IJWR

Consistent Responses to Paraphrased Questions as Evidence Against Hallucination: A Study on Hallucinations in LLMs

Tara Zare[a], Mehrnoush Shamsfard[b]*

[a] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran; tara99zare@gmail.com

[b] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran; m-shams@sbu.ac.ir

ABSTRACT

The increasing adoption of large language models (LLMs) has intensified concerns about hallucinations—outputs that are syntactically fluent but factually incorrect. In this paper, we propose a method for detecting such hallucinations by evaluating the consistency of model responses to paraphrased versions of the same question. The underlying assumption is that if a model produces consistent answers across different paraphrases, the output is more likely to be accurate. To test this method, we developed a system that generates multiple paraphrases of each question and analyzes the consistency of the corresponding responses. Experiments were conducted using two LLMs—GPT-4O and LLaMA 3–70B Chat—on both Persian and English datasets. The method achieved an average accuracy of 99.5% for GPT-4O and 98% for LLaMA 3–70B, indicating the effectiveness of our approach in identifying hallucination-free outputs across languages. Furthermore, by automating the consistency evaluation using an instruction-tuned language model, we enabled scalable and unbiased detection of semantic agreement across paraphrased responses.

Keywords— ***Large Language Models, Hallucination of Large Language Models, Inconsistency Detection, Paraphrasing***

1. Introduction

Large language models are models designed for modeling human languages using self-supervised learning [1] and employing special training methods to improve the performance of unseen tasks and better adhere to natural language instructions [2]. These models can be applied to various tasks, including question answering [3], code generation [4], and solving mathematical problems [5].

Despite the many strengths and useful applications of large language models, they have significant weaknesses in generating content that appears correct and fluent but is actually incorrect. Generating seemingly correct but factually or conceptually inaccurate content is referred to as hallucination. The reason for this naming goes back to the definition of the term "hallucination" in psychology. Bloom [6] defines hallucination as "a perception experienced by an awake individual in the absence of an appropriate external stimulus." In simpler terms, hallucination is an unreal perception that appears real. The generation of incorrect, meaningless, or source-unfaithful information by language models shares a similar trait with such psychological hallucinations. Just as a delusional person is unaware of their hallucination, believing the events they experience to be real, a large language model does not produce incorrect information with prior intent and awareness. In fact, the content of these incorrect outputs did not appear during the model's training and are the inventions of the model itself. Given the rapid advancement of large language models, there is considerable concern about their tendency to generate hallucinations, which results in seemingly acceptable and fluent content without a correct source [7].

This definition of hallucination aligns with other research that considers hallucination as meaningless content or unfaithful to the provided input [8].

doi http://dx.doi.org/10.22133/ijwr.2025.524365.1283

Citation T. Zare, M. Shamsfard, " Consistent Responses to Paraphrased Questions as Evidence Against Hallucination: A Study on Hallucinations in LLMs", *International Journal of Web Research*, vol.8, no.4,pp.15-23, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.524365.1283.

*Corressponding Author

Article History: Received: 18 May 2025; Revised: 15 September 2025; Accepted: 27 September 2025.

The hallucinations present in the output of large language models are classified into two types: Factuality Hallucination and Faithfulness Hallucination [9]. Factuality Hallucination refers to the inconsistency between the generated content and verifiable facts in the real world. For example, if a language model is asked, "Who is the first female president of Iran?" and it produces any name in the output, we have a factuality hallucination. Faithfulness hallucination refers to the divergence of the generated content from the user's instructions or the context provided by the input. For instance, suppose we provide a text in Persian or another language to a large language model and ask it to translate it into another language. Any error in the output of this request would indicate a Faithfulness Hallucination.

In other sources, alternative terms have been used instead of factuality hallucination and faithfulness hallucination. For example, some sources [10] categorize hallucinations caused by large language models into two types: open-domain and closed-domain. Open-domain hallucinations are essentially false claims about the world around us generated by large language models. Closed-domain hallucinations involve the model's output deviating from a specific source or reference text. For instance, if a language model makes a mistake in summarizing a text and produces information that is contradictory and inconsistent with the input source, this incorrect information falls into the category of closed-domain hallucinations.

Given the diverse and widespread applications of large language models in various fields, we are witnessing a rapid increase in their use in different applications. It is clear that if the output of these models is prone to hallucinations, these errors will propagate throughout these applications and undermine user trust. Therefore, to improve systems or conversational agents that directly or indirectly rely on these models, it is better to detect these errors at the output stage of the large language models and prevent them from entering other applications.

The importance of detecting these hallucinations primarily lies in not losing user trust. Additionally, if we do not prevent this incorrect information from propagating into downstream systems, it can lead to more significant errors in more sensitive areas such as medicine and healthcare. Imagine a language model hallucinating while translating a patient's medication instructions. These hallucinations generated by the large language model could pose a life-threatening risk to the patient if they follow these instructions.

So far, we have highlighted the importance of detecting hallucinations and preventing the propagation of erroneous outputs into downstream systems. To address this, we aim to develop a method that can determine whether the output of an LLM is hallucinatory. The purpose of this study is to systematically examine whether consistency across paraphrased questions can serve as a reliable signal for detecting hallucinations in large language models. This study serves as an extended version of our previous work [11], in which the evaluation was conducted exclusively on the Persian language. Moreover, while our previous approach relied on manual semantic consistency judgments, we automated this process, enabling scalable, annotation-free evaluation over response sets**.** These additions enabled a more thorough and efficient investigation of the accuracy and reliability of model outputs across languages and model types.

The contributions of this study are summarized as follows:

1. Proposing a consistency-based framework for hallucination detection in large language models
2. Extending our previous framework from Persian to English, enabling a multilingual evaluation.
3. Evaluating the method on multiple LLMs (GPT-4O and LLaMA 3–70B), demonstrating robustness across models.
4. Automating the semantic consistency evaluation process, making the method scalable and efficient.

In the next sections, we will review related work, describe our proposed method, and discuss experiments and results. The final section will present the conclusion and future work.

2. Related Work

The research on hallucinations generated by large language models is categorized into several areas: benchmark creation, detecting hallucinations, and reducing hallucinations in large language models. Here, we review studies on benchmark creation, and on detecting and reducing or eliminating hallucinations.

Although many benchmarks designed for evaluating LLMs can also be used to detect hallucinations —particularly in assessing factual consistency— some benchmarks have been specifically developed for hallucination detection or are more commonly used in this context. Among these, the TruthfulQA benchmark, introduced in 2022, aims to evaluate LLMs by testing their responses [12]. This benchmark includes 817 diverse questions across 37 categories such as health, law, finance, politics, and more. Among these questions, 380 are non-adversarial, and 437 are adversarial.

Here, "adversarial" indicates that these questions are designed to mislead the language models and challenge their responses. Another benchmark introduced in 2023 is SelfCheckGPT-Wikibio [13]. During its construction, the WikiBio dataset [14], which contains Wikipedia-based biographical information, was used. A set of 238 relatively long articles was selected, and GPT-3 was utilized to generate Wikipedia-style statements in the same selected topics. These generated statements were then manually annotated with one of three labels: entirely false, partially false, or accurate. This resulted in a dataset of 1,908 sentences, creating a valuable resource for evaluating a model's ability to detect hallucinations in large language models. An important benchmark that gained attention in 2023 is HaluEval, designed to evaluate the tendency of large language models, like ChatGPT, to produce hallucinations [15]. This benchmark consists of an extensive set of examples generated either automatically or with human assistance. It includes 35,000 samples, covering 5,000 general questions with ChatGPT responses and 30,000 questions in more specialized tasks such as question answering, text summarization, knowledge-based conversation, and more. Another benchmark dataset, introduced in 2024, is ANAH, which provides fine-grained annotation of hallucinations in large language models. ANAH includes around 12,000 sentence-level annotations across more than 4,300 responses on over 700 topics. Each sentence is annotated with reference fragments, hallucination types (e.g., contradictory, unverifiable, or lacking fact), and corrections, creating a highly detailed resource for training and evaluating hallucination detection models.

In reviewing studies on detecting hallucinations in large language models, one notable approach is the SCALE architecture, which focuses on hallucination detection in longer texts [16]. This approach involves breaking down documents into smaller, more manageable sections and using a natural language inference model to detect inconsistencies within each section to identify hallucinations. Another study utilizes the SelfCheckGPT-Wikibio benchmark [12], which we previously discussed. In this study, each instruction is given multiple times to the language model, resulting in *N* outputs for the same prompt [12]. These *N* responses are then analyzed using various methods, such as BERTscore or N-grams. It's important to note that the initial *N* instructions were identical and unchanging. A recent approach, InterrogateLLM, leverages a method where the model's response is repeatedly queried to reconstruct the initial prompt, allowing inconsistencies in generated queries to reveal potential hallucinations [17]. This technique does not rely on external data, making it versatile and effective compared to methods like SelfCheckGPT [13]. Another recent method, LLM-Check, detects hallucinations from a single model response by exploiting internal model signals such as eigenvalue-based statistics over self-attention and hidden states, as well as token-level uncertainty measures like perplexity and logit-entropy [18]. This design makes the approach computationally efficient since it avoids generating multiple responses for the same input. However, its effectiveness depends on access to internal representations or on the availability of a suitable proxy model through teacher-forcing, which defines its main limitation. Another approach, Lookback Lens, targets contextual hallucinations by analyzing attention maps during generation [19]. It introduces a simple feature called the lookback ratio, measuring the proportion of attention weights allocated to the provided context versus newly generated tokens. Using these ratios, a lightweight classifier can detect hallucinations and even be integrated into decoding to mitigate them. While efficient and transferable across tasks and models, the method depends on access to attention patterns and annotated examples for training.

In the area of reducing hallucinations, recent methods, such as TruthX introduced in 2024, employ techniques to edit internal model representations in real-time [20]. By aligning these representations with a "truthful space", TruthX enhances the model's accuracy and reduces hallucinations by approximately 20% on benchmarks like TruthfulQA (2024.acl-long.178).

3. Proposed Methods

3.1 Dataset and Data Collection

To evaluate the proposed method, we constructed a test set of 200 factoid questions with unique, verifiable answers, balanced across two languages: 100 in Persian and 100 in English. The questions cover five categories—(1) History, Politics, and Culture, (2) Music and Cinema, (3) Basic Sciences, (4) Arts and Literature, and (5) Technology—with 20 questions per category in each language.

For each original question, we generated ten paraphrases using GPT-4O in a separate interaction conducted prior to evaluation. From these, four paraphrases with the highest semantic similarity and grammatical correctness were manually selected, yielding five versions per question (the original plus four paraphrases). In total, the dataset comprises 1,000 inputs (200 questions × 5 versions).

During evaluation, each input was submitted to the target language models (GPT-4O and LLaMA-3–70B Chat), and the corresponding outputs were stored for analysis. Figure 1 provides an overview of the pipeline: paraphrase generation, response collection, and consistency analysis across the five versions of each question.

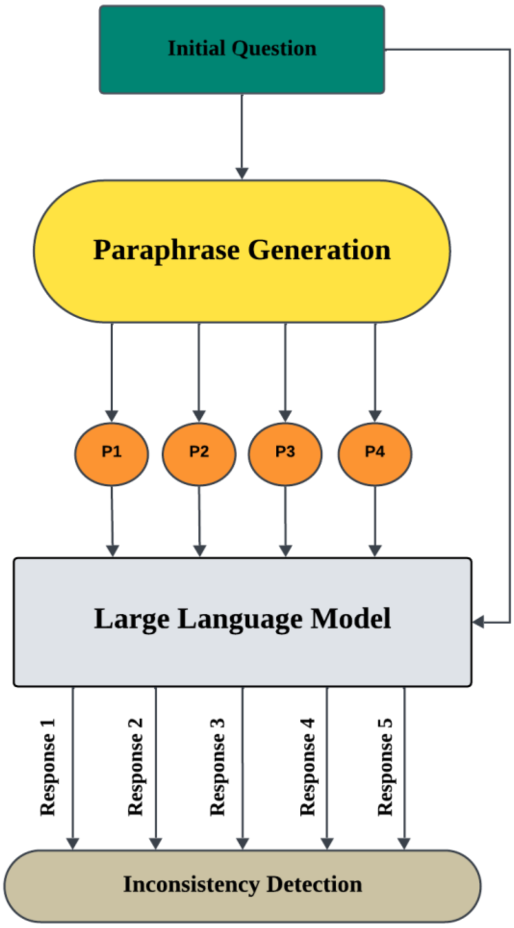

Figure. 1. Overall architecture of the proposed algorithm for testing the initial claim, involving the generation of multiple paraphrases for each original question, storing all corresponding responses, and analyzing the consistency or inconsistency among

3.2 Method Overview

Based on an analysis of the results of language models, a claim was proposed: if large language models know the answer to a specific question, their responses to that question will remain consistent and uniform, even when the question is phrased differently.

To verify this claim, each factoid question was presented in multiple paraphrased forms, and the model's responses were analyzed for agreement. The details of dataset construction are already provided in Section 3.1.

At this stage, it is necessary to examine the presence or absence of inconsistencies among the different responses produced by the LLM for various questions within a single set. For inconsistency, we refer to its simplest and most fundamental meaning: two statements are contradictory if the truth of one statement implies the falsehood of the other. In other words, the generated responses are mutually exclusive. For example, if a large language model is asked, "Who won the Nobel Prize in Literature in 2020?" we expect it to produce only the name "Louise Glück", because in that year, the Nobel Prize in Literature was awarded to her alone. Any other name that does not refer to her would be inconsistent and considered an incorrect answer. However, if the question posed is, "Who won the Nobel Prize in 2020?" the answer is not unique, as the Nobel Prize is awarded in various fields to different individuals. In such cases, the model might produce names corresponding to different fields, and this cannot be classified as a contradiction. Therefore, our question selection enforces uniqueness of the correct answer as any inequality in the responses to their paraphrased versions would be interpreted as inconsistency.

It is important to note that, during consistency checking, we did not rely solely on exact string matching between responses. Instead, we adopted a more semantic approach, where responses were considered consistent if they referred to the same real-world entity or concept, even if their wording differed. While we did not perform formal entity linking—i.e., mapping textual mentions to entries in a structured knowledge base—our method was inspired by its core idea: treating different expressions that refer to the same underlying entity as equivalent. For example, if one response mentioned "Paris" and another stated "the capital of France," both were interpreted as consistent because they referred to the same real-world entity. By applying an entity-level semantic equivalence criterion, we aimed to avoid misclassifying semantically valid but lexically diverse responses as inconsistent. This approach enabled a more accurate and fair evaluation of the consistency of large language model outputs.

In the next step, we analyze the results against our initial claim, using the correct answers that were predefined for each question set based on verified factual sources. According to the initial claim, if a large language model knows the answer to a question, it should produce consistent and non-contradictory outputs for all the paraphrased versions of that question. In other words, if the language model produces a unique answer for a set of questions, it can be concluded that the response is most probably free from hallucination.

The outputs of our evaluation can fall into three categories. In the first case, the model generates a correct and consistent response aligned with factual information. In the second case, the model produces incorrect or hallucinatory information. In the final case, the output of the large language model is free from hallucination but incomplete compared to the correct answer. While it is evident that this category does not involve hallucination, due to the incompleteness of the response, it is not considered part of the first category, i.e., correct and complete answers.

In our previous work, we examined the validity of this claim using the large language model GPT-3.5 [11]. Here, we extend that study by evaluating the claim across two languages and multiple LLMs (GPT-4O and LLaMA-3–70B Chat), with category coverage and data scale specified in Section 3.1, enabling finer-grained analysis across languages, models, and domains.

4. Experimental Results

We used 1,000 initial inputs, which were provided to the two targeted large language models, GPT-4O and LLaMA 3–70B Chat. The outputs were recorded. Based on the explanations provided earlier regarding the type of questions and the uniqueness of answers within each set, Table 1 shows the QA accuracy of the GPT-4O and LLaMA 3–70B Chat models in generating correct answers across categories and subcategories, for both Persian and English. As evident, the best performance belongs to the GPT-4O model in English, while the lowest performance is observed for LLaMA 3–70B Chat in Persian. Furthermore, on average across both languages and models, the highest QA accuracy was achieved in the basic sciences category, whereas the music and cinema category consistently demonstrated the lowest accuracy.

These variations can be explained by differences in both model capacity and data availability. GPT-4O consistently demonstrates fewer hallucinations compared to LLaMA 3–70B, which can be attributed to its larger scale and training on more diverse and extensive datasets. Likewise, the higher accuracy observed in English compared to Persian is a natural outcome of the greater abundance and reliability of English resources, while the relative scarcity of high-quality Persian data makes hallucination more frequent in that language. The contrast across domains follows a similar pattern: scientific questions, which rely on well-established and standardized knowledge, tend to yield more consistent and accurate responses, whereas domains such as cinema and music involve more dynamic, less standardized, and culturally diverse information, leading to lower accuracy.

We next evaluate the validity of our initial claim and first note a challenge observed during evaluation.

One challenge concerned the granularity of information in certain specific topics. For instance, the question asked was, "Which country has the longest national anthem?" and the GPT-4O model consistently answered "Greece" for all responses. Upon further investigation, we discovered that the answer to this question can vary depending on the level of granularity considered. This is because the concept of the "longest national anthem" can be divided into two distinct interpretations: the longest musical duration of a national anthem and the longest text of a national anthem. Currently, the official national anthem with the longest musical duration belongs to Uruguay. This distinction arises because Greece, despite having an exceptionally long anthem text, has chosen to shorten its rendition by singing only selected verses. In contrast, Uruguay has not made such modifications, and as a result, the longest official national anthem in terms of music duration is attributed to Uruguay. Armed with this nuanced understanding, we reformulated the question to explicitly incorporate such granularity. However, despite this adjustment, the model continued to consistently answer "Greece" for all questions in this category.

Given these considerations, and the limitations of our knowledge and the discrepancies in the information provided by various sources, it is plausible that a more accurate answer exists but remains inaccessible to us or does not appear in our searches. Considering all these limitations, we can ultimately observe the accuracy of our claim in Table 2. Based on the benchmark answers available to us, the claim was found to be correct in 98.75% of cases on average.

This finding indicates that consistency across paraphrased questions serves as a highly reliable signal of factual accuracy. In other words, when a model provides the same answer to different phrasings of a question, the probability that this answer is hallucination-free is extremely high. For example, if the GPT-4O language model provides consistent answers to all different questions within a category, 99% of the time, this consistent response is correct and free from hallucination. Additionally, according to the results presented in Table 2, it can be observed that our approach achieved notably better outcomes when evaluated specifically with the GPT-4O language model.

Compared with our previous study [11], which validated the hypothesis only on GPT-3.5 and Persian data, the present work extends the evaluation to multiple advanced LLMs (GPT-4O and LLaMA-3–70B) and to both Persian and English. While the earlier study already reported about 99% validity of the hypothesis, our new experiments confirm that this result generalizes across models and languages, highlighting the robustness of the proposed framework.

Now that we have assessed the validity of the stated claim, we aim to explore whether, in cases where not all responses to questions within a category are identical, the most frequent response can still lead us to the correct and accurate answer. For this purpose, two types of response recurrence can be considered:

1. Majority Response: A response that constitutes the majority of answers provided.

Table 1. QA accuracy (Percentage of correct, partially correct, and hallucinatory outputs generated by large language models GPT-4O and LLaMA 3–70B Chat across various question sets within different categories in Persian and English.)

	LLM	***Language***	***Correct***	***Partially Correct***	***Hallucination***
Overall	GPT-4O	Persian	78%	4%	18%
	GPT-4O	English	92%	1%	7%
	LLaMA 3–70B	Persian	65%	0	34%
	LLaMA 3–70B	English	87%	0	13%
Cinema and Music	GPT-4O	Persian	70%	5%	25%
	GPT-4O	English	80%	0	20%
	LLaMA 3–70B	Persian	50%	0	50%
	LLaMA 3–70B	English	75%	0	25%
Art and Literature	GPT-4O	Persian	75%	0	25%
	GPT-4O	English	95%	0	5%
	LLaMA 3–70B	Persian	40%	0	60%
	LLaMA 3–70B	English	80%	0	20%
Culture, History and Politics	GPT-4O	Persian	75%	0	25%
	GPT-4O	English	90%	5%	5%
	LLaMA 3–70B	Persian	70%	0	30%
	LLaMA 3–70B	English	95%	0	5%
Science	GPT-4O	Persian	85%	0	15%
	GPT-4O	English	95%	0	5%
	LLaMA 3–70B	Persian	80%	0	20%
	LLaMA 3–70B	English	100%	0	0
Technology	GPT-4O	Persian	85%	15%	0
	GPT-4O	English	100%	0	0
	LLaMA 3–70B	Persian	85%	0	10%
	LLaMA 3–70B	English	85%	0	15%

2. Most Frequent Response: A response with the highest frequency among all responses, but not necessarily forming a majority.

By analyzing these two types of recurrence, we aim to determine if the most frequent or majority response can reliably guide us toward the correct answer in cases of inconsistency. As shown in Table 3 we observe how the consideration of "Majority" and "Most Frequent" differs across large language models for each language. For LLaMA 3–70B Chat, frequency analysis did not reliably indicate the correct answer. The largest benefit of this heuristic was observed with the GPT-4O model in Persian, where, as indicated, the "Majority" and "Most Frequent" heuristics coincided.

Approaches for evaluating response agreement: exact string matching and semantic equivalence-based evaluation. In the exact match method, any lexical variation between answers—even if they were semantically identical—was treated as inconsistency. In contrast, our semantic approach considered answers consistent if they referred to the same real-world entity or concept, even when expressed differently. The results of this comparison are presented in Table 4. As shown, the semantic evaluation method significantly reduced the number of false inconsistency detections, leading to higher consistency evaluation accuracy in identifying truly consistent answer sets.

In addition to accuracy, we also report recall, which measures the proportion of truly consistent answer sets correctly identified by the evaluation method. Precision, however, is not included, because under exact string matching the number of false positives is always zero—that is, identical strings are never mistakenly labeled as inconsistent—making precision trivially equal to 100%. Table 4 therefore highlights improvements in both accuracy and recall when adopting the semantic evaluation method.

To automate the semantic consistency evaluation, we utilized the LLaMA 3–3B Instruct (Turbo) model. For each pair of answers generated from paraphrased versions of the same factoid question, the model was prompted with a fixed instruction in Persian and English, asking whether the two answers could both serve as valid and equivalent ressponses to a single factual question. The model returned a binary decision (0 or 1), indicating inconsistency or equivalence, respectively. This approach enabled fast, scalable, and language-aware consistency checking without human supervision.

Table 5 summarizes the agreement between the model's automatic consistency judgments and human-verified labels. As shown, the model achieves high alignment across both languages, indicating that it can reliably distinguish between semantically consistent and inconsistent answers.

For human labeling, the annotations were performed by the author following the same definition of inconsistency described in Section 3. Specifically, the responses generated for each paraphrased question were compared against the response to the original question, and whenever a contradiction was detected, the pair was labeled as inconsistent; otherwise, it was labeled as consistent. This set of human annotations was treated as the gold standard. The automatic evaluation was then carried out using the same definition, and the model’s binary decisions were compared against these gold labels. Agreement was measured as the percentage of cases where the model’s decision matched the human annotation.

Table 2. Claim accuracy across GPT-4O and LLaMA 3–70B Chat in Persian and English.

LLM / *Language*	***GPT-4O***	***LLaMA 3–70B Chat***
Persian	99%	100%
English	99%	97%

Table 3. Impact of "majority" and "most frequent" response counting on identifying the correct answer in GPT-4O and LLaMA 3–70B Chat models across Persian and English languages.

LLM	***Language***	***Majority***	***Most Frequent***
GPT-4O	Persian	61.1%	61.1%
	English	42.8%	57.1%
LLaMA 3–70B Chat	Persian	17.6%	29.4%
	English	30.7%	30.7%

Table 4. Improvement in consistency evaluation accuracy and Recall achieved by semantic equivalence evaluation compared to exact matching.

Language	***LLM***	***Accuracy Improvement***	***Recall Improvement***
Persian	GPT-4O	14%+	15.3%+
English		12.25%+	12.7%+
Persian	LLaMA 3–70B Chat	10.5%+	27.6%+
English		7.75%+	8%+

Table 5. Agreement of automatic consistency judgments by LLaMA-3.3-70B-Instruct-Turbo with human-verified labels.

Language	***LLM***	***Agreement with Human Labels (%)***
Persian	GPT-4O	97.5%
English		98.75%
Persian	LLaMA 3-70B Chat	96%
English		99%
Overall		97.8%

5. Conclusion and Further Work

This study validated the claim that the consistency of a large language model’s responses to paraphrased versions of the same question serves as a strong indicator of the factual correctness of its outputs. Through extensive experiments conducted on both Persian and English datasets, and by utilizing two advanced large language models—GPT-4O and LLaMA 3–70B Chat—we found that in 99% of the cases where the model's responses were consistent, the outputs were free from hallucinations. These results confirm that consistency checking is an effective, language-agnostic approach for identifying hallucination-free outputs. Furthermore, by employing a semantic-based evaluation rather than relying solely on exact string matching, the methodology demonstrated greater robustness in recognizing semantically correct but lexically diverse responses. This improvement was reflected in a measured increase of 11.13% in consistency detection accuracy, highlighting the tangible benefits of adopting a semantic approach. This enhances the practicality and fairness of consistency-based hallucination detection across languages and models.

Compared to our previous study, which was limited to a single model and the Persian language, this work expanded the evaluation to include both Persian and English, larger dataset, and multiple LLMs. The results clearly demonstrate that our method is not tied to any specific model or language. Instead, it can be readily applied to other languages and architectures, as evidenced by the consistent improvements observed across diverse settings in this study. This establishes the extendibility of the proposed approach beyond the initial scope, showing that it generalizes well to broader contexts.

In this study, efforts were made to address additional limitations of the previous research [11], such as designing more challenging questions and automating the consistency evaluation process. By leveraging the LLaMA-3.3-70B-Instruct-Turbo model and providing it with a controlled prompt in Persian and English, we automated pairwise consistency judgments between paraphrased responses without the need for human intervention. Our evaluation showed that this automatic process achieved an average agreement of 97.8% with human-verified labels, demonstrating both reliability and scalability.

One of these challenges is the complete automation of the inconsistency detection process. Due to the scarcity of resources in low-resource languages (such as Persian), more efficient solutions must be sought. Additionally, it is essential to develop a method to identify hallucination in the outputs of paraphrased questions in cases where the outputs are inconsistent, and it is not possible to determine the correct answer among the responses. This is because, as observed in the results section, it is not always feasible to deduce the correct answer by identifying the most frequently occurring response in all large language models. Consequently, there is a need for a solution that, in the first step, predicts whether each output is hallucinatory, and in the second step, identifies the correct answer with a higher success rate.

Finally, one of the most significant and persistent challenges encountered in both series of studies has been determining the factual answer to each question. This challenge arises due to the abundance of resources and their variations across many topics, as well as the scarcity of resources in certain categories, making this issue highly contentious.in

In addition to these directions, future work will also focus on developing a predictive mechanism that directly determines whether a given output is hallucinatory. Another promising line of work is to identify the correct answers among multiple paraphrased responses in cases where inconsistencies arise, thereby moving beyond mere detection to answer validation. Future work may also extend beyond factoid questions with unique answers to cases where multiple correct answers exist, where the challenge shifts from hallucination detection to managing incompleteness. We further plan to expand our dataset in both size and topical diversity, covering a broader set of domains and questions, and to evaluate the framework on an even larger set of LLMs. These extensions will enable more comprehensive evaluations and enhance the generalizability of the proposed approach.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

TZ: Study design, data collection, data analysis, interpretation of results, drafting the manuscript.

MS: Study design, interpretation of results, revision of the manuscript, supervision.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] T. Brown *et al.*, “Language Models Are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] V. Sanh *et al.*, “Multitask Prompted Training Enables Zero-Shot Task Generalization,” in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Apr. 2022. https://openreview.net/pdf?id=9Vrb9D0WI4

[3] R. Thoppilan *et al.*, “LaMDA: Language Models for Dialog Applications,” *arXiv preprint arXiv:2201.08239*, 2022. https://doi.org/10.48550/arXiv.2201.08239

[4] M. Chen *et al.*, “Evaluating Large Language Models Trained on Code,” *arXiv preprint arXiv:2107.03374*, 2021. https://doi.org/10.48550/arXiv.2107.03374

[5] A. Lewkowycz *et al.*, “Solving Quantitative Reasoning Problems with Language Models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 3843–3857, Dec. 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html

[6] J. D. Blom, *A Dictionary of Hallucinations*. New York, NY: Springer, 2010. https://doi.org/10.1007/978-1-4419-1223-7

[7] Y. Bang *et al.*, “A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity,” in *Proceedings of the 13th International Joint Conference on Natural Language Processing (IJCNLP-AACL 2023)*, Bali, Indonesia, Nov. 2023, pp. 675–718. https://doi.org/10.48550/arXiv.2302.04023

[8] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Comput. Surv.*, vol. 55, no. 12, p. 248:1-248:38, Mar. 2023, https://doi.org/10.1145/3571730

[9] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Art. no. 28, pp. 1–55, 2025, https://doi.org/10.1145/3703155

[10] R. Friel and A. S. Sanyal, "ChainPoll: A High Efficacy Method for LLM Hallucination Detection," *arXiv preprint arXiv:2310.18344*, 2023, https://doi/org/10.48550/arXiv.2310.18344

[11] T. Zare and M. Shamsfard, "Detecting Hallucinations Generated by Large Language Models Using Paraphrasing Technique," in *Proceedings of the 10th International Web Research Conference (ICWR)*, Tehran, Iran, Apr. 2024, pp. 1–6. https://www.sid.ir/paper/1147671/en

[12] 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. https://doi.org/10.18653/v1/2022.acl-long.229

[13] P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

[14] T. Liu, K. Wang, L. Sha, B. Chang, and Z. Sui, "Table-to-Text Generation by Structure-Aware Seq2seq Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Art. no. 1, Apr. 2018, https://doi.org/10.1609/aaai.v32i1.11925

[15] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, Dec. 2023, pp. 6449–6464, https://doi.org/10.18653/v1/2023.emnlp-main.397

[16] B. M. Lattimer, P. H. Chen, X. Zhang, and Y. Yang, "Fast and Accurate Factual Inconsistency Detection Over Long Documents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, Dec. 2023, pp. 1691–1703, https://doi.org/10.18653/v1/2023.emnlp-main.105

[17] Y. Yehuda, I. Malkiel, O. Barkan, J. Weill, R. Ronen, and N. Koenigstein, "InterrogateLLM: Zero-Resource Hallucination Detection in LLM-Generated Answers," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9333–9347. https://doi.org/10.18653/v1/2024.acl-long.506

[18] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "LLM-Check: Investigating Detection of Hallucinations in Large Language Models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34188–34216, Dec. 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/3c1e1fdf305195cd620c118aaa9717ad-Abstract-Conference.html

[19] Y. S. Chuang, L. Qiu, C. Y. Hsieh, R. Krishna, Y. Kim, and J. R. Glass, "Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y. N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1419–1436. https://doi.org/10.18653/v1/2024.emnlp-main.84

[20] S. Zhang, T. Yu, and Y. Feng, "TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 8908–8949. https://doi.org/10.18653/v1/2024.acl-long.483

Tara Zare is currently an M.Sc. student in Artificial Intelligence and Robotics at Shahid Beheshti University, Tehran, Iran, and a member of the NLP Research Laboratory of the university since 2022. She received her B.Sc. degree in Computer Engineering from Isfahan University of Technology. Her research interests include machine learning, deep learning, and particularly natural language processing (NLP).

Dr. Mehrnoush Shamsfard received her B.Sc. and M.Sc. degrees in Computer Software Engineering from Sharif University of Technology, and her Ph.D. in Computer Engineering–Artificial Intelligence from Amirkabir University of Technology, Tehran, Iran. She has been with Shahid Beheshti University since 2004, where she is currently an Associate Professor at the Faculty of Computer Science and Engineering and the head of the NLP Research Laboratory. Her research interests include natural language processing, knowledge and ontology engineering, text mining, and the semantic and intelligent web.