




Article

ChatGPT in Occupational Medicine: A Comparative Study with Human Experts

Martina Padovan ¹, Bianca Cosci ¹, Armando Petillo ¹, Gianluca Nerli ¹, Francesco Porciatti ¹, Sergio Scarinci ¹, Francesco Carlucci ¹, Letizia Dell'Amico ¹, Niccolò Meliani ¹, Gabriele Necciari ¹, Vincenzo Carmelo Lucisano ¹, Riccardo Marino ¹, Rudy Foddis ^{1,*} and Alessandro Palla ²

¹ Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, 56126 Pisa, Italy; padovan.martina@gmail.com (M.P.); coscibianca@gmail.com (B.C.); armando.petillo@virgilio.it (A.P.); g.nerli@studenti.unipi.it (G.N.); f.porciatti@studenti.unipi.it (F.P.); scasers@gmail.com (S.S.); f.carlucci@studenti.unipi.it (F.C.); letizia.dellamico@gmail.com (L.D.); n.meliani1@studenti.unipi.it (N.M.); gabriele.necciari@gmail.com (G.N.); riccardo.marino@med.unipi.it (R.M.)

² Intel Corporation, Santa Clara, CA 95054, USA; alessandro.palla@intel.com

* Correspondence: rudy.foddis@unipi.it

Abstract: The objective of this study is to evaluate ChatGPT's accuracy and reliability in answering complex medical questions related to occupational health and explore the implications and limitations of AI in occupational health medicine. The study also provides recommendations for future research in this area and informs decision-makers about AI's impact on healthcare. A group of physicians was enlisted to create a dataset of questions and answers on Italian occupational medicine legislation. The physicians were divided into two teams, and each team member was assigned a different subject area. ChatGPT was used to generate answers for each question, with/without legislative context. The two teams then evaluated human and AI-generated answers blind, with each group reviewing the other group's work. Occupational physicians outperformed ChatGPT in generating accurate questions on a 5-point Likert score, while the answers provided by ChatGPT with access to legislative texts were comparable to those of professional doctors. Still, we found that users tend to prefer answers generated by humans, indicating that while ChatGPT is useful, users still value the opinions of occupational medicine professionals.

Keywords: artificial intelligence; ChatGPT; occupational health and safety; health promotion; digital health; large language model



Citation: Padovan, M.; Cosci, B.; Petillo, A.; Nerli, G.; Porciatti, F.; Scarinci, S.; Carlucci, F.; Dell'Amico, L.; Meliani, N.; Necciari, G.; et al. ChatGPT in Occupational Medicine: A Comparative Study with Human Experts. *Bioengineering* **2024**, *11*, 57. <https://doi.org/10.3390/bioengineering11010057>

Academic Editor: Ramin Ranjbarzadeh

Received: 7 December 2023

Revised: 1 January 2024

Accepted: 4 January 2024

Published: 6 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Artificial intelligence (AI) has become an increasingly popular topic in healthcare, as proven by the rapidly increasing number of studies [1–5] and reviews [6–9] on this topic. More specifically, the use of AI chatbots like ChatGPT in healthcare is frequently discussed. ChatGPT is a publicly available chatbot created by OpenAI, a non-profit AI research and development company. It was released in November 2022 and can be accessed at chat.openai.com. AI chatbots typically consist of a chat interface and a large language model (LLM) based on the transformer architecture [10].

LLMs are powerful artificial neural networks trained on vast amounts of data to understand and generate natural languages. They can be used for translation, text creation, automated responses, and much more [11]. ChatGPT belongs to the Generative Pre-trained Transformer (GPT) [12] model topology and was fine-tuned from the GPT3.5 model using a massive corpus of chat conversations and trained on instruction-based following [13] and reinforcement learning from human feedback [14]. In March 2023, Open AI released GPT-4, an updated model of the GPT architecture [15]. At the time of our study, ChatGPT, as a standalone tool, did not have access to the internet but is now able to access a limited set of external data sources thanks to OpenAI plugins. To use ChatGPT, a prompt in a

natural language is inputted, and the chatbot replies in the same language. The resulting experience closely resembles a real conversation between two individuals. Prompts are typically queries or instructions and can also incorporate diverse data inputs like research papers, mathematical equations, and spreadsheets [16,17].

Some healthcare-related uses of ChatGPT, including medical education, scientific research, medical writing, and diagnostic decision-making, are being explored [18,19]. More specifically, ChatGPT can generate scientific articles with appropriate vocabulary and a range of tones spanning from colloquial to highly technical [20]. ChatGPT could also be of valuable assistance to physicians in their clinical decision-making [21] by generating accurate lists of differential diagnoses [22] and providing insights for cancer screening decisions [23]. Integrating ChatGPT or using medical document templates can greatly streamline medical documentation, saving time and effort for healthcare professionals [24,25].

ChatGPT can be a valuable means of implementing an intelligent question–answering tool in the healthcare sector [19], as demonstrated by numerous studies evaluating the potential of GPT based chatbots in question–answer scenarios within specialized fields like microbiology [26], ophthalmology [27], radiology [28], and neurosurgery [29]. Most ChatGPT studies in medicine, while very promising, have only been tested through a multiple-choice simulated model. In contrast, the chatbot’s ability to provide healthcare professionals with more complex argumentative assistance remains much less investigated.

At the time of this analysis, there were no studies in the medical literature analyzing ChatGPT’s performance in the context of occupational medicine question–answering, while a study [30] already explored how it can provide a range of customized features to address challenges in the field of occupational health and safety. Our study aims to evaluate ChatGPT’s effectiveness in answering open-ended questions on the fundamental concepts of occupational medicine as established by general safety and occupational health principles and local sector regulations. Also, the study explores the potential implications and limitations of the use of ChatGPT in occupational health medicine and provides recommendations for future research.

2. Materials and Methods

The study involved a pool of twelve physicians, consisting of eight residents and four specialist physicians in occupational medicine, who were divided into two groups for blind review. In each group, there were two specialists and four residents. The work was supervised by a full professor of occupational medicine.

2.1. Questions and Reference Answers Generation

The two groups of physicians were tasked with creating questions and their corresponding answers on the primary topics of occupational medicine. Each physician was assigned a specific topic and was instructed to refer to the most recent version of the legislative decree (D.lgs) 81/08, the Italian framework for occupational safety and health [31], for their questions. The topics covered the main hazards present in workplaces, including safety, biological, physical, ergonomic, chemical, and work organization hazards. The selected topics were workplace safety and use of work equipment, manual handling of loads, use of video terminals, physical hazards, chemical hazards, asbestos, carcinogenic and mutagenic agents, and biological and explosive atmospheres in the workplace. Each team member was tasked with generating approximately 40 questions and their respective answers, which were then uploaded on the dedicated Google Form platform. Overall, 433 questions and their respective reference answers were generated.

2.2. Bot-Generated Answers

We used the OpenAI ChatGPT API (Application Programming Interface) to reply to the physicians’ questions. The model used is gpt-3.5-turbo-0301, and the answers were generated using an automatic script on the 14th and 15th of April 2023 by interfacing directly with the OpenAI API. The method employed two distinct approaches. In the

first approach, the questions generated by the two groups were fed directly into ChatGPT (an approach labeled ChatGPT). To mimic a real-use scenario, questions were submitted in Italian.

ChatGPT's system message serves as the primary instructions for the model, and it can be tailored to include various information about the system's role. This can include a brief overview of the assistant, its personality traits, specific instructions or guidelines that you want the assistant to adhere to, or relevant data or information that the model should have, such as frequently asked questions. It is possible to personalize the system role to suit specific use cases. Although the system role/message is not mandatory, we found out that high-quality system messages improve the overall quality of the generated replies [32]. We perform prompt engineering [33] to optimize the system message on a small subset (10%) of the overall questions and reference answers and measure its effectiveness using automatic evaluation methods such as ROUGE [34] and LLM-based evaluation. We used automatic evaluation methods for prompt engineering to quickly iterate through different system configurations to find the most effective one. In addition, the prompt was fine-tuned to minimize hallucinations and non-relevant replies. More information about prompt engineering and implementation details can be found in Appendix A.

In the second approach, the reference legislative context from D.lgs 81/08 was also passed alongside the questions (approach labeled ChatGPT + context). To retrieve the context, we used semantic search with embeddings. Text embeddings are a technique employed in natural language processing (NLP) and machine learning that converts sentences into numerical vectors [35]. As text embeddings outperform traditional keyword-based searches and try to understand the meaning of the query and the context being searched, we used this technique to perform a semantic search to retrieve the relevant section from the D.lgs 81/80. The extracted legislative context was then passed alongside the original question to ChatGPT to generate the final reply. More information about this method and technical details can be found in Appendix A.

2.3. Answers Evaluation

The evaluation of generated responses involved a qualitative assessment of the accuracy, precision, completeness, usability, and relevance of each question and its three generated response options (ChatGPT, ChatGPT + context, and human answer generated during the question creation phase). Each question with its three response options was blindly evaluated by each physician, following pre-defined metrics. The evaluation criteria used to assess the quality of the generated responses are briefly described as follows:

- *Accuracy*: the answer provides correct and precise information without containing errors or inaccuracies;
- *Precision*: the answer is concise and to the point, without being overly verbose or ambiguous;
- *Completeness*: the answer provides all the necessary information to address the question;
- *Usability*: the answer is easy to understand and use for the target audience, which may consist of workers, managers, or safety professionals;
- *Relevance*: was the answer appropriate and relevant to the question asked.

In the evaluation phase, the user is prompted to choose the answer that performs better in each of the five metrics.

Additionally, a Likert 5-point scale was used to measure the absolute value of the accuracy and completeness of each generated response. This scale allowed for a more fine-grained evaluation of the quality of the responses by providing a numerical value to indicate the extent to which the response was accurate and complete. The Likert scale ranged from 1 (not accurate/complete at all) to 5 (completely accurate/complete), with intermediate values indicating varying degrees of accuracy and completeness.

This approach allowed for a more objective and quantitative evaluation of the quality of the generated responses, in addition to the qualitative evaluation based on accuracy, precision, usability, and relevance. The use of both qualitative and quantitative evaluation methods provided a comprehensive assessment of the quality of the generated responses and allowed for a more robust analysis of the results.

2.4. Evaluation Process and Criteria

To facilitate the evaluation process, a website was created where each physician could log in and access the questions and response options for evaluation. Information about the website used for the evaluation phase is presented in Appendix B.

In the pursuit of maximizing comprehensiveness and impartiality in the evaluations, every physician from both groups scrutinized about 100 questions and their corresponding answers exclusively from the opposing group, with the evaluation sequence randomized across various users. This approach was implemented to minimize potential crosstalk between evaluations and ensure that all questions were covered as thoroughly as possible. For the same reason, the order of the answers (users', ChatGPT, ChatGPT + context) is randomized in all evaluations to avoid any implicit bias.

2.5. Error Analysis in ChatGPT + Context Answers

After the evaluation phase, a questionnaire survey was conducted to investigate the specific errors encountered in the responses generated by ChatGPT + context. The questionnaire was administered alongside the poorly rated questions, which scored 2 or less on the Likert accuracy scale, to gain insights into the nature of these errors.

The questionnaire also asked whether specific errors, such as typographical mistakes, grammatical errors, offensive or inappropriate language, internal contradictions, incorrect acronyms, or the misuse of Roman numerals, were present in the chatbot's responses.

Additionally, participants were asked to assess the comprehensibility of the response and provide feedback if it was found to be unclear. Furthermore, respondents were encouraged to elaborate on their reasons for rating the question poorly. The questionnaire is available in Appendix C.

2.6. Error Analysis in ChatGPT + Context Answers

The statistical analysis was performed using Python 3.10 software and Pandas (version 2.0.0), NumPy (version 1.24.2), SciPy (version 1.10.1), Matplotlib (version 3.7.0), and Seaborn (version 0.12.2) libraries. Numerical values are expressed as averages and standard errors. The statistical differences between several groups were assessed using the two-sided Mann–Whitney U test. The statistical significance was accepted at $p < 0.01$. The source data is available upon request.

3. Results

Figure 1 shows the relative Likert score of ChatGPT, ChatGPT + context, and user-generated answers. Each set of questions, along with three alternative answers (users', ChatGPT, ChatGPT + context), was evaluated by multiple physicians. We gathered a total of 1259 assessments, with each combination (question + answer) receiving an average of approximately 2.9 evaluations. In terms of accuracy, the users' answers generally performed best, followed by the ChatGPT + context and the ChatGPT ones.

The accuracy scores for the responses generated by the physicians were significantly higher than those generated by the ChatGPT model, with an average score of 4.042 ± 0.032 for the physicians, 3.361 ± 0.035 for ChatGPT + context, and 3.091 ± 0.035 for the ChatGPT model ($p < 0.001$).

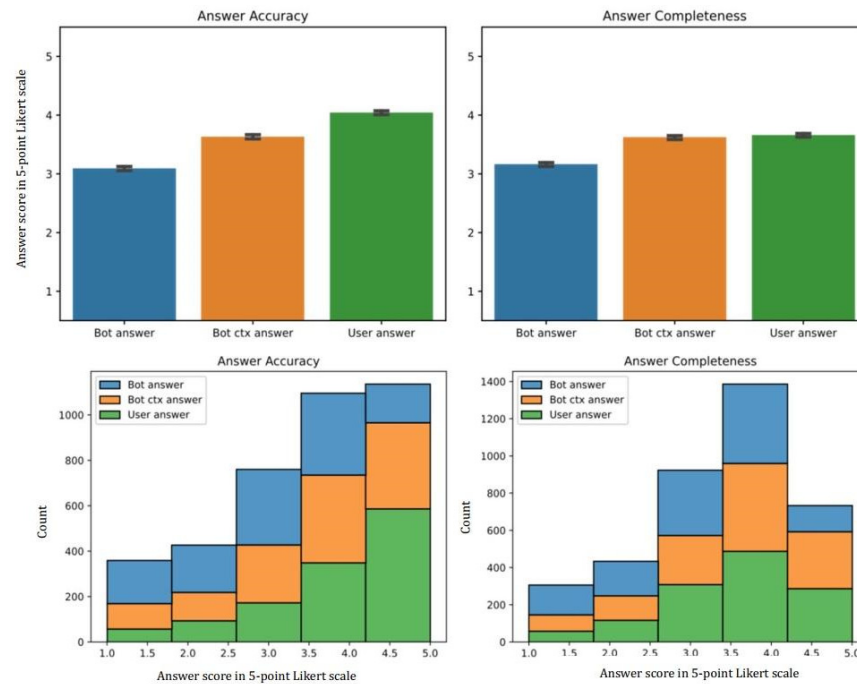


Figure 1. Accuracy and completeness Likert score and distribution for ChatGPT, ChatGPT + context, and users’ generated questions.

The completeness scores for the responses generated by the physicians were significantly higher than those generated by the ChatGPT model, with an average score of 3.658 ± 0.030 for the physicians and 3.159 ± 0.033 for the ChatGPT model ($p < 0.001$). Although the completeness score was higher in physicians than in ChatGPT + context responses, the difference was not statistically significant ($p = 0.862$). All results are summarized in Table 1. Overall, the results suggest that the ChatGPT + context model can generate responses that are reasonably complete and accurate.

Table 1. Comparative scores and p -values for Physicians’, ChatGPT, and ChatGPT + context.

Metric	Physicians vs. ChatGPT		Physicians vs. ChatGPT + Ctx		ChatGPT vs. ChatGPT + Ctx	
	Values	p -Value	Values	p -Value	Values	p -Value
Compl.	3.658 vs. 3.159	<0.001	3.658 vs. 3.618	0.862	3.159 vs. 3.618	<0.001
Accuracy	4.042 vs. 3.091	<0.001	4.042 vs. 3.631	<0.001	3.091 vs. 3.631	<0.001

In terms of accuracy, ChatGPT garnered 17.5% preferences, ChatGPT + context received 34.8%, and users demonstrated the highest accuracy with 47.7% preferences. Similarly, regarding precision, ChatGPT achieved 18.0% preferences, ChatGPT + context had 37.4%, and users scored 44.6%. Completeness showed a similar trend, with ChatGPT at 18.2%, ChatGPT + context at 38.7%, and users at 43.1%. In terms of usability, ChatGPT obtained 16.6%, ChatGPT + context had 32.2%, and users showed the highest preference at 51.2%. Lastly, for relevance, ChatGPT received 18.7%, ChatGPT + context had 38.5%, and users scored 42.8%. Importantly, all differences were statistically significant across comparisons, with a p -value < 0.001 for all five metrics.

The results of the evaluation showed that ChatGPT + context responses were generally preferred to the ones generated from the model without context in all five metrics. However, users were found to have a stronger preference for human-generated responses in every metric. Results are shown in Figure 2.

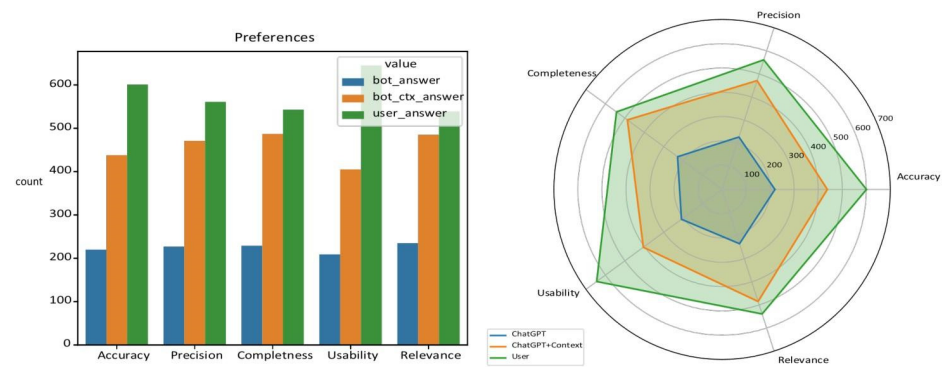


Figure 2. Results of the evaluation of the generated response options (ChatGPT, ChatGPT + context, human answer) for each question in terms of accuracy, precision, completeness, usability, and relevance.

Following the evaluation, a questionnaire was dispatched to users to determine the nature of errors in ChatGPT + context responses. The questionnaire was sent together with the poorly rated questions (2 or less on the Likert accuracy scale). The results are depicted in Figure 3. The most common cause of error turned out to be wrong content (a wrong reply by the bot), followed by improper context (an off-topic response caused by a failure in the context passed to the bot). Upon re-analyzing the generated responses, it was observed that, in most cases, the error originated from an incorrect reference context provided to ChatGPT by the semantic search process. We also found that ChatGPT quality degrades if the user provides a question with Roman numbers or acronyms, which are often mistaken for something else.

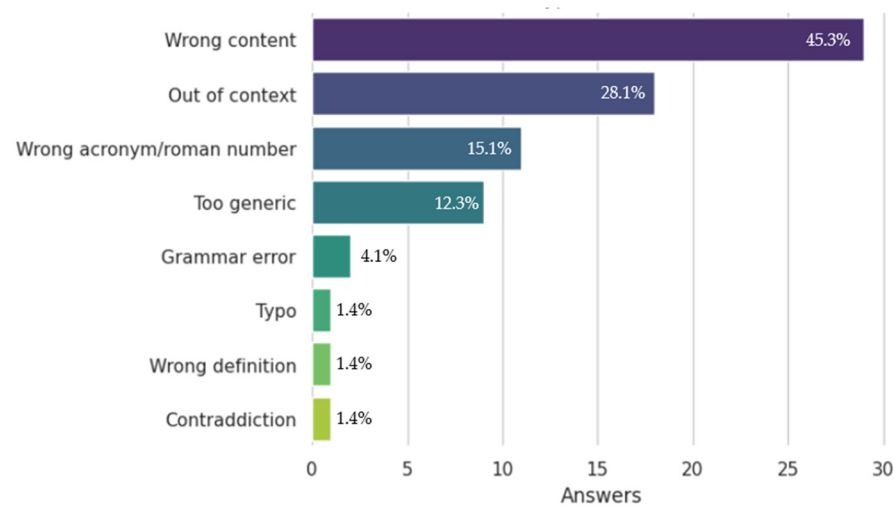


Figure 3. The results of the questionnaire were conducted to investigate the most frequent specific errors encountered in the answers generated by ChatGPT + context.

4. Discussion

The results obtained from this study represent the first evaluation of ChatGPT’s performance in generating responses to complex questions based on the regulatory context of Italian occupational medicine. In the study, it was found that the completeness of responses provided by ChatGPT with context is like that of responses given by doctors. However, when it comes to accuracy, although ChatGPT with context is better than the version without context, it does not reach the level of accuracy achieved by physicians. Even when comparing preferences for all five metrics, ChatGPT with context outperforms the version without context but falls short of the quality of responses generated by physicians, which were generally preferred by the evaluators. In our study, it is also noteworthy to

observe that even human-generated responses did not consistently achieve a perfect score in terms of accuracy and comprehensiveness. This is because users can also make errors or may not have analyzed all the necessary information to respond accurately to the query. Also, the evaluation of human responses is often influenced by subjectivity and variability among different assessors. Evaluations may be based on personal opinions, cognitive biases, prior experiences, and individual interpretations. This can lead to variable scores in the assessment of human-generated responses as well.

In the literature, few studies have assessed ChatGPT's performance in responding to open-ended medical questions. Johnson et al. [36] assessed the precision of cancer-related information delivered by ChatGPT when compared to the responses from the National Cancer Institute (NCI). They evaluated the accuracy (accurate: yes vs. no) and carried out a comparative analysis between the anonymized responses from NCI and ChatGPT. Their study indicates that ChatGPT delivers accurate information regarding common cancer-related information. In another study, Johnson D. et al. [37] assessed the accuracy and comprehensiveness of medical queries generated by healthcare professionals using ChatGPT. Subsequently, the physicians assigned ratings to the responses generated by ChatGPT based on their accuracy, measured on a 6-point Likert scale ranging from 1 (completely incorrect) to 6 (completely correct), and completeness, assessed on a 3-point Likert scale ranging from 1 (incomplete) to 3 (complete with additional context). The results indicated that across all the questions (in total, 284), the median accuracy score was 5.5 (indicating responses between almost completely and completely correct), with a mean score of 4.8 (signifying responses between mostly and almost completely correct). Furthermore, the median completeness score was three (reflecting complete and comprehensive responses), with a mean score of 2.5. Our study stands out as the first to test ChatGPT on workplace health and safety legislation. Furthermore, the metrics used in our study allow for an assessment not only of the accuracy and completeness of the chatbot-generated responses but also of the potential usability and relevance of these responses in the day-to-day practice of occupational health professionals. In addition, the errors found in responses that received low scores enabled the analysis of ChatGPT's performance limitations. We also evaluated not only ChatGPT but also its integrated version with selected context, which demonstrates an increase in ChatGPT's response performance.

Despite its promising performance in responding to open-ended questions, ChatGPT undoubtedly still has some important limitations. As ChatGPT training data goes only as far as the end of 2021, the chatbot cannot keep up with the latest trends and news, even though the introduction of plugins partially mitigates this time constraint. ChatGPT can also generate inconsistent or contradictory responses that do not align with medical guidelines [38]. In evaluating ChatGPT's performance in responding to open-ended questions, the phenomenon of hallucinations must be taken into consideration. Hallucinations are instances where LLMs assertively incorporate incorrect details into the generated responses. Recognizing hallucinations is essential, but understanding how to mitigate their occurrence is equally crucial [39], while their mitigation remains an open question and is subject to numerous field studies [40,41]. However, it is a well-documented fact [42] that enhancing semantic research improves ChatGPT's performance, something that aligns with the findings of our study, where the context-integrated version outperforms the one without context. Therefore, our future approach will continue to focus on refining the semantic research process to reduce text hallucinations. Additionally, through a survey sent to study participants, we tried to understand why the ChatGPT + context version received low accuracy scores in the answers to certain questions. It turned out that ChatGPT often misinterprets acronyms and Roman numerals and frequently makes references to a context that is unrelated to the question's requirements. Consequently, by addressing these system limitations and coupling them with ongoing expert evaluations, we may enhance the quality of responses generated by ChatGPT. As a reference, we included some samples of responses from ChatGPT + context alongside the reference answers from users in Appendix D. There we reported an example of a reply evaluated highly by physicians

and three hallucinated replies caused by a wrong reply (Table A4), an incorrect semantic search process (Table A5), and an incorrect identification of an acronym (Table A6).

There are also concerns about the ethical implications to consider regarding conversational AI in medical practice, as well as the potential for bias and errors in the data used to train AI algorithms. The legal implications of using these technologies cannot be downplayed. For instance, determining liability in the case of an inevitable mistake by an AI physician is yet to be established. The use of ChatGPT in medical practice can raise various legal and liability issues. If a medical error or negligence is attributed to the use of a system like ChatGPT, healthcare professionals and organizations may be held legally accountable [43]. Therefore, it is crucial to establish clear procedures and technology usage protocols.

Based on the findings of this study, several recommendations can be made for the future use of ChatGPT and LLMs in occupational health medicine. Firstly, it is important to ensure that the language model has full access to the legislative context to mitigate potential hallucinations and errors in its responses. ChatGPT accuracy greatly depends on the quality of the context data and the reliability of the semantic search process. This is particularly true in medicine, where LLMs' errors and hallucinations may have potentially severe repercussions on people's health. To mitigate these downsides, it is essential to ensure the use of accurate and up-to-date data from reliable sources (guidelines, regulations, etc.). However, because the study shows that even an occupational health LLM with full access to the legislative context is still prone to errors, human supervision is always necessary, and we do not suggest its direct and unfiltered interaction with patients who are unable to detect potential hallucinations.

Additionally, ChatGPT should be used as a tool to support, rather than replace, human expertise and should not be currently relied upon as the sole source of information. ChatGPT cannot replace the experience, knowledge, and skills of a professional medical doctor in the diagnosis and treatment of medical conditions, as well as in establishing a doctor-patient relationship. Human medical professionals play an indispensable role in the decision-making process.

Furthermore, it is important to implement appropriate safeguards to protect the privacy and security of user data when using ChatGPT, especially if the model is deployed in the cloud and patients' information is sent to a third party. Open-source LLMs like Llama [44], Alpaca [45], and Vicuna [46] are now approaching the level of quality of proprietary models like ChatGPT and do not need to send data in the cloud, as they run fully on the user's PC and thus eliminate any privacy concerns. However, these models require system memory and computational power well beyond the reach of the average user's laptop. Nevertheless, next-generation Central Processing Units (CPU) from all major manufacturers will have an AI accelerator, so running an LLM locally on a user's laptop will become accessible to everyone.

The use of ChatGPT and similar AI can bring numerous benefits to the field, including round-the-clock assistance to occupational physicians, increased efficiency, cost reduction, health monitoring, and personalized services. With ChatGPT, occupational physicians could ask questions about workplace safety, workers' rights, safety protocols, and more at any time and receive immediate responses. Its impact on health and safety, as well as societal values, requires critical reflection. However, based on the ethical and legal limitations and implications, a responsible, ethical, and reflective approach is necessary for their development and use. It is important that developers, practitioners, and policymakers work together to ensure that the use of such tools is ethical, transparent, and accountable and that appropriate safeguards are in place to protect both users and professionals in the field of occupational health.

Several limitations of this study can be identified that we plan to overcome in future work. One possible limitation of this work is the small sample size of employed physicians and the fact that this is a monocentric study. Moreover, questions were created solely based on legislative decree 81/08, which represents the reference legislation in the field of occupa-

tional medicine. However, there are also numerous other laws or decrees that regulate the protection of workers' health and safety in the workplace, for example, legislative decree n. 66/2003 about night work or legislative decree n.151/2001 on the protection of working mothers.

5. Conclusions

Our study revealed that ChatGPT + context outperformed ChatGPT in responding to complex open-ended questions based on the regulatory context of Italian occupational medicine, even though it did not reach the level of accuracy and completeness of the questions generated by physicians. This study not only underscores ChatGPT's potential as a valuable resource within the field of occupational health but also emphasizes the need to acknowledge the current limits of AI.

To assess chatbots' potential in occupational medicine as virtual assistants for medical inquiries in workplace health and safety, we plan to improve both the quality of the reference questions and answers and their size by employing a more diverse, multi-center, and wider sample of occupational health professionals, including medical residents, academics, occupational physicians, and other professionals working in the field of workplace health and safety.

Our team also aims to conduct a follow-up study to compare the performance of the new version of ChatGPT with GPT-4 and Google's Bard AI. Further studies should be encouraged to explore other possible applications of ChatGPT and other large language models in occupational medicine.

Author Contributions: Conceptualization, A.P. (Alessandro Palla), M.P., R.F. and R.M.; data curation, M.P., B.C. and A.P. (Armando Petillo), G.N. (Gianluca Nerli), F.P., S.S., F.C., L.D., N.M., G.N. (Gabriele Necciari), V.C.L. and R.M.; methodology and statistics, A.P. (Alessandro Palla) and M.P.; software, A.P. (Alessandro Palla); formal analysis, A.P. (Alessandro Palla); writing—original draft preparation, A.P. (Alessandro Palla) and M.P.; writing—review and editing, A.P. (Alessandro Palla), M.P., R.F. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available as their release is subject to ongoing collaboration and further analysis to derive additional insights.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

We used the OpenAI ChatGPT API to reply to the physicians' questions. The method employed two distinct approaches.

Appendix A.1. ChatGPT-Generated Answers

In the first approach, the questions generated by the two groups were fed directly into ChatGPT. To mimic a real-use case scenario, both questions and answers were submitted in Italian. An example of a prompt, question, and answer generated by ChatGPT is shown in Table A1 with the original question and answer that have been translated for a wider audience.

Table A1. Example of ChatGPT question and answer in Italian and English.

Language	Question	Answer
Italian	Secondo il D.lgs 81/08 chi è il lavoratore video terminalista?	Il lavoratore che utilizza un'attrezzatura munita di videotermini, in modo sistematico o abituale, per venti ore settimanali, dedotte le interruzioni.
English	According to legislative decree 81/08, who is the video terminal worker?	A worker who systematically or habitually uses video terminal equipment for twenty hours per week, excluding breaks.

Appendix A.2. ChatGPT + Context-Generated Answers

In the second approach, ChatGPT + context, the reference legislative context from D.lgs 81/08 was also passed alongside the questions. To retrieve the context, we used a semantic search with embeddings. Text embeddings refer to a technique used in natural language processing (NLP) and machine learning that involves the conversion of sentences into numerical vectors. Text embeddings encode the semantics of words, sentences, and entire texts into vectors in a high-dimensional vector space. They are used to measure how related two different strings of text are. For example, the text embedding model we used, OpenAI 'text-embedding-ada-002', can take a sentence up to 8192 tokens and generate a single vector of dimensionality 1536. A typical application of text embeddings is to perform a semantic search to retrieve the relevant documents and paragraphs from a vast corpus from a user's query written in natural language. Models that produce text embedding are trained such that a pre-defined distance metric (Euclidean distance, cosine distance, etc.) is minimized for semantically similar text strings and maximized for the others. This behavior can be exploited for semantic search by selecting from the corpus the paragraphs with a smaller distance from the user's query.

We use semantic search to retrieve the relevant legislative context from D.lgs 81/08 for each question by executing the following steps:

- Parse the original document into chunks of text. We parsed the document using the 'UnstructuredPDF' library (version 0.5.11) to extract raw text from the document. We then split the documents into several chunks of N tokens. The chunks' size is N, a system hyperparameter to optimize later;
- Create text embeddings for the documents in the search corpus. Each document in the corpus is represented as a numerical vector using text embeddings. We used the OpenAI Embeddings API using the 'text-embedding-ada-002' model;
- Create a query embedding: the user's search query is also converted as before and represented as a numerical vector using text embeddings;
- Calculate the similarity between the query embedding and the document embeddings. We used cosine similarity, which is often used to measure the similarity between two embedding vectors. For each document in the corpus, the cosine similarity between the query embedding and the document embedding is calculated;
- Rank the documents by similarity: The documents are ranked in descending order of similarity to the query. The most similar documents are returned as the search results.

We then concatenate together the highest score-retrieved paragraphs until the length of the question + context is smaller than 2048 tokens. That ensured plenty of tokens were left to generate a high-quality, complete, and precise answer to the question while at the same time providing the most possible context to the AI. The generation of the vector database and the architecture of the ChatGPT + context system are explained in Figure A1.

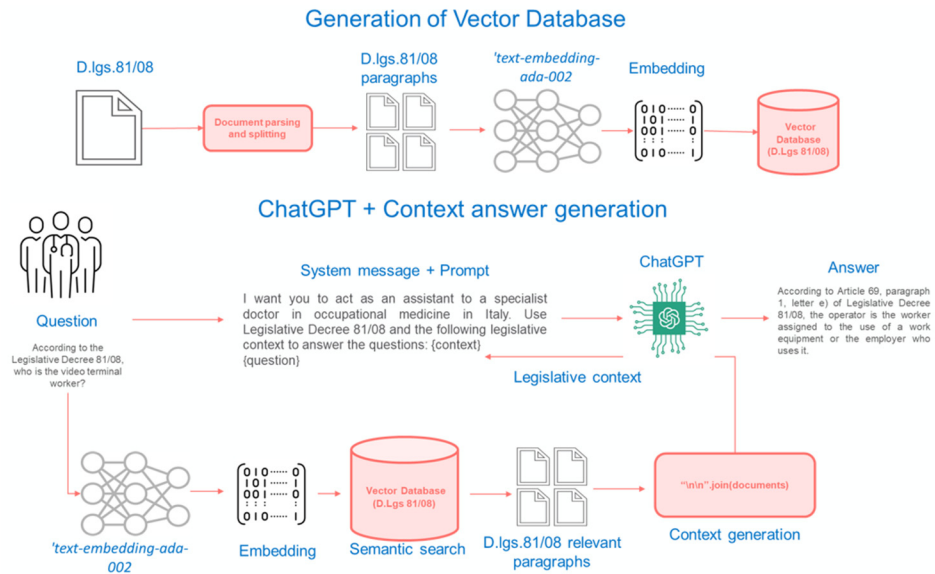


Figure A1. Generation of a vector database and implementation of the ChatGPT + context system.

Appendix A.3. ChatGPT System Message

The ChatGPT system message used is described in Table A2. The ChatGPT system message serves as the primary instructions for the model, and it can be tailored to include various information for the system role. This can comprise a brief overview of the assistant, its personality traits, specific instructions, guidelines that you want the assistant to adhere to, or relevant data or information that the model should have, such as frequently asked questions. It is possible to personalize the system role to suit specific use cases. Although the system role/message is not mandatory, we found that high-quality system messages improve the overall quality of the generated replies.

Table A2. Example of a question and answer in Italian and English. The text {context} is being replaced with the actual legislative context from the search before submitting the text to ChatGPT.

Language	ChatGPT	ChatGPT + Context
Italian	Voglio che tu agisca come un assistente di un medico specialista in medicina del lavoro in Italia. Usa il decreto legislativo 81/08 per rispondere alle domande.	Voglio che tu agisca come un assistente di un medico specialista in medicina del lavoro in Italia. Usa il decreto legislativo 81/08 ed il seguente contesto normativo per rispondere alla seguente domanda. Contesto normativo: {context}
English	I want you to act as an assistant to a specialist doctor in occupational medicine in Italy. Use legislative decree 81/08 to answer the questions.	I want you to act as an assistant to a specialist doctor in occupational medicine in Italy. Use legislative decree 81/08 and the following legislative context to answer the questions. Legislative context: {context}

Appendix A.4. Question/Answering Optimization and Prompt Engineering

Because of the limited context size of language models, to deal with long pieces of text, it is necessary to break the legislative context down into smaller chunks/paragraphs. Although this may seem simple, it can be quite complex for an automatic algorithm as we wanted to keep semantically related pieces of text together, which can vary depending on the type of text.

At a high level, text splitters operate in the following manner:

- Split the text into smaller, semantically meaningful chunks (usually sentences or paragraphs);

- Combine these smaller chunks into larger chunks until a certain size is reached;
- Make that chunk its own piece of text once the size threshold is reached, and then create a new chunk with some overlap to maintain context between the chunks.

This process is very text-dependent and, in general, quite hard to find the optimal choice of chunk size, system prompt, and text processing that maximizes results. Usually, the process is performed by iterating through multiple different parameter configurations and using human evaluators on the generated replies to assess the parameter combination that gives the best result. However, we decided against this approach since we were in need of quickly iterating through multiple combinations of parameters while designing the system, and using human evaluation would introduce an unacceptable latency in the development cycle.

To circumvent the human evaluator bottleneck, we can use ChatGPT itself as a close approximation to human judgment to drive the chunk size and system prompt decision by using the LLM itself to score the generated answers based on their accuracy and completeness. In this way, it is possible to optimize our QA pipeline with a combination of system parameters such as split type, length, and system message that maximize our metrics. This approach has the clear advantage that it is much faster and scales better, allowing it to quickly iterate over multiple parameter combinations.

The evaluator was implemented using prompt engineering in the LangChain framework. In a nutshell, we gave as input to the model the question, the reference answer from physicians, and the ChatGPT + context-generated answer. Then, we asked ChatGPT to provide an evaluation score for both accuracy and completeness and then rank each parameter combination considering its average score on both metrics. We then selected the parameter combination that maximized our metrics. The ChatGPT system message is shown in Table A3. We pass it alongside the ChatGPT + context-generated answer, and the evaluator returns the accuracy and completeness metrics as instructed. Despite the potential misalignment between human and ChatGPT evaluators, this approach provides a fast way to iterate through many system parameters and optimize/fine-tune an LLM Question/Answering system given the multiple design choices and possible system prompts. Once the system has been fine-tuned, we use it to generate new answers for the final human evaluation phase.

Table A3. System message for using ChatGPT as an evaluation for QA pipeline finetuning and prompt engineering.

Language	Evaluator System Message
Italian	<p>Data la seguente domanda: {question}, la risposta corretta è {reference}. Misura in una scala Likert che va da 1 a 5 la completezza ed accuratezza della risposta dell'utente. Le risposte possono essere non completamente uguali a quelle di riferimento ed essere più complete ed accurate ed in quel caso non penalizzarle. Il formato della tua risposta deve essere un JSON parsabile nel seguente formato:</p> <p>Esempio in caso di risposta accurate e completa {"completezza": 5, "accuratezza": 5}</p> <p>Esempio in caso di risposta accurata ma non completa {"completezza": 2, "accuratezza": 5}</p> <p>Esempio in caso di risposta errata {"completezza": 1, "accuratezza": 1}</p>
English	<p>Given the following question: {question}, the correct answer is {reference}. Measure the completeness and accuracy of the user's response on a Likert scale ranging from 1 to 5. The answers may not be completely the same as the reference ones and may be more complete and accurate; in that case, do not penalize them. The format of your response must be a parsable JSON in this format.</p> <p>For example, in the case of an accurate and complete response. {"completeness": 5, "accuracy": 5}</p> <p>For example, in the case of an accurate but incomplete response. {"completeness": 2, "accuracy": 5}</p> <p>For example, in the case of an incorrect answer. {"completeness": 1, "accuracy": 1}</p>

Appendix B. Details of the Web Interface Used for Evaluation

The website provided a user-friendly interface that allowed the doctors to evaluate the responses based on the predefined criteria and using the Likert 5-point scale. The layout of the website was carefully crafted to ensure that the questions and their corresponding three options were always visible to the users. This approach was implemented to simplify the decision-making process for the evaluators.

The website was designed in such a way that participants could only proceed to the next evaluation after completing all fields. This feature ensured that all evaluations were completed in full and that there were no missing data.

To provide the participants with a comprehensive understanding of the evaluation procedures, an external document was provided separately to the users. This document contained detailed explanations of all evaluation procedures and protocols, which ensured that the participants were fully aware of what was expected of them during the evaluation process.

Appendix C. Evaluation Questionnaire

You have received a PDF file containing five questions for which you rated the accuracy of ChatGPT’s response along with its context as low. You are now asked to provide context for your rating for each of the five questions. Therefore, you will need to submit the questionnaire five times, once for each of the five questions received via email.

* REQUIRED

1. Which question are you about to evaluate? *

- question 1
- question 2
- question 3
- question 4
- question 5

2. Are there any spelling or grammar errors in the ChatGPT + context answer? If yes, please specify. *

3. Does the ChatGPT + context answer in an inappropriate or offensive tone? *

- no
- yes

4. Did you notice any inconsistencies in ChatGPT + context answer? If yes, please specify which ones *

5. Are there any acronyms or Roman numerals in the question that were not understood by ChatGPT + context? *

- no
- yes

6. Was the answer provided by ChatGPT + context understandable? If not, please specify why. *

7. Why did you score the answer low? *

Appendix D. Examples of Good/Bad/Hallucinating Replies from ChatGPT

Table A4. Example of a good reply from ChatGPT + context.

Language	Italian	English
Question	Quale è la definizione di operatore secondo il D. Lgs 81/08?	What is the definition of an operator according to legislative decree 81/08?
Reference answer	L’operatore è definito come il lavoratore incaricato dell’uso di una attrezzatura di lavoro o il datore di lavoro che ne fa uso.	The operator is defined as the worker responsible for the use of work equipment or the employer who uses it.
ChatGPT + context answer	Secondo l’art. 69, comma 1, lettera (e) del D.Lgs 81/08, l’operatore è il lavoratore incaricato dell’uso di una attrezzatura di lavoro o il datore di lavoro che ne fa uso.	According to Article 69, paragraph 1, letter (e) of legislative decree 81/08, the operator is the worker in charge of using work equipment or the employer who uses it.

Table A4. *Cont.*

Language	Italian	English
Average ChatGPT + context accuracy		5.0
Number of evaluations		4

ChatGPT + context gave a correct definition with the proper reference to the D.Lgs 81/08 article.

Table A5. Hallucinating reply from ChatGPT + context.

Language	Italian	English
Question	A quale classe appartiene il virus dell'Epatite B secondo la classificazione prevista dal D.lgs 81/08?	To which class does the hepatitis B virus belong according to the classification provided by legislative decree 81/08?
Reference answer	Secondo l' allegato XLVI—Elenco degli agenti biologici classificati, il virus dell'epatite B è classificato come agente biologico di classe 3.	According to Annex XLVI—List of classified biological agents, the hepatitis B virus is classified as a biological agent of class 3.
ChatGPT + context answer	Il virus dell'Epatite B appartiene alla classe D (rischio alto) secondo la classificazione prevista dal D.lgs 81/08.	The hepatitis B virus belongs to class D (high risk) according to the classification provided by legislative decree 81/08.
Average ChatGPT + context accuracy		1.0
Number of evaluations		2

ChatGPT + context is hallucinating a class (Class D) that does not exist.

Table A6. Hallucinating reply from ChatGPT + context caused by incorrect semantic search.

Language	Italian	English
Question	Quale titolo del D.Lgs 81/08 definisce i criteri per la valutazione del rischio da atmosfere iperbariche e microclima?	Which title of legislative decree 81/08 defines the criteria for assessing the risk of hyperbaric atmospheres and microclimates?
Reference answer	Non esiste un titolo che definisce i criteri per la valutazione del rischio da atmosfere iperbariche e microclima, tuttavia al titolo VIII del D.Lgs 81/08 si parla di valutazione del rischio da agenti fisici ivi compresi microclima ed atmosfere iperbariche.	No title defines the criteria for assessing the risk of hyperbaric atmospheres and microclimates. However, Title VIII of legislative decree 81/08 addresses the risk assessment of physical agents, including microclimates and hyperbaric atmospheres.
ChatGPT + context answer	Il titolo X del D.Lgs 81/08, "Lavori in ambienti iperbarici e rischio da microclima", definisce i criteri per la valutazione del rischio da atmosfere iperbariche e microclima.	Title X of legislative decree 81/08, "Work in Hyperbaric Environments and Risk from Microclimate," defines the criteria for assessing the risk of hyperbaric atmospheres and microclimate.
Average ChatGPT + context accuracy		1.0
Number of evaluations		3

ChatGPT uses an incorrect reference, as Chapter X refers to biological risk. Furthermore, D.Lgs 81/08 does not define criteria for hyperbaric atmosphere and microclimate. This is a clear example of a hallucinating response that uses an incorrect reference context from the semantic search process (Chapter X) and gives an incorrect reply.

References

1. Fogel, A.L.; Kvedar, J.C. Artificial intelligence powers digital medicine. *NPJ Digit. Med.* **2018**, *1*, 5. [[CrossRef](#)] [[PubMed](#)]
2. Moassefi, M.; Faghani, S.; Khosravi, B.; Rouzrokh, P.; Erickson, B.J. Artificial Intelligence in Radiology: Overview of Application Types, Design, and Challenges. *Semin. Roentgenol.* **2023**, *58*, 170–177. [[CrossRef](#)] [[PubMed](#)]
3. Raghunath, S.; Pfeifer, J.M.; Ulloa-Cerna, A.E.; Nemani, A.; Carbonati, T.; Jing, L.; Vanmaanen, D.P.; Hartzel, D.N.; Ruhl, J.A.; Lagerman, B.F.; et al. Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation-Related Stroke. *Circulation* **2021**, *143*, 1287–1298. [[CrossRef](#)] [[PubMed](#)]
4. Chen, D.; Liu, J.; Zang, L.; Xiao, T.; Zhang, X.; Li, Z.; Zhu, H.; Gao, W.; Yu, X. Integrated Machine Learning and Bioinformatic Analyses Constructed a Novel Stemness-Related Classifier to Predict Prognosis and Immunotherapy Responses for Hepatocellular Carcinoma Patients. *Int. J. Biol. Sci.* **2022**, *18*, 360–373. [[CrossRef](#)] [[PubMed](#)]
5. Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. *Sensors* **2021**, *21*, 2852. [[CrossRef](#)] [[PubMed](#)]
6. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [[CrossRef](#)] [[PubMed](#)]
7. Haug, C.J.; Drazen, J.M. Artificial Intelligence and Machine Learning in Clinical Medicine. *N. Engl. J. Med.* **2023**, *388*, 1201–1208. [[CrossRef](#)]
8. Aung, Y.Y.; Wong, D.C.; Ting, D.S. The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare. *Br. Med. Bull.* **2021**, *139*, 4–15. [[CrossRef](#)]
9. Rajpurkar, P.; Lungren, M.P. The Current and Future State of AI Interpretation of Medical Images. *N. Engl. J. Med.* **2023**, *388*, 1981–1990. [[CrossRef](#)]
10. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *arXiv* **2023**, arXiv:2307.06435.
11. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. arXiv:1706.03762.
13. Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-tuning language models from human preferences. *arXiv* **2019**, arXiv:1909.08593.
14. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
15. Open, A.I. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
16. 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*, 1233–1239. [[CrossRef](#)] [[PubMed](#)]
17. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv* **2023**, arXiv:230211382.
18. Dahmen, J.; Kayaalp, M.E.; Ollivier, M.; Pareek, A.; Hirschmann, M.T.; Karlsson, J.; Winkler, P.W. Artificial intelligence bot ChatGPT in medical research: The potential game changer as a double-edged sword. *Knee Surg. Sports Traumatol. Arthrosc.* **2023**, *31*, 1187–1189. [[CrossRef](#)]
19. Liu, J.; Wang, C.; Liu, S. Utility of ChatGPT in Clinical Practice. *J. Med. Internet Res.* **2023**, *25*, e48568. [[CrossRef](#)]
20. Gordijn, B.; Have, H.T. ChatGPT: Evolution or revolution? *Med. Health Care Philos.* **2023**, *26*, 1–2. [[CrossRef](#)]
21. Rao, A.S.; Pang, M.; Kim, J.; Kamineni, M.; Lie, W.; Prasad, A.K.; Landman, A.; Dryer, K.; Succi, M.D. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv* **2023**. 2023-02. [[CrossRef](#)]
22. Hirose, T.; Harada, Y.; Yokose, M.; Sakamoto, T.; Kawamura, R.; Shimizu, T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3378. [[CrossRef](#)] [[PubMed](#)]
23. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Assessing the value of ChatGPT for clinical decision support optimization. *medRxiv* **2023**. 2023-02. [[CrossRef](#)]
24. Chintagunta, B.; Katariya, N.; Amatriain, X.; Kannan, A. Medically aware GPT-3 as a data generator for medical dialogue summarization. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Virtual, 21 October 2021; pp. 354–372.
25. Joshi, A.; Katariya, N.; Amatriain, X.; Kannan, A. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv* **2020**, arXiv:2009.08666.
26. Sivasubramanian, J.; Shaik Hussain, S.M.; Virudhunagar Muthuprakash, S.; Periadurai, N.D.; Mohanram, K.; Surapaneni, K.M. Analysing the clinical knowledge of ChatGPT in medical microbiology in the undergraduate medical examination. *Indian J. Med. Microbiol.* **2023**, *45*, 100380. [[CrossRef](#)] [[PubMed](#)]
27. Antaki, F.; Touma, S.; Milad, D.; El-Khoury, J.; Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol. Sci.* **2023**, *3*, 100324. [[CrossRef](#)]
28. Patil, N.S.; Huang, R.S.; van der Pol, C.B.; Larocque, N. Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment. *Can. Assoc. Radiol. J.* **2023**, *14*, 8465371231193716. [[CrossRef](#)] [[PubMed](#)]

29. Guerra, G.A.; Hofmann, H.; Sobhani, S.; Hofmann, G.; Gomez, D.; Soroudi, D.; Hopkins, B.S.; Dallas, J.; Pangal, D.J.; Cheek, S.; et al. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. *World Neurosurg.* **2023**, *18*. [[CrossRef](#)]
30. Sridi, C.; Brigui, S. The use of ChatGPT in occupational medicine: Opportunities and threats. *Ann. Occup. Environ. Med.* **2023**, *35*, e42. [[CrossRef](#)]
31. Amato FDF Gianfranco. Decreto Legislativo 81/08: Test Unico Sulla Salute e Sicurezza Sul Lavoro. Available online: <https://www.ispettorato.gov.it/files/2023/03/TU-8108-Ed-Gennaio-2023.pdf> (accessed on 1 May 2023).
32. Jones, E.; Palangi, H.; Simões, C.; Chandrasekaran, V.; Mukherjee, S.; Mitra, A.; Awadallah, A.; Kamar, E. Teaching Language Models to Hallucinate Less with Synthetic Tasks. *arXiv* **2023**, arXiv:2310.06827.
33. Sisaengsuwanchai, K.; Nananukul, N.; Kejriwal, M. How does prompt engineering affect ChatGPT performance on unsupervised entity resolution? *arXiv* **2023**, arXiv:2310.06174.
34. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
35. Neelakantan, A.; Xu, T.; Puri, R.; Radford, A.; Han, J.M.; Tworek, J.; Yuan, Q.; Tezak, N.; Kim, J.W.; Hallacy, C.; et al. Text and code embeddings by contrastive pre-training. *arXiv* **2022**, arXiv:2201.10005.
36. Johnson, S.B.; King, A.J.; Warner, E.L.; Aneja, S.; Kann, B.H.; Bylund, C.L. Using ChatGPT to evaluate cancer myths and misconceptions: Artificial intelligence and cancer information. *JNCI Cancer Spectr.* **2023**, *7*. [[CrossRef](#)] [[PubMed](#)]
37. Johnson, D.; Goodman, R.; Patrinely, J.; Stone, C.; Zimmerman, E.; Donald, R.; Chang, S.; Berkowitz, S.; Finn, A.; Jahangir, E.; et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res. Sq.* **2023**. [[CrossRef](#)]
38. Alkaiissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179. [[CrossRef](#)] [[PubMed](#)]
39. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [[CrossRef](#)]
40. Rebuffel, C.; Roberti, M.; Soulier, L.; Scoutheeten, G.; Cancelliere, R.; Gallinari, P. Controlling hallucinations at word level in data-to-text generation. *Data Min. Knowl. Discov.* **2022**, *36*, 318–354. [[CrossRef](#)]
41. Wang, J.; Zhou, Y.; Xu, G.; Shi, P.; Zhao, C.; Xu, H.; Ye, Q.; Yan, M.; Zhang, J.; Zhu, J.; et al. Evaluation and Analysis of Hallucination in Large Vision-Language Models. *arXiv* **2023**, arXiv:2308.15126.
42. Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Dou, Z.; Wen, J.R. Large Language Models for Information Retrieval: A Survey. *arXiv* **2023**, arXiv:2308.07107.
43. Maliha, G.; Gerke, S.; Cohen, I.G.; Parikh, R.B. Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. *Milbank Q.* **2021**, *99*, 629–647. [[CrossRef](#)]
44. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
45. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Alpaca: A strong, replicable instruction-following model. *Stanf. Cent. Res. Found. Models* **2023**, *3*, 7. Available online: <https://crfm.stanford.edu/2023/03/13/alpaca.html> (accessed on 7 December 2023).
46. Peng, B.; Li, C.; He, P.; Galley, M.; Gao, J. Instruction tuning with gpt-4. *arXiv* **2023**, arXiv:2304.03277.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.