TITLE

ChatGPT's performance in dentistry and allergy-immunology assessments: a comparative study

Accepted for publication: October 4, 2023

AUTHOR NAMES AND INSTITUTIONAL AFFILIATIONS

Alexander Fuchs[1], Tina Trachsel[2], Roland Weiger[1], Florin Eggmann[1]

[1]Department of Periodontology, Endodontology, and Cariology, University Center for Dental Medicine Basel UZB, University of Basel, Basel, Switzerland

[2]Division of Allergy, University Children's Hospital Basel, Basel, Switzerland

CORRESPONDENCE

Dr. med. dent. Florin Eggmann, Klinik für Parodontologie, Endodontologie und Kariologie, Universitäres Zentrum für Zahnmedizin Basel UZB, Universität Basel, Mattenstrasse 40, CH-4058 Basel, Schweiz. Tel. +41 61 267 26 80; E-Mail: florin.eggmann@unibas.ch

Acknowledgments

Alexander Fuchs contributed to this work as part of the requirements for his Master of Dental Medicine degree at the University of Basel.

Conflicts of interest

The authors declare no financial or non-financial conflicts of interest related to this work.

TITLE

ChatGPT’s performance in dentistry and allergy-immunology assessments: a comparative study

ABSTRACT

Large language models (LLMs) such as ChatGPT have potential applications in healthcare, including dentistry. Priming, the practice of providing LLMs with initial, relevant information, is an approach to improve their output quality. This study aimed to evaluate the performance of ChatGPT 3 and ChatGPT 4 on self-assessment questions for dentistry, through the Swiss Federal Licensing Examination in Dental Medicine (SFLEDM), and allergy and clinical immunology, through the European Examination in Allergy and Clinical Immunology (EEAACI). The second objective was to assess the impact of priming on ChatGPT’s performance. The SFLEDM and EEAACI multiple-choice questions from the University of Bern’s Institute for Medical Education platform were administered to both ChatGPT versions, with and without priming. Performance was analyzed based on correct responses. The statistical analysis included Wilcoxon rank sum tests (α=0.05). The average accuracy rates in the SFLEDM and EEAACI assessments were 63.3% and 79.3%, respectively. Both ChatGPT versions performed better on EEAACI than SFLEDM, with ChatGPT 4 outperforming ChatGPT 3 across all tests. ChatGPT 3's performance exhibited a significant improvement with priming for both EEAACI (p=0.017) and SFLEDM (p=0.024) assessments. For ChatGPT 4, the priming effect was significant only in the SFLEDM assessment (p=0.038). The performance disparity between SFLEDM and EEAACI assessments underscores ChatGPT’s varying proficiency across different medical domains, likely tied to the nature and amount of training data available in each field. Priming can be a tool for enhancing output, especially in earlier LLMs. Advancements from ChatGPT 3 to 4 highlight the rapid developments in LLM

technology. Yet, their use in critical fields such as healthcare must remain cautious owing to LLMs' inherent limitations and risks.

KEY WORDS

Allergology, artificial intelligence, dental education, clinical immunology, machine learning, medical informatics applications

Introduction

Machine learning applications have brought about significant advancements in medicine, including dentistry (DUCRET ET AL. 2022, SCHWENDICKE ET AL. 2022, HAUG & DRAZEN 2023). Among these advancements, a notable development has been the emergence of large language models (LLMs) with a conversational interface, such as ChatGPT, Bard, Baidu's Ernie Bot, Claude 2, Llama 2, and the chatbot function of the revamped Bing search engine.

These LLMs, underpinned by deep learning transformer architectures, are trained on vast amounts of tokenized text data (VASWANI ET AL. 2017). This training allows them to generate fluent, contextually pertinent responses based on the input they receive. Their capabilities span a wide range of tasks, from answering questions, summarizing texts, translating languages, to writing computer code.

Priming, the practice of providing LLMs with initial, contextually relevant information, is a useful approach to enhance their output quality (RAFFEL ET AL. 2020). By initiating a conversation with strategically chosen words, phrases, or longer text excerpts, users can guide LLMs to produce more accurate and contextually congruous responses.

LLMs have many potential use cases in healthcare, including dentistry (EGGMANN ET AL. 2023). For instance, healthcare professionals could soon leverage LLMs to streamline routine administrative tasks and improve patient education. However, LLMs come with a set of significant risks and some inherent limitations (MELLO & GUHA 2023). Many LLMs operate with knowledge cutoffs, which means they lack up-to-date information (DASHTI ET AL. 2023). Determining the reliability of their response sources can be difficult, if not impossible (WALKER ET AL. 2023). Moreover, LLMs sometimes produce answers that seem plausible but are incorrect, underscoring the importance of human oversight (DASHTI ET AL. 2023). Given these limitations, there are serious

concerns regarding the utility and safety of LLMs, especially in high-stakes fields of application such as healthcare (BEAM ET AL. 2023).

In light of the potential implications of LLMs for healthcare, rigorous evaluation of their outputs is paramount. By assessing LLMs' performance against external benchmarks—including reasoning, coding, and knowledge tests—one can discern their strengths and weaknesses (KUNG ET AL. 2023). Such evaluations can then inform strategies to enhance LLMs' performance and guard against incautious use.

A prime resource for such evaluations is the University of Bern's Institute for Medical Education (IML). The IML hosts a digital platform offering a vast array of self-assessment questions tailored for dental and medical students and healthcare professionals (https://self-assessment.measured.iml.unibe.ch/). Among its offerings are multiple-choice questions designed for dental students preparing for the Swiss Federal Licensing Examination in Dental Medicine (SFLEDM) and allergists and immunologists preparing for the European Examination in Allergy and Clinical Immunology (EEAACI). These curated question banks present an ideal tool for assessing and comparing the performance of ChatGPT across distinct medical fields.

Considering the importance of examining the output accuracy of LLMs, this study pursues two objectives. First, it aims to compare the performance of ChatGPT 3 and ChatGPT 4 in responding to the SFLEDM and EEAACI self-assessment questions. Second, it seeks to evaluate the impact of priming on ChatGPT's performance in these assessments.

Materials and Methods

Input sources

The SFLEDM and EEAACI self-assessment questions were obtained from the IML platform on February 13, 2023. While SFLEDM questions were translated from German to English, EEAACI questions were already available in English. Any questions with images or illustrations were excluded. The questions were of two multiple-choice formats:

- A-type questions: These comprised a stem (either a question or a case scenario) followed by potential answers. The task was to identify the single most appropriate answer. Within the SFLEDM and EEAACI self-assessments, these questions had four and five options, respectively.
- Kprim-type questions: These also started with a stem, succeeded by four related statements or answers. The task was to determine the correctness of each of these statements or answers.

The study used 32 SFLEDM questions, comprising 22 A-type and 10 Kprim-type questions. In total, 28 EEAACI questions were used, comprising 19 A-type and 9 Kprim-type questions. The terms of service of the IML platform restrict the dissemination of these self-assessment questions, even though they are publicly accessible at https://self-assessment.measured.iml.unibe.ch/ (last accessed on October 3, 2023). They are therefore not featured in this report.

Priming

The primers, utilized to provide context for the questions, encompassed details about the respective test, main subject information with relevant keywords, as well as information about the question format and response guidelines. They offered a thorough overview of the examination, including insights into the organizing body,

exam purpose, and covered topics, while also instructing the use of scientific reasoning and adherence to general guidelines of the respective field for answering questions. Designed to be analogous in length, structure, style, and content, the primers for the SFLEDM and EEAACI self-assessments underwent several optimization rounds using ChatGPT 3, adhering to principles of effective prompt design. Each trial for the primed groups consistently utilized the same primer.

Conversely, the non-primed groups received a prompt that only supplied basic information about the question format and response guidelines, deliberately omitting context about the examination or topics to maintain succinctness and avoid priming. Supplementary Table S-I details the input texts used for both primed and non-primed groups prior to administering the multiple-choice questions.

Administering questions to ChatGPT

The tests involving ChatGPT 3 and ChatGPT 4 took place on February 19, 2023, and March 25, 2023, respectively. For each group, 20 trials were conducted. Before initiating each trial, the entire chat history was cleared. A new chat window was then opened to eliminate any potential context carryover. For the non-primed groups, the input prompt contained brief instructions on answering the questions, followed by either the A-type or Kprim-type questions. In contrast, for the primed groups, the primer was introduced before presenting the questions.

Performance assessment

An unblinded investigator recorded ChatGPT's responses in a pilot-tested spreadsheet. For A-type questions, a score of 1 point was given for correct answers and 0 points for incorrect ones. For Kprim-type questions, correctly answering all four related statements or answers earned 1 point. If three out of the four statements or answers were evaluated correctly, 0.5 points were given. A score of 0 points was assigned if fewer than three statements or answers were correctly evaluated.

Statistical analysis

For each trial, the attained points were presented as a percentage of the maximum possible points. This percentage was chosen as a performance metric since the maximum points varied between the SFLEDM and EEAACI self-assessments. Descriptive statistics, including mean, standard deviation, median, and interquartile range, were computed for each group. Analysis of the distribution within each group revealed a non-normal distribution, verified using a graphical method (normal probability plot). Performance was analyzed between primed and non-primed groups for both the SFLEDM and EEAACI self-assessments. This comparison was made within each ChatGPT version, as well as across the ChatGPT 3 and ChatGPT 4 subsets. The Wilcoxon rank sum test was used for this analysis.

To assess the impact of priming, the improvement due to priming was calculated for both the SFLEDM and EEAACI self-assessments within the ChatGPT 3 and ChatGPT 4 subsets. To quantify the improvement, trials—both without and with priming—were ranked within their respective groups based on the percentage of the maximum points attained. These ranks were paired before subtracting the values of the non-primed groups from the values of the primed group, producing 20 improvement values within each group. Analysis utilizing a normal probability plot confirmed a non-normal distribution of data. Consequently, the Wilcoxon rank sum test was used for group comparisons. The level of significance was set at $\alpha=0.05$. The statistical analyses were performed by an unblinded investigator using R software (version 4.2.2, R Core Team, R Foundation for Statistical Computing, Vienna, Austria). The dataset generated and analyzed in this study is available in an open repository (FUCHS ET AL. 2023).

Results

Table I and Figure 1 present the detailed results. Both ChatGPT 3 and ChatGPT 4 exhibited superior performance in the EEAACI compared with the SFLEDM assessment ($p<0.001$). Overall, ChatGPT 4 scored higher than ChatGPT 3 across all groups ($p<0.001$). The performance gap between ChatGPT 4 and ChatGPT 3 was wider in the EEAACI assessment than in the SFLEDM assessment. In the SFLEDM assessment, without and with priming, the average percentage point increases for ChatGPT 4 over ChatGPT 3 were 5.1 and 4.1, respectively. In contrast, for the EEAACI assessment, these increases were 18.2 (without priming) and 15.0 (with priming).

Priming significantly enhanced the performance of ChatGPT 3 in both the SFLEDM ($p=0.012$) and EEAACI ($p=0.001$) assessments. For ChatGPT 4, while there was a significant performance increase in the SFLEDM assessment due to priming ($p=0.03$), priming had no significant effect on the performance in the EEAACI assessment ($p=0.221$).

As shown in Table II, with ChatGPT 3, priming enhanced the performance for EAACI more than for SFLEDM ($p=0.037$). Conversely, when using ChatGPT 4, priming improved the performance for SFLEDM performance more than for EEAACI ($p=0.002$).

Discussion

This study compared ChatGPT 3's and ChatGPT 4's performance on SFLEDM and EEAACI self-assessment questions. These multiple-choice questions served to benchmark and contrast the LLMs' proficiency in the field of dentistry and allergy and immunology. The results showed that both versions performed better on the EEAACI, with ChatGPT 4 surpassing ChatGPT 3 in all tests. Priming notably improved ChatGPT 3's performance in both tests, but only impacted ChatGPT 4 in the SFLEDM assessment.

The observed performance disparity between the EEAACI and SFLEDM assessments suggests that ChatGPT's proficiency may vary across all medical specialties. One plausible explanation for this disparity may lie in the nature of the data the LLM has been trained on (PATCAS ET AL. 2022, BORNSTEIN 2023, WALKER ET AL. 2023). Most of the medical literature, discussions, and queries available in open sources focus on broader medical fields, with allergy and immunology being more extensively represented than smaller branches of medicine such as dentistry. Furthermore, in dentistry, diagnoses and treatments frequently rely heavily on physical examinations and imaging, aspects that textual models such as ChatGPT are not adept at grasping. By contrast, allergy and immunology, being more systemic and often reliant on patient history and laboratory results, are better suited for textual analysis and understanding by LLMs.

This study, while specifically pertaining to the SFLEDM and EEAACI assessments, may offer broader implications for the application of LLMs in other medical domains. The observed performance disparities and the impact of priming across different assessments suggest that the effectiveness of LLMs can be significantly influenced by the subject matter. Extending this to other domains, it becomes pivotal to consider the availability and specificity of training data, as well as the inherent characteristics of the

medical field in question. For instance, medical specialties that heavily rely on textual information and have abundant data available might observe better LLM performance, akin to the results seen in the EEAACI assessment. Conversely, fields that depend more on visual or practical elements may present additional challenges for LLMs, as seen in the SFLEDM assessment. Further research is warranted to explore these dynamics, identifying patterns and strategies to optimize LLMs' performance across diverse medical specialties.

ChatGPT's performance has been studied across various medical knowledge examinations, with the accuracy rates demonstrating considerable variation among different tests and medical disciplines. A recent systematic review and meta-analysis revealed that the performance range for ChatGPT 3.5 in these evaluations spanned from 40% to 100%, with an average accuracy rate of 61.1% (LEVIN ET AL. 2023). The mean performance of 63.3% in the SFLEDM assessment aligns with this average across medical domains. In contrast, ChatGPT's performance in the EEAACI assessment yielded a higher average score of 79.3%, placing it at the top range compared with results from other studies (LEVIN ET AL. 2023). It is noteworthy that ChatGPT 4, when primed, exceeded the commonly used passing threshold of 60% in the SFLEDM assessment. This level of performance was observed in the EEAACI assessment across the two examined ChatGPT iterations, regardless of priming.

In dental and medical education, LLM chatbots hold potential for supplementing learning materials and providing interactive learning opportunities (ALI ET AL. 2023). However, it is important to emphasize that while ChatGPT 3 and ChatGPT 4 showed promise in answering self-assessment questions from the SFLEDM and EEAACI, they should not be relied on for exam preparation. Despite their capabilities, these LLMs frequently provide inaccurate or misleading information (MELLO & GUHA 2023). Relying on ChatGPT for exam preparation, especially in critical fields like healthcare, could

lead to misconceptions and an incomplete understanding of the subject matter (LEVIN ET AL. 2023, SAAD ET AL. 2023). Therefore, it is crucial to always approach their outputs with caution and cross-reference with trusted educational resources. Today, LLMs should serve merely as supplements to more traditional methods of information seeking (MELLO & GUHA 2023).

The noticeable performance improvement from ChatGPT 3 to ChatGPT 4, available exclusively to subscription fee payers, underscores the rapid advancements in LLM development within a short period. Comparable enhancements in response quality from ChatGPT 4 have been noted for queries related to dermatology and myopia (LEWANDOWSKI ET AL. 2023, LIM ET AL. 2023). Moreover, while ChatGPT 3 operated solely with text, ChatGPT 4 is multimodal, allowing it to accept and produce text and image inputs and outputs. This shift to multimodality represents a substantial enhancement in ChatGPT's functionality. The increasing adaptability of LLMs suggests that they might soon serve as additional tools in specific use cases in healthcare (VAIRA ET AL. 2023).

However, on the road to artificial general intelligence, LLMs underpinned by the next-token-prediction paradigm are likely an off-ramp (MARCUS 2022). Their capabilities, based on brute statistics, are impressive, but their genuine understanding remains shallow (THIRUNAVUKARASU 2023). Medical professionals, including allergists, immunologists, and dentists, are therefore not predicted to face major changes due to the widespread adoption of LLM applications (THIRUNAVUKARASU 2023).

Priming and adept prompt design serve as strategic tools to guide LLMs towards generating more contextually congruous responses (RAFFEL ET AL. 2020). The results of this study are in line with this assertion, particularly with ChatGPT 3, where priming significantly enhanced its performance in both the SFLEDM and EEAACI assessments. However, whereas priming exhibited a significant impact on ChatGPT

4's performance in the SFLEDM assessment, its influence was negligible in the EEAACI assessment. This difference underscores the evolving nature of LLMs and suggests that as these models become more advanced, the relative impact of priming may vary depending on the complexity and specificity of the task at hand.

This study has several limitations that warrant careful consideration. First, the questions from the IML platform, specifically the SFLEDM and EEAACI self-assessments, represented only a narrow spectrum of knowledge within dentistry, allergy, and immunology. This limits the generalizability of the findings.

Second, tasks like answering board examination questions or retrieving information from medical records have only a tangential connection to real-world care decisions (MELLO & GUHA 2023). This means that assessments using such tasks as benchmarks offer limited insight into a LLM's usefulness for clinical decision support (MELLO & GUHA 2023).

Third, the translation of SFLEDM questions from German to English introduced potential biases, as nuances in language might affect the LLM's comprehension and response accuracy.

Fourth, the exclusion of questions with images or illustrations omits a significant aspect of medical assessments, which often rely on visual diagnostics and the interpretation of data charts and graphs.

Fifth, an unblinded evaluator recorded and graded ChatGPT’s responses to the multiple-choice questions. Since the answer key for these questions was objective and definitive, allowing no room for interpretation or discretion, calibration procedures, evaluator blinding, and employment of multiple evaluators were foregone. Nonetheless, to guard against potential biases inherent in unblinded assessments—even when utilizing unequivocal answer keys—future investigations should consider implementing evaluator calibration and blinding.

Sixth, by focusing solely on two versions of ChatGPT, the study did not capture the full range of LLM capabilities across various models or iterations. These limitations emphasize the critical need for additional research to thoroughly evaluate the performance and potential impact of LLMs in medical disciplines.

Conclusion

Within the constraints of this study, the following conclusions were drawn:

- ChatGPT 3 and ChatGPT 4 both demonstrated stronger performance on the EEAACI compared with the SFLEDM assessment. This performance disparity highlights ChatGPT's varying proficiency across different medical domains, likely influenced by the type and volume of training data available in each field.
- Priming improved ChatGPT 3's performance across both assessments. For ChatGPT 4, while priming influenced results in the SFLEDM assessment, its effect was negligible for the EEAACI. This underscores the nuanced influence of priming as LLMs become more advanced.
- The progress from ChatGPT 3 to ChatGPT 4 reveals rapid advancements in LLM development, including the shift to multimodality. Yet, their enhanced capabilities notwithstanding, LLMs have major inherent limitations and risks, emphasizing the need for cautious use in high-stakes fields such as healthcare.

Table I Results of the performance assessments

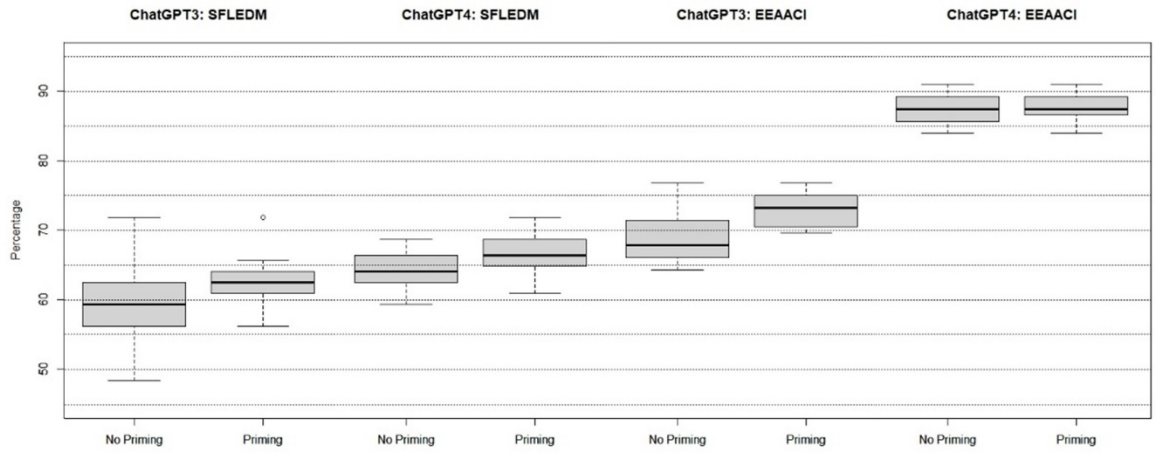
Assessment	LLM	Priming	N	Mean	SD	Median	IQR
SFLEDM	ChatGTP 3	None	20	59.3%	5.2%	59.4%	5.5%
		Yes	20	62.6%	3.3%	62.5%	3.1%
	ChatGTP 4	None	20	64.4%	2.8%	64.1%	3.5%
		Yes	20	66.7%	3.2%	66.4%	3.5%
EEAACI	ChatGTP 3	None	20	69.0%	3.7%	67.9%	5.4%
		Yes	20	72.9%	2.6%	73.2%	4.0%
	ChatGTP 4	None	20	87.2%	1.9%	87.5%	3.6%
		Yes	20	87.9%	1.9%	87.5%	2.2%

EEAACI, European Examination in Allergy and Clinical Immunology; *IQR,* interquartile range; *LLM,* large language model; *SD,* standard deviation; *SFLEDM,* Swiss Federal Licensing Examination in Dental Medicine

Table II Performance improvement through priming

Assessment	LLM	N	Mean	SD	Median	IQR
SFLEDM	ChatGTP 3	20	3.3%	2.2%	3.1%	3.1%
	ChatGTP 4	20	2.3%	0.9%	1.6%	1.6%
EEAACI	ChatGTP 3	20	3.9%	1.5%	3.6%	1.2%
	ChatGTP 4	20	0.7%	0.9%	0.0%	1.8%

EEAACI, European Examination in Allergy and Clinical Immunology; *IQR,* interquartile range; *LLM,* large language model; *SD,* standard deviation; *SFLEDM,* Swiss Federal Licensing Examination in Dental Medicine



334

335 **Figure 1** Box plots depicting the distribution of assessment scores, represented by
336 correct response rates, across the eight distinct groups.

337

338 **Supplementary Table S-I** Input texts used for the primed and non-primed groups

339 before administering the multiple-choice questions to ChatGPT

Assessment	Multiple-choice format	Priming	Input text
EEAACI	A-type questions	None	Please answer the following single-choice questions. For each question, please clearly indicate which answer (A, B, C, D, or E) is correct. If uncertain, please clearly indicate the most probable answer. List the correct answers in a table. Number the questions as given.
		Yes	The European Academy of Allergy & Clinical Immunology (EAACI) has been conducting the European Examination in Allergology and Clinical Immunology annually since 2008. The exam is designed to test candidates' knowledge on a wide range of topics related to allergology, including allergens, dermatology, respiratory and pediatric allergy, anaphylaxis, venom hypersensitivity, drug and food hypersensitivity, as well as relevant issues such as pregnancy and allergology, occupational allergies, eosinophilic disorders, mastocytosis, and CI-INH deficiency. The exam also covers basic immunology and clinical immunology, including autoimmunity and immune deficiency. To aid allergists and immunologists in preparing for the exam, training questions are available for practice. Please answer the following single-choice questions. For each question, please clearly indicate which answer (A, B, C, D, or E) is correct. Use scientific reasoning and general guidelines for allergology and immunology to answer the questions correctly. If uncertain, please clearly indicate the most probable answer. List the correct answers in a table. Number the questions as given.
	Kprim-type questions	None	Please answer the following multiple-choice questions. For each question, please clearly indicate for each answer (A, B, C, and D) whether the answer is correct or incorrect. List the answers in a table. Number the questions as given.
		Yes	The European Academy of Allergy & Clinical Immunology (EAACI) has been conducting the European Examination in Allergology and Clinical Immunology annually since 2008. The exam is designed to test candidates' knowledge on a wide range of topics related to allergology, including allergens, dermatology, respiratory and pediatric allergy, anaphylaxis, venom hypersensitivity, drug and food hypersensitivity, as well as relevant issues such as pregnancy and allergology, occupational allergies, eosinophilic disorders, mastocytosis, and CI-INH deficiency. The exam also covers basic immunology and clinical immunology, including autoimmunity and immune deficiency. To aid allergists and immunologists in preparing for the exam, training questions are available for practice. Please answer the following multiple-choice questions. For each question, please clearly indicate for each answer (A, B, C, and D) whether the answer is correct or incorrect. Therefore, each question can have 0, 1, 2, 3, or 4 correct answers. Use scientific reasoning and general guidelines for allergology and immunology to answer the questions correctly. List the answers in a table. Number the questions as given.
SFLEDM	A-type questions	None	Please answer the following single-choice questions. For each question, please clearly indicate which answer (A, B, C, or D) is correct. If uncertain, please clearly indicate the most probable answer. List the correct answers in a table. Number the questions as given.
		Yes	You will be asked questions from the Swiss Federal Licensing Examination in dental medicine. The exam is designed to test candidates' knowledge on a wide range of topics related to dental medicine, including preventive dentistry, stomatology, oral health, cariology, restorative dentistry, endodontics, periodontics, dental implantology, prosthodontics, esthetic dentistry, pediatric dentistry, orthodontics, dental radiology, geriatric dentistry, special needs dentistry, and communication in the dentist-patient relationship. Additionally, the catalogue of learning objectives demands that candidates know the most common medical issues and corresponding treatment approaches in the following medical specialties: infectiology, internal medicine, oral and maxillofacial surgery, dermatology/allergy, psychiatry, geriatrics, and care for patients with special needs. To aid dental students in preparing for the Swiss Federal Licensing Examination in dental medicine, training questions are available for practice. Please answer the following single-choice questions. For each question, please clearly indicate which answer (A, B, C, or D) is correct. Use scientific reasoning and general guidelines for dentistry to answer the questions correctly. If uncertain,

		please clearly indicate the most probable answer. List the correct answers in a table. Number the questions as given.
Kprim-type questions	None	Please answer the following multiple-choice questions. For each question, please clearly indicate for each answer (A, B, C, and D) whether the answer is correct or incorrect. List the answers in a table. Number the questions as given.
	Yes	You will be asked questions from the Swiss Federal Licensing Examination in dental medicine. The exam is designed to test candidates' knowledge on a wide range of topics related to dental medicine, including preventive dentistry, stomatology, oral health, cariology, restorative dentistry, endodontics, periodontics, dental implantology, prosthodontics, esthetic dentistry, pediatric dentistry, orthodontics, dental radiology, geriatric dentistry, special needs dentistry, and communication in the dentist-patient relationship. Additionally, the catalogue of learning objectives demands that candidates know the most common medical issues and corresponding treatment approaches in the following medical specialties: infectiology, internal medicine, oral and maxillofacial surgery, dermatology/allergy, psychiatry, geriatrics, and care for patients with special needs. To aid dental students in preparing for the Swiss Federal Licensing Examination in dental medicine, training questions are available for practice. Please answer the following multiple-choice questions. For each question, please clearly indicate for each answer (A, B, C, and D) whether the answer is correct or incorrect. Therefore, each question can have 0, 1, 2, 3, or 4 correct answers. Use scientific reasoning and general guidelines for dentistry to answer the questions correctly. List the answers in a table. Number the questions as given.

340 *EEAACI,* European Examination in Allergy and Clinical Immunology; *SFLEDM,* Swiss Federal Licensing
341 Examination in Dental Medicine

342

343

344

345

346

347

348

349

350

Zusammenfassung

Einleitung

Anwendungen der künstlichen Intelligenz (KI) können dem Gesundheitspersonal, einschliesslich Zahnärzten, verschiedene Vorteile bieten. Grosse Sprachmodelle (GSM) sind KI-Anwendungen, die mit grossen Mengen von Textdaten trainiert werden und verschiedene sprachbezogene Aufgaben durchführen können. ChatGPT, ein GSM mit einer Konversationsschnittstelle, wurde im November 2022 auf den Markt gebracht und ist online verfügbar. Trotz seiner beeindruckenden Fähigkeiten hat ChatGPT erhebliche Einschränkungen und Unzulänglichkeiten. Beispielsweise gibt ChatGPT teilweise fehlerhafte Antworten oder stellt Fehlinformationen als Fakten dar. Vor der Anwendung GSM in medizinischen Disziplinen ist es von grosser Bedeutung, die Fähigkeiten und Grenzen von GSM zu verstehen. Ein interessanter Ansatz ist das "Priming", bei einem GSM vorab relevante Informationen gegeben werden, um die Qualität seiner Antworten zu verbessern. Diese Studie konzentriert sich auf die Bewertung der Leistung von ChatGPT Versionen 3 und 4 in den medizinischen Bereichen Zahnmedizin sowie Allergologie und klinische Immunologie, unter besonderer Berücksichtigung des Priming-Effekts.

Material und Methoden

Zur umfassenden Evaluation von ChatGPT wurden Multiple-Choice-Fragen zur Selbstbewertung in Zahnmedizin («Swiss Federal Licensing Examination in Dental Medicine» [SFLEDM]) und Allergologie sowie klinischer Immunologie («European Examination in Allergy and Clinical Immunology» [EEAACI]) vom Institut für Medizinische Lehre der Universität Bern zusammengestellt. ChatGPT 3 und 4 wurden unter zwei Bedingungen getestet: mit Priming und ohne Priming. Das Hauptkriterium für die Leistungsbewertung war die Genauigkeitsrate, gemessen an der Anzahl korrekt

beantworteter Fragen. Die statistischen Analysen erfolgten mittels Wilcoxon-Rangsummentests mit einem Signifikanzniveau von $\alpha = 0,05$.

Resultate

Im SFLEDM-Bereich betrug die durchschnittliche Genauigkeitsrate 63,3%. Im Gegensatz dazu zeigte ChatGPT im EEAACI-Bereich mit einer durchschnittlichen Genauigkeit von 79,3% eine überlegene Leistung. Beide ChatGPT-Modelle zeigten im EEAACI-Bereich bessere Leistungen als im SFLEDM-Bereich. Bemerkenswert ist, dass ChatGPT 4 durchgehend bessere Leistungen als ChatGPT 3 in beiden Bereichen zeigte. In Bezug auf das Priming zeigte ChatGPT 3 sowohl bei den Fragen aus dem EEAACI Bereich (p=0,001) als auch im SFLEDM Bereich (p=0,012) eine deutliche Verbesserung bei Verwendung von Priming. Im Gegensatz dazu verbesserte sich die Leistung durch Priming bei ChatGPT 4 nur im SFLEDM-Bereich signifikant (p=0,03).

Diskussion

Die unterschiedliche Leistung von ChatGPT in der Beantwortung von Multiple-Choice-Fragen aus dem SFLEDM und EEAACI Bereich weist auf eine unterschiedliche Kompetenz von GSM in verschiedenen medizinischen Bereichen hin. Diese unterschiedliche Kompetenz könnte durch die Art und das Volumen der verfügbaren Trainingsdaten für jeden Bereich beeinflusst werden. Priming erweist sich als vorteilhafte Methode zur Leistungsverbesserung von GSM, besonders bei älteren Versionen wie ChatGPT 3. Der signifikante Leistungszuwachs von ChatGPT 3 zu 4 unterstreicht die rasanten Entwicklungen in der GSM-Technologie. Dennoch ist beim Einsatz von GSM im Gesundheitssektor, einschliesslich der Zahnmedizin, höchste Sorgfalt und Umsicht angebracht, denn GSM weisen weiterhin zahlreiche Limitationen und Risiken auf.

Résumé

Introduction

Les applications d'intelligence artificielle (IA) peuvent offrir divers avantages aux professionnels de la santé, y compris aux dentistes. Les modèles de langage de grande taille (abrégé LLM de l'anglais *large language model*) sont des applications d'IA entraînées avec de grandes quantités de données textuelles et capables d'effectuer différentes tâches liées à la langue. ChatGPT, un LLM doté d'une interface conversationnelle, a été lancé en novembre 2022 et est disponible en ligne. Malgré ses capacités impressionnantes, ChatGPT présente des limitations et des insuffisances importantes. Par exemple, ChatGPT donne parfois des réponses erronées ou présente des informations erronées comme des faits. En raison de la nature critique des disciplines médicales, il est très important de comprendre les capacités et les limites du LLM. Une approche intéressante est le "priming", qui consiste à donner au LLM des informations pertinentes à l'avance afin d'améliorer la qualité de ses réponses. Cette étude se concentre sur l'évaluation des performances de ChatGPT versions 3 et 4 dans les domaines médicaux de la dentisterie ainsi que de l'allergologie et de l'immunologie clinique, en accordant une attention particulière à l'effet d'amorçage.

Matériels et methodes

Pour une évaluation complète de ChatGPT, des questions à choix multiples d'auto-évaluation en médecine dentaire («Swiss Federal Licensing Examination in Dental Medicine» [SFLEDM]) et en allergologie et immunologie clinique («European Examination in Allergy and Clinical Immunology» [EEAACI]) ont été compilées par l'Institut d'enseignement médical de l'Université de Berne. ChatGPT 3 et 4 ont été testés dans deux conditions : avec et sans amorçage. Le principal critère d'évaluation des performances était le taux de précision, mesuré par le nombre de questions

auxquelles il a été répondu correctement. Les analyses statistiques ont été effectuées à l'aide de tests de répartition des rangs de Wilcoxon avec un niveau de signification de $\alpha = 0,05$.

Résultats

Dans le domaine SFLEDM, le taux de précision moyen était de 63,3%. En revanche, ChatGPT a montré une performance supérieure dans le domaine EEAACI, avec une précision moyenne de 79,3%. Les deux modèles ChatGPT ont montré de meilleures performances dans le domaine EEAACI que dans le domaine SFLEDM. Il est à noter que ChatGPT 4 a montré des performances systématiquement meilleures que ChatGPT 3 dans les deux domaines. En ce qui concerne l'amorçage, ChatGPT 3 a montré une nette amélioration lors de l'utilisation de l'amorçage, tant pour les questions du domaine EEAACI (p=0,001) que pour le domaine SFLEDM (p=0,012). En revanche, la performance de ChatGPT 4 ne s'est améliorée de manière significative par l'amorçage que dans le domaine SFLEDM (p=0,03).

Discussion

La différence de performance de ChatGPT dans les réponses aux questions à choix multiples des domaines SFLEDM et EEAACI pourrait indiquer une différence de compétence des LLMs dans différents domaines médicaux. Cette différence de compétence pourrait être influencée par le type et le volume des données d'entraînement disponibles pour chaque domaine. L'amorçage s'avère être une méthode avantageuse pour améliorer les performances du LLM, en particulier pour les anciennes versions comme ChatGPT 3. L'augmentation significative des performances de ChatGPT 3 à 4 souligne les développements rapides de la technologie LLM. Toutefois, l'utilisation du LLM dans le secteur de la santé, y compris la médecine dentaire, requiert la plus grande prudence et le plus grand soin. En effet, les LLMs présentent encore de nombreuses limites et risques.

References

ALI K, BARHOM N, TAMIMI F, DUGGAL M: ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ (2023). doi: 10.1111/eje.12937.

BEAM A L, DRAZEN J M, KOHANE I S, LEONG T-Y, MANRAI A K, RUBIN E J: Artificial intelligence in medicine. N Engl J Med 388: 1220–1221 (2023)

BORNSTEIN M M: Artificial intelligence and personalised dental medicine - just a hype or true game changers? Br Dent J 234: 755 (2023)

DASHTI M, LONDONO J, GHASEMI S, MOGHADDASI N: How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? J Prosthet Dent (2023). doi: 10.1016/j.prosdent.2023.05.023.

DUCRET M, MÖRCH C-M, KARTEVA T, FISHER J, SCHWENDICKE F: Artificial intelligence for sustainable oral healthcare. J Dent 127: 104344 (2022)

EGGMANN F, WEIGER R, ZITZMANN N U, BLATZ M B: Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent (2023). doi: 10.1111/jerd.13046.

FUCHS A, TRACHSEL T, WEIGER R, EGGMANN F: ChatGPT's performance in dentistry and allergy-immunology assessments: a comparative study (Version 1) [Data set]. Zenodo (2023). https://doi.org/10.5281/zenodo.8331147

HAUG C J, DRAZEN J M: Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 388: 1201–1208 (2023)

KUNG T H, CHEATHAM M, MEDENILLA A, SILLOS C, DE LEON L, ELEPAÑO C, MADRIAGA M, AGGABAO R, DIAZ-CANDIDO G, MANINGO J, TSENG V: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2: e0000198 (2023). doi: 10.1371/journal.pdig.0000198.

LEVIN G, HORESH N, BREZINOV Y, MEYER R: Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG (2023). doi: 10.1111/1471-0528.17641.

LEWANDOWSKI M, ŁUKOWICZ P, ŚWIETLIK D, BARAŃSKA-RYBAK W: An original study of ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the dermatology specialty certificate examinations. Clin Exp Dermatol (2023). doi: 10.1093/ced/llad255.

LIM Z W, PUSHPANATHAN K, YEW S M E, LAI Y, SUN C-H, LAM J S H, CHEN D Z, GOH J H L, TAN M C J, SHENG B, CHENG C-Y, KOH V T C, THAM Y-C: Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 95: 104770 (2023)

MARCUS G: Deep learning is hitting a wall. Accessed on October 3, 2023. https://nautil.us/deep-learning-is-hitting-a-wall-238440/. (2022)

MELLO M M, GUHA N: ChatGPT and physicians' malpractice risk. JAMA Health Forum 4: e231938 (2023)

PATCAS R, BORNSTEIN M M, SCHÄTZLE M A, TIMOFTE R: Artificial intelligence in medico-dental diagnostics of the face: a narrative review of opportunities and challenges. Clin Oral Investig 26: 6871–6879 (2022)

RAFFEL C, SHAZEER N, ROBERTS A, LEE K, NARANG S, MATENA M, ZHOU Y, LI W, LIU P J: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21: 5485–5551 (2020)

SAAD A, IYENGAR K P, KURISUNKAL V, BOTCHU R: Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon (2023). doi: 10.1016/j.surge.2023.07.001.

503 SCHWENDICKE F, CEJUDO GRANO DE ORO, J, GARCIA CANTU A, MEYER-LUECKEL H,
504 CHAURASIA A, KROIS J: Artificial intelligence for caries detection: value of data and
505 information. J Dent Res 101: 1350–1356 (2022)
506 THIRUNAVUKARASU A J: Large language models will not replace healthcare
507 professionals: curbing popular fears and hype. J R Soc Med 116: 181–182 (2023)
508 VAIRA L A, LECHIEN J R, ABBATE V, ALLEVI F, AUDINO G, BELTRAMINI G A, BERGONZANI
509 M, BOLZONI A, COMMITTERI U, CRIMI S, GABRIELE G, LONARDI F, MAGLITTO F,
510 PETROCELLI M, PUCCI R, SAPONARO G, TEL A, VELLONE V, CHIESA-ESTOMBA C M,
511 BOSCOLO-RIZZO P, SALZANO G, DE RIU G: Accuracy of ChatGPT-generated
512 information on head and neck and oromaxillofacial urgery: a multicenter
513 collaborative analysis. Otolaryngol Head Neck Surg (2023). doi: 10.1002/ohn.489.
514 VASWANI A, SHAZEER N, PARMAR N, USZKOREIT J, JONES L, GOMEZ A N, KAISER Ł,
515 POLOSUKHIN I: Attention is all you need. Advances in Neural Information Processing
516 Systems 30: 1–11 (2017)
517 WALKER H L, GHANI S, KUEMMERLI C, NEBIKER C A, MÜLLER B P, RAPTIS D A, STAUBLI S
518 M: Reliability of medical information provided by ChatGPT: assessment against
519 clinical guidelines and patient information quality instrument. J Med Internet Res 25:
520 e47479 (2023)
521