Quantifying Hallucination of Large Language Models via Simple Memory Consistency Test

1st Given Name Surname
Xi'an Jiaotong University
Xi'an, China
email address or ORCID

***Abstract*—The emergent abilities of large language models (LLMs) give rise to an intriguing phenomenon: their erroneous generation behaviors have become increasingly subtle. Ultimately, these distinguished behaviors are referred to as hallucination and have attracted much dedicated research. In this study, we investigate LLM hallucination through the lens of memory consistency and divide it into two categories: internal hallucination and external hallucination. This viewpoint provides a valuable framework for future research into the development of quantitative methods for evaluating and demystifying LLM hallucination. Within this framework, we introduce two simple yet effective evaluation methods for both types of hallucination and apply them to three prevalent LLMs. For external hallucination, we assess a LLM's ability to generate consistent responses across various transformations of a single query, as well as the relevance of those responses to the original query. Regarding internal hallucination, we measure a LLM's accuracy in associating simple knowledge pairs, thereby evaluating the robustness of its internal memory. We observe that the performance of all LLMs deteriorates as the number of knowledge pairs increases, even though these models have well acquired each individual knowledge.**

***Index Terms*—large language model, hallucination, memory consistency, memory robustness.**

I. Introduction

THE performance of large language models (LLMs) has continuously improved as their parameters have grown from millions to trillions [1]. Nevertheless, in tandem with their rapid advancement, there are growing concerns about these powerful models for their inclination to generate hallucinatory content. As LLMs are increasingly applied to various critical domains, including software development, healthcare, and legal systems, where there is a stringent demand for faithful and factual generated content, it is imperative to delve into this mysterious hallucination phenomenon, to unveil the underlying causes and accurately evaluate the severity of model hallucination. Indeed, recent research [2] has show that generating hallucinatory responses poses a practical challenge for language models. To make it worse, larger models exhibit more difficulties in managing conceptual knowledge.

While there are numerous research analyze the factors contributing to hallucinations across the entire spectrum of LLMs' capacity acquisition process, they can only yield tentative causal connections between these factors and hallucination, primarily due to the absence of precise hallucination definitions. For instance, from a data preparation aspect, training LLMs on factually incorrect data may inadvertently amplify these inaccuracies, potentially leading to hallucinations [3]. From the training aspect, GPT-like unidirectional models exclusively utilize context from a single direction, which hinders their ability to capture intricate contextual dependencies, potentially increasing risks for the emergence of hallucination [4]. Furthermore, from the inference aspect, prevailing decoding strategies provide randomness to enhance LLMs' creativity, but may also lead to unfactual responses, thereby increasing the risk of hallucinations [5].

In addition to prior work's broad factors contributing to hallucination, the prior evaluations for hallucination are one-sided and inefficient. For instance, TruthfulQA [6], HalluQA [7] and ChineseFactEval [8] are designed to assess the truthfulness of LLMs and they require manual or a GPT-like model for evaluation, which is costly and inefficient. Besides, Med-HALT [9] evaluates hallucination through responses' accuracy on test questions, but it only focuses on the medical domain.

This paper addresses both of the above open questions by proposing an actionable definition for LLMs' hallucination from the memory consistency perspective and a straightforward yet effective evaluation based on this definition.

We first define LLMs' hallucination as the phenomenon that model's responses are inconsistent with the model's memory. We categorize hallucinations into two distinct types, internal hallucination and external hallucination, based on two types of memory, pre-trained memory and augment memory, which could help to explore the specific factors contributing to hallucination.

Then, we construct a corresponding evaluation methods for two types of hallucination, which are very simple and can induce hallucination stably. For internal hallucination, we first construct a set of data pairs confirmed to be held by models. By combining and shuffling them randomly, we check whether the models are still able to correctly match the data pairs. This allows us to measure internal hallucination across different models, data pairs of varying sizes and different shuffling strategies. We observe that different models performs varies under the same situation, but as the size of data pairs larger, all models perform worse. For external hallucination, we use the dataset from summarization tasks which containing enough knowledge for LLMs' responses. By rephrasing them semantically consistent, we check the difference of the model's responses of these rephrasing questions between normal responses' distribution. This allows us to observe external hallucination across different models.

II. Related Work

A. Hallucination in LLMs

The issue of hallucination in LLMs has gained significant attention due to its negative impact on performance and the risks it introduces in various NLP tasks, such as machine translation [10], summarization [11], dialogue generation [12], and question answering [13]. Recent surveys [14] [15] have highlighted the importance of addressing this issue.

1) Hallucination Causes: Previous research has explored a lot about hallucination causes. While scaling up pre-training data enhances the capabilities of LLMs, the duplicated data [16] and data with certain biases [17] like gender and nationality are tied to hallucinations. Besides, because model relies on its own generated tokens during inference, if a erroneous token is generated, it will have a cascading effect on the subsequent sequence [18], resulting in hallucination. However, all of these are broad causes that impact the capabilities of LLMs, but they are not have a tight connection with the hallucination phenomenon.

2) Hallucination Evaluations: Numerous research efforts have been made on the evaluation of hallucination. REALTIMEQA [19] offers real-time open-domin multiple-choice questions to validate LLMs' factuality and evaluate it through a multiple-choice format assessed by accuracy. FreshQA [20] evaluate LLMs' capability to identify questions with false premises with 600 hand-crafted and get results through human annotations. Additionally, SAC3 [21] and RealHall [22] concentrate on question-answering tasks, categorizing them into closed and open groups based on the availability of a reference text in the prompt, and assign labels to responses through human annotation. Although the above methods perform well in the scenarios they are designed for, their evaluation of hallucination is one-sided or costly.

B. Consistency in LLMs

An essential characteristic of logically valid intelligent systems is self-consistency, which entails that no two statements provided by the system contradict each other. Self-consistency of a LLM is defined as the invariance of responses across various types of semantics-preserving prompt transformations [23]. More researches [24] [25] have further enriched this definition, which demonstrates that self-consistency can significantly enhance the chain of thought reasoning in LLMs. Recent research has employed self-consistency to detect and evaluate hallucination, based on an instruction-tuned LLM [26].

C. Memorization in LLMs

Extensive prior work has shown that LLMs can memorize parts of their training data, which allows us to consider hallucination from the perspective of memory. For instance, 600 memorized training examples could be identified by querying the GPT-2 language model [27]. The following research [28] has observed that larger models could memorize much more than smaller ones and examples duplicated in datasets are more likely to being memorized. Specifically, prior work has demonstrated extraction attacks that recover memorized data, including URLs, phone numbers, and other personal information [29].

III. Methodology

A. Definition of Hallucination

To begin, we first propose a definition for hallucination of LLM:

Definition 1: Given a set of questions $\hat{Q} = \{\hat{q}_1, \hat{q}_2, ..., \hat{q}_n\}$, they are derived from the same original question q_0 and satisfy $C(\hat{q}_p, q) > T_{cq}, p = 1, ..., n$. C is a function to measure the semantic consistency between arguments. After the large language model f generate the corresponding set of responses $R = \{r_1, r_2, ..., r_n\}$ for $\hat{Q}$, a evaluating function $H(q, \hat{Q}, R)$ is employed to measure the semantic consistency and relevance between them. If $H(q, \hat{Q}, R) \leq T_{cs}$, this phenomenon is defined as hallucination.

In more detail, if hallucinated responses are inconsistent with the *pre-trained memory*, which means the knowledge embedded in the weights of the model during the pre-training phase, such as the response, "Hamlet is a tragedy written by this phenomenon is defined as hallucination.", we define this situation as *internal hallucination*. If hallucinated responses are inconsistent with the *augment memory*, which means the knowledge encapsulated within the questions, like the response, "Today is a sunny day", while today is rainy day. We define this situation as *external hallucination*.

Typically, hallucination of LLMs is defined as generated content that is nonsensical or unfaithful to the provided source content [21]. These hallucinations are further categorized into closed-domain hallucination and open-domain hallucination [30], depending on their contradiction with the source content. While this category is shared among various language generation tasks, the existence of task-specific variations makes it unable to guide quantitative evaluation in all tasks. Hence, we prefer ours in this paper as it is more actionable.

B. Quantifying Internal Hallucination via Pre-trained Memory Consistency Test

Having given a definition, we next describe our quantifying procedure, a high-level overview of which is provided in Fig.1. To utilize pre-trained memory consistency to evaluate internal hallucination, the first challenge is to check which is pre-trained memory when we can not have access to the model's training dataset. To overcome this, we choose elementary knowledge as our candidate evaluation dataset D_{pc}, such as the author of poetry and the capitals of countries,which has a high probability of being included in LLMs' training dataset. Furthermore, we query the model for each pair of data in D_{pc}, only the pairs correctly answered by the model can be included in the evaluation dataset D_p.

Contrary to the existing methods that evaluate LLM hallucination through math tasks, which may be influenced by the logic capabilities of LLM, we propose a simple matching

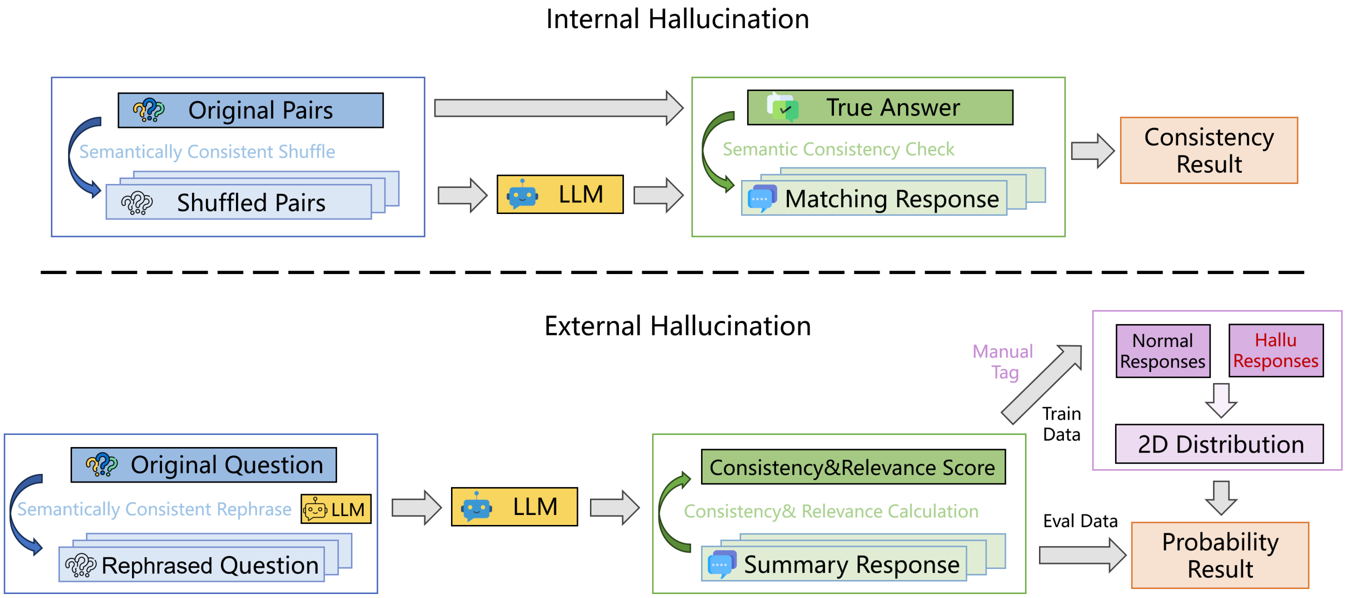

Fig. 1. Overview of hallucination exploring method.

task from the perspective of memory consistency. Firstly, we sample some pairs of data from D_p to generate the original question: $q_0 = \{\{a_1, b_1\}, \{a_2, b_2\}, ..., \{a_k.b_k\}\}$, where k is the number of sampled pairs and each pair of data consists of two parts a_i, b_i. Then, we just shuffle the order of q to generate semantically consistent questions:

$$\begin{aligned} \hat{q}_p &= \{\{a_1, b_{i_1}\}, \{a_2, b_{i_2}\}, ..., \{a_k, b_{i_k}\}\}, \\ p &= 1, ..., n, \;\; o = 1, ..., k, \;\; i_o \in \{1, ..., k\}, \end{aligned} \tag{1}$$

where n is the number of questions generated and i_o are the position of shuffled data. In this way, we could ensure that all the knowledge in the questions is contained in the models' pre-trained memory and test its consistency simply. Besides, since we just shuffle the order of pairs, $\hat{q}_p$ is naturally semantically consistent with q_0:

$$C_{in}(\hat{q}_p, q_0) = 1, \;\; p = 1, ..., n, \tag{2}$$

where C_{in} is the semantic consistency estimator function for this part. Also, it is obvious that q is the true answer for each $\hat{q}_p$.

After collecting LLM's responses $R = \{r_1, r_2, ..., r_n\}$ to $\hat{Q}$, we define H_{in} as:

$$H_{in}(q_0, \hat{Q}, R) = \sum_{p=1}^{n} h_{in} \tag{3}$$

$$h_{in} = \begin{cases} 1 & \text{if } r_p = q_0 \\ 0 & \text{if } r_p \neq q_0 \end{cases}, \;\; p = 1, ..., n, \tag{4}$$

where h_{in} is the judgment function if the response is true.

C. Quantifying External Hallucination via Augment Memory Consistency Test

Since augment memory is within the questions, data in the evaluation dataset need to contain enough knowledge for LLM to generate true response. Thus, we choose the dataset for the summarization task as our evaluation dataset D_a. To test evaluate external hallucination through augment memory consistency, we first need to ensure the consistency of the knowledge contained in the questions. Hence, we propose an efficient way to rephrase the questions with semantic consistency to ensure it.

Contrary to existing techniques that assess semantic consistency through entailment or paraphrasing, we leverage advances in LLM prompting to rephrase the input question. Starting with a question q_0, we acquire a set of n semantically consistent inputs $\hat{Q} = \{\hat{q}_1, \hat{q}_2, ..., \hat{q}_n\}$ through the prompt: "You are a linguistic expert who specializes in text analysis and writing. Please help me rephrase the original text enclosed within the symbol {}, retaining the original text information as much as possible, and try to be consistent with the semantics while maintaining the same meaning." for a couple of time. Additionally, to ensure that $\hat{q}_p$ are not identical, we utilize $\{\hat{q}_1, ...\hat{q}_{p-1}\}$ as examples in prompt and add additional prompt as follows: "Please output text different from examples, the examples are enclosed within symbols []." Finally, to ensure the quality of the generated questions in this step, we utilize BERTSCORE [31], which is an automatic similarity evaluation metric for text generation to check the semantic consistency between $\hat{q}_p$ and q_0, and only $\hat{q}_p$ with high enough score over the threshold T_{cq} can be adopted:

$$\begin{aligned} C_{ex}(\hat{q}_p, q_0) = & BERTSCORE(\hat{q}_p, q_0), \\ & p = 1, ..., n, \end{aligned} \tag{5}$$

For the responses evaluation stage, let r denote the response from a LLM f based on a given question q. Our objective is to judge whether s is hallucinated by calculating its probability within the normal responses distribution G_{norm}. More specifically, we define it as a Gaussian distribution of two variables, semantic consistency and relevance. We add a relevance metric to complement the consistency metric because LLMs may generate consistently hallucinated responses that are not relevant with the question. Consequently, we define evaluating function for external hallucination as follows:

$$H_{ex}(q_0, \hat{Q}, R) = PDF_{G_{norm}}(h_{consis}(R), h_{rele}(\hat{Q}, R)) \tag{6}$$

$$h_{consis}(R) = \frac{\sum_{p=1}^{n} BERTSCORE(r_1, r_p)}{n} \tag{7}$$

$$h_{rele}(\hat{Q}, R) = \frac{\sum_{p=1}^{n} BM25(\hat{q}_p, r_p)}{n} \tag{8}$$

where $PDF_{G_{norm}}$ means probability distribution function for G_{normal}, h_{consis} measures semantic consistency within the set of responses R and h_{rele} measure semantic relevance between the response r_p and the question $\hat{q}_p$. In detail, we still use BERTSCORE to check the semantic consistency, but for the relevance metric, we choose BM25 [32], a ranking function used by search engines to estimate the relevance of documents to a given search query.

IV. Experiments

A. Experimental Setup

1) Data Preparation: For internal hallucination, we choose 366 most common poems as our candidate evaluation dataset from the most complete database of classical Chinese poetry collections, ”chinese-poetry” [33], which contains over 300,000 Chinese classical collections. After querying evaluation models, we adopt 273 of them in our evaluation dataset, the author of which could be correctly answered by models. For external hallucination, we use CNewSum [34], a Chinese news summarization dataset which consists of 304,307 documents and human-written summaries for the news feed. We choose 300 documents from it as our evaluation dataset, which contains news from multiple fields, such as politic news, economic news, sports news, social news, etc.

2) Target Models: We use glm-4 model from ZhipuAI, ERNIE-4.0-8K model from Baidu and qwen-max model from Aliyun, which are the prevalent Chinese LLMs.

3) Implementation Details: The evaluation is conducted using API services of each company. For all the models, we set the temperature to 0.95 to balance the creativity and stability of responses. For internal hallucination, we use the prompt ”You are an expert specializing in ancient Chinese poetry. Please match the following poems with the poets who wrote them. Please strictly follow the following rules for the answer format: the answer only consists of the poet’s names, separated by commas and does not contain any other content.” to guide models’ responses. For external hallucination, we set $T_{cq} = 0.73$ to balance the quality of generated questions and time consumption. Also, we use prompt as following to guide the responses: ”You are a linguistics expert who specializes in text analysis and writing. Please summarize the following text in one sentence, condensing the information as much as possible. The output contains only the summarized text, please do not add additional responses or explanations”. We use the probability of generating hallucination responses to evaluate the degree of hallucination in target models. We execute all experiments on 1 NVIDIA 3090 24G GPU.

B. Internal Hallucination Evaluation Results

1) More Internal Hallucination with Larger k: We begin by considering the impact of the number of sample pairs k on internal hallucination. In this case, we set the number of generated questions n as 100. TABLE 2 compares the severity of internal hallucination for different models with different k. We observe that as k becoming **larger**, all the models perform **worse** on internal hallucination. With $k = 10$, glm-4 and qwen-max even can not generate normal answer. Surprisingly, ERNIE-4.0-8K performs much better than other target models in all situations. For example, with $k = 5$, glm-4 and qwen-max can only response normally with a probability less than 15%, but ERNIE-4.0-8K performs well with a probability of 67%.

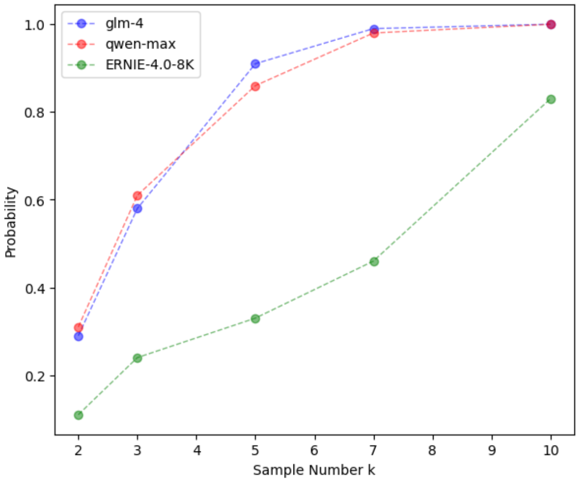

Fig. 2. Probability of Internal Hallucination Responses with Different Sample Numbers

2) Question Generating Strategies Have No Effect: We further explore the impact of different question generating strategies on evaluation results. We experiment two different strategies on model glm-4. The first strategy is just shuffle once and generate one question for all the combination of data pairs sampled out. The other one is to generate questions for all permutations of the combination of data pairs sampled out, $k = 2$ example of which is shown in TABLE I. Surprisingly, we observe that the model performs similarly across different strategies. TABLE II presents the performance of glm-4 under different strategies. Hence, we can conclude that different strategies have no effect on pre-trained memory consistency.

TABLE I
DIFFERENT QUESTION GENERATING STRATEGIES

Sampled Pairs	[”白日依山尽，黄河入海流。”, ”王之涣”] [”昔闻洞庭水，今上岳阳楼。”, ”杜甫”]
1^{st} **Strategy**	[”白日依山尽，黄河入海流。”, ”杜甫”] [”昔闻洞庭水，今上岳阳楼。”, ”王之涣”]
2^{ed} **Strategy**	[”白日依山尽，黄河入海流。”, ”王之涣”] [”昔闻洞庭水，今上岳阳楼。”, ”杜甫”]
	[”白日依山尽，黄河入海流。”, ”杜甫”] [”昔闻洞庭水，今上岳阳楼。”, ”王之涣”]

TABLE II
PROBABILITY OF INTERNAL HALLUCINATION RESPONSES WITH DIFFERENT QUESTION GENERATING STRATEGIES

Sample	**Question Generating Strategies**	
Number k	1^{st} **Strategy**	2^{ed} **Strategy**
2	0.29	0.31
3	0.58	0.56
5	0.91	0.97

C. External Hallucination Evaluation Results

1) Two-dimensional Distribution Better than Single Parameter: To estimate the normal responses distribution of models, we assume it is a two-dimensional Gaussian distribution(2D distribution): semantic consistency and relevance. To fit the 2D distribution, we tag 56 model responses, which contains 46 normal responses and 10 hallucination responses, shown in Fig.3. We further compare the area under the receiver operating characteristic curve (AUROC) to evaluate the performance between the 2D distribution and the single semantic metric. Through Fig.4, we observe that the classifier based on the 2D distribution performs much better, with a high AUROC of 0.91.

2) Models' External Hallucination Performance Similar to Internal One's: After got the 2D distribution, we evaluate target models' performance for external hallucination, with hyperparameters $n = 60$ and $k = 4$. As shown in TABLE 5, *ERNIE-4.0-8K* still performs much better than other two models, with a hallucination probability of 20%.

V. CONCLUSION

Our paper presents a more actionable definition of hallucination in LLM through the perspective of memory consistency, categorizing it into internal and external hallucination based on two types of memory: pre-trained memory and augment memory. We have demonstrated two straightforward evaluation methods for each type of hallucination, yielding the following insights.

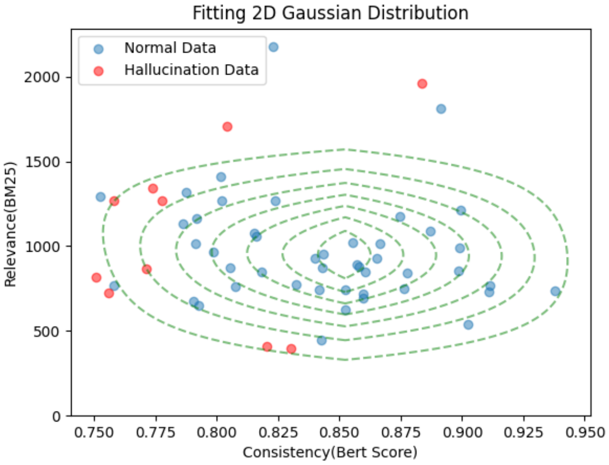

Fig. 3. Data Points and Two-dimensional Gaussian Distribution

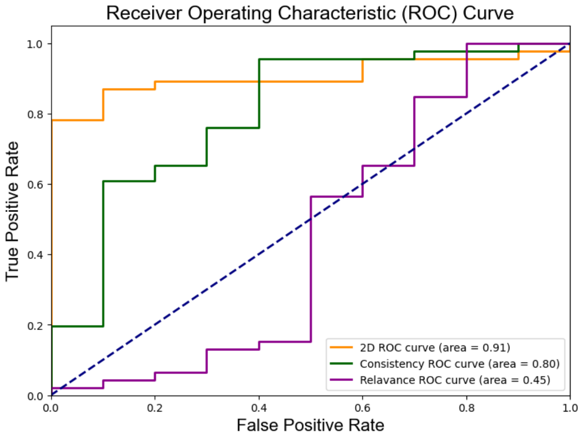

Fig. 4. AUROC for Different Evaluation Metrics

For the evaluation of internal hallucination, we have shown that simply shuffling data pairs containing elementary knowledge governed by LLMs can induce internal hallucination. In particular, by simply containing more data pairs in one question, we can induce more hallucination responses, and when the number of data pairs is large enough, all the test LLMs are not able to generate normal responses. It is therefore important to carefully analyze the causes of internal hallucination, as LLMs' tendency to generate responses inconsistent with its pre-trained memory is much worse than our expectation.

For the evaluation of external hallucination, our work employs a two-dimension Gaussian distribution with parameters of semantic consistency and relevance to model the normal response distribution of LLMs in the text summarization task. Our findings demonstrate the effectiveness of this distribution in detecting hallucination responses and evaluating the degree of external hallucination across different LLMs. This approach holds promise for future evaluation research, providing a

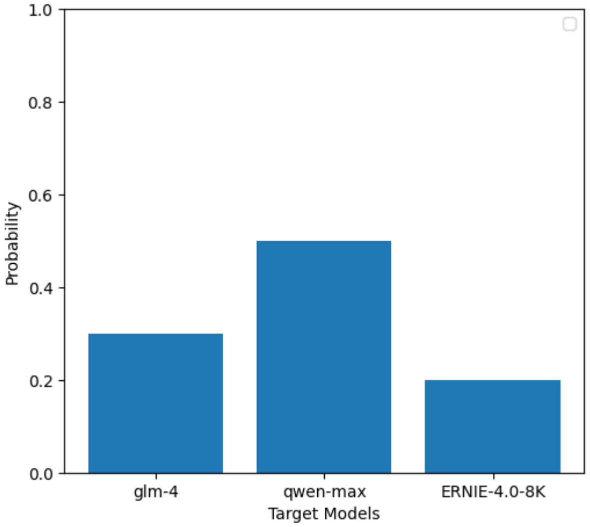

Fig. 5. Probability of External Hallucinationpng

valuable method to assess external hallucination phenomena.

References

[1] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[2] N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro, "Factuality enhanced language models for open-ended text generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 586–34 599, 2022.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big? ‘‘‘‘," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610‘‘623. [Online]. Available: https://doi.org/10.1145/3442188.3445922

[4] Z. Li, S. Zhang, H. Zhao, Y. Yang, and D. Yang, "Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer," *arXiv preprint arXiv:2307.00360*, 2023.

[5] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," *arXiv preprint arXiv:2309.03883*, 2023.

[6] 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

[7] Q. Cheng, T. Sun, W. Zhang, S. Wang, X. Liu, M. Zhang, J. He, M. Huang, Z. Yin, K. Chen *et al.*, "Evaluating hallucinations in chinese large language models," *arXiv preprint arXiv:2310.03368*, 2023.

[8] B. Wang, E. Chern, and P. Liu, "Chinesefacteval: A factuality benchmark for chinese llms," 2023.

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

[10] C. Zhou, G. Neubig, J. Gu, M. Diab, P. Guzman, L. Zettlemoyer, and M. Ghazvininejad, "Detecting hallucinated content in conditional neural sequence generation," *arXiv preprint arXiv:2011.02593*, 2020.

[11] M. Cao, Y. Dong, and J. Cheung, "Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3340–3354. [Online]. Available: https://aclanthology.org/2022.acl-long.236

[12] S. Das, S. Saha, and R. K. Srihari, "Diving deep into modes of fact hallucinations in dialogue systems," *arXiv preprint arXiv:2301.04449*, 2023.

[13] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," *arXiv preprint arXiv:2309.11495*, 2023.

[14] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[15] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[16] D. Hernandez, T. Brown, T. Conerly, N. DasSarma, D. Drain, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume *et al.*, "Scaling laws and interpretability of learning from repeated data," *arXiv preprint arXiv:2205.10487*, 2022.

[17] F. Ladhak, E. Durmus, M. Suzgun, T. Zhang, D. Jurafsky, K. McKeown, and T. Hashimoto, "When do pre-training biases propagate to downstream tasks? a case study in text summarization," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 3206–3219. [Online]. Available: https://aclanthology.org/2023.eacl-main.234

[18] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," *arXiv preprint arXiv:2305.13534*, 2023.

[19] J. Kasai, K. Sakaguchi, R. Le Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, K. Inui *et al.*, "Realtime qa: What's the answer right now?" *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le *et al.*, "Freshllms: Refreshing large language models with search engine augmentation," *arXiv preprint arXiv:2310.03214*, 2023.

[21] J. Zhang, Z. Li, K. Das, B. A. Malin, and S. Kumar, "Sac ˇ3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency," *arXiv preprint arXiv:2311.01740*, 2023.

[22] R. Friel and A. Sanyal, "Chainpoll: A high efficacy method for llm hallucination detection," *arXiv preprint arXiv:2310.18344*, 2023.

[23] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg, "Measuring and improving consistency in pretrained language models," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1012–1031, 2021.

[24] M. Jang, D. S. Kwon, and T. Lukasiewicz, "Becel: Benchmark for consistency evaluation of language models," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 3680–3696.

[25] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[26] N. Mündler, J. He, S. Jenko, and M. Vechev, "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation," *arXiv preprint arXiv:2305.15852*, 2023.

[27] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[28] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying memorization across neural language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=TatRHT_1cK

[29] A. Ziegler, "Github copilot: Parrot or crow," 2021.

[30] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," 2020.

[32] S. Robertson, H. Zaragoza *et al.*, “The probabilistic relevance framework: Bm25 and beyond,” *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[33] jackeyGao, HongzhiW, FleetingWang *et al.*, “https://github.com/chinese-poetry/chinese-poetry,” 2023.

[34] D. Wang, J. Chen, X. Wu, H. Zhou, and L. Li, “Cnewsum: a large-scale chinese news summarization dataset with human-annotated adequacy and deducibility level,” *arXiv preprint arXiv:2110.10874*, 2021.