Halwasa: Quantify and Analyze Hallucinations in Large Language Models: Arabic as a Case Study

Hamdy Mubarak[1], **Hend Al-Khalifa**[2], **and Khaloud Alkhalefah**[3]
[1]**Qatar Computing Research Institute (QCRI), HBKU, Qatar;**
[2]**King Saud University, KSA;** [3]**Imam Mohammad Ibn Saud Islamic University, KSA**
hmubarak@hbku.edu.qa, Hendk@ksu.edu.sa, Kskhalifah@imamu.edu.sa

Abstract

Large Language Models (LLMs) have shown superb abilities to generate texts that are indistinguishable from human-generated texts in many cases. However, sometimes they generate false, incorrect, or misleading content, which is often described as "hallucinations". Quantifying and analyzing hallucinations in LLMs can increase their reliability and usage. While hallucination is being actively studied for English and other languages, and different benchmarking datasets have been created, this area is not studied at all for Arabic. In our paper, we create the first Arabic dataset that contains 10K generated sentences by LLMs and annotate it for factuality and correctness. We provide a detailed analysis of the dataset to analyze factual and linguistic errors. We found that 25% of the generated sentences are factually incorrect. We share the dataset with the research community.

1 Introduction

The advent of Large Language Models (LLMs) has unfolded a new chapter in the field of artificial intelligence (AI), particularly in natural language processing (NLP). These models have shown remarkable proficiency in text generation, translation, and other linguistically driven tasks across various languages. However, a concerning phenomenon known as "hallucination" shadows the reliability of these models, where they generate text that is factually incorrect, nonsensical, or misleading (Ji et al., 2023a; Li et al., 2023). This issue is not confined to English or other widely studied languages but extends to Arabic, a Semitic language with unique phonetic, morphological, syntactic, and semantic characteristics (Shaalan et al., 2019).

Earlier studies have dived into the hallucination issue in LLMs, shedding light on its manifestations such as the generation of incorrect information, nonsensical text, and conflation of sources (McKenna et al., 2023). Furthermore, hallucinations have been studied in the context of machine translation between specific language pairs, with rates exceeding 10% in some cases (Guerreiro et al., 2023). However, a focused examination of hallucinations in the Arabic language through LLMs remains largely uncharted.

The need for more investigation into hallucinations in LLMs for Arabic is underlined by the linguistic complexity and the significance of the language. Addressing hallucinations in LLMs for the Arabic language not only enhances the reliability and applicability of these models but also holds potential implications for a wide array of applications including information retrieval, sentiment analysis, and machine translation.

Therefore, the motivation behind our work is twofold: (1) to enable a rigorous evaluation of the factuality and reliability of LLMs when generating Arabic text, and (2) to pave the path for developing techniques to mitigate hallucinations in Arabic LLMs.

This paper aims to bridge the existing knowledge gap by conducting a case study on hallucinations in LLMs for the Arabic language. The two LLMs we experimented with are GPT 3.5 (aka ChatGPT) and GPT-4. The research questions will aim to answer the following:

I) To what extent can we rely on the factual output of LLMs in the Arabic language?

II) What are the types of factual errors and linguistic errors in the generated output from LLMs?

III) Do Arabic-centric LLMs suffer from hallucinations, and how much?

The contributions of our work are as follows:

- We developed the first Arabic dataset for studying hallucination in LLMs in Arabic. The dataset comprises 10K sentences generated equally by ChatGPT and GPT-4 and

LREC-COLING 2024, pages 8008–8015

judged by a large number of human annotators. The annotation covers factuality, correctness, linguistic errors, and reference links. We make the dataset publicly available for research purposes.

- We prepared detailed guidelines that cover problematic cases during annotation.
- We performed a detailed analysis of factual errors and linguistic errors in ChatGPT and GPT-4.

The rest of the paper is organized as follows: In Section 2, we provide a brief overview of previous work regarding hallucination. We discuss the dataset in detail in Section 3, and the annotation process and guidelines in Section 3.3. In Section 4, we present an in-depth analysis of different aspects like model accuracy, reasons for errors, etc. Finally, we conclude and mention possible future directions of our work in Section 5.

2 Related Work

Hallucination refers to the generation of factually inaccurate or ungrounded information by LLMs, posing challenges to the reliability and trustworthiness of the generated content (Ji et al., 2023a). This literature review aims to encapsulate the current state of research on hallucination in LLMs, including efforts toward understanding, evaluating, and mitigating hallucination, with a particular emphasis on multilingual scenarios.

Recent studies have ventured into unraveling the underpinnings of hallucination in LLMs. For instance, a survey about hallucination in Large Foundation Models (LFMs) explores the models' tendency to produce inaccurate content. It categorizes these inaccuracies, suggests evaluation criteria, reviews mitigation strategies, and points to future research directions (Rawte et al., 2023b). Similarly, a survey conducted in the domain of Natural Language Generation shed light on the metrics, mitigation methods, and task-specific research progress concerning hallucinations (Ji et al., 2023b). On the other hand, (Rawte et al., 2023a) categorized hallucination by degree, orientation, and type, introducing the HallucInation eLiciTation (HILT) dataset and the Hallucination Vulnerability Index (HVI) for evaluating LLMs. They aimed to guide AI policy-making and provide mitigation solutions.

Evaluation frameworks have also been proposed to assess the extent of hallucination in LLMs. Notably, a benchmark named FELM was introduced to evaluate the factuality of text from Large Language Models (LLMs), highlighting the need for accurate evaluations across diverse domains beyond just world knowledge. This benchmark provides detailed annotations, error types, and reference links to support or challenge statements, emphasizing the challenges current LLMs face in reliably detecting factual inaccuracies (Chen et al., 2023). Likewise, a benchmark and dataset named Med-HALT was introduced to specifically evaluate hallucination in the medical domain, underscoring the serious consequences of incorrect information generation in healthcare applications (Pal et al., 2023).

Various strategies have been proposed to curb hallucination in LLMs. A recent paper introduced an uncertainty-aware in-context learning framework aimed at enhancing the reliability of LLMs by empowering the model to enhance or reject its output, thereby addressing the issue of hallucination (Yang et al., 2023). Moreover, the Chain-of-Verification (CoVe) method was proposed to enable LLMs to deliberate on their responses, correcting mistakes, thereby reducing hallucination (Dhuliawala et al., 2023).

The issue of hallucination is not confined to monolingual models but extends to multilingual models as well. A study explored hallucinations in large multilingual translation models, though primarily focusing on small bilingual models trained on English-centric high-resource language pairs (Guerreiro et al., 2023). Despite the insights provided, the study leaves a gap in understanding hallucination in low-resource languages or non-English-centric settings. Another study addresses hallucination within a multi-modal context, which also encompasses multilingual scenarios given the multimodal nature of communication across languages (Liu et al., 2023).

A recent study (Dale et al., 2023) presents the first large-scale dataset with human-annotated hallucinations across 18 translation directions covering diverse language resources and scripts. Their dataset contains naturally generated translations from a public model. Analyzing this, they show for low-resource languages, model internal methods outperform external ones for hallucination detection. Previous conclusions about detection

methods do not transfer well when evaluated across many languages. This work enables reliable multilingual research on mitigating hallucination beyond English-centric evaluations.

As we can see from the previous work, the body of literature on hallucination in LLMs is growing, with researchers exploring various dimensions including understanding the causes, developing evaluation benchmarks, and proposing mitigation strategies. However, the exploration of hallucination in multilingual and non-English-centric scenarios, especially Arabic, remains an area warranting further investigation.

3 Dataset

3.1 Background

Arabic is the official language in 25 countries in the Arab region. Typically, Arabic is divided into three varieties; namely: Modern Standard Arabic -used for formal speeches and books, Classical Arabic - the language used in historical books and literature, and Dialectal Arabic -the spoken language used in each country for daily communications and on social media (Mubarak and Darwish, 2014).

Arabic has a rich and complex morphology. Sentences can start with nouns or verbs, and they have a relatively free word order. Typically nouns and adjectives have gender markers such as Taa Marbouta letter "ة" as a feminine (f) suffix, and in case of absence, they can be considered masculine (m).

Nouns, verbs, adjectives, and some articles can be singular, dual, or plural. Also, nouns and adjectives can be preceded by a definite article as a prefix "ال" as in البيت (the+home: the home)

There should be an agreement on gender, number, and definiteness between the nouns and their describing adjectives, the verbs and their subjects, among other rules.

3.2 Data Collection

In order to evaluate the factuality of the sentences generated by the GPT models, we asked ChatGPT and GPT-4[1] to generate factual sentences that can be checked and verified. To have diverse sentences, we chose 1,000 random words from SAMER Arabic readability lexicon (Al Khalil et al., 2020) and we asked the models to generate factual sentences having each word. As the models sometimes generate duplicate sentences, we started with ten sentences from each model, then we took random unique five sentences for each word[2]. The final data set has 10,000 sentences (half from ChatGPT and half from GPT-4).

We opt to choose more than one factual sentence that has each word to be close to the cases where users ask the model to generate not only one fact but more facts about a certain topic.

For each word in SAMER corpus (n=26,578), we have its stem, part-of-speech, translation, frequency in different corpora, and readability level assigned by language experts (from Level 1; beginner to Level 5; Specialist). We opt to use SAMER corpus because it has rich linguistic information, and the relation between factuality and linguistic correctness can be explored in our paper and in future research.

The used prompt is: *"Give exactly TEN Arabic complete and diverse factual sentences having the following word: <word>. These sentences should have facts that can be checked and verified. Write the sentences separated by a new line without translation and without numbering"*

3.3 Annotation

We hired and trained extensively 50 students from Al-Imam University[3] in Saudi Arabia studying in their last year to participate in data annotation. Each student annotated 200 random sentences generated by ChatGPT or GPT-4. The large number of annotators can reduce biases due to the subjective nature of this job.

For quality control, the authors of the paper independently annotated 50 randomly generated sentences as test questions, and inserted them randomly in the range of each student. These test questions did not appear in the data to be annotated. The initial agreement between the authors was 94% on average before conducting a consolidation session to have full agreement.

We showed only sentences to students without any other information and asked them to provide the following labels:

[1]We used OpenAI models gpt-3.5-turbo-0301 and gpt-4-0314 (released on 2023-06-13).

[2]We skipped cases where the models give errors or generate less than five sentences.

[3]https://imamu.edu.sa/

- **factual**: whether the sentence has factual information that can be verified. Values are 1 and 0.
- **correct**: whether the factual sentence is correct or not. Values are 1 and 0.
- **linguistic error**: whether the sentence has linguistic errors or not. Values are 1 and 0.
- **corrected sentence**: the sentence after correcting its factual and linguistic errors, if any.
- **reference link** : the link used to verify whether the factual sentence is correct or not.

Table 1 shows examples from the annotated data.

3.4 Guidelines

Judging the factuality and correctness of the generated sentences is challenging in some cases. We prepared detailed guidelines with examples and we share them with the dataset in addition to the test questions used in quality control. We list here some important items.

I) AI models (ex: ChatGPT) cannot be used for verification. Search for keywords on search engines instead, and consider Arabic search pages first. This is to reduce biases in non-Arabic pages in understanding the Arabic culture.

II) If the sentence has more than one fact, check them all. if any fact is incorrect, then the "correct" field should be 0.

III) Be tolerant and accept minor differences in numbers in cases when they are approximated (e.g. distance between cities), or when numbers represent changing facts (e.g. a country's population).

IV) For sentences that have time adverbs (ex: now, this year, etc.), assume the date is September 2021 -the date of models training data. This case applies also to temporal information, e.g. the sentence "Karim Benzema plays for Real Madrid", should be labeled as correct according to the date of the training data.

3.5 Annotation quality

All the test questions are factual sentences and half of them are correct. The average agreement between the students and the test questions on the factuality field was 91%. We list the correctness agreement for students, ChatGPT and GPT-4, in Table 2.

As shown in Table 2, all the models struggle to verify the correctness of the test questions. This can demonstrate the limited ability of these models to verify factual sentences they already generated. The best model is GPT-4 with an accuracy of 66% compared to 87% for human annotation.

Here, we give examples of the model errors. For the following wrong sentence سابع أكبر دولة في العالم هي أستراليا (The seventh largest country in the world is Australia), all the models (ChatGPT and GPT-4) considered it to be correct. In the human reference, we provide the correct information which is "India". This sentence is another example of the hallucination: سافر المهندس العالمي أنور السادات في أول رحلة عربية إلى الفضاء عام ١٩٨٥. (International engineer Anwar Sadat traveled on the first Arab flight into space in 1985.) where the information about the person's name and the event is all fabricated.

It is worth mentioning that the models sometimes generate responses that show their inability to understand Arabic, as in the following response from ChatGPT for the last sentence:

"I'm sorry, but I cannot determine the accuracy of the information in the sentence as it is written in Arabic and I am not programmed to understand the Arabic language".

We release our dataset "**Halwasa**", which means "hallucination" in Arabic, for research purposes on the following link[4].

Note: We merged "factual" and "correct" labels in one label as follows:

- **FC**: factual and correct.
- **FI**: factual and incorrect.
- **NF**: not factual.

4 Analysis

4.1 Model Accuracy

Table 3 shows the models accuracy for the following aspects:

I) Factuality: Although we asked the models to generate factual sentences that can be verified, around two-thirds and one-fifth of the sentences generated by ChatGPT and GPT-4 in order cannot be verified.

[4] https://alt.qcri.org/resources/ArabicLLMsHallucination.zip

Sentence	F	C	E	CS	Ref
يبلغ طول نهر النيل حوالي ٦٦٥٠ كيلومتراً (The Nile River is about 6,650 kilometers long)	1	1	0	-	ar.wikipedia.org/wiki/Nile
هو الجمال يعتبر أسرع حيوان على اليابسة بسرعة ١٢٠ كيلومتراً في الساعة (*It's camels considered the fastest animal on land at a speed of 120 km/hour)	1	0	1	الفهد يعتبر.. (Cheetah is ..)	natgeotv.com/... (National Geographic)
الصبر يساعد على تحقيق الأهداف / أخرج الطالب كتابه من حقيبته (Patience helps achieve goals/The student took his book out of his bag)	0	0	0		

Table 1: Annotation examples. F: Factual, C: Correct, E: Linguistic Error, CS: Corrected Sentence, Ref: Ref. Link

Model	Accuracy%
Reference: Human Annotation	**87**
Random Guess	50
ChatGPT	54
GPT-4	**66**

Table 2: Annotation quality for correctness of the test questions

II) Correctness: Out of the sentences that can be verified from each model, around **one-quarter of them is incorrect**. The percentages of the correct sentences out of the total sentences are about half and two-thirds for ChatGPT and GPT-4 in order. From these results, we can deduce that a sizable portion of the "factual" sentences generated by ChatGPT and GPT-4 are either not factual or are incorrect.

III) Linguistic Errors: Surprisingly, ChatGPT produced more linguistically correct sentences than GPT-4. ChatGPT errors are two-thirds that of GPT-4. This can be attributed to the observation in our dataset that ChatGPT tends to generate shorter sentence -hence less chance to have linguistic errors- than GPT-4. The average sentence length for ChatGPT and GPT-4 in our dataset were 10 and 13 words in order.

We analyzed random 100 errors from ChatGPT and GPT-4. We can classify linguistic errors into these main types:

i) str: sentence structure error, ex: عن الممثل الأمريكي، حصل على جائزة الأوسكار (About the American actor, he won the Oscar)

ii) vocab: usage of wrong vocabulary or odd synonyms, ex: أرقام درجة الحرارة القياسية تصل معلوم ٥٠ درجة (Temperature records reach a high *known of 50 degrees).

iii) agr: wrong agreement (on gender or number) between verbs and their subjects, ex: نالت الفيلم الكوري جائزة أفضل فيلم (The Korean film(m.) won(f.) the Best Film award)

iv) prep: wrong usage of prepositions, ex: يعتبر النفط مورداً هاماً عن الإيرادات. (Oil is an important source *from revenue.)

v) dial: usage of dialectal words, ex: راح يتم افتتاح أكبر مصنع (The largest factory *gonna be opened).

Please refer to the Arabic background section 3.1 for details.

Table 4 shows the percentages of different types of linguistic errors.

4.2 Reasons for Incorrect Facts

As GPT-4 is the best model that generated the correct factual sentences in our dataset, we analyzed random 200 GPT-4 errors as shown in Table 5 to understand their reasons.

Most errors are due to generating incorrect facts (e.g. using terms like "the largest" as opposed to "the second largest", or "the first" vs "second"), using incorrect dates, and fabrication of non-existing named entities like person or city names.

Many human annotation errors are due to the confusion between factual errors and linguistic errors. It is worth mentioning that some human errors are due to temporal information. for example, the sentence تم تأسيس منطقة اليورو في عام ١٩٩٩ ويضم ١٩ دولة من الاتحاد الأوروبي (The Eurozone was established in 1999 and includes 19 European Union countries) is labeled as incorrect by the annotators and they wanted to correct it to 20 countries. In fact, the Eurozone currently has 20 countries after Ukraine joined in 2022, and the generated sentence was correct on the date of the GPT-4 training (September 2021).

Although we asked the annotators to be tolerant

Model	Total	Factual	Factual/Total	Correct	Correct/Factual	Correct/Total	Ling. Error	Ling. Error/Total
ChatGPT	5,000	3,317	66%	2,506	**76%**	50%	387	**07.7%**
GPT-4	5,000	4,163	**83%**	3,072	74%	**61%**	586	11.7%

Table 3: Model Accuracy

Model	str	vocab	agr	prep	dial
ChatGPT	39	36	16	9	0
GPT-4	37	36	21	2	4

Table 4: Analysis of Linguistic Errors for 100 Sample Errors

and accept minor differences in numbers, they reported cases where some sentences were considered incorrect if any sort of approximation was applied. For example, this sentence الرقم القياسي العالمي لأسرع رجل في العالم ٩٫٦٣ ثانية (The world record for the fastest man in the world is 9.63 seconds) is wrong and the number is corrected to 9.58 seconds.

4.3 Verification Websites

For each sentence that needs verification, we asked the annotators to provide a link to the web page that supports their annotation. We recommend using trusted websites. The annotators used 800+ diverse websites as references. In Table 6, we list the top five websites used in our dataset (accounted for 64% of all sources). Other websites were used individually in 1% of the cases or less. According to our data, the Arabic Wikipedia is the most trusted source of information followed by famous news websites and online content publishers.

4.4 Factuality versus Readability

We want to see whether there is a correlation between word readability and the correctness of the factual sentences generated by the models. Normally, complex words (e.g. in readability levels 4 and 5) are less frequent than simple words in lower levels, and we hypothesize that they will appear less frequently in the models' training data. This may affect the generated facts having such complex words.

Figure 1 shows that there is no clear relation between the readability levels and percentages of correct factual sentences generated by ChatGPT and GPT-4[5]. We plan to investigate this in more depth in the future.

[5]Words with readability level 5 are rare in SAMER corpus, and they didn't appear in our random sample.

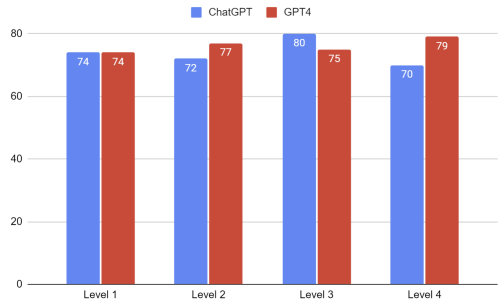

Figure 1: Percentage of Factual/Correct Sentences for Different Readability Levels

5 Conclusion

In this paper, we present "Halwasa", the first Arabic dataset to benchmark LLMs; namely ChatGPT and GPT-4. We asked the models to generate diverse factual sentences and annotated them manually. We described the data collection and annotation processes in details. Also, we analyzed the factual errors and linguistic errors in the generated sentences.

Several future directions emerged from our work. Evaluating ChatGPT and GPT-4 on English equivalents of the Arabic words could enable cross-lingual comparisons, while prompting them to fact-check English translations of incorrect Arabic sentences could yield insights into cross-lingual hallucination patterns. Further analysis of GPT-4's dialectical errors, where it underperformed compared to ChatGPT, is warranted and could inform developing more robust, dialect-aware Arabic LLMs, as indiscriminate data scaling may not be optimal for multi-dialectal languages like Arabic. Additionally, we plan to increase the size of our dataset, include more Arabic LLMs like Jais (Sengupta et al., 2023) and BloomZ in our evaluation, use the human-corrected sentences as a new test set, and employ LLMs to verify their own results. Moreover, we will build a machine learning model to predict the factuality of the generated sentences and attempt to correct their mistakes. We will also explore the change in performance when translating the generated Arabic sentences to English and evaluating English LLMs on them. Exploring these directions would advance our understanding of Arabic LLM hallucinations and facilitate techniques to

Error Type	Example	%	Comment
Fact	في المملكة العربية السعودية يوجد *أكبر احتياطي نفط في العالم	22	ثاني أكبر
(ex: ordering)	(In Saudi Arabia there is the *largest oil reserve in the world)		second largest
Human Error	ازدهرت الفلسفة والطب والرياضيات في العالم الإسلامي *خلا العصور الوسطى	20	Confusion
	(Philosophy, medicine, and mathematics flourished in the Islamic world *durin the Middle Ages)		
Date	تأسس مقر الإنتربول الدولي في العام * ١٩٤٩ بمدينة ليون الفرنسية	20	في ١٩٢٣
	(Interpol's international headquarters was established in *1949 in Lyon, France)		in 1923
Name	*دل يعتبر عاصمة دولة الهند وتأسست في القرن السادس عشر	10	دلهي
	(*Del is considered the capital of India and was founded in the sixteenth century)		Delhi
Area	مساحة الهرم الأكبر في مصر * ٢٠ فدان	8	١٣ فدان
(Quantities)	(The area of the Great Pyramid in Egypt is *20 acres)		13 acres
Fractions	مساحة الصين * ٩,٥ مليون كيلومتر مربع	6	٩,٦
	(China's area is estimated at about *9.5 million square kilometers)		9.6

Table 5: Common Reasons of Factual Errors in GPT-4

Website	Description	%
ar.wikipedia.org	**Arabic Wikipedia**	**55.6**
aljazeera.net	Qatari Aljazeera Media Network	3.4
mawdoo3.com	Jordanian content publisher	2.3
bbc.com	British broadcaster	1.5
un.org	The United Nations	1.5

Table 6: Verification Websites and their Usage

mitigate them, thereby enhancing the reliability of Arabic language technologies.

Ethical Concern and Social Impact

User Privacy We asked the annotators to report any case where any private information has been leaked in the generated sentences by LLMs. No case was reported. We believe that our work can lead to enhancing the reliability of LLMs such that more users can trust their outputs with confidence.

Biases and Limitations We asked annotators to mark sentences that have any kind of biases. Very few cases were reported, for example يوجد خلاف بين الطبقات الاجتماعية عند العرب (There is a conflict/dispute between social classes among Arabs). Any biases found in our dataset are unintentional as the text is fully generated by LLMs and corrected by humans. In our study, we tried to remove biases in data collection by choosing a large number of annotators and giving them detailed guidelines and training sessions. We acknowledge that our statistics and results may not represent the performance of LLMs in other factuality benchmarks. Further, there are some annotation errors. Therefore, the statistics presented in our paper provide an estimate of the whole picture.

References

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for standard arabic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hal-

lucination in large multi-modal models via robust instruction tuning.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks.

Hamdy Mubarak and Kareem Darwish. 2014. Using twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023a. The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023b. A survey of hallucination in large foundation models.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Khaled Shaalan, Sanjeera Siddiqui, Manar Alkhatib, and Azza Abdel Monem. 2019. Challenges in arabic natural language processing. In *Computational linguistics, speech and image processing for arabic language*, pages 59–83. World Scientific.

Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. 2023. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning.