JMIR Publications
The Leading eHealth Publisher

Assessing the Performance of Plugin-Integrated ChatGPT-4 in the German Medical Board Examination: An Experimental Study on the Advancements and Limitations of Modern AI Modelling Approaches

Julian Madrid, Philipp Diehl, Mischa Selig, Bernd Rolauffs, Felix Patricius Hans, Hans-Jörg Busch, Tobias Scheef, Leo Benning

Submitted to: JMIR Medical Education
on: March 14, 2024

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript **5**
Supplementary Files **35**
..... **35**
Multimedia Appendixes **36**
Multimedia Appendix 1 ..... **37**
Multimedia Appendix 2 ..... **37**

Preprint
JMIR Publications

Assessing the Performance of Plugin-Integrated ChatGPT-4 in the German Medical Board Examination: An Experimental Study on the Advancements and Limitations of Modern AI Modelling Approaches

Julian Madrid[1] MD; Philipp Diehl[1] MD, PhD; Mischa Selig[2, 3] PhD; Bernd Rolauffs[2, 3] MD, PhD; Felix Patricius Hans[4, 2] MD, MSc; Hans-Jörg Busch[4, 2] MD, PhD; Tobias Scheef[2, 5*] MD; Leo Benning[4, 2*] MD, MPH

[1]Ortenau Klinikum Department of Cardiology, Pneumology, Angiology, Acute Geriatrics and Intensive Care Lahr DE
[2]University of Freiburg Faculty of Medicine Freiburg DE
[3]University of Freiburg Department of Orthopedics and Trauma Surgery G.E.R.N. Research Center for Tissue Replacement, Regeneration and Neogenesis Freiburg DE
[4]University of Freiburg Medical Center University Emergency Center Freiburg DE
[5]University of Freiburg Medical Center Department of Diagnostic and Interventional Radiology Freiburg DE
[*]these authors contributed equally

Corresponding Author:
Julian Madrid MD
Ortenau Klinikum
Department of Cardiology, Pneumology, Angiology, Acute Geriatrics and Intensive Care
Klosterstrasse 18
Lahr
DE

Abstract

Background: The Generative Pre-trained Transformer (GPT-4) is a large language model (LLM) trained and fine-tuned on an extensive dataset. After the public release of its predecessor in November 2022, the use of LLMs has seen a significant spike in interest, and a multitude of potential use cases have been proposed. In parallel, however, important limitations have been outlined. Particularly, current LLM encounters limitations, especially in symbolic representation and accessing contemporary data. The recent version of GPT-4, alongside newly released plugin features, has been introduced to mitigate some of these limitations.

Objective: Before this background, this work aims to investigate the performance of GPT-3.5, GPT-4, GPT-4 with plugins, and GPT-4 with plugins using pre-translated English text on the German medical board examination. Recognizing the critical importance of quantifying uncertainty for LLM applications in medicine, we furthermore assess this ability and develop a new metric termed 'confidence accuracy' to evaluate it.

Methods: We employed GPT-3.5, GPT-4, GPT-4 with plugins, and GPT-4 with plugins and translation to answer questions from the German medical board examination. Additionally, we conducted a thorough analysis to assess how the models justify their answers, the accuracy of their responses, and the error structure of their answers. Bootstrapping and confidence intervals were utilized to evaluate the statistical significance of our findings.

Results: This study demonstrated that available GPT models, as LLM examples, exceeded the minimum competency threshold established by the German medical board for medical students to obtain board certification to practice medicine. Moreover, the models could assess the uncertainty in their responses, albeit exhibiting overconfidence. Additionally, this work unraveled certain justification and reasoning structures that emerge when GPT generates answers.

Conclusions: The high performance of GPTs in answering medical questions positions it well for applications in academia and, potentially, clinical practice. Its capability to quantify uncertainty in answers suggests it could be a valuable AI agent within the clinical decision-making loop. Nevertheless, significant challenges must be addressed before AI agents can be robustly and safely implemented in the medical domain.

(JMIR Preprints 14/03/2024:58375)
DOI: https://doi.org/10.2196/preprints.58375

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Original article submitted to JMIR Medical Education

Assessing the Performance of Plugin-Integrated ChatGPT-4 in the German Medical Board Examination: An Experimental Study on the Advancements and Limitations of Modern AI Modelling Approaches

[1] Julian Madrid, [1]Philipp Diehl, [5]Mischa Selig, [5]Bernd Rolauffs, [2,3]Felix Patricius Hans,

[2,3]Hans-Jörg Busch, [4]Tobias Scheef, [2,3]Leo Benning

[1] Department of Cardiology, Pneumology, Angiology, Acute Geriatrics and Intensive Care Medicine, Ortenau Klinikum, Lahr, Germany

[2] University Emergency Center, Medical Center - University of Freiburg, Freiburg, Germany

[3] Faculty of Medicine, University of Freiburg, Freiburg, Germany

[4] Department of Diagnostic and Interventional Radiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[5] G.E.R.N. Research Center for Tissue Replacement, Regeneration & Neogenesis, Dept. of Orthopedics and Trauma Surgery

Word count: Abstract 324 words; Main text 6466 words (7 Tables, 2 Figures, 1 Supplementary File)

Correspondence and Reprint Requests:

Julian Madrid, MD
Department of Cardiology, Pneumology, Angiology, Acute Geriatrics and Intensive Care Medicine, Ortenau Klinikum, Lahr, Germany
Address: Klostenstraße 19, 77933 Lahr/Schwarzwald, Germany
Telephone: +49 7821932403
Email: julian.madrid@ortenau-klinikum.de
ORCID-ID: 0000-0001-5135-6873

Leo Benning, MD, MPH
Email: leo.benning@uniklinik-freiburg.de

Preprint JMIR Publications

Abstract:

Background: The Generative Pre-trained Transformer (GPT-4) is a large language model (LLM) trained and fine-tuned on an extensive dataset. After the public release of its predecessor in November 2022, the use of LLMs has seen a significant spike in interest, and a multitude of potential use cases have been proposed. In parallel, however, important limitations have been outlined. Particularly, current LLM encounters limitations, especially in symbolic representation and accessing contemporary data. The recent version of GPT-4, alongside newly released plugin features, has been introduced to mitigate some of these limitations.

Objective: Before this background, this work aims to investigate the performance of GPT-3.5, GPT-4, GPT-4 with plugins, and GPT-4 with plugins using pre-translated English text on the German medical board examination. Recognizing the critical importance of quantifying uncertainty for LLM applications in medicine, we furthermore assess this ability and develop a new metric termed 'confidence accuracy' to evaluate it.

Methods: We employed GPT-3.5, GPT-4, GPT-4 with plugins, and GPT-4 with plugins and translation to answer questions from the German medical board examination. Additionally, we conducted a thorough analysis to assess how the models justify their answers, the accuracy of their responses, and the error structure of their answers. Bootstrapping and confidence intervals were utilized to evaluate the statistical significance of our findings.

Results: This study demonstrated that available GPT models, as LLM examples, exceeded the minimum competency threshold established by the German medical board for medical students to obtain board certification to practice medicine. Moreover, the models could assess the uncertainty in their responses, albeit exhibiting overconfidence. Additionally, this work unraveled certain justification and reasoning structures that emerge when GPT generates answers.

Conclusion: The high performance of GPTs in answering medical questions positions it well for applications in academia and, potentially, clinical practice. Its capability to quantify uncertainty in answers suggests it could be a valuable AI agent within the clinical decision-making loop. Nevertheless, significant challenges must be addressed before AI agents can be robustly and safely implemented in the medical domain.

Keywords:

Medical education; artificial intelligence; generative pre-trained transformer; ChatGPT; Large language models; medical AI application

Introduction:

The Generative Pre-trained Transformer (GPT) –recently updated to its fourth iteration (GPT-4)– is a generative and autoregressive large language model (LLM). It is pre-trained on a vast corpus of internet text and fine-tuned on a labeled dataset using a transformer architecture [1–3]. With billions of parameters, GPT's scale is notably extensive [4]. GPT generates coherent and contextually appropriate text. It likely discovered a semantic grammar of language (i. e., semantic regularities), enabling it to construct semantically and syntactically correct sentences. However, GPT does not perform meaningful computations on symbolic representations [5–9].

The Wolfram language, a Turing-complete computational language, in contrast, represents the world in a precise computational and symbolic manner to compute answers to problems. GPT and the Wolfram language likely cover two different aspects of human cognition [5, 10, 11]. While GPT follows a statistical approach, reconstituting human-written internet text without creating new knowledge, the Wolfram language follows a symbolic approach, aiming to compute and potentially create new knowledge (i.e. computational creativity) [5].

Combining these features, particularly when computation and symbolic representations are needed, represents a significant step towards general AI. This combination has already been successfully used to examine contradictions in Einstein’s Special Theory of Relativity equations [12].

One limitation of earlier GPT models, most notably GPT-3.5, is their training on data only up to 2021, meaning they lack access to any more recent information. To overcome this limitation, other plugins have been developed and connected to GPT-4, enabling it to browse scholarly literature and the internet (as facilitated by Scholar AI and web requests).

In this light, LLMs show increasing promise in supporting medical practice and acting as educational agents. However, the models must acquire an in-depth and accurate representation of medical knowledge to be utilized in these sensitive domains. A medical board examination exemplifies these domains well, as it determines the qualification of medical students to obtain their license to practice medicine. In order to assess the model’s ability to achieve the minimum required score for passing both written parts of the examination, we submitted the German medical board examination to GPT. This represents our primary outcome. This task probably poses a different challenge to an LLM than medical board examinations in the English language [4, 13], as the performance of such models in other languages and in combination with more recent GPT versions remains unclear. Furthermore, the additional translation of non-English text into English before inputting it into GPT could represent a strategy to access the broader body of knowledge available online in English.

Especially in the medical field, where mistakes can have harmful consequences, assessing the amount of uncertainty is of paramount importance [14]. Furthermore, it is crucial to gain insights into the depth and structure the LLMs have of the medical knowledge representation and where the limitations lie [15]. For these reasons, we assessed the following secondary outcomes: we recorded the correct answer rates, the presence of logical justification of the answer, the presence of information internal to the question, the presence of information external to the question, the confidence GPT displays in its answers, the difficulty of the question, information errors, logical errors, reasoning errors, and the correctness of a second try answer when the first answer was wrong.

Methods:

Medical Board Examination Dataset

The German medical board examination consists of three steps. The first board examination, taken after two years of study, primarily covers basic natural sciences. It comprises 320 questions, which students answer over two consecutive days. The second board examination takes place after six years of study and serves as the final test before becoming a physician. It consists of 320 medical questions, which students answer over three consecutive days. The third board examination, also after six years of study, is an oral examination and was, hence, excluded from this study. The German medical board examination takes place biannually, once in spring and once in fall. As a representative sample, we used the medical board examination from spring 2023. We excluded questions the medical board examination committee deemed inconsistent with the medical literature in the regular post-exam review of the content. Additionally, we did not include questions displaying images, as current GPT models cannot analyze them. All questions and outcomes were provided by Amboss GmbH, a German medical education content creator and service provider.

GPT Models and Prompt Engineering

We evaluated several GPT models with varying characteristics using OpenAI's web interface. The models tested included GPT-3.5, GPT-4, GPT-4 integrated with the Wolfram, ScholarAI, and Web Request plugins, and GPT-4 integrated with the Wolfram, ScholarAI, Web Request plugins, and an additional feature for translating German inputs into English. We did not investigate earlier versions of GPT as they demonstrated lower performance in a similar study on the American medical board examination [13].

Creating a precise and adequate context is crucial for generating expected results [16][17]. Thus, we aimed to be as specific as possible, simulating the context of a medical student taking the medical board examination. The prompts hence included the request to answer each respective question with five possible answers, where only one answer was correct. We asked the models to justify their choices based on the provided patient case information, and to estimate their confidence in the answer’s accuracy as a percentage of maximal confidence (i.e. 100%). If the selected answer was incorrect, the GPT models were asked to explain their mistake in a second attempt. For the GPT-4 model with plugin integration, we asked the model to utilize the available plugins (Wolfram, ScholarAI, and Web Request). For the GPT-4 model with plugin integration and English translation, we first asked the model to translate the input into English language, and then to use the translated text to perform the abovementioned tasks. All utilized prompts are available in Supplementary File 1.

Model Testing and Outcome Parameters

For each GPT model, we used the appropriate prompt followed by the question and the possible answers. The investigators then analyzed the GPT's answer to assess the defined primary and secondary outcomes, which were either binary or proportions. In cases of uncertainty, the investigators (JM, TS, LB) convened to resolve the issue.

First, the correctness of the answer was recorded (binary variable), followed by the presence of logical justification, the presence of information internal to the question, and the presence of information external to the question (binary variables).

Next, we recorded the model's confidence in its answer (proportion), and the difficulty of the question, derived from the number of students who answered correctly on the Amboss platform (proportion).

To enhance our understanding of where GPT models falter, we sought to classify potential errors. As literature on error types is limited, we conducted a formal analysis to determine distinctive error types and established a formal definition. We propose a classification into three categories: information error, logical error, and reasoning error.

The GPT response can be formalized as 'answer A' is given ‘link’ because of ‘information B’. There are only three possibilities for errors: 1) 'Answer A' is incorrect because 'information B' is incorrect—termed an information error; 2) 'Answer A' is incorrect while 'information B' is correct, but the link between them is incorrect—termed a logical error; 3) 'Answer A' is incorrect, 'information B' is correct, and the link between 'answer A' and 'information B' is correct—termed a reasoning error. (See Figure 1). If the answer provided was incorrect, the investigator informed the GPT of its faulty answer, recorded whether it understood its mistake, and provided the correct answer in a second attempt. In the models with integrated plugin use, the active use of plugins was documented for Wolfram, ScholarAI, and Web Requests (binary variables).

Data Analysis

Summary statistics were calculated for the outcome variables (Tables 1, 2). Dichotomous variables were represented by frequency and proportions with 95% confidence intervals, while continuous variables were expressed as mean values with 95% confidence intervals. Uncertainty calculations displayed as 95% confidence intervals were computed via bootstrapping [18].

The primary outcome was determined by comparing the performance of the GPT-4 model, integrated with the Plugin and the English translation, to the required passing score for the medical board examination, which is 60%. The difference of proportions was calculated with 95% CI using bootstrapping (Table 3).

Subsequently, secondary outcomes were calculated: The final exam rate for each GPT model was compared to both chance and the required passing score for the medical board examination. The difference of proportions was calculated with 95% CI using bootstrapping (Table 3).

The proportions of logical justification within the answer, information internal to the answer, and information external to the answer were compared between correct and incorrect responses. The difference of proportions was calculated with 95% CI using bootstrapping (Table 4).

The model's confidence in its answers was compared between correct and incorrect responses. Additionally, the relationship between the model's confidence in its answers and

the difficulty of the question was assessed. Cohen's d values and 95% CI were computed using a linear regression model and bootstrapping (Tables 5, 6).
To evaluate the accuracy of the model's confidence in its answers, we developed a parameter termed Confidence Accuracy (CA). It is defined as follows:

CA = (confidence of correct answers in % – confidence of incorrect answers in %)/100

The difficulty of the question was assessed using real correct response proportions from students available on the Amboss platform. The difficulty was assessed as follows:

Difficulty = 1 – correct answer proportion

Then, the difficulty of the question was compared between correct and incorrect answers, with Cohen's d calculated using a linear regression model (Table 6).

Furthermore, we compared the proportion of correct answers between models (Table 7).

We compared the proportion of correct answers in the GPT4 models with the proportion of correct answers in the answers where a plugin has been used. We compared the proportion of plugin usage in GPT models with German and English input. We compared the confidence of the model when using plugins to the confidence of the model overall. We compared the proportion of correct answers when averaging the four different models to each model in particular.

In instances where questions were accompanied by images, GPT models sometimes responded by describing the image, although the models could not access the respective images. This phenomenon is known as a type of hallucination [19]. Therefore, we compared the proportion of hallucinations present in each model when answering questions, including image questions. We calculated the proportion of correct answers for each model when keeping the questions with pictures.

We compared the different error proportions between different models. We compared the proportion of logical errors when using the Wolfram plugin to the proportion of errors when using the entire model. We compared correct second-try answers between different models. These supplementary analyses are available in the supplementary files (Supplementary File 2).

The 95% confidence intervals were calculated using bootstrapping. Where necessary, parametric assumptions were tested using QQ plots for normality and Levene's tests for the homogeneity of variances. The independence of question answers was assumed. All statistical analyses were performed in Rstudio (Version 2023.06.0+421, Rstudio, Inc.). The significance level for all tests was set a priori at 95% CI.

Results:

All tests were performed on the 541 questions of the German medical board examination from spring 2023. Sub analyses were performed on other sub-groups, the respective sample sizes are indicated in the appropriate tables. All results for GPT3.5, GPT4, GPT4 + Plugin (GPT4P), and GPT4 + Plugin + Translation (GPT4PT) are listed in full detail in the tables and the supplementary materials. To ensure legibility, only relevant results are addressed in the result section.

Descriptive statistics with confidence intervals for the first board examination, second board examination, and the overall examination are displayed in Tables 1 and 2.
All models performed significantly better than chance: GPT3.5 (0.49 95%CI 0.45; 0.53), GPT4 (0.71 95%CI 0.69; 0.73), GPT4P (0.71 95%CI 0.69; 0.74), GPT4PT (0.70 95%CI 0.67; 0.72). Furthermore, all GPT models were significantly better than the required proportion to pass the final medical board examination. The difference between the GPT model score and the required score was as follows: GPT 3.5 (0.09 95% CI 0.05; 0.13), GPT 4 (0.31 95% CI 0.29; 0.34), GPT 4 + Plugin (0.31 95% CI 0.29; 0.34), GPT4 + Plugin + Translation (0.30 95% CI 0.27; 0.32).

All GPT models had a significantly higher proportion of providing a logical justification for correct answers compared to incorrect answers: GPT3.5 (0.37 95%CI 0.30; 0.45), GPT4 (0.31 95%CI 0.19; 0.44), GPT4P (0.25 95%CI 0.13; 0.38), GPT4PT (0.26 95%CI 0.15; 0.38). There was no statistical significance for the proportion of used internal information for correct and incorrect answers: GPT3.5 (-0.01 95%CI -0.04; 0.02), GPT4 (0 95%CI -0.01; 0), GPT4P (0 95%CI -0.01; 0), GPT4PT (0 95%CI -0.01; 0). Similarly, there was no statistical significance for the proportion of used external information for correct and incorrect answers: GPT3.5 (0.01 95%CI 0.; 0.04), GPT4 (0 95%CI -0.01; 0), GPT4P (0 95%CI 0; 0), GPT4PT (0 95%CI 0; 0).

All models had a confidence mean which was significantly higher for correct answers than incorrect answers; this was assessed using Cohen's d: GPT3.5 (-0.34 95%CI -0.52; -0.16), GPT4 (-0.93 95%CI -1.24; -0.63), GPT4P (-0.69 95%CI -0.99; -0.39), GPT4PT (-1.0 95%CI -1.29; -0.72). This is reflected in confidence accuracy values significantly different from zero: GPT3.5 (0.028 95%CI 0.011; 0.048), GPT4 (0.041 95%CI 0.023; 0.062), GPT4P (0.037 95%CI 0.021; 0.053), GPT4PT (0.043 95%CI 0.028; 0.059).

Questions which the models answered correctly, were less difficult for the medical students taking the respective exam as well, as shown by Cohen's d GPT3.5 (0.58 95%CI 0.40; 0.77), GPT4 (0.68 95%CI 0.39; 0.98), GPT4P (0.57 95%CI 0.27; 0.86), GPT4PT (0.92 95%CI 0.63; 1.20).

However, there was no statistically significant correlation between confidence of the model and question's difficulty, as shown by Pearson's r GPT3.5 (-0.008 95%CI -0.078; 0.061), GPT4 (-0.0816 95%CI -0.168; 0.0096), GPT4P (-0.0874 95%CI -0.176; 0.004), GPT4PT (-0.055 95%CI -0.138; 0.032).

The GPT4-based models all performed better than the GPT 3.5 model in providing correct answers as reflected in the difference of correct answer proportions: GPT3.5 vs GPT4 (0.22 95%CI 0.18; 0.27), GPT3.5 vs GPT4P (0.22 95%CI 0.18; 0.27), GPT3.5 vs GPT4PT (0.21

95%CI 0.16; 0.26). However, no GPT4-based model was better than another GPT4-based model, as reflected in the difference of correct answer proportions: GPT4 vs GPT4P (0.00 95%CI -0.03; 0.03), GPT4 vs GPT4PT (-0.013 95%CI -0.05; 0.02), GPT4P vs GPT4PT (-0.01 95%CI -0.05; 0.02).

The proportion of correct answers was not higher in answers where plugins were used compared to GPT4P (0.01 95%CI -0.067; 0.10) and GPT4PT (0.026 95%CI -0.07; 0.13). The proportion of correct answers was not higher in answers where the Wolfram plugin was used compared to GPT4P (0.026 95%CI -0.07; 0.13) and GPT4PT (0.026 95%CI -0.07; 0.13).

The proportion of plugin usage was significantly higher for GPT4P, where the input was German, compared to GPT4PT, where the input was English (0.094 95%CI 0.052; 0.14). The mean confidence of models was not higher when using plugins for GPT4P (0.004 95% CI -0.03; 0.02) and GPT4PT (0.004 95%CI -0.03; 0.02). Whereas the proportion of correct answers was significantly higher for an average model of all four models compared to GPT3.5 (0.23 95%CI 0.19; 0.28), GPT4 (0.01 95%CI -0.02; 0.04), GPT4P (0.01 95%CI -0.02; 0.04) and GPT4PT (0.02 95%CI -0.01; 0.06) showed no significant difference with the average model. GPT3.5 had significantly more hallucinations than GPT4 (0.24 95%CI 0.10; 0.37), GPT4P (0.20 95%CI 0.07; 0.34) and GPT4PT (0.24 95%CI 0.10; 0.37). However, there were not significantly more hallucinations between GPT4-based models: GPT4 vs GPT4P (-0.03 95%CI -0.18, 0.11), GPT4 vs GPT4PT (0.00 95%CI -0.14, 0.15), GPT4P vs GPT4PT (0.033 95%CI -0.01, 0.09). The proportions of correct answers when keeping questions with pictures that the models can’t currently analyze are displayed with confidence intervals in supplementary 2.

From all models, only GPT4P made significantly more reasoning errors than logical errors (0.37 95%CI 0.125; 0.60). All models made significantly more reasoning errors than information errors: GPT3.5 (0.21 95%CI 0.11; 0.30), GPT4 (0.44 95%CI 0.27; 0.60), GPT4P (0.52 95%CI 0.31; 0.71), GPT4PT (0.40 95%CI 0.20; 0.58). All models but GPT4P made significantly more logical errors than information errors: GPT3.5 (0.14 95%CI 0.029; 0.26), GPT4 (0.27 95%CI 0.10; 0.44), GPT4PT (0.22 95%CI 0.05; 0.38). GPT4 (0.12 95%CI 0.05; 0.22) and GPT4P (0.12 95%CI 0.02; 0.22) made significantly less information errors than GPT3.5.

The second try correct answer proportion was better than chance for all GPT models: GPT3.5 (0.29 95%CI 0.21; 0.36), GPT4 (0.42 95%CI 0.29; 0.54), GPT4P (0.50 95%CI 0.38; 0.63), GPT4PT (0.35 95%CI 0.22; 0.48). Only GPT4P had a higher second-try correct answer proportion than GPT3.5 (-0.21 95%CI -0.35; -0.07).

Discussion:

Primary Outcome

The GPT-4 models excelled in the medical board examination by not only exceeding the minimum required score of 60%, but also by outperforming most students in the given examinations. Specifically, for the first board examination, all GPT-4 models performed better than 98.6% of students. For the second board examination, they surpassed 95.8% of students, as detailed in the records of the examining body [20]. Furthermore, there was a significant gap between GPT-3.5 and the GPT-4 models. The larger models, with substantially more parameters and the capacity to remember longer prompts, appear to increase the accuracy of responses. However, we observed no additional benefit when GPT-4 models were paired with plugins.

During our study, we noted that the Wolfram Plugin was frequently utilized for more complex calculations. Yet, in the context of medical questions, complex mathematical procedures are typically not required and the use of symbolic language is usually not required. Thus, employing the Wolfram Alpha plugin is likely more beneficial for questions that involve extensive computations or advanced mathematical problems requiring symbolic representations. The Scholar AI plugin was activated for complex informational queries, but the resulting papers were not consistently useful. Surprisingly, the Internet Access plugin was the least utilized. This may be because answering medical questions typically demands expert-level knowledge, and general internet searches do not provide sufficiently specific information. Moreover, since the model has been trained on a vast amount of internet data, it likely already encompasses the knowledge available online within its parameters.

We speculated that posing questions in German might hinder the model's access to the broader body of knowledge available in English. However, this was not the case; the GPT model equipped with translation capabilities did not outperform the GPT-4 models without translation features. The GPT model likely abstracts high-level concepts and is not impeded by the language of the queries. This aligns with the LLMs' transformer architecture, which accesses higher-level concepts prior to translating text into another language [21].

Interestingly, the GPT-4 model with translation invoked plugins less frequently than the model without translation. We hypothesize that plugin calls occur at a lower level in the neural network, making them less necessary in English due to the larger available language corpus. In German, the model might need to delve deeper into the latent representation of concepts not tied to a specific language. However, this remains speculative and warrants further research.

The use of plugins did not yield a higher proportion of correct answers than the standard model. It is possible that GPT-4 already achieves a very high rate of accuracy, resulting in a ceiling effect. Hence, the addition of plugins may not offer a significant advantage for general-purpose questions.

Secondary Outcomes

While all models provided a very high proportion of logical justification for correct answers, it was significantly less extensive for incorrect answers. However, upon further analysis, we

did not detect a significant difference in the proportion of internal information from the question in the answer or in the use of external information not contained in the question between correct and incorrect answers. Thus, it appears that when LLMs provide justifications for their answers, the likelihood of the answer being correct is greater, regardless of the information content of the answer. One study already assessed the presence of logical justification in answers to USMLE questions, both correct and incorrect [13]. However, since all answers exhibited logical justification regardless of their accuracy, this metric could not be used as a discriminator for correctness.

We were unable to demonstrate a significant correlation between the model's confidence in an answer and the difficulty level of the question for humans. This suggests that the model's interpretation of question difficulty differs from that of humans. However, like humans, the model showed improved performance on easier questions compared to more challenging ones. Thus, it appears that the representation of question difficulty is distinct between LLMs and humans.

Conceptual Implications

Use for Medical Education

This performance suggests that LLMs like GPT could assume a greater role in medical education. Additionally, their integration could significantly change the conventional approach to medical education, which has traditionally emphasized the acquisition of knowledge through the consumption and memorization of medical literature. Yet the available medical literature has expanded exponentially in recent decades, reaching volumes far beyond human memorization capabilities [22–24]. The emergence of AI agents with superior information retention, therefore, prompts a reevaluation of our educational focus. In this light, teaching methodologies could shift towards navigating and structuring available information with digital tools, including AI agents. This approach could address the challenges posed by the era of big data, suggesting that our academic systems may not be adequately preparing students for the future of medical practice. The emphasis could shift from retaining information to learning how to efficiently access information and deeply understand these systems, along with their benefits and drawbacks.

Use in Clinical Practice

The utility of LLMs is not limited to educational settings but also extends to clinical practice. Since board examinations aim to replicate authentic medical scenarios, assessing the performance of LLMs on such exams can indicate their value as AI support tools in everyday clinical use. Although LLMs may not be as effective in highly specialized tasks where dedicated machine learning algorithms excel—for instance, XGBoost in identifying pulmonary embolisms [25–27] — LLMs are highly proficient in text processing and information integration from diverse algorithms [28]. This positions them as intelligent medical assistants, capable of transforming complex data into narratives that are comprehensible in a human context. Currently, clinicians have a limited understanding of AI agents and their functions. Yet, this knowledge is crucial for making safe and informed decisions with AI assistance in the future. Clinicians must, therefore, gain a thorough understanding of how various AI agents function, including their strengths and weaknesses. Additionally, a consensus on methodological and practical frameworks is needed to simplify the complexity of applied AI systems in a medical context. Indeed, user simplicity is a

prerequisite for the integration of complex AI systems into the stressful clinical environment. Moreover, validated guidelines and clinical algorithms need to incorporate AI meaningfully, not to replace but to enhance their efficiency. Yet, ongoing research is vital to fully delineate the potential of LLMs in the medical field.

Nonetheless, there is a risk inherent in blindly following an AI agent's guidance without fully understanding its operations [29, 30]. Due to the inherent complexity of LLMs, which often function as a black box, we can only partially monitor their operations at varying levels of complexity and behavior [29]. Given the marginal uncertainty intrinsic to such complex models, the AI agent should not supplant clinicians in decision-making, but rather provide additional informed perspectives. However, in this study, we attempted to analyze the task-oriented functions of LLMs, along with their inherent strengths and weaknesses in a medical context. Thus, it is essential for clinicians to incorporate these insights when using LLMs in a medical context.

In clinical practice, the level of confidence a physician has in a diagnosis is crucial for the quality and efficiency of care [31, 32]. The key attribute enabling this evaluation is the ability to quantify uncertainty, a trait humans are presumed to possess [14]. Yet, it has not been determined whether LLMs can quantify uncertainty, which raises concerns about the reliability of AI agent outputs. To address this, a standardized measure is needed to gauge the confidence in an AI agent's output. For binary outcomes like healthy/diseased, metrics such as specificity, sensitivity, and area under the curve are effective. For more complex queries with multiple potential answers –as managed by LLMs– traditional measures like sensitivity and specificity are inadequate. We therefore developed a new metric called 'Confidence Accuracy' (CA) which correlates the confidence assigned to an answer with its empirical accuracy. This allows for the quantification of uncertainty, crucial for clinical decision-making. This parameter ranges from -1 to 1, where 1 accurately reflects the model's uncertainty, 0 indicates no ability to quantify uncertainty, and -1 suggests incorrect quantification. Bayesian methods are preferred for quantifying uncertainty [33, 34], and while LLMs might implicitly use Bayesian principles to generate outputs [35], we hypothesized that LLMs could quantify uncertainty. Indeed, our work showed that all GPT models have the ability to quantify uncertainty. However, the expression in percentage does not seem to reflect the actual confidence for any specific decision (i.e., the models were overall largely overconfident). Although statistically different from zero, confidence accuracy values were consistently close to zero. New LLM methodologies aim to enhance this by incorporating uncertainty estimation [36]. Future AI agents should be fine-tuned using the confidence accuracy metric in order to improve uncertainty quantification, a critical objective for implementing AI as a supportive tool for physicians in clinical environments.

Identified Errors

Recognizing the essential need for clinicians to understand LLMs in the context of medical data, our study has examined the technical facets of posing medical queries to LLMs. We aimed to understand the ways in which LLMs process information, including their strengths and weaknesses, and to distinguish their reasoning from human thought processes through the analysis of their responses and justifications to medical data inquiries. We observed that GPT models commit different types of errors, particularly reasoning errors. Reasoning errors typically occur in situations where multiple options are correct, but one is more critical than the other. GPT models overproportionately make reasoning errors likely because this skill is acquired through human experience and is challenging to learn from text-based web sources, the primary training source of GPT. The second most common error type in GPT

models was logical errors. Since LLMs use a statistical approach to reconstruct human-written text, we anticipated difficulties with logic and mathematics, which require formal symbolic representation [5–9]. We hypothesized that the Wolfram plugin, utilizing the Wolfram Language, would mitigate these challenges. Yet, using the Wolfram plugin did not reduce the number of logical errors. Finally, fewer information errors were observed compared to other error types across all GPT models. This likely reflects the strength of these LLMs, which have assimilated a vast corpus of knowledge, greatly exceeding the reading and memorization capabilities of humans. In addition to the three error types derived from the informational and logical structure of GPT's answers, there are two error types that arise prior to answer generation. Firstly, due to the stochastic nature of token generation, there is likely a stochastic error inherent in all GPT responses. Secondly, due to in-context generation conditioned by the prompting strategy, a systematic error probably occurs as well. We attempted to mitigate the stochastic error by averaging the results from all models and selecting the most common outcome. However, the performance of such averaged models did not surpass that of the GPT-4 models. This may be because the signal was relatively strong compared to the noise, rendering the noise-to-signal ratio too small to be significant. Nevertheless, it would be interesting to further explore the extent of stochastic error (e.g., by repeating the same question multiple times) and systematic error (e.g., by varying the prompting strategy) present in LLMs like GPT, especially in the context of medical applications.

To assess whether the GPT models could recognize and correct their own mistakes, we prompted them to attempt another answer after providing incorrect responses. In most instances, the model would acknowledge the mistake and provide the correct answer along with a new explanation. This phenomenon could likely be attributed to the differing mechanics of forward and backward reasoning in LLMs. With forward reasoning, the LLM calculates the probability of the next token without a specific reasoning goal [37]. In contrast, backward reasoning enables the LLM to better contextualize the information. It is crucial to note, however, that we did not request the model to immediately reassess the answer; instead, we informed it of the answer’s incorrectness before asking for a reevaluation [37]. Future studies could further investigate the model's ability to self-correct without prior notification of its errors. To illustrate such errors, it is worthwhile to discuss the following example: Sometimes, GPT models correctly described the answer but inadvertently chose the wrong answer option from the five answer options. When asked to confirm its certainty due to the answer being incorrect, the model recognized the error, corrected itself, and apologized. This behavior demonstrates a surprisingly human-like trait.

In instances where questions were accompanied by images (i.e., the model did not have access to the images), GPT models, particularly GPT-3.5, often responded by describing the image that the model had not actually seen. This unexpected information error, known as a hallucination [19], persisted in the GPT-4 models, albeit at a significantly reduced frequency compared to GPT-3.5. Nevertheless, the propensity for overconfidence in entirely fabricated information remains a challenge for the latest generation of LLMs and is a phenomenon not fully understood [38]. However, new strategies to curb the hallucinations of LLMs appear promising [38, 39].

Limitations

Technological Limitations of LLMs

Although the results were impressive with GPT outperforming most students in the German medical board examination, it is crucial to remember that these models still possess significant limitations. During our study, GPT-4 was incapable of interpreting medical images, such as chest X-rays or histological samples. This is a considerable drawback, given that medical information is inherently multimodal, and the ability to integrate multimodal data will be essential for the adoption of such models in academic and clinical settings. Nonetheless, even when including questions that GPT could not answer due to image-related content, the top-performing model achieved a score of 83%. It is anticipated that future GPT iterations and other LLMs will be fully multimodal, which necessitates additional research to evaluate their performance across a more diverse array of questions.
A further concern relates to the stochastic nature of token generation, meaning that answers may vary slightly when questions are posed multiple times [40]. However, if the noise-to-signal ratio is very low, it becomes negligeable compared to a human agent output. Even if the noise-to-signal ratio is very low, problems persist in the framework trying to regulate these models.
A further concern pertains to the prompt sensitivity of LLMs. This trait can be advantageous as it allows the incorporation of context into the generation of meaningful output and may contribute to the models' Bayesian characteristics [35]. However, prompt sensitivity also increases the risk of systematic errors with repetitive use of the same prompt. Prompt engineering is a discipline that emerged in trying to minimize systematic errors [41, 42].
Within the extensive volume of data available online, there are significant risks of bias. Given that LLMs are trained on vast datasets, there is an inherent risk of adopting biases from the underlying data structures, which are beyond our control. However, fine-tuning through supervised learning on labeled data can help mitigate these risks [43, 44].

Limitations of the use of LLMs in a medical context

Despite the immediate promise of using LLMs in both educational and clinical contexts, the current ethical and regulatory environment needs to be considered to advance the use of these novel technologies safely and accountably.

As the representation of medical information of an LLM must not be confused with medical knowledge from a medical professional, it remains crucial to enable students and medical professionals alike to identify LLM-generated outputs as such in order to interpret them very carefully. Different to, for example, a senior medical colleague providing guidance for a clinical decision, a LLM-generated output is neither based on clinical knowledge, nor experience. The risk of such confusion has been described as anthropomorphic projection and efforts for advancing these novel technologies in the medical field need to simultaneously foster the awareness of such phenomena. This differentiation resonates with the provisions of the EU on a risk-based assessment approach [45] and, more recently, with the Bletchley Declaration [46]. The latter emphasizes the risks at the “frontier” of AI, at which we operate with the presented project.

While the concerns discussed in the context of medical education – and, more widely, training – are mainly within the realm of AI ethics, more specific limitations apply to the clinical use of these technologies. As of today, no commercially available LLM in the

European Union (EU) – including the GPT versions assessed in this work – have an assigned intended medical use, a basic regulatory prerequisite for their use in a clinical context. Without such intended medical use, the Medical Device Regulation (MDR), the regulatory framework for medical devices in the EU, is not applicable. Hence, such a device would not be a medical device in the regulatory sense and could, therefore, not be used in a clinical context without irresponsible safety and liability risks. While it is not the user (e.g. researchers, clinicians), but the provider (e.g. OpenAI for the ChatGPT models) to provide an intended medical use – which itself comes with further regulatory requirements -, the clinical use of the currently available and mostly all-purpose LLMs remains challenging.

Yet, even developing a LLM with an intended medical use and fulfilling all adjacent regulatory requirements would – as of now – not necessarily resolve the challenge centering around the clinical use of such program, as a key requisite for software as a medical device (SaMD) outlined in the MDR (“Devices that incorporate electronic programmable systems, including software, or software that are devices in themselves, shall be designed to ensure repeatability, reliability and performance in line with their intended use.” MDR Annex I, Rule 17.1 [47]) is currently considered to be violated, although this question remains subject to the legal and scholarly debate.

However, the rapid development of technological advances and the concurrent establishment of respective regulations should not be perceived as a “race to get grips with AI” [48], but should be viewed as a co-evolution to eventually yield the best population-wide benefit from these advances. In this light, the emphasis of a “pro-innovation and proportionate governance”, as proposed in the Bletchley Declaration, is equally crucial as the implementation of regulatory frameworks.

Limitations of the study

Lastly, there are limitations to the study. We utilized a specific German medical board examination as a sample to represent the general distribution of medical questions. While it is acknowledged that questions evolve over time and may introduce bias, the objective of the medical board examination is to maintain a consistent level of difficulty, reflecting the minimum required knowledge to become a certified physician. Furthermore, the distribution of student grades has remained relatively stable over time, leading us to believe that this potential bias is minimal. In the model with translation, we employed GPT to translate the questions before applying them to the model. Although we did not observe any, it is possible that translation errors occurred, potentially acting as a confounder in the study. In the context of the medical board examination, multiple-choice questions are posed to elicit clear answers that can be quantitatively assessed. By contrast, in a clinical setting, questions tend to be open-ended, which introduces a different dynamic. Nevertheless, we asked the model to justify its answers to glean insight into its computational process, thus rendering the questions more comparable to open-ended inquiries.

Conclusion

The performance of GPT models in the German medical board examination have surpassed those of most students, indicating a significant stride forward in the application of

AI in the fields of education and potentially clinical practice.

While GPT appears to possess a latent representation of uncertainty, it currently exhibits a significant degree of overconfidence. We have developed a new metric known as ‘Confidence Accuracy’ (CA), which could facilitate the appropriate measurement and fine-tuning of models to improve this aspect.

When analyzing the properties of correct and incorrect answers, there are strategies that can enhance the accuracy of responses due to the technical aspects of LLMs. However, there are numerous limitations that clinicians should be aware of.

Upon analysis, GPT models resemble humans in certain respects and differ in others.
Challenges such as hallucinations, the stochastic nature of token generation, and prompt sensitivity are highlighted, indicating areas for further research and development. Further, we see remaining open questions regarding the ethical and regulatory use of LLMs in the educational and clinical context, which need to be addressed on a policy level.

Overall, while GPT models have demonstrated a high degree of proficiency, their full integration into medical practice requires careful consideration of their limitations, a deeper understanding of their decision-making processes, and further refinement to enhance their multimodal capabilities and reduce inherent biases. The future of AI in medicine hinges on the balanced development of these models, ensuring they serve as reliable adjuncts to human expertise in clinical environments.
LLMs have ushered medicine in a new era, the surface of which we are only just beginning to scratch, revealing a new frontier of complexity. What will be discovered beneath remains a thrilling prospect.

Preprint JMIR Publications

Acknowledgements.
JM participated in the conceptualization, data acquisition, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing – original draft, writing – review & editing, and should be considered the first author.
PD, MS, BR, FPH and HJB participated in the methodology and review editing and should be considered as second authors.
LB and TS participated in the conceptualization, data acquisition, formal analysis, investigation, methodology, validation, writing – review & editing, and should be considered last authors.

Corresponding authors:
Correspondence should be addressed to Julian Madrid and Leo Benning

Conflicts of interest:
Authors declare no conflicts of interest

Preprint
JMIR Publications

References:

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg U Von, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.
2. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. npj Digit Med 2021 41. 2021;4:1–3.
3. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023.
4. Nori H, King N, Mckinney SM, Carignan D, Horvitz E, Openai M 2. Capabilities of GPT-4 on Medical Challenge Problems. 2023.
5. Wolfram S. What Is ChatGPT Doing... and Why Does It Work? Stephen Wolfram; 2023.
6. Traylor A, Feiman R, Pavlick E. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. ACL-IJCNLP 2021 - 59th Annu Meet Assoc Comput Linguist 11th Int Jt Conf Nat Lang Process Proc Conf. 2021;2:158–67.
7. Misra K, Rayz J, Ettinger A. COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. EACL 2023 - 17th Conf Eur Chapter Assoc Comput Linguist Proc Conf. 2022;:2920–41.
8. Kim N, Linzen T. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. EMNLP 2020 - 2020 Conf Empir Methods Nat Lang Process Proc Conf. 2020;:9087–105.
9. Ettinger A. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Trans Assoc Comput Linguist. 2020;8:34–48.
10. Goertzel B, Singularitynet *. Generative AI vs. AGI: The Cognitive Strengths and Weaknesses of Modern LLMs. 2023.
11. Vzorin G, Bukinich A, Sedykh A, Vetrova I, Sergienko E. Emotional Intelligence of GPT-4 Large Language Model. 2023. https://doi.org/10.20944/PREPRINTS202310.1458.V1.
12. Bryant S. Assessing GPT-4's Role as a Co-Collaborator in Scientific Research: A Case Study Analyzing Einstein's Special Theory of Relativity. Artiicial Intell. 2023. https://doi.org/10.21203/RS.3.RS-2808494/V2.
13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9e45312 https//mededu.jmir.org/2023/1/e45312. 2023;9:e45312.
14. Cosmides L, Tooby J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition. 1996;58:1–73.
15. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388:1233–9.
16. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt Engineering for Healthcare: Methodologies and Applications. 2023.
17. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What Makes Good In-Context Examples for GPT-3? DeeLIO 2022 - Deep Learn Insid Out 3rd Work Knowl Extr Integr Deep Learn Archit Proc Work. 2021;:100–14.
18. Haukoos JS, Lewis RJ. Advanced Statistics: Bootstrapping Confidence Intervals for Statistics with "Difficult" Distributions. Acad Emerg Med. 2005;12:360–5.
19. Wang J, Zhou Y, Xu G, Shi P, Zhao C, Xu H, et al. Evaluation and Analysis of Hallucination in Large Vision-Language Models. arXiv. 2023;:arXiv:2308.15126.
20. Archiv Medizin. 2024. https://www.impp.de/pruefungen/medizin/archiv-medizin.html. Accessed 4 Jan 2024.
21. Belinkov Y, Glass J. Analysis Methods in Neural Language Processing: A Survey. Trans

Assoc Comput Linguist. 2019;7:49–72.
22. Austin C, Kusumoto F. The application of Big Data in medicine: current implications and future directions. J Interv Card Electrophysiol. 2016;47:51–9.
23. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. Front Med. 2019;6 MAR:414018.
24. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inform. 2017;98:22–32.
25. Ryan L, Maharjan J, Mataraso S, Barnes G, Hoffman J, Mao Q, et al. Predicting pulmonary embolism among hospitalized patients with machine learning algorithms. Pulm Circ. 2022;12:e12013.
26. Dua R, Ronald Wallace G, Chotso T, Francis Densil Raj V. Classifying Pulmonary Embolism Cases in Chest CT Scans Using VGG16 and XGBoost. Lect Notes Data Eng Commun Technol. 2023;131:273–92.
27. Ding R, Ding Y, Zheng D, Huang X, Dai J, Jia H, et al. Machine Learning-Based Screening of Risk Factors and Prediction of Deep Vein Thrombosis and Pulmonary Embolism After Hip Arthroplasty. Clin Appl Thromb. 2023;29.
28. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. 2023.
29. Verdicchio M, Perin A. When Doctors and AI Interact: on Human Responsibility for Artificial Risks. Philos Technol. 2022;35:1–28.
30. Xu J. Overtrust of Robots in High-Risk Scenarios. AIES 2018 - Proc 2018 AAAI/ACM Conf AI, Ethics, Soc. 2018;:390–1.
31. Zwaan L, Hautz WE. Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. BMJ Qual Saf. 2019;0:1–4.
32. Borracci RA, Arribalzaga EB. The Incidence of Overconfidence and Underconfidence Effects in Medical Student Examinations. J Surg Educ. 2018;75:1223–9.
33. Park I, Amarchinta HK, Grandhi R V. A Bayesian approach for quantification of model uncertainty. Reliab Eng Syst Saf. 2010;95:777–85.
34. Kwon Y, Won JH, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput Stat Data Anal. 2020;142:106816.
35. Xie SM, Raghunathan A, Liang P, Ma T. An Explanation of In-context Learning as Implicit Bayesian Inference. ICLR 2022 - 10th Int Conf Learn Represent. 2021.
36. Sankararaman KA, Wang S, Fang H. BayesFormer: Transformer with Uncertainty Estimation. 2022.
37. Jiang W, Shi H, Yu L, Liu Z, Zhang Y, Li Z, et al. Forward-Backward Reasoning in Large Language Models for Mathematical Verification. 2023.
38. Yao J-Y, Ning K-P, Liu Z-H, Ning M-N, Yuan L. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. 2023.
39. Zhang X, Guo Y, Stepputtis S, Sycara K, Campbell J. Explaining Agent Behavior with Large Language Models. 2023.
40. Bender EM, Gebru T, Mcmillan-Major A, Shmitchell S, Shmitchell S-G. On the dangers of stochastic parrots: Can language models be too big?. dl.acm.orgEM Bender, T Gebru, A McMillan-Major, S ShmitchellProceedings 2021 ACM Conf fairness, accountability, and, 2021•dl.acm.org. 2021;:610–23.
41. Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-shot Performance of Language Models. 2021;:12697–706.
42. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large Language Models Are Not Robust Multiple Choice Selectors. arXiv. 2023;:arXiv:2309.03882.
43. Jin X, Barbieri F, Kennedy B, Davani AM, Neves L, Ren X. On Transferability of Bias

Mitigation Effects in Language Model Fine-Tuning. NAACL-HLT 2021 - 2021 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Proc Conf. 2020;:3770–83.
44. Chu T, Song Z, Yang C. Fine-tune Language Models to Approximate Unbiased In-context Learning. 2023.
45. Regulation of the European Parliament. 2021. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF. Accessed 4 Jan 2024.
46. The Bletchley Declaration. 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Accessed 4 Jan 2024.
47. ANNEX I Medical Device Regulation. https://www.medical-device-regulation.eu/2019/07/23/annex-i-general-safety-and-performance-requirements/. Accessed 4 Jan 2024.
48. Regulating the machine. 2023. https://www.politico.eu/article/regulate-europe-race-artificial-intelligence-ai-drugs-medicines/. Accessed 4 Jan 2024.

Preprint JMIR Publications

Figures:

Figure 1 Formal Definition of Error Types; We propose a classification into three categories: information error, logical error, and reasoning error. The GPT response can be formalized as 'answer A' is given ‘link’ because of ‘information B’. There are only three possibilities for errors: 1) 'Answer A' is incorrect because 'information B' is incorrect—termed an information error; 2) 'Answer A' is incorrect while 'information B' is correct, but the link between them is incorrect—termed a logical error; 3) 'Answer A' is incorrect, 'information B' is correct, and the link between 'answer A' and 'information B' is correct—termed a reasoning error.

Table 1 Performance, Information Content, Confidence and Plugin Usage of GPT Model Answers

Performance of GPT Models (Step 1)								
	Correct Answer (Proportion +/- 95% CI)	Logical Justification (Proportion +/- 95% CI)	Presence of Internal Information (Proportion +/- 95% CI)	Presence of External Information (Proportion +/- 95% CI)	Confidence Mean (+/- 95% CI)	Use of Plugin Wolfram (Proportion +/- 95% CI)	Use of Plugin ScholarAI (Proportion +/- 95% CI)	Use of Plugin Web Requests (Proportion +/- 95% CI)
GPT 3.5 (N=278)	193 (0.69 +/- 0.64; 0.75)	252 (0.91 +/- 0.87; 0.94)	270 (0.97 +/- 0.95; 0.99)	277 (1 +/- 0.99; 1)	0.946 (0.936; 0.953)	NA	NA	NA
GPT 4 (N=278)	264 (0.95 +/- 0.92; 0.97)	270 (0.97 +/- 0.95; 0.99)	277 (1 +/- 0.99; 1)	277 (1 +/- 0.99; 1)	0.956 (0.952; 0.961)	NA	NA	NA
GPT 4 + Plugin (N=278)	261 (0.94 +/- 0.91; 0.96)	272 (0.98 +/- 0.96; 0.99)	278 (1 +/- 1; 1)	278 (1 +/- 1; 1)	0.943 (0.938; 0.948)	38 (0.14 +/- 0.10; 0.18)	34 (0.12 +/- 0.09; 0.16)	2 (0.01 +/- 0.00; 0.02)
GPT 4 + Plugin + Translation (N=278)	261 (0.94 +/- 0.91; 0.96)	271 (0.97 +/- 0.96; 0.99)	275 (0.99 +/- 0.97; 1)	278 (1 +/- 1; 1)	0.938 (0.933; 0.943)	26 (0.09 +/- 0.06; 0.13)	22 (0.08 +/- 0.05; 0.11)	6 (0.02 +/- 0.01; 0.04)

Performance of GPT Models (Step 2)								
	Correct Answer (Proportion +/- 95% CI)	Logical Justification (Proportion +/- 95% CI)	Presence of Internal Information (Proportion +/- 95% CI)	Presence of External Information (Proportion +/- 95% CI)	Confidence Mean (+/- 95% CI)	Use of Plugin Wolfram (Proportion +/- 95% CI)	Use of Plugin ScholarAI (Proportion +/- 95% CI)	Use of Plugin Web Requests (Proportion +/- 95% CI)
GPT 3.5 (N=263)	180 (0.68 +/- 0.63; 0.74)	227 (0.86 +/- 0.82; 0.90)	251 (0.95 +/- 0.93; 0.98)	261 (0.99 +/- 0.98; 1)	0.876 (0.867; 0.884)	NA	NA	NA
GPT 4 (N=263)	229 (0.87 +/- 0.83; 0.91)	256 (0.97 +/- 0.95; 0.99)	260 (0.99 +/- 0.97; 1)	263 (1 +/- 1; 1)	0.919 (0.913; 0.924)	NA	NA	NA
GPT 4 + Plugin (N=263)	232 (0.88 +/- 0.84; 0.92)	257 (0.98 +/- 0.96; 0.99)	259 (0.98 +/- 0.97; 0.99)	263 (1 +/- 1; 1)	0.894 (0.888; 0.900)	12 (0.05 +/- 0.02; 0.07)	73 (0.28 +/- 0.22; 0.33)	0 (0 +/- 0.0; 0.0)
GPT 4 + Plugin + Translation (N=263)	225 (0.86 +/- 0.81; 0.90)	256 (0.97 +/- 0.95; 0.99)	262 (1 +/- 0.99; 1)	263 (1 +/- 1; 1)	0.899 (0.894; 0.903)	21 (0.08 +/- 0.05; 0.11)	25 (0.1 +/- 0.06; 0.13)	19 (0.07 +/- 0.04; 0.10)
Performance of GPT Models (All Questions)								
	Correct	Logical Justific	Presence of	Presence of	Confidence	Use of Plugin	Use of Plugin	Use of Plugin

	Answer (Proportion +/- 95% CI)	ation (Proportion +/- 95% CI)	Internal Information (Proportion +/- 95% CI)	External Information (Proportion +/- 95% CI)	Mean (+/- 95% CI)	Wolfram (Proportion +/- 95% CI)	ScholarAI (Proportion +/- 95% CI)	Web Requests (Proportion +/- 95% CI)
GPT 3.5 (N=541)	373 (0.69 +/- 0.65; 0.73)	479 (0.89 +/- 0.86; 0.91)	521 (0.96 +/- 0.95; 0.98)	538 (0.99 +/- 0.99; 1)	0.912 (0.904; 0.918)	NA	NA	NA
GPT 4 (N=541)	493 (0.91 +/- 0.89; 0.93)	526 (0.97 +/- 0.96; 0.98)	537 (0.99 +/- 0.98; 1)	540 (1 +/- 0.99; 1)	0.938 (0.934; 0.942)	NA	NA	NA
GPT 4 + Plugin (N=541)	493 (0.91 +/- 0.89; 0.94)	529 (0.98 +/- 0.96; 0.99)	537 (0.99 +/- 0.98; 1)	541 (1 +/- 1; 1)	0.919 (0.915; 0.924)	50 (0.09 +/- 0.07; 0.12)	107 (0.20 +/- 0.16; 0.23)	2 (0.003 +/- 0.0; 0.01)
GPT 4 + Plugin + Translation (N=541)	486 (0.90 +/- 0.87; 0.92)	527 (0.97 +/- 0.96; 0.99)	537 (0.99 +/- 0.98; 1)	541 (1 +/- 1; 1)	0.919 (0.915; 0.923)	47 (0.09 +/- 0.06; 0.11)	47 (0.09 +/- 0.06; 0.11)	25 (0.05 +/- 0.03; 0.06)

Table 2 Question's Difficulty and Error Structure of GPT Model Answers

Performance of GPT Models (Step 1)						
	Question's Difficulty Mean (+/- 95% CI)	Error overall (Proportion +/- 95% CI)	Information Error (Proportion +/- 95% CI)	Logical Error (Proportion +/- 95% CI)	Reasoning Error (Proportion +/- 95% CI)	Correct Answer in Second Try (Proportion +/- 95% CI)
GPT 3.5 (N=278)	0.300 (0.279; 0.324)	85 (0.31 +/- 0.25; 0.36)	34 (0.4 +/- 0.29; 0.51)	35 (0.41 +/- 0.31; 0.52)	17 (0.2 +/- 0.12; 0.29)	44 (0.52 +/- 0.41; 0.62)
GPT 4 (N=278)	0.300 (0.279;	14 (0.05 +/- 0.03;	3 (0.21 +/- 0.0;	9 (0.64 +/- 0.36;	3 (0.21 +/- 0.0;	11 (0.79 +/- 0.57;

	0.324)	0.08)	0.43)	0.86)	0.43)	1.0)
GPT 4 + Plugin (N=278)	0.300 (0.279; 0.324)	17 (0.06 +/- 0.036; 0.09)	5 (0.29 +/- 0.12; 0.53)	5 (0.29 +/- 0.12; 0.53)	7 (0.41 +/- 0.18; 0.65)	15 (0.88 +/- 0.71; 1.0)
GPT 4 + Plugin + Translation (N=278)	0.300 (0.279; 0.324)	17 (0.06 +/- 0.036; 0.09)	6 (0.35 +/- 0.12; 0.59)	10 (0.59 +/- 0.35; 0.82)	1 (0.06 +/- 0.0; 0.18)	11 (0.65 +/- 0.41; 0.88)
Performance of GPT Models (Step 2)						
	Question's Difficulty Mean (+/- 95% CI)	Error overall (Proportion +/- 95% CI)	Information Error (Proportion +/- 95% CI)	Logical Error (Proportion +/- 95% CI)	Reasoning Error (Proportion +/- 95% CI)	Correct Answer in Second Try (Proportion +/- 95% CI)
GPT 3.5 (N=263)	0.274 (0.256; 0.294)	83 (0.32 +/- 0.26; 0.37)	3 (0.04 +/- 0.0; 0.08)	26 (0.31 +/- 0.22; 0.41)	55 (0.66 +/- 0.55; 0.76)	46 (0.55 +/- 0.45; 0.66)
GPT 4 (N=263)	0.274 (0.256; 0.294)	34 (0.13 +/- 0.09; 0.17)	2 (0.06 +/- 0.0; 0.15)	9 (0.26 +/- 0.12; 0.41)	23 (0.68 +/- 0.5; 0.82)	21 (0.62 +/- 0.44; 0.76)
GPT 4 + Plugin (N=263)	0.274 (0.256; 0.294)	31 (0.12 +/- 0.08; 0.16)	0 (0 +/- 0.0; 0.0)	7 (0.23 +/- 0.10; 0.39)	23 (0.74 +/- 0.58; 0.87)	21 (0.68 +/- 0.52; 0.84)
GPT 4 + Plugin + Translation (N=263)	0.274 (0.256; 0.294)	38 (0.14 +/- 0.10; 0.19)	1 (0.03 +/- 0.0; 0.08)	9 (0.24 +/- 0.11; 0.37)	28 (0.74 +/- 0.58; 0.87)	22 (0.58 +/- 0.42; 0.74)
Performance of GPT Models (All Questions)						
	Question's Difficulty Mean (+/- 95% CI)	Error overall (Proportion +/- 95% CI)	Information Error (Proportion +/- 95% CI)	Logical Error (Proportion +/- 95% CI)	Reasoning Error (Proportion +/- 95% CI)	Correct Answer in Second Try (Proportio

						n +/- 95% CI)
GPT 3.5 (N=541)	0.288 (0.272; 0.303)	168 (0.31 +/- 0.27; 0.35)	37 (0.22 +/- 0.16; 0.29)	61 (0.36 +/- 0.29; 0.43)	72 (0.42 +/- 0.36; 0.51)	90 (0.54 +/- 0.46; 0.61)
GPT 4 (N=541)	0.288 (0.272; 0.303)	48 (0.09 +/- 0.07; 0.11)	5 (0.1 +/- 0.02; 0.19)	18 (0.38 +/- 0.25; 0.52)	26 (0.54 +/- 0.40; 0.69)	32 (0.67 +/- 0.52; 0.79)
GPT 4 + Plugin (N=541)	0.288 (0.272; 0.303)	48 (0.09 +/- 0.06; 0.11)	5 (0.1 +/- 0.02; 0.20)	12 (0.25 +/- 0.125; 0.375)	30 (0.63 +/- 0.48; 0.75)	36 (0.75 +/- 0.63; 0.88)
GPT 4 + Plugin + Translation (N=541)	0.288 (0.272; 0.303)	55 (0.10 +/- 0.08; 0.13)	7 (0.13 +/- 0.05; 0.22)	19 (0.35 +/- 0.22; 0.47)	29 (0.53 +/- 0.40; 0.65)	33 (0.6 +/- 0.47; 0.73)

Table 3 Correct Answers of GPT Models Compared with Required and Random Scores

Correct Answers of GPT Models Compared with Required and Random Scores			
	Correct Answer	Compared with Required Score (0.6)	Compared with Random Score (0.2)
	Correct Answer (Proportion +/- 95% CI)	Difference (+/- 95% CI)	Difference (+/- 95% CI)
GPT 3.5 (N=541)	373 (0.69 +/- 0.65; 0.73)	0.09 (0.05; 0.13)	0.49 (0.45; 0.53)
GPT 4 (N=541)	493 (0.91 +/- 0.89; 0.93)	0.31 (0.29; 0.34)	0.71 (0.69; 0.73)
GPT 4 + Plugin (N=541)	493 (0.91 +/- 0.89; 0.94)	0.31 (0.29; 0.34)	0.71 (0.69; 0.74)
GPT 4 + Plugin + Translation (N=541)	486 (0.90 +/- 0.87; 0.92)	0.30 (0.27; 0.32)	0.70 (0.67; 0.72)

Table 4 Comparison of GPT Models Justifications between Correct and Incorrect Answers

Comparison of GPT Models Justifications between Correct and Incorrect Answers									
GPT 3.5 (N=54	All Correct	All Incorrect		All Correct	All Incorrect		All Correct	All Incorrect	

1)	Answers (n= 373)	Answers (n= 168)		Answers (n= 373)	Answers (n= 168)		Answers (n= 373)	Answers (n= 168)	
	Logical Justification (Proportion +/- 95% CI)	Logical Justification (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	External Information (Proportion +/- 95% CI)	External Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)
	373 (1 +/- 1; 1)	106 (0.63 +/- 0.55; 0.70)	0.37 (0.30; 0.45)	358 (0.96 +/- 0.94; 0.98)	163 (0.97 +/- 0.94; 0.99)	-0.01 (- 0.04; 0.02)	373 (1 +/- 1; 1)	165 (0.98 +/- 0.958; 1.00)	0.01 (0; 0.04)
GPT 4 (N=541)	All Correct Answers (n= 493)	All Incorrect Answers (n= 48)		All Correct Answers (n= 493)	All Incorrect Answers (n= 48)		All Correct Answers (n= 493)	All Incorrect Answers (n= 48)	
	Logical Justification (Proportion +/- 95% CI)	Logical Justification (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	External Information (Proportion +/- 95% CI)	External Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)
	493 (1 +/- 1; 1)	33 (0.69 +/- 0.56; 0.81)	0.31 (0.19; 0.44)	489 (0.99 +/- 0.983; 0.998)	48 (1 +/- 1; 1)	0 (- 0.01; 0)	492 (0.99 +/- 0.99; 1.0)	48 (1 +/- 1; 1)	0.00 (- 0.01; 0.00)
GPT 4 + Plugin (N=541)	All Correct Answers (n= 493)	All Incorrect Answers (n= 48)		All Correct Answers (n= 493)	All Incorrect Answers (n= 48)		All Correct Answers (n= 493)	All Incorrect Answers (n= 48)	

PREPRINT JMIR Publications

	Logical Justification (Proportion +/- 95% CI)	Logical Justification (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	External Information (Proportion +/- 95% CI)	External Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)
	493 (1 +/- 1; 1)	36 (0.75 +/- 0.63; 0.88)	0.25 (0.13; 0.38)	489 (0.99 +/- 0.983; 998)	48 (1 +/- 1; 1)	0 (-0.01; 0)	493 (1 +/- 1; 1)	48 (1 +/- 1; 1)	0.0 (0.0; 0.0)
GPT 4 + Plugin + Translation (N=541)	All Correct Answers (n= 486)	All Incorrect Answers (n= 55)		All Correct Answers (n= 486)	All Incorrect Answers (n= 55)		All Correct Answers (n= 486)	All Incorrect Answers (n= 55)	
	Logical Justification (Proportion +/- 95% CI)	Logical Justification (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Internal Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)	External Information (Proportion +/- 95% CI)	External Information (Proportion +/- 95% CI)	Difference in Proportions (+/- 95% CI)
	486 (1 +/- 1; 1)	41 (0.75 +/- 0.62; 0.85)	0.26 (0.15; 0.38)	482 (0.99 +/- 0.984; 0.998)	55 (1 +/- 1; 1)	0 (-0.01; 0)	486 (1 +/- 1; 1)	55 (1 +/- 1; 1)	0.0 (0.0; 0.0)

Table 5 Confidence of GPT Models Compared Between Correct and Incorrect Answers

Confidence of GPT Models Compared Betwwen Correct and Incorrect Answers				
GPT 3.5 (N=541)	All Correct Answers (n=373)	All Incorrect Answers (n=168)		
	Confidence	Confidence	Cohen's d (+/-	Confidence

	Mean (+/- 95% CI)	Mean (+/- 95% CI)	95% CI)	Acuracy (+/- 95% CI)
	0.920 (0.914; 0.927)	0.893 (0.873; 0.908)	-0.34 (-0.52; -0.16)	0.028 (0.011; 0.048)
GPT 4 (N=541)	All Correct Answers (n=493)	All Incorrect Answers (n=48)		
	Confidence Mean (+/- 95% CI)	Confidence Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)	Confidence Accuracy (+/- 95% CI)
	0.942 (0.938; 0.946)	0.901 (0.880; 0.918)	-0.93 (-1.24; -0.63)	0.041 (0.023; 0.062)
GPT 4 + Plugin (N=541)	All Correct Answers (n=493)	All Incorrect Answers (n=48)		
	Confidence Mean (+/- 95% CI)	Confidence Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)	Confidence Accuracy (+/- 95% CI)
	0.923 (0.918; 0.928)	0.886 (0.870; 0.901)	-0.69 (-0.99; -0.39)	0.037 (0.021; 0.053)
GPT 4 + Plugin + Translation (N=541)	All Correct Answers (n=486)	All Incorrect Answers (n=55)		
	Confidence Mean (+/- 95% CI)	Confidence Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)	Confidence Accuracy (+/- 95% CI)
	0.923 (0.920; 0.927)	0.880 (0.866; 0.895)	-1.00 (-1.29; -0.72)	0.043 (0.028; 0.059)

Table 6 Relationship between Question's difficulty, Performance and Confidence of GPT Model Answers

Comparison of Question's Difficulty of GPT Models Between Correct and Incorrect Answers			
GPT 3.5 (N=541)	All Correct Answers (n=373)	All Incorrect Answers (n=168)	
	Question's Difficulty Mean (+/- 95% CI)	Question's Difficulty Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)
	0.257 (0.239; 0.273)	0.358 (0.329; 0.386)	0.58 (0.40; 0.77)
GPT 4 (N=541)	All Correct Answers (n=493)	All Incorrect Answers (n=48)	
	Question's Difficulty Mean (+/- 95% CI)	Question's Difficulty Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)

	0.277 (0.262; 0.293)	0.399 (0.345; 0.454)	0.68 (0.39; 0.98)
GPT 4 + Plugin (N=541)	All Correct Answers (n=493)	All Incorrect Answers (n=48)	
	Question's Difficulty Mean (+/- 95% CI)	Question's Difficulty Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)
	0.279 (0.263; 0.295)	0.379 (0.327; 0.438)	0.57 (0.27; 0.86)
GPT 4 + Plugin + Translation (N=541)	All Correct Answers (n=486)	All Incorrect Answers (n=55)	
	Question's Difficulty Mean (+/- 95% CI)	Question's Difficulty Mean (+/- 95% CI)	Cohen's d (+/- 95% CI)
	0.272 (0.257; 0.288)	0.431 (0.377; 0.488)	0.92 (0.63; 1.20)
Correlation of Confidence and Question's Difficulty for all Answers			
	Confidence Mean (+/- 95% CI)	Question's Difficulty Mean (+/- 95% CI)	Pearson's r (+/- 95%CI)
GPT 3.5 (N=541)	0.912 (0.906; 0.919)	0.288 (0.274; 0.304)	-0.008 (-0.078; 0.061)
GPT 4 (N=541)	0.939 (0.935; 0.942)	0.288 (0.273; 0.302)	-0.0816 (-0.168; 0.0096)
GPT 4 + Plugin (N=541)	0.920 (0.916; 0.924)	0.288 (0.273; 0.304)	-0.0874 (-0.176; 0.004)
GPT 4 + Plugin + Translation (N=541)	0.919 (0.916; 0.923)	0.288 (0.273; 0.302	-0.055 (-0.138; 0.032)

Table 7 Comparison of Correct Answers Between GPT Models

Comparison of Correct Answers Between GPT Models			
	Correct Answer Rate (Proportion +/- 95%CI)	Correct Answer Rate (Proportion +/- 95%CI)	Difference in Proportions (+/- 95% CI)
	(N=541)	(N=541)	
GPT 3.5 vs GPT 4	373 (0.69 +/- 0.65, 0.73)	493 (0.91 +/- 0.89, 0.94)	0.22 (0.18; 0.27)
GPT 3.5 vs GPT 4	373 (0.69 +/-	493 (0.91 +/- 0.89,	0.22 (0.18; 0.27)

+ Plugin	0.65, 0.73)	0.94)	
GPT 3.5 vs GPT 4 + Plugin + Translation	373 (0.69 +/- 0.65, 0.73)	486 (0.90 +/- 0.87, 0.92)	0.21 (0.16; 0.26)
GPT 4 vs GPT 4 + Plugin	493 (0.91 +/- 0.89, 0.94)	493 (0.91 +/- 0.89, 0.94)	0.00 (-0.03; 0.03)
GPT 4 vs GPT 4 + Plugin + Translation	493 (0.91 +/- 0.89, 0.94)	486 (0.90 +/- 0.87, 0.92)	-0.013 (-0.05; 0.02)
GPT 4 + Plugin vs GPT 4 + Plugin + Translation	493 (0.91 +/- 0.89, 0.94)	486 (0.90 +/- 0.87, 0.92)	-0.01 (-0.05; 0.02)

Supplementary 1 Prompting Strategies for different GPT Models
(See separate file)

Supplementary 2 Supplementary analysis of GPT models Answers (statistically significant results are highlighted in blue, statistically non-significant results are highlighted in brown)
(See separate file)

Preprint JMIR Publications

Supplementary Files

Preprint JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

Prompting Strategies for different GPT Models.
URL: http://asset.jmir.pub/assets/56abe2a9e577340634ae83aace3d6b3a.docx

Supplementary analysis of GPT models Answers (statistically significant results are highlighted in blue, statistically non-significant results are highlighted in brown).
URL: http://asset.jmir.pub/assets/aef6457aef1807f3928a549007866130.xlsx

Preprint JMIR Publications