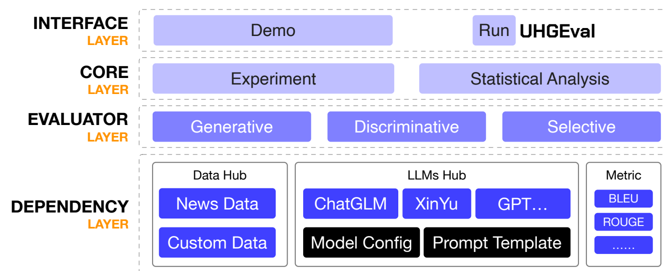

Fig. 7. Evaluation Framework

The framework comprises four ascending layers: the dependency layer, the evaluator layer, the core layer, and the interface layer. The dependency layer delineates the requisite underlying modules for the evaluation framework, encompassing datasets, LLM hubs, and diverse metrics. Notably, all underlying modules are extensible; datasets may be supplanted with customized versions, LLMs sourced from APIs or platforms such as Hugging Face[3], and metrics tailored individually. The evaluator layer, constituting the second tier, 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 module facilitates experiments involving multiple LLMs, evaluators, and processes, whereas the latter module is tasked with the statistical analysis of experimental outcomes. The interface layer, constituting the final tier, orchestrates the user's interaction with $\mathbb{UHG}$Eval. A concise 20-line demonstration is provided to expedite user initiation, complemented by run.py capable of initiating experiments via the command line.

$\mathbb{UHG}$Eval is both intuitive and secure for users, offering efficient usage while concurrently ensuring the integrity of

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

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.

D. Experimental Setup

To establish a robust experimental framework, our configuration includes prompt engineering, ensuring equilibrium between positive and negative examples, optimizing hyperparameters, and configuring evaluators.

Prompt engineering. The prompt engineering technique employed is "intent + instruction + 3-shot (explainable) prompting." Intent delineates the LLM's role, instruction outlines the task for the LLM to execute, and the prompt incorporates three examples to aid the LLM's few-shot learning [1]. 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, regardless of the impact on evaluation speed and cost. In discriminative and selective evaluations, it is indiscernible whether the model is conjecturing the outcome or discerning the presence of hallucinations. Consequently, the use of explainable prompting enables the validation of the model's confidence through the analysis of experimental results.

Balancing positive and negative examples. To guarantee the reliability of experimental outcomes for all LLMs, we meticulously balance examples in discriminative and selective evaluations. Specifically, the LLM under evaluation will encounter an equal number of examples with and without hallucinations. This approach addresses the tendency of some models to learn patterns from the three examples in the prompts and produce conjectural rather than reasoned responses when making judgments. Such a tendency can introduce a considerable bias towards certain outcomes. An imbalance could complicate the analysis of experimental outcomes.

Hyperparameter settings. Managing parameters for heterogeneous LLMs is a multifaceted endeavor, as different LLMs feature unique interface designs, and the same parameters can have varying implications across LLMs. For example, the level of determinism influenced by the temperature parameter varies. 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 configured parameters including temperature, top_p, top_k [1], and random seed. To ensure output determinism and improve reproducibility, we set the temperature to 0.1. Considering that OpenAI models advise against adjusting temperature and top_p simultaneously, we minimally altered top_p, setting it at 0.9. We set top_k to 5, which is effective for certain models. To further enhance reproducibility, we established a seed for random number generators, setting it at 22.

Evaluator Settings. Discriminative evaluation encompasses assessments at two levels of granularity: sentence-level and keyword-level. Prompt design for both levels utilizes the "intent + instruction + 3-shot (explainable) prompting" approach. Furthermore, we maintain a balanced representation of positive and negative examples at both levels. For discriminative evaluation, accuracy serves as the metric. Selective evaluation adheres to the identical prompt design. Each evaluated LLM is presented with one positive and one negative example for every news item. To uphold the integrity of the evaluation, the order of positive and negative examples is randomly alternated with a 50% chance. Accuracy is also employed as the evaluation metric. The generative evaluation's prompt design adheres to the principle of $\mathbb{UHG}$. Evaluation metrics comprise 4-gram BLEU (BLEU-4), longest common subsequence-based ROUGE (ROUGE-L), kwPrec, and BERTScore.

E. Results and Analysis

Results are presented in Table IV, Table V, and Table VI.

Discriminative evaluation. Initially, the GPT series models' performance is notably superior. In the keyword-level assessment, GPT4-0613 and GPT3.5-Turbo respectively achieve the top two rankings. At the sentence level, GPT4-0613 and GPT4-1106 respectively attain the first and second spots. As previously hypothesized, discriminative evaluation requires robust foundational capabilities from LLMs, such as knowledge recall, utilization, and judgment. The GPT series models markedly surpass other models, showcasing their formidable foundational capabilities. 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. This is consistent with our prior conjecture that furnishing LLMs with more contrasting information alleviates the demand on the model's fact recall, thus diminishing the challenge of selective evaluation. Therefore, we posit that selective evaluation is comparatively simpler for LLMs. Thirdly, a decline is observed in discriminative evaluation outcomes from GPT4-0613 to GPT4-1106, whereas selective evaluation outcomes register a notable increase of around 5%. This substantiates the "seesaw phenomenon," wherein certain capabilities are enhanced while others may

TABLE IV
DISCRIMINATIVE (KEYWORD AND SENTENCE LEVEL) AND SELECTIVE EVALUATION RESULTS

	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	<u>53.72%</u>	3.08	4183	50.02%	5039	49.03%	5103
GPT4-0613	**70.04%**	3.07	4100	**57.42%**	5024	55.20%	5047
GPT4-1106	69.48%	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

Note: In the table, #kws denotes the number of keywords and #valid denotes number of valid evaluations. In the same column of values, optimal values are bolded and suboptimal values are underlined. The same as below.

TABLE V
GENERATIVE EVALUATION RESULTS

	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 VI
EVALUATION RESULTS BY DIFFERENT TYPES

	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%**	<u>51.06%</u>	47.63%	47.85%
GPT4-0613	**59.87%**	<u>55.99%</u>	51.93%	55.73%
GPT4-1106	**68.73%**	60.19%	54.77%	<u>62.04%</u>
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%

Note: Read by row. In the same row of values, optimal values are bolded and suboptimal values are underlined.

regress, in tandem with the model's upgrade [30]. This suggests that the decision to either enhance a single capability individually or to balance multiple capabilities is critical.

Generative evaluation. Firstly, InternLM-20B secures two top spots, one runner-up position, and boasts the longest average generation length. This reflects the model's superior credibility in content generation. However, its kwPrec score is modest, indicating potential for enhancement in keyword-level information generation. Secondly, Xinyu2-70B captures one top spot, two runner-up positions, and has the second-longest average generation length, underscoring its strong credibility in content generation. Its sole underperformance is in the ROUGE metric, which is recall-oriented. Conversely, BLEU and kwPrec are precision-oriented, suggesting the model is adept at delivering consistent output yet faces challenges with factual recall. Thirdly, Aquila-34B achieves the pinnacle in kwPrec scoring, signaling a notable edge in generation quality. However, this could be attributed to its comparatively shorter average generation length. kwPrec assesses the coverage of extended tokens (i.e., keywords), allowing for brief continuations with limited keywords to secure higher keyword coverage in relation to reference information. Fourthly, Baichuan2-53B registers a high ROUGE score, indicative of its proficiency in fact recall from the parameters, demonstrating accurate factual retrieval. Fifthly, the GPT series exhibits subpar performance, owing to the insubstantial Chinese data in its training corpus. For example, the Chinese data incorporated in GPT's training from the Common Crawl corpus comprises less than 5%[4].

Evaluations by Type. Given the categorization of news into four types, we can proceed with an in-depth analysis. We focus on selective evaluation results and perform a comprehensive breakdown analysis of these across the four types, as illustrated in Table VI. Initially, the majority of LLMs demonstrate enhanced accuracy for knowledge-intensive and document-

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

intensive news. This observation is consistent with the general consensus that 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, potentially leading to a deficiency in this factual domain within the model parameters. Regarding number-intensive news, it poses a considerable challenge for most LLMs, given that encoding identical numbers with varied historical meanings is complex. Lastly, GPT4-1106 attains especially high scores in the demanding number-intensive news, which might be attributed to its sophisticated parameterization for numerical data handling.

F. Discussion

Each of the three evaluation methods possesses distinct advantages and drawbacks. Discriminative evaluation is often the method of choice for a range of standard benchmarks [6], [24]. 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 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. 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, which suits models with minimal fine-tuning. 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 factual information, thereby increasing the challenge.

Moreover, various evaluations cater to different application contexts. Should the objective be to solely improve the model's capacity for reliable continuation, generative evaluation would suffice. In the training of a dependable chatbot, selective and discriminative evaluations prove suitable. When aiming to train a reward model, selective evaluation is beneficial, offering evaluation for positive and negative instances. If the goal is to enhance the model's ability to recall and apply knowledge, discriminative evaluation emerges as the demanding option.

IV. Related Works

A. 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 [1]. 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 [17], Tsinghua University's GLM [12], and Alibaba's Qwen [14], while closed-source models feature OpenAI's GPT [20], Baidu's ERNIE Bot [31], and Anthropic's Claude [5], 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, with the goal of improving 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.

B. 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 [5]. 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 [2]. Regarding the causes of hallucinations, several factors can be responsible [5]. 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

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

TABLE VII
HALLUCINATION EVALUATION BENCHMARKS SORTED BY NAME

Benchmark (Released Year)	Generation Method	Annotation	Metric	Granularity	Lang.
ChineseFactEval'23 [32]	Manual	Manual	Acc	Sentence	CN
CSK-PN'23 [33]	Direct: Common KGs	No Need	Acc	Word	EN
FACTOR'23 [10]	CHG: Wiki, News	Auto	FACTOR Acc	Sentence	EN
FActScore'23 [9]	CHG: Wiki	No Need	FActScore by Human	Short Sentence	EN
HaLoCheck'23 [34]	CHG	No Need	HaLoCheck, selfcheckGPT	Sentence	EN
FactualityPrompts'22 [**?**]	Direct: Wiki	Auto	NE Error, Entailment	Document, Sentence	EN
HADES'22 [**?**]	CHG: Wiki	Manual	Acc, G-Mean, BSS, AUC, etc.	Word	EN
HalluQA'23 [24]	CHG, Manual: TruthfulQA, Wiki	Manual, Auto	Non-hallucination Rate	Sentence	CN
HaluEval'23 [6]	CHG: Alpaca, HotpotQA, etc.	Manual, Auto	Acc	Document	EN
HILT'23 [2]	CHG: NYT, Politifact	Manual	HVI	Word	EN
KoLA-KC'23 [35]	Direct: Wiki, evolving dataset	Auto	BLEU, ROUGE	Document	EN
Med-HALT'23 [36]	Direct: MedMCQA, PubMed, etc.	No Need	Acc, Pointwise Score	All	EN
PHD'23 [8]	CHG: Wiki	Manual	F1, Acc, Prec, Reca	Document	EN
SelfAware'23 [37]	CHG: Quora, HowStuffWorks	Manual	F1, Acc	Sentence	EN
STSN'23 [38]	UHG	Manual	Acc, Prec, Reca	Sentence, Concept	EN
TruthfulQA'22 [28]	Manual	Manual	Acc by Human or GPT-judge	Sentence	EN
$\mathbb{UHG}$Eval (Ours)	UHG: Xinhua News	Auto, Manual	Acc, kwPrec, BERTScore, etc.	Sentence, Keyword	CN
XSum Hallu'20 [**?**]	UHG: XSum	Manual	ROUGE, BERTScore, Acc, etc.	Word, Document	EN

Note: Generation Method column provides the approach, and the base dataset if used. In this 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.

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.

C. Hallucination Evaluation Benchmarks

To more effectively tackle the issue of hallucinations, constructing evaluation benchmarks is essential. In this context, numerous outstanding contributions have surfaced. This section reviews existing contributions regarding the development of benchmark datasets, their characteristics, and the particular methodologies for evaluation. Basic information about these benchmarks is presented in Table VII.

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 [10], Alpaca [6], PubMed [36], Quora [37] and so on. Secondly, hallucinations are generated usually by LLMs such as GPT3.5-Turbo, and most works uses constrained hallucination generation (CHG) paradigm [10], [9], [34], [6], [2], [8], [37]. STSN [38] and XSum Hallu [**?**] 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 requires annotation, which is mostly done by human involvement. There are also works using automatic machine labeling [10], [**?**], [24], [6], [35]. 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 [32], HADES [**?**], TruthfulQA [28], etc.

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. With respect to the topic, the majority of datasets cover the general domain, while some benchmarks target specific domains; for instance, HaLoCheck [34] focuses on the NBA, Med-HALT [36] on medicine, and our $\mathbb{UHG}$Eval on news. Concerning language, most evaluation datasets are in English. To our knowledge, the only two Chinese benchmarks, ChineseFactEval [32] and HalluQA [24], 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.

Evaluation scheme. Existing works use a variety of ways to measure hallucinations. However, due to cost and time constraints, building automatic metrics for evaluation is still dominant, and a small proportion of works use human evaluation [9], [28], [**?**]. In terms of specific evaluation metrics, most works adopt common classification metrics, e.g., F1, accuracy, precision, recall. some other works construct their own calculation methods, e.g., FACTOR [10], FActScore [9], HaLoCheck [34], HVI [2], 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 [35] evaluates knowledge creation (KC) using BLEU and ROUGE to measure the degree of overlap between the output and the reference, TruthfulQA [28] evaluates hallucinations using a specially trained classifier, GPT-judge, and FactualityPrompts [**?**] simultaneously employs a hallucinated named entity error based on n-gram coverage and a semantic-based entailment ratio.

V. Conclusion

LLMs are experiencing a rapid evolution, 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 benchmark dataset using unconstrained hallucination generation, comprising a dataset specifically curated for hallucinated news continuation, which encompasses in excess of 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. Concurrently, on the benchmarking front, we aspire to augment our datasets to encompass a more diverse spectrum of domains and linguistic variations, thereby broadening the applicability and relevance of our benchmarks.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S. Tonmoy, A. Chadha *et al.*, "The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations," *arXiv preprint arXiv:2310.04988*, 2023.

[3] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang *et al.*, "Survey on factuality in large language models: Knowledge, retrieval and domain-specificity," *arXiv preprint arXiv:2310.07521*, 2023.

[4] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922*, 2023.

[5] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu *et al.*, "Siren's song in the ai ocean: A survey on hallucination in large language models," *arXiv preprint arXiv:2309.01219*, 2023.

[6] 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," *arXiv preprint arXiv:2305.11747*, 2023.

[7] T. Liu, Y. Zhang, C. Brockett, Y. Mao, Z. Sui, W. Chen *et al.*, "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)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6723–6737. [Online]. Available: https://aclanthology.org/2022.acl-long.464

[8] S. Yang, R. Sun, and X. Wan, "A new benchmark and reverse validation method for passage-level hallucination detection," *arXiv preprint arXiv:2310.06498*, 2023.

[9] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh *et al.*, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *arXiv preprint arXiv:2305.14251*, 2023.

[10] D. Muhlgay, O. Ram, I. Magar, Y. Levine, N. Ratner, Y. Belinkov *et al.*, "Generating benchmarks for factuality evaluation of language models," *arXiv preprint arXiv:2307.06908*, 2023.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[12] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang *et al.*, "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)*, 2022, pp. 320–335.

[13] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023.

[14] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[15] InternLM, "Internlm: A multilingual language model with progressively enhanced capabilities," https://github.com/InternLM/InternLM, 2023.

[16] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao *et al.*, "Crosslingual generalization through multitask finetuning," *arXiv preprint arXiv:2211.01786*, 2023.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[20] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[21] M.-C. de Marneffe and J. Nivre, "Dependency grammar," *Annual Review of Linguistics*, vol. 5, no. 1, pp. 197–218, 2019. [Online]. Available: https://doi.org/10.1146/annurev-linguistics-011718-011842

[22] BAAI, "Aquila2," https://github.com/FlagAI-Open/Aquila2, 2023.

[23] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[24] Q. Cheng, T. Sun, W. Zhang, S. Wang, X. Liu, M. Zhang *et al.*, "Evaluating hallucinations in chinese large language models," *arXiv preprint arXiv:2310.03368*, 2023.

[25] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen *et al.*, "Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization," *arXiv preprint arXiv:2306.05087*, 2023.

[26] J. Novikova, O. Dušek, A. Cercas Curry, and V. Rieser, "Why we need new evaluation metrics for NLG," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2241–2252. [Online]. Available: https://aclanthology.org/D17-1238

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[28] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. [Online]. Available: https://aclanthology.org/2022.acl-long.229

[29] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[30] S. Zheng, Y. Zhang, Y. Zhu, C. Xi, P. Gao, X. Zhou *et al.*, "Gpt-fathom: Benchmarking large language models to decipher the evolutionary path towards gpt-4 and beyond," *arXiv preprint arXiv:2309.16583*, 2023.

[31] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[32] B. Wang, E. Chern, and P. Liu, "Chinesefacteval: A factuality benchmark for chinese llms," https://GAIR-NLP.github.io/ChineseFactEval, 2023.

[33] J. Chen, W. Shi, Z. Fu, S. Cheng, L. Li, and Y. Xiao, "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)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9890–9908. [Online]. Available: https://aclanthology.org/2023.acl-long.550

[34] M. Elaraby, M. Lu, J. Dunn, X. Zhang, Y. Wang, and S. Liu, "Halo: Estimation and reduction of hallucinations in open-source weak large language models," *arXiv preprint arXiv:2308.11764*, 2023.

[35] J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-Li, X. Lv *et al.*, "Kola: Carefully benchmarking world knowledge of large language models," *arXiv preprint arXiv:2306.09296*, 2023.

[36] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-halt: Medical domain hallucination test for large language models," *arXiv preprint arXiv:2307.15343*, 2023.

[37] Z. Yin, Q. Sun, Q. Guo, J. Wu, X. Qiu, and X. Huang, "Do large language models know what they don't know?" in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8653–8665. [Online]. Available: https://aclanthology.org/2023.findings-acl.551

[38] 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.