ASOS at Arabic LLMs Hallucinations 2024: Can LLMs detect their Hallucinations :)

Serry Sibaee, Samar Ahmed, Omer Nacar,
Abdullah I. Alharbi, Lahouri Ghouti, Anis Kouba

Robotics and Internet-of-Things Lab, Prince Sultan University
Faculty of Computing and Information Technology Rabigh, King Abdulaziz University
Riyadh 12435 Saudi Arabia, Jeddah 22254 Saudi Arabia
{ssibaee, onajar, lghouti, akoubaa}@psu.edu.sa
aamalharbe@kau.edu.sa, Samar.sass6@gmail.com

Abstract

This research investigates hallucination detection in Large Language Models (LLMs) using datasets in the Arabic language. As LLMs gain widespread application, they tend to produce hallucinations—grammatically coherent but factually inaccurate content—posing substantial challenges. We participated in the OSACT 2024 Shared-task, which focuses on the Detection of Hallucination in Arabic Factual Claims Generated by ChatGPT and GPT-4. Our approach evaluates several methods for detecting and mitigating hallucinations, employing models such as GPT-4, Mistral, and Gemini within an innovative experimental framework. Our findings demonstrate significant variability in the models' ability to categorize claims as Fact-Claim (FC), Fact-Improvement (FI), and Non-Fact (NF), highlighting the challenges of dealing with hallucinations in morphologically complex languages. The results underline the necessity for more sophisticated modelling and training strategies to improve the reliability and factual accuracy of the content generated by LLMs. This study lays the foundation for future work on reducing the risks of hallucinations. Notably, we achieved an F1 score of 0.54 in detecting hallucinations with the GPT-4 model.

Keywords: Large Language Models(LLMs), Hallucination Detection, and Arabic Text Classification

1. Introduction

LLMs have experienced a rapid increase in popularity and application since the introduction of GPT in 2021. Capable of producing diverse forms of content including text, code, images, and videos, these advanced models have revolutionized neural natural language generation (NLG) systems. Their enhanced realism in text generation has proven beneficial across a variety of real-world applications such as question-answering, summarization, translation, and paraphrasing. However, alongside these advancements, LLMs face a significant challenge: the phenomenon of hallucination.

Hallucination, as defined by (Ji et al., 2023), is the generation of text or responses that, while grammatically accurate and coherent, deviate from the source inputs in terms of faithfulness or factual accuracy. Essentially, it results in the production of misaligned or factually incorrect information, posing substantial risks to the deployment of LLMs in sensitive real-world applications. With the demand for integrating LLMs into various domains to streamline operations, addressing hallucinations has become a critical concern.

Research to combat this issue generally adopts two main strategies: hallucination detection and mitigation. Hallucination Detection, as explored in (Luo et al., 2024), entails identifying potential hallucinations within LLM-generated responses, at both token and sentence levels, to flag content that significantly diverges from the input. Hallucination Mitigation, on the other hand, aims to reduce the occurrence of hallucinations by enhancing the factual accuracy and reliability of generated content, with methods including the integration of knowledge graphs and retrieval systems.

This study seeks to build upon existing research on Hallucination Detection. The paper is organized as follows: Section 2 reviews related work, Section 3 presents our proposed methodology, Section 4 discusses our experimental results, and Section 5 concludes the paper with a summary of our key findings.

2. Related Work

Prior studies have focused on the detection of hallucinations in LLMs. The research conducted by Snyder et al. (2023) aimed to answer factual questions while examining outputs from three models: OpenLLaMA, OPT, and Falcon. A variety of techniques, including integrated gradient token attribution, SoftMax probabilities, self-attention scores, and fully connected activations, were utilized to distinguish between hallucinated and non-hallucinated generations. While input attribution sometimes per-

OSACT 2024 Workshop, pages 130–134

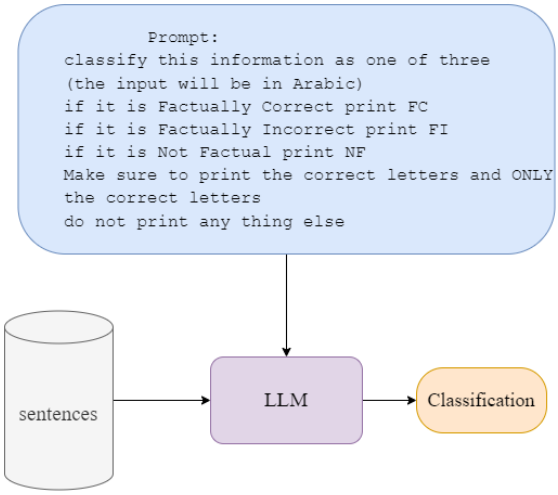

Figure 1: General Framework for our proposed system and the prompt used for the task

formed only marginally better than random chance across different datasets, other techniques demonstrated superior performance on certain datasets. Li et al. (2023) introduced HaluEval, a two-stage framework designed to generate hallucinated samples and conduct high-quality hallucination filtering to evaluate LLMs' performance in recognizing hallucinations. This framework incorporates strategies such as knowledge retrieval, Chain-of-Thought (CoT) reasoning, and sample contrast, enhancing LLMs' abilities to recognize hallucinations and analyze their informational blind spots.

Varshney et al. (2023) proposed an approach for detecting and mitigating hallucinations, focusing on the text generation process. Utilizing GPT-3.5, their study showcased the effectiveness of detection and mitigation techniques, achieving an 88% recall rate and successfully mitigating 57.6% of detected hallucinations without introducing new ones. Liang et al. (2024) emphasized the importance of self-awareness in LLMs for mitigating factual hallucinations. They proposed DreamCatcher, an automated tool designed to evaluate the extent of hallucinations in LLM outputs, classify them by factual accuracy, and provide data for refining LLMs to reduce factual hallucinations. Additionally, the Reinforcement Learning from Knowledge Feedback (RLKF) training framework aims to enhance the factuality and honesty of LLM outputs.

In a comprehensive survey, Tonmoy et al. (2024) discussed the issue of hallucination in LLMs and its impact on their real-world deployment. They highlighted the importance of mitigating hallucinations through prompt engineering and model development techniques. Furthermore, they provided a taxonomy of hallucinations in text generation tasks, analyzed the theoretical aspects of hallucinations in LLMs, and presented existing detection and improvement methods, proposing future research directions in this area. This study aims to contribute to the understanding and mitigation of hallucinations in LLMs.

3. Methodology

In this section, we provide a detailed description of the dataset released by the organizers of the shared task, followed by an explanation of the task itself. We then describe the methods we employed, including the models we experimented with in this study.

3.1. Data and Task Definition

The task involves working with datasets in the Arabic language for Subtask A and Subtask B. For our study, we participated exclusively in Subtask A. In this subtask, participants are required to utilize only the "claim" and "label" columns. The data is tab-separated and includes columns for "claim ID," "word position," "readability," "model," "claim text," and "label." The labels—FC (Factually Correct), FI (Factually Incorrect), and NF (Non-factual)—are used to classify claims into these categories based on their factual accuracy. While Subtask B permits the use of all columns in the dataset, our focus remained solely on Subtask A. Participants are provided with training, development, and testing datasets.

3.2. Models

In our initial experiments, we attempted to use Arabic pre-trained models, such as AraBERT, and fine-tuned them on the provided training data. Unfortunately, this approach did not yield promising results,

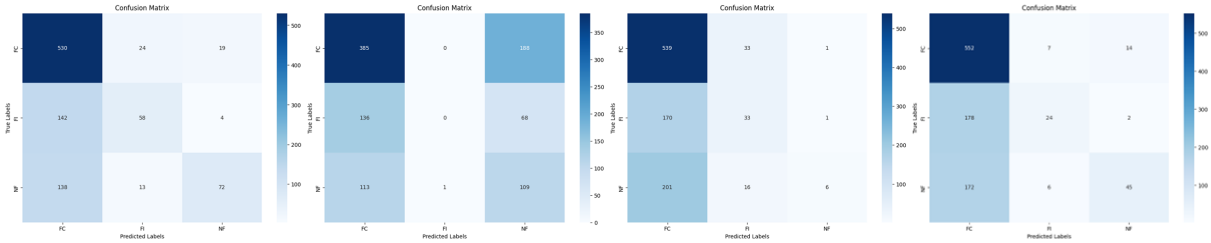

Figure 2: The confusion matrices for the three voter models: a) GPT-4, b) Mistral, c) Gemini and d) ensemble (Majority Voting) for all three models, respectively.

with the maximum F1 score achieved being 44%. Another approach was to represent the input as a one sentence to try to learn the distribution (was was assume to be one that generate correct and wrong sentences) of the data $\wp$ (this assumption was extremely hard to implement). The model used to represent the sentence (input) was "distiluse base multilingual cased" from sentence transformers library (Reimers and Gurevych, 2019) that has a 512 dimensions then was forwarded to a neural network. Despite trying multiple architectures, this approach did not produce encouraging results, prompting us to explore alternative methods.

The main idea of this experiment is to test LLMs ability to detect hallucinations or classify given information as either factually correct, factually incorrect, or non-factual (data that is not declarative). This was achieved by forwarding the text, wrapped in a comprehensive prompt, to control the output format. The LLMs used were GPT-4 and Gemini, both capable of handling Arabic text directly, and Mistral 7B, which was used with a pipeline approach due to its training on English. For Mistral 7B, inputs were translated to English using the Google Translate API before being fed into the model, which was accessed through the Hugging Face Transformers library.

- GPT-4 (OpenAI et al., 2024): GPT-4, the latest iteration in OpenAI's Generative Pre-trained Transformer series, marks a significant leap in natural language processing. With a larger model size and enhanced architecture, GPT-4 excels in tasks like text generation, comprehension, and translation. Its adaptability across various linguistic domains and improved fine-tuning capabilities make it versatile for applications such as conversational agents and sentiment analysis. Despite its technical prowess, GPT-4 prioritizes ethical AI development, focusing on bias mitigation and safety measures. Overall, GPT-4 represents a milestone in NLP, offering unprecedented sophistication and ethical considerations for human-computer interaction and communication.

- Gemini (Team and Rohan Anil, 2023): Gemini, a multimodal AI model by Google, comprehends text, code, and figures, allowing it to read vast scientific literature, reason across disciplines, and answer complex questions. This empowers researchers to conduct faster literature reviews, generate novel hypotheses, and gain insights from complex datasets, ultimately accelerating scientific discovery.

- Mistral : (Jiang et al., 2023) Mistral 7B is a high-performing language model with 7 billion parameters designed for superior efficiency. It surpasses even larger models like Llama 2 (13 billion parameters) across various benchmarks. Mistral 7B particularly outshines in reasoning, mathematics, and code generation compared to Llama 1 (34 billion parameters). The model employs grouped-query attention (GQA) for faster inference and sliding window attention (SWA) to handle sequences of any length efficiently. Additionally, a fine-tuned version, Mistral 7B – Instruct, excels in following instructions, outperforming Llama 2 13B – chat model in both human and automated benchmarks. Overall, Mistral 7B demonstrates outstanding performance and efficacy in natural language processing tasks.

4. Experiments and Results

In this section, we detail the procedure adopted to tackle the problem, beginning with the development of an effective and comprehensible prompt for the used LLMs.

4.1. Experimental Setup

The final prompt, arrived at after several iterations, is depicted in Figure 1. This prompt was utilized with GPT-4, Gemini, and Mistral. For the Mistral model, sentences were translated to English using the Google Translate API before being fed into the model, due to its English-centric training.

Model	Precision	Recall	F1-score
GPT-4	0.67	0.51	0.54
Gemini	0.58	0.38	0.34
Mistral	0.67	0.43	0.42

Table 1: Results for Subtask A on Dev set.

Model	Precision	Recall	F1-score
GPT-4	0.663	0.495	0.516

Table 2: Test Results for Subtask A on test set.

4.2. Results

Our study on the detection of hallucination in Arabic statements generated by LLMs revealed how different models, including GPT-4, Mistral, and Gemini, performed in classifying claims into FC, FI, and NF. The outcomes, presented in Table 1 and through confusion matrices in Figure 2, demonstrate varied model performances. GPT-4 showed overall strong performance but faltered with FI claims, highlighting a deficiency in grasping nuanced content. Mistral had limited success, especially with FI claims, which revealed its difficulty with complex classifications. Gemini, while accurate with NF claims, showed a low recall rate, indicating a potential overemphasis on specific claim types. The use of a Majority Voting technique improved the recall for FC claims but did not significantly improve the classification of FI and NF claims. This highlights the complex nature of nuanced text classification and the need for improved modelling and training approaches to handle the intricacies of languages such as Arabic effectively.

The variance in the performance of different models in various categories highlights the importance of carefully selecting models and utilizing ensemble methods in downstream tasks. The consistent challenge faced with FI claims across all models calls for further investigation into the models' ability to identify and categorize subtle factual changes. In addition, the partial success of the Majority Voting method suggests that combining model outputs does not entirely solve the nuanced classification challenges, which indicates a potential focus for future research in model architecture or training data refinement. Ultimately, we submitted our final results based on the findings obtained from GPT-4, as detailed in Table 2.

5. Conclusion

Identifying and categorizing sentences as factual, non-factual, or uncertain is a challenging task. This challenge arises from the need for models to interpret and extract factual meaning, which is not always a straightforward task. In our research, we introduced a structured prompt designed to utilize LLMs as a tool for factual verification. We tested several models, including GPT-4, Gemini, and Mistral, and found that GPT-4 was the most effective, achieving a Macro F1 Score of 0.54. In future work, we plan to investigate the optimization of Arabic LLMs, with a particular focus on models like Jais, AceGPT, AraGPT, and ArabianLLM, to enhance further their capabilities in verifying factual content.

6. Acknowledgements

The authors thank Prince Sultan University for their support.

7. Bibliographical References

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand,

Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *arXiv preprint arXiv:2401.15449*.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*.

OpenAI, :, and Josh Achiam etal. 2024. Gpt-4 technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2023. On early detection of hallucinations in factual question answering. *arXiv preprint arXiv:2312.14183*.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Gemini Team and etal Rohan Anil. 2023. Gemini: A family of highly capable multimodal models.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

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

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.