Exploring Hallucination in Large Language Models

A Comparative Study of GPT-4o and GPT-4o-mini in Medical Domains

Nesreen M. Alharbi[1], Thoria Alghamdi[2], Raghda M. Alqurashi[3] ,Reem Alwashmi[4], Amal Babour[5]*, Entisar Alkayal[6]
Department of Computer Science-Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabia[1]
THE LIGHT-Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabia[1, 6]
Department of Information Systems-Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia[2, 5]
Department of Computer Science-Jamoum University College, Umm Al-Qura University, Al Jumum 25371, Saudi Arabia[3]
Department of Information Technology-Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabi[4]

Abstract*—Large Language Models such as GPT-4o and GPT-4o-mini have shown significant promise in various fields. However, hallucination, when models generate inaccurate information, remains a critical challenge, especially in domains that require high accuracy, such as the healthcare field. This study investigates hallucinations in two different LLMs, focusing on the healthcare domain. Four different experiments were defined to examine the two models' memorization and reasoning abilities. For each experiment, a dataset with 193,155 multiple-choice medical questions from postgraduate medical programs was prepared by splitting it into 21 subsets according to medical topics. Each subset has two versions: one with the correct answers included and one without them. Accuracy and compliance were evaluated for each model. Models' adherence to requirements in prompts was assessed. Also, the correlation between size and accuracy was tested. The experiments were repeated to evaluate the models' stability. Finally, the models' reasoning was evaluated by human experts who assessed the models' explanations for correct answers. The results revealed poor rates of accuracy and compliance for the two models, with rates below 70% and 75%, respectively, in most datasets; yet, both models showed low uncertainty (3%) in their responses. The findings showed that the accuracy was not affected by the size of the dataset provided to the models. Also, the results indicated that GPT-4o-mini demonstrates greater performance stability compared to GPT-4o. Furthermore, the two models provided acceptable justifications for choosing the correct answer in most cases, according to 68.8% of expert questionnaire participants who agreed with both models' justifications. According to these results, both models cannot be relied upon when accuracy is critical, even though GPT-4o-mini slightly outperformed GPT-4o in providing the correct answers. The findings highlight the importance of improving LLM accuracy and reasoning to ensure reliability in critical fields like healthcare.

Keywords*—ChatGPT; GPT-4o; GPT-4o-mini; hallucination; healthcare; large language models

I. Introduction

One development in artificial intelligence (AI) applications is large language models (LLMs), which include several technologies such as deep learning and natural language processing (NLP), reinforcement learning, and transformers to produce new content from available material. Today, there are many LLMs that have been developed by different vendors. These models follow the same process and start with training the LLM on a large amount of data from the Internet to produce new content. However, these models differ in their training methods and the technologies used to generate and predict output data. Most of these models rely on language general context, and linguistic rules to predict subsequent words based on the preceding ones using technology [1].

LLM models have been applied in many fields to generate content ranging from answering questions, executing commands and instructions, summarizing, translating, and conducting long conversations with users [2]. One well-known company in this field is OpenAI. In November 2022, it made its LLM, ChatGPT, available to the public. Following that, OpenAI launched several versions of ChatGPT. For instance, in March 2023, GPT-4 was introduced, followed by GPT-4o in May 2024. These new releases had a high impact on NLP technology. The accuracy of understanding the context of input data and generating the response has enhanced in comparison to the old models. In addition, GPT-4o shows strong performance across many benchmarks [3], and supports multi-modal content generation.

LLMs have been widely used in several fields, including education, healthcare, industry, business, and marketing, to enhance productivity and quality as well as reducing effort and saving time [1], [4], and [5]. However, the use of LLMs may involve some ethical issues such as bias and hallucinations which can affect data accuracy and the model reliability [6], [7], and [8]. Hallucinations may occur when the models been trained with invalid and insufficient data, thus LLM generates inaccurate or false outputs [9], and [10]. Several studies have identified factors to minimize the possibility of hallucination. These factors include improving the quality of training data, developing methods to detect and correct errors, and enhancing the models' ability to think and reason to distinguish accurate information from false [10].

The problem of hallucination is that it looks like a piece of meaningful and valuable information, but, in fact, it is inaccurate and invalid. Generally, hallucinations can occur for several reasons, including a lack of adequate information, which leads to inaccurate results during analysis. In addition, poor data

*Corresponding author.

quality will generate models that rely on low-quality data, reducing result accuracy and increasing the likelihood of hallucinations [9]. Moreover, some instructions or questions given to the LLM models may be unclear or can result in multiple interpretations according to different contexts. Furthermore, most LLM models lack reasoning, critical thinking, and analytical abilities, although companies have focused on improving these issues in the current updated versions [10].

This research contributes by examining two models of ChatGPT: GPT-4o and GPT-4o-mini and comparing the hallucination of the two models in the healthcare field. The motivation behind this work comes from the hallucination issue in LLMs, which is particularly critical in fields such as healthcare, where accurate responses are essential and any misinformation could have severe or even fatal consequences. Reducing hallucination in LLMs is essential for ensuring their reliability in critical fields. Thus, the objective of this research is to evaluate the two models' responses to medical questions. Different sets of medical exam questions in different medical topics for postgraduate medical programs were prepared by including or excluding the correct answers to test the models' ability to reason and distinguish the absence of correct answers. Experiments were designed to evaluate GPT-4o and GPT-4o-mini models on memorization and reasoning, where accuracy, compliance, uncertainty, and reasoning were measured for each experiment. Also, this research examines the relationship between the size of the dataset and the models' accuracy, as well as the impact of repeating the test on the models' accuracy.

The rest of the research is organized as follows: Section II reviews the literature. In Section III, the methodology is addressed. A detailed explanation of the different experiments and their results is presented in Section IV and Section V, respectively. Section VI discusses the findings of the study. Concluding remarks and a brief of future works are given in Section VII.

II. Literature Review

Models of large languages like LLama 3 [1], and GPT-4o [11] have the capability to generate loads of information as needed in several domains, including professional and crucial [12], [4]. However, in some cases, hallucination phenomena may occur in LLMs [2] when providing information, which can vary, ranging from minor fabrications to major misconceptions.

Regardless of the effectiveness of the prompting techniques, researchers have found that LLMs can produce erroneous responses that mimic actual statements but contain unsubstantiated information [13]. Others have confirmed that when hallucinations occur in LLMs, the resulting errors can negatively impact the user experience, leading to confusion, distrust, and decreased enthusiasm for using these AI models [7] and [14]. Ji et al. [7] define hallucinations in LLMs as the natural generation of information; however, they may lack meaning or be inconsistent with the actual content of the source. Zhang et al. [15] have standardized the definition of hallucinations in LLMs into three categories: input-conflicting hallucinations, context-conflicting hallucinations, and fact-conflicting hallucinations. With input-conflicting hallucinations, the content that LLMs produce conflicts with what the user has input. With context-conflicting hallucinations, the content produced varies between two different attempts. Lastly, fact-conflicting hallucinations occur when the information produced contradicts fundamental knowledge. Several factors can significantly cause hallucination in LLMs, including inadequate data samples in the training dataset or utilizing algorithms with extremely uncertain sampling [7] and [16].

Therefore, dependence on language generative technologies, where hallucinations in LLMs may occur, raises serious concerns. Given these challenges, it is particularly concerning when LLMs are applied, particularly in critical domains like healthcare. This is due to the significant impact on people's lives [17], including incorrect clinical decision-making and delayed or improper treatment [8]. It is possible to train LLMs to perform remarkably well on a range of medical and healthcare tasks [11], [4], pass medical examinations [5], [18], [19], and generate relevant texts about medical topics. However, in a question-answering task, LLMs may provide an answer that appears reasonable but is factually inaccurate, while in content-generation tasks, the model may also generate cohesive narratives or explanations based on fictitious facts or events in content generation tasks. For example, in response to the question, "What are the common side effects of metformin?" an LLM might say, "Medication side effects include nausea as well as trouble breathing", which is slightly incorrect [20]. The reliability of LLMs is a major concern for both healthcare organizations and patients, especially if the model hallucinates and produces incorrect answers to medical questions. Any incorrect information can significantly impact the patient's health and the healthcare organization's reputation. Thus, hallucinations in LLMs might be extremely challenging in the medical domain [21], [22], where integrity and dependability are crucial [23] to maintain both patient satisfaction and trustworthiness while receiving healthcare.

There are several studies that highlighted the performance of different LLM models, including MCQ examination in the domain of health care [24, 25, 26, 27]. Each research has evaluated the performance of hallucinations based on one or more LLM models like ChatGPT3, ChatGPT 4 [24, 25, 26] while [25] compares ChatGPT-4 with three other different models, namely ChatGPT, QWen 2.1, and Ernie 4.0. Nevertheless, [27] has examined the performance of ChatGPT solely. In addition, the dataset language used in each research varies between English and Non-English, aimed to measure both the model's accuracy and reasoning.

Moreover, [24] focused on evaluating ChatGPT-4 performance on the Japanese medical licensing examination dataset. Four aspects of the model were examined: accuracy, category sub-score, effect of translation, and error analysis. The model scored 82.7% and 77.2% for the essential questions and the basic and clinical questions, respectively. On the other hand, [25] has tested the performance of the GPT-4.0, ChatGPT, QWen 2.1, and Ernie 4.0 models on a Chinese dataset that has been evaluated by experts; the models' accuracy, performance, and reasoning were measured. The research measures model reasoning by calculating the rates of fact hallucinating, fact fabrication, Instruction Inconsistency, and logical inconsistency. For all models, Ernie 4.0 achieved better performance compared to GPT-4.0.

Furthermore, [26] compared the performance of ChatGPT-4 in answering medical MCQ questions with the answers coming from medical students. This study showed that ChatGPT-4 accuracy outperformed the students with 73.7%, 66.7% respectively. On the other hand, 78% of the students achieved at least 90% accuracy, which is higher than the accuracy rate of ChatGPT-4. Finally, in [27], ChatGPT's accuracy and reasoning were evaluated using MCQ datasets from the United States Medical Licensing Examination (Steps 1 and 2), using two different datasets with and without a provided hint to the model. Having four different versions of the question sets, the study revealed accuracy over datasets for the first dataset (Step1) exam as 44%, (Step2) as 42% and the second dataset (Step1) as 64.4%, and (Step2) as 57.8%.

Although these studies tested the abilities of different LLMs, none of them evaluated the ability of the model while excluding the correct answers. Also, none of these studies have tested hallucination on Chat-GPT-4mini and measured the correlation between the model's accuracy and the size of the dataset, nor tested the model's stability while answering medical questions.

III. Materials and Methods

The dataset for this research [28] consists of multiple-choice questions (MCQs) with fine-grained human-labeled classes in different medical fields at the graduate level, designed as a prerequisite for admission to postgraduate medical programs. These questions are designed to assess the skills of medical professionals. The dataset samples contain questions with answer options, correct answers, and explanations of the solutions. To ensure that all questions could be answered by text input, a few steps were taken to clean the data. First, the questions and options were proofread. In addition, to improve the quality of the dataset, the content was supervised by humans. Some tools were used to prepare the data, e.g., a spellchecker to identify and correct some cases, such as extra white spaces or missing options. Any question with an inconsistent format or questions without the correct answer option was removed. In addition, questions without a null candidate or those requiring external information were deleted. Questions containing specific keywords, such as "equation" or "India", were removed. In addition, all duplicate questions were removed. The final version of the dataset contained 193,155 questions.

To enable the generalization and re-usability of the models, the dataset was split by exams rather than by the questions given. The training set consisted of collected mock and online test series, while the test set comprised MCQs from the AIIMS PG exams (years 1991–the date at the time of when you compiled the dataset). The NEET PG exam MCQs (years 2001–the date at the time of when you compiled the dataset) were used to make the development set, aiming to approximate a real exam evaluation. The dataset consisted of 183,000 training examples, 4000 in the test set, and 6000 in the development set. To avoid overlapping between questions from the training, test, and development datasets, it was ensured that the test and development sets contained questions distinct from the training data. By calculating the Levenshtein distance between each pair of questions in the dataset, a question was excluded from the development and test sets if its similarity to any other question exceeded 0.9. For this research, the dataset was split into 21 smaller subsets, each corresponding to specific topics covered by the questions. These topics were Surgery, Social and Preventive, Skin, Radiology, Psychiatry, Physiology, Pharmacology, Pediatrics, Pathology, Orthopedics, Ophthalmology, Microbiology, Medicine, Gynecology and Obstetrics, Forensic Medicine, ENT, Dental, Biochemistry, Anatomy, Anesthesia, and Unknown. The Unknown dataset included all the questions that could not be categorized into any of the other 20 topics. After splitting, the dataset was divided into 21 sub-datasets, and 10% of the questions from each sub-dataset were randomly selected. This reduced the number of questions used to evaluate the targeted models to 100,000. Then, to prepare the data, all unnecessary records were dropped, and only the question, the four answer options, and the correct answer were retained. These datasets were used to examine the targeted models in the set of experiments where the correct answer is provided to the model with the set of possible answers. The final structure of the questions in this dataset is shown in Fig. 1.

```
Question: {Question}
Opt0: {Option1}
Opt1: {Option2}
Opt2: {Option3}
Opt3: {Option4}
Answer: {Correct Choice (0-3)}
Opt0, 1, 2, or 3 includes a correct choice
```

Fig. 1. Structure of question set with correct answer provided.

Then, new datasets were generated from the previous datasets to conduct another set of experiments where the correct answer was not included in the possible choices. Instead, the correct choice was replaced by (NONE). The final structure of the questions in these datasets is shown in Fig. 2.

```
Question: {Question}
Opt0: {Option1}
Opt1: {Option2}
Opt2: {Option3}
Opt3: {Option4}
Answer: {Correct Choice (0-3)}
Opt0, 1, 2, or 3 includes a NON choice, no
    correct answewr provided.
```

Fig. 2. Structure of question set with NO correct answer provided.

IV. Experiments

In order to test hallucination in LLMs, two different models were tested, focusing on two key capabilities: memorization and reasoning. This study examined two well-known models: GPT-4o and GPT-4o-mini. The experiments conducted in this research were implemented using Google Collab, with Python 3 and a CPU hardware accelerator. Four separate experiments were designed to evaluate the models: two for GPT-4o and two for GPT-4o-mini. Each model was tested for both memorization and reasoning abilities. In addition, two of the experiments were repeated to examine the ability of the two models to improve their levels of memorization and reasoning.

A. GPT-4O Hallucination Evaluation

1) Analyzing Memorization and Reasoning with Correct Answer (MWRCA)

The objective of this experiment was to evaluate the hallucination of the GPT-4o model at two different levels:

memorization and reasoning. For this purpose, a benchmark consisting of multiple-choice questions, each with four answer options, was used that covered 21 different domains (see Section II). Among the four options for each question, one was the correct answer. The evaluation method (GPT-4o: MWRCA) began by feeding the model this set of questions, then asking it to select the correct answer and provide a reason for its selection. Additionally, the model was asked to explain why the other options were incorrect. After collecting the model's outputs, its memorization ability was evaluated based on the number of correct answers it chose, while its reasoning ability was assessed by analyzing the justifications the model provided for selecting the correct answer and avoiding the incorrect ones. Fig. 3 shows the prompt that asked the model to comply with a specific format to generate the output response. In this test, the model was allowed to respond "(IDK)" if it was uncertain of the correct answer. However, the expectation in the GPT-4o: MWRCA experiment was that the model would choose the correct answer rather than selecting an incorrect option or responding with "IDK". The results produced by the model were then saved in a CSV file.

```
You are a knowledgeable and meticulous specialist in the medical field.
Your role is to analyze multiple-choice questions along with their
provided options. Given a multiple-choice question with four possi-
ble answers (0, 1, 2, 3), determine the correct answer and ‘ deliver a
comprehensive explanation for its accuracy.
Additionally, combine the explanations for why the remaining options are
incorrect into a single column labeled “Incorrect Explanations”.
The output must strictly follow this structure:
- Column 1: Question Header (the question text or ID).
- Column 2: Correct Option Index (e.g., 9, 1, 2, 3, or IDK or IDK).
- Column 3: Correct Explanation (a single, clear sentence ezolaning the
correct answer).
- Column 4: Incorrect Explanations (all incorrect option explanations
combined lnto one clear sentence, separated by semicolons).
Ensure that:
 1. Each column is clearly separated by a single comma (…).
 2. No explanation spills over into another column. Keep explanations and
limited to their respective columns.
 3. If uncertaln about the correct answer, write ‘IDK’ in Column 2, leave
Column 3 blank, and provide reasoning in Column 4
Provide only one row of output in the format:“Question ‘Expalanations”
Strict adherencce to this format is required  Any deviation is unacetable.
```

Fig. 3. Prompt instruction to ChatGPT-4o.

To calculate the accuracy of the model, its output answers were compared to the actual correct answers. For each correct response, the model received a score of one. The total score was then divided by the number of questions in each topic, as shown in Formula (1):

$$\text{Accuracy} = \frac{\text{TotalNumberofCorrectAnswers}}{\text{TotalNumberofQuestions}} \times 100 \quad (1)$$

To calculate the compliance of the model, the model's output was compared to the correct structure of the output format (question, correct option index, correct explanation, incorrect explanation). The model received a score of one if it followed the specific format; otherwise, it received a score of zero. The total score of the model was then divided by the total number of questions in each topic, as shown in Formula (2):

$$\text{Compliance} = \frac{\text{ScoreofCompliance}}{\text{TotalNumberofQuestions}} \times 100 \quad (2)$$

Finally, the uncertainty of the model was calculated based on the number of IDK responses it provided. The total number of IDK responses was then divided by the total number of questions to calculate the percentage of uncertainty, as shown in Formula (3):

$$\text{Uncertainty} = \frac{\text{TotalNmuberofIDKresponses}}{\text{TotalNumberofQuestions}} \times 100 \quad (3)$$

2) Analyzing Memorization and Reasoning with No Correct Answer (MWRNCA)

The GPT-4o: MWRNCA experiment followed the same structure as the GPT-4o: MWRNCA experiment in terms of constructing the prompt. The key difference between the two experiments is the structure of the benchmark used to evaluate the models. In this experiment, the set of questions presented to the model had no correct answer in the available choices for each question. Instead, options such as "Blank", "NONE", or "NONE OF THE ABOVE" were included to represent "No Correct Answer Provided". Fig. 4 shows an example of a question from the dataset. The expectation for the results was that the model would choose "Blank", "NONE", or "NONE OF THE ABOVE" as an answer for each question. To run GPT-4o: MWRNCA, the model was not given any hints about removing the correct answers from the dataset. The model was expected to know this by itself.

```
1 Question: {Post operative muscle ache is
      caused by}
2 Opt1: {d-TC}
3 Opt2: {None}
4 Opt3: { Gallamine}
5 Opt4: { Pancuronium}
```

Fig. 4. Question sample from the anaesthesia dataset.

The GPT-4o responses were evaluated based on accuracy, compliance, and uncertainty. Accuracy was calculated based on the number of "NONE" responses provided by the model, relative to the total number of questions [see Formula (4)]. Further, the percentage of the model compliance and uncertainty was calculated following Formula (2) and (3), respectively.

$$\text{Accuracy} = \frac{\text{TotalNmuberofNONEresponses}}{\text{TotalNumberofQuestions}} \times 100 \quad (4)$$

B. GPT-4O-Mini Hallucination Evaluation

1) Analyzing Memorization and Reasoning with Correct Answer (MWRCA)

In this experiment, GPT-4o-mini hallucination in memorization and reasoning was evaluated. Following the methodology of the GPT-4o: MWRCA experiment, questions from 21 different medical topics were passed to the model to answer. The same prompt defined in Section II was used with GPT-4o-mini. The model's responses were evaluated based on the model accuracy, compliance, and uncertainty, following Formula (1), (2), and (3), respectively.

2) Analyzing Memorization and Reasoning with No Correct Answer (MWRNCA)

In this experiment, the hallucination behavior of GPT-4o-mini was evaluated at both the memorization and reasoning levels. The prompt provided to the model was similar to the one in Section I, with the only change the model's name, i.e., GPT-4o-mini instead of GPT-4o. The dataset used in this experiment was the one that included no correct answers in the provided

answer choices. The model was expected to respond with "NONE"; any other response reduces its accuracy according to Formula (4). Compliance and uncertainty percentages are calculated using Formula (2) and (3), respectively.

V. Results

In each of the four experiments, the compliance rate, accuracy, and uncertainty were measured for each dataset. Additionally, the correlation between the size of the dataset and the accuracy was analyzed, as well as the learning rate after repeating the experiments. Lastly, the models were prompted to provide reasoning for their answers, and these justifications were sent to medical experts for evaluation of the models' reasoning.

A. Compliance Rate

The compliance rate is defined as the number of compliant responses divided by the total number of responses. As shown in Fig. 5, the compliance rate of both models slightly decreased across all datasets when the correct answer was replaced with "NONE" in the given options; however, this drop is insignificant. GPT-4o-mini outperformed GPT-4o in terms of compliance in nearly all experiments. The only exception was observed on the ENT dataset, where GPT-4o complied with the requested response format on about 99.8% of the questions when the correct answer was given in the options. On average, the compliance rates for the GPT-4o: MWRCA, GPT-4o: MWRNCA, GPT-4o-mini: MWRCA, and GPT-4o-mini: MWRNCA were 60.98%, 57.92%, 76.30%, and 72.94%, respectively.

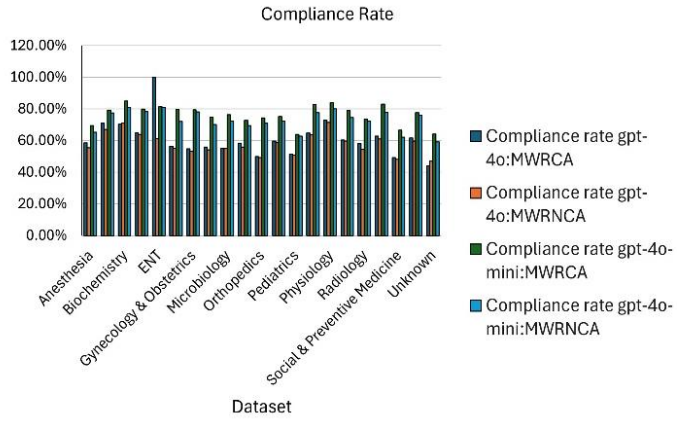

Fig. 5. The compliance rate of each experiment on each dataset.

B. Accuracy

The accuracy of each experiment across all datasets is shown in Fig. 6. Although GPT-4o-mini slightly outperforms GPT-4o on most datasets, both models perform poorly with accuracy rates below 70% on all datasets, except the ENT dataset, where GPT-4o achieves an accuracy of 81%. Both models show a significant drop in performance when the correct answer is replaced with "NONE", with average decreases of 75% for GPT-4o and 66% for GPT-4o-mini. The average accuracy for GPT-4o: MWRNCA, GPT-4o- mini: MWRCA, and GPT-4o-mini: MWRNCA are 48.60%, 11.90%, 53.13%, and 17.38%, respectively. These results indicate that both models are highly hallucinatory when applied to the given medical datasets.

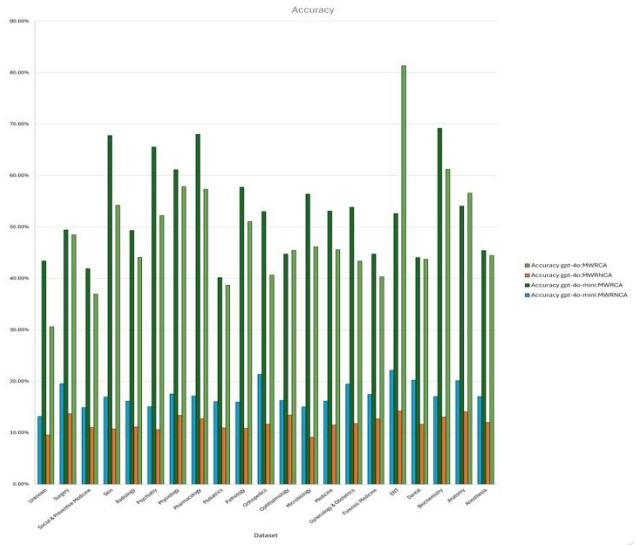

Fig. 6. The accuracy of each experiment on each dataset.

C. Uncertainty

When the correct answer is listed within the options, the models had less than 3% uncertainty on all datasets, as seen in Fig. 7. Additionally, GPT-4o reported a higher uncertainty than GPT-04-mini on all datasets. Notably, despite the low accuracy of all models as mentioned above, they all reported low uncertainties, averaging 0.88%, 26.63%, 0.59%, and 13.15% for GPT-4o: MWRCA, GPT-4o: MWRNCA, GPT-4o-mini: MWRCA, and GPT-4o-mini: MWRNCA, respectively.

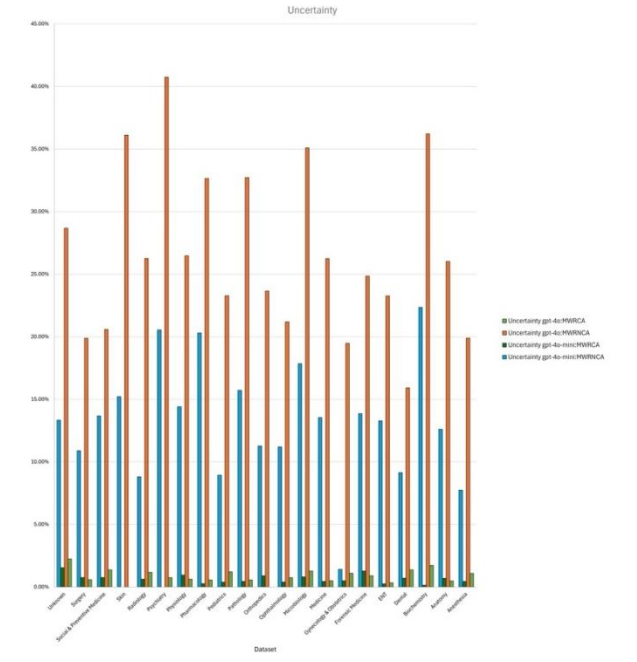

Fig. 7. The uncertainty of each experiment on each dataset.

D. Correlation

The correlation factors between the experiments' accuracy and the dataset sizes are all below 0.05, as seen in Fig. 8. The results indicate no clear correlation between dataset size and model performance.

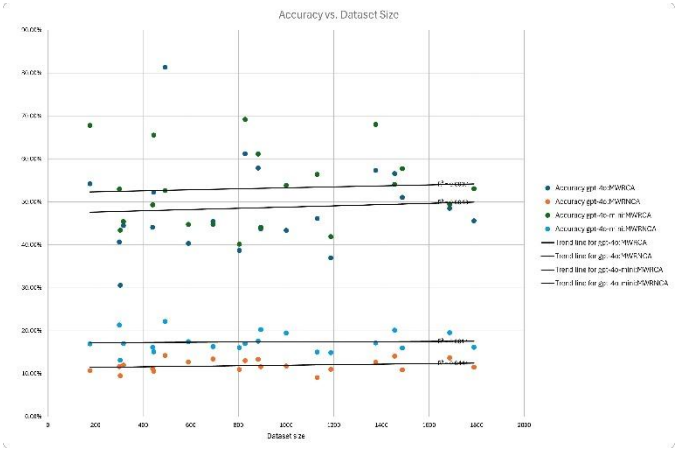

Fig. 8. The correlation between the dataset size and the accuracy.

E. Model Stability

To measure the models' performance stability, each of the four experiments was rerun on the following 10 selected datasets: Pathology, Pediatrics, Pharmacology, Physiology, Psychiatry, Radiology, Skin, Social and Preventive Medicine, Surgery, and Unknown. The computed difference between the accuracies of the two iterations of each of the four experiments, as in Formula (5), was 30.74% for GPT-4o with the correct answer, 1.75% for GPT-4o-mini with the correct answer, 3.72% for GPT-4o with "NONE" instead of the correct answer, and 1.64% GPT-4o-mini with "NONE'" replacing the correct answer. These results indicate that GPT-4o-mini demonstrates greater performance stability compared to GPT-4o.

$$\text{ChangeInAccuracy} = \text{SecondRunAccuracy} - \text{FirstRunAccuracy} \quad (5)$$

F. Reasoning

An electronic questionnaire was completed by 22 medical doctors. The questionnaire consisted of six sections, each representing a different medical field. Each section contained randomly selected case-based questions that both models answered correctly. Along with each question were two justifications, one from GPT-4o and the other from GPT-4o mini, along with the model's selected answer and its justification.

Participants were asked to evaluate the output of each model by selecting one of four possible options for each question which are:

- Agree with both GPT-4o and GPT-4o mini justifications.
- Agree with GPT-4o justification but disagree with GPT-4o mini justification.
- Disagree with both GPT-4o and GPT-4o mini justifications.
- Disagree with GPT-4o justification but agree with GPT-4o mini justification.

The results of this questionnaire revealed that the first choice had the highest percentage (68.8%) among all options. The third-choice percentage was the lowest (3.3%). The fourth option received 15.5%, while the second option received 12.4% of participants' selections. These results indicate that the justifications provided by both GPT-4o and GPT-4o-mini were considered reasonable for most of the questions.

Fig. 9 summarizes the average compliance, accuracy, and uncertainty for each experiment across all datasets. While GPT-4o-mini slightly outperformed GPT-4o in answering multiple-choice medical questions, both models frequently hallucinated and were unreliable for medical decision-making. Hallucination rates increased further when the correct answer was replaced with "NONE" as an option. Despite low overall accuracy, the models reported very low uncertainty when the correct answer was explicitly provided among the choices. However, substituting the correct answer with "NONE" reduced accuracy while increasing reported uncertainty. In terms of correlation, no relationship was found between dataset size and model accuracy, suggesting that repeated exposure to a topic does not improve performance. Additionally, retesting revealed that GPT-4o-mini's results remained stable across experiments, whereas GPT-4o's performance varied in the experiment with the correct answers explicitly included.

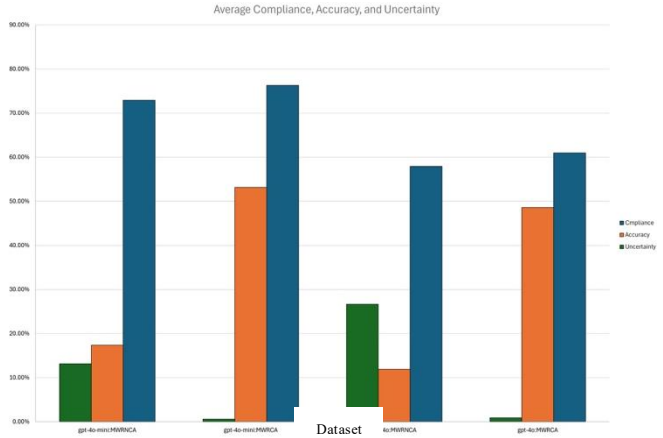

Fig. 9. The overall average of compliance, accuracy, and uncertainty.

VI. Discussion

The results of this study highlight significant findings regarding the occurrence of hallucinations in two large language models, GPT-4o and GPT-4o-mini, in the context of healthcare-related medical questions. Both models demonstrated notable challenges in their ability to reliably generate correct answers, with accuracy rates falling below 70% in most datasets. Although GPT-4o-mini performed better than GPT-4o in terms of compliance and accuracy, it still exhibited considerable hallucinatory behavior and weak adherence to a given task instruction. On the other hand, the accuracy of the two models dropped when the correct answer was replaced with a "NONE" option. This drop indicates that both models could not recognize the absence of the correct answer and generate unreliable answers without real understanding which highlights the limitations of these LLMs when response correctness is a key factor. Moreover, the high level of compliance combined with the drop in accuracy suggests that these LLMs cannot be relied upon, especially if used in high-stakes contexts such as healthcare. Furthermore, the study's experiments demonstrated a lack of correlation between the size of the question set sent to the model and its performance, showing no relation between

changing the size of the data and the model's accuracy which indicated that model hallucination is not sensitive to a given task size. Finally, re-running the experiments on smaller and specific datasets showed that GPT-4o-mini has minimal accuracy differences between runs compared to GPT-4o. This indicates that GPT-4o-mini may be a more stable model in real-world applications, although it remains at risk of hallucinations.

Comparing the results of this study with previous studies conducted on LLMs hallucination in the medical field, the presented results are consistent with the previous studies [24, 25, 26, 27] that asserted that the LLMs, including different GPT-versions exhibit hallucination and show low accuracy. In the context of evaluating model hallucination in different languages, two studies [24,25] conducted in non-English languages showed similar results to the studies conducted in English [26,27] and to this study, indicating that LLMs exhibit hallucination despite the language used. However, this research is distinguished from the previous studies in the variety of evaluation criteria and the medical topics. In this study, multiple tests were designed to measure specific aspects of the models, providing deeper insight into their performance. In addition to recall and reasoning, the study assessed models' stability and their correlation with the size of the question set. Also, the study explores the models' behavior in the absence of correct answers. Whereas previous studies tested the medical dataset ranged from one to four topics, this study's experiments were conducted on 21 medical topics to provide a deeper insight into hallucination and to avoid the limitation of a broad analysis.

In conclusion, this study concluded that GPT-4o and GPT-4o-mini exhibit hallucination in the medical field, highlighting the need for further study to test different LLMs and apply on different datasets. Finally, this research recommends improving the LLMs and using them cautiously, considering the hallucination in their responses.

VII. Conclusion and Future Work

The main findings of this study present the results of the compliance, accuracy, and uncertainty of each experiment across all datasets. Although GPT-4o-mini slightly outperformed GPT-4o in correctly answering multiple-choice questions, both models strongly hallucinated and cannot be relied upon for medical questions. Furthermore, the models hallucinated even more when the correct answer was replaced with "NONE" as an option. Despite the low accuracy, the models reported extremely low uncertainty rates for questions where the correct answer was explicitly listed among the options. Replacement of the correct answer with "NONE" in the options reduced the accuracy of the models but increased their uncertainty. Moreover, regarding correlation, there was no relationship between the dataset size and the models' accuracy, indicating that the models do not improve as they are questioned more on a particular topic. Additionally, retesting each model on the same questions from a sample of the datasets showed the stability of the GPT-4o-mini model compared to GPT-4o's results. Future work will focus on refining model training to reduce hallucinations and exploring methods to enhance reasoning abilities in specialized domains like healthcare.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., ''A survey on evaluation of large language models,'' ACM transactions on intelligent systems and technology, vol. 15, no. 3, pp. 1–45, 2024.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., ''Language models are few-shot learners,'' Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[3] B. T. Bicknell, D. Butler, S. Whalen, J. Ricks, C. J. Dixon, A. B. Clark, O. Spaedy, A. Skelton, N. Edupuganti, L. Dzubinski, et al., ''Chatgpt-4 omni performance in usmle disciplines and clinical skills: Comparative analysis,'' JMIR Medical Education, vol. 10, no. 1, p. e63430, 2024.

[4] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, ''Chatgpt in healthcare: a taxonomy and systematic review,'' Computer Methods and Programs in Biomedicine, p. 108013, 2024.

[5] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., ''Perfor- mance of chatgpt on usmle: potential for ai-assisted medical education us- ing large language models,'' PLoS digital health, vol. 2, no. 2, p. e0000198, 2023.

[6] 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,'' ACM Transactions on Information Systems, 2023.

[7] 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.

[8] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi, ''The ethics of ai in health care: a mapping review,'' Social Science & Medicine, vol. 260, p. 113172, 2020.

[9] S. Rogulsky, N. Popovic, and M. Färber, ''The effects of hallucinations in synthetic training data for relation extraction,'' Proceedings of the Knowl- edge Base Construction from Pre-Trained Language Models Workshop, 2024.

[10] R. Emsley, ''Chatgpt: these are not hallucinations–they're fabrications and falsifications,'' Schizophrenia, vol. 9, no. 1, p. 52, 2023.

[11] J. Gallifant, A. Fiske, Y. A. Levites Strekalova, J. S. Osorio-Valencia, R. Parke, R. Mwavu, N. Martinez, J. W. Gichoya, M. Ghassemi, D. Demner-Fushman, et al., ''Peer review of gpt-4 technical report and systems card,'' PLOS digital health, vol. 3, no. 1, p. e0000417, 2024.

[12] H. Ye, J. Xu, D. Huang, M. Xie, J. Guo, J. Yang, H. Bao, M. Zhang, and C. Zheng, ''Assessment of large language models' performances and hallucinations for chinese postgraduate medical entrance examination,'' Discover Education, vol. 4, no. 1, pp. 1–10, 2025.

[13] M. Kumar, U. A. Mani, P. Tripathi, M. Saalim, and S. Roy, ''Artificial hallucinations by google bard: think before you leap,'' Cureus, vol. 15, no. 8, 2023.

[14] N. Gillespie, S. Lockey, C. Curtis, J. Pool, and A. Akbari, ''Trust in artificial intelligence: A global study,'' The University of Queensland and KPMG Australia, vol. 10, 2023.

[15] K. Zhang, G. Li, H. Zhang, and Z. Jin, ''Hirope: Length extrapolation for code models using hierarchical position,'' in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13615–13627, 2024.

[16] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. Martins, ''Hallucinations in large multilingual translation models,'' Transactions of the Association for Computational Linguistics, vol. 11, pp. 1500–1517, 2023.

[17] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, ''A comprehensive survey of hallucination mitigation techniques in large language models,'' Findings of the Association for Computational Linguis- tics: EMNLP 2024, p. 11709–11724, November 12-16, 2024.

[18] T. Watari, S. Takagi, K. Sakaguchi, Y. Nishizaki, T. Shimizu, Y. Yamamoto, and Y. Tokuda, ''Performance comparison of chatgpt-4 and japanese med- ical residents in the general medicine in-training examination: comparison study,'' JMIR medical education, vol. 9, p. e52202, 2023.

[19] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, et al., ''Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset,'' Advances in Neural Information Processing Systems, vol. 36, 2024.

[20] L. Verlingue, C. Boyer, L. Olgiati, C. B. Mairesse, D. Morel, and J.-Y. Blay, ''Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice,'' The Lancet Regional Health– Europe, vol. 46, 2024.

[21] A. Pal, L. K. Umapathi, and M. Sankarasubbu, ''Med-halt: Medical domain hallucination test for large language models,'' in the 27th Conference on Computational Natural Language Learning (CoNLL), p. 314–334, Asso- ciation for Computational Linguistics, 2023.

[22] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al., ''Trustllm: Trustworthiness in large language models,'' in Proceedings of the 41st International Conference on Machine Learning, ICML'24, 2024.

[23] P. Puchert, P. Poonam, C. van Onzenoodt, and T. Ropinski, ''Llmmaps – a visual metaphor for stratified evaluation of large language models,'' 2023.

[24] T. Yudai, T. Nakata, K. Aiga, T. Etani, R. Muramatsu, S. Katagiri, H. Kawai, et al., "Performance of generative pretrained transformer on the national medical licensing examination in Japan," PLOS Digital Health, vol. 3, no. 1, 2024.

[25] H. Ye, et al., "Assessment of large language models' performances and hallucinations for Chinese postgraduate medical entrance examination," Discover Education, vol. 4, no. 1, pp. 1-10, 2025.

[26] A. Bharatha, N. Ojeh, A.M. Fazle Rabbi, et al., "Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy," Advances in Medical Education and Practice, pp. 393-400, 2024.

[27] A. Gilson, et al., "How Does ChatGPT Perform on the United States Medical Licensing Examination (USML)? The Implications of Large Language Models for Medical Education and Knowledge Assessment," JMIR medical education, vol. 9, no. 1, 2023.

[28] A. Pal, L. K. Umapathi, and M. Sankarasubbu, ''Medmcqa: A large- scale multi-subject multi-choice dataset for medical domainquestion answering,'' in Conference on health, inference, and learning, pp. 248–260, PMLR, 2022.