





Brief Report

Mirror, mirror on the wall, who is the best of them all? Artificial intelligence versus gastroenterologists in solving clinical problems

Felice Benedicenti ^{1,2,†}, Tommaso Pessarelli^{1,2,†}, Mattia Corradi^{1,2}, Marco Michelson^{1,2}, Nicoletta Nandi^{1,2}, Pietro Lampertico^{1,3}, Maurizio Vecchi ^{1,2}, Lucia Scaramella² and Luca Elli^{2,*,†}

¹Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy

²Gastroenterology and Endoscopy Unit, Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

³Division of Gastroenterology and Hepatology, CRC "A. M. and A. Migliavacca" Center for the Study of Liver Disease, Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

*Corresponding author. Gastroenterology and Endoscopy Unit, Foundation IRCCS Ca' Granda Policlinico Hospital of Milan, Via Francesco Sforza 35, Milan 20122, Italy. Email: luca.elli@policlinico.mi.it

[†]The authors wish it to be known that, in their opinion, F.B. and T.P. should be regarded as joint First Authors.

Introduction

Artificial intelligence (AI) is a concept that commonly refers to machines mimicking human cognitive behavior during learning and problem-solving [1]. Chatbots are a typical example of an AI system, capable of interacting with humans [2]. ChatGPT (OpenAI, San Francisco, CA, USA) [3] is a new-generation chatbot that captures the context and relationship between words in input sequences through multiple layers of self-attention and feed-forward neural networks. Then, it predicts the most likely "token" to succeed the previous one based on patterns in its training data. Therefore, it is a self-contained system that does not copy existing information [4].

Considering the important impact AI has on radiology and endoscopy, gastroenterology has become a major field of AI application and special interest has already been focused on several areas [5]. To date, however, a possible use of AI in gastroenterological clinical problem-solving has not been addressed [6].

In this setting, the aim of this study was to compare the accuracy of AI, through ChatGPT, with that of a group of gastroenterologists in solving gastroenterological clinical problems and thereby assess the potential usefulness of ChatGPT in improving clinicians' diagnostic workflow.

Methods

Study design

A set of 20 clinical gastroenterological and hepatological vignettes with 5 multiple-choice answers was independently created by 3 experts in gastroenterology and hepatology, and subsequently revised by the 3 authors. The vignettes were successively moved into a Google form questionnaire and submitted to 25 residents and 31

specialists in gastroenterology and hepatology, mainly working at a tertiary referral center in Northern Italy, the Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, University of Milan. The proposed questions concerned hepatological and gastroenterological clinical cases. The questions had five possible answers (A–E), of which only one was correct (Supplementary Table 1). To prevent any external consultation, the questionnaire was submitted simultaneously to all participants without warning and with external supervision to secure the smooth running of the completion of the questionnaires.

The generated vignettes were directly copied into ChatGPT [4] with the same multiple-choice format. All ChatGPT model outputs were collected from ChatGPT 3 version on January 9, 2023 and then on May 9, 2023. To avoid the influence of answers of other vignettes on the model output, a new ChatGPT session was initiated for each vignette. The answers of the chatbot were considered to be correct if they most likely referred to the single correct alternative. This evaluation was carried out by two independent observers (F.B. and T.P.). In the event of discordance between the observers, an agreement was reached through verbal discussion.

Statistical analysis

All statistical analyses were performed using SPSS version 28.0.1.0 (IBM, Armonk, NY, USA). Continuous data are presented as mean \pm standard deviation or as median \pm interquartile range (IQR), with the corresponding 95% confidence interval. Categorical data are expressed as frequency and percentage. Continue variables were analysed using Student's *t*-test or Mann–Whitney *U* test according to data distribution. Correlation between two continuous variables was analysed through Bivariate Correlation and Pearson correlation coefficient. Statistical significance was set at $P < 0.05$.

Received: 18 May 2023. Revised: 15 July 2023. Accepted: 02 August 2023

© The Author(s) 2023. Published by Oxford University Press and Sixth Affiliated Hospital of Sun Yat-sen University

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

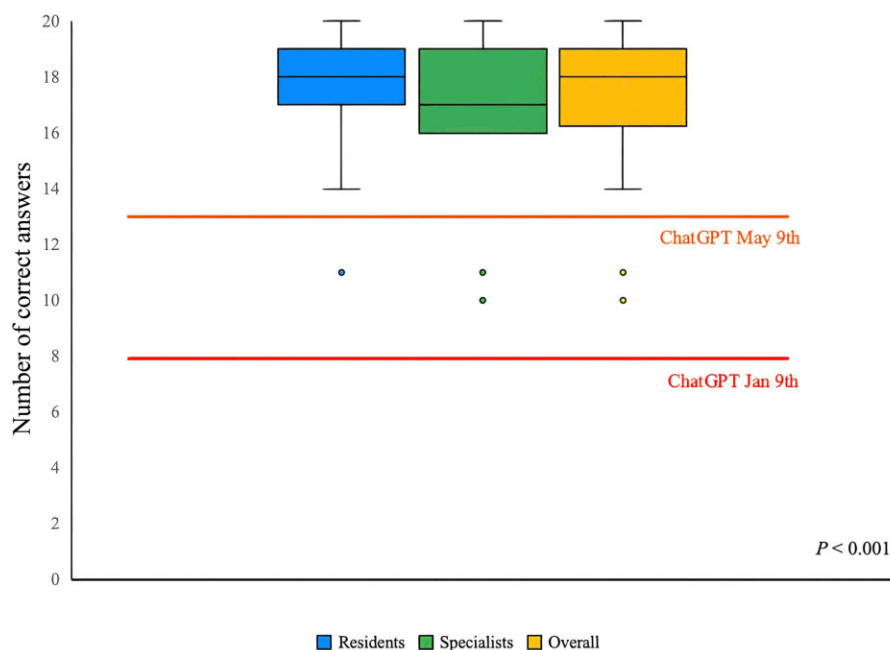


Figure 1. Number of correct answers achieved by residents, specialists, and overall participants compared with those achieved by ChatGPT on January 9 and May 9 2023.

Results

A total of 56 gastroenterologists were included in the study (48.2% females) with a mean age of 38.9 ± 15.2 years (Supplementary Table 2). The group of specialists was composed of 31 gastroenterologists, with 51.6% females and a mean age of 47.9 ± 15.6 years; they worked in university hospitals (38.7%), public hospitals (22.6%), or private facilities (38.8%). Of the 25 residents (44.0% females) with a mean age of 27.7 ± 1.6 years, 64% were in their first 2 years of residency training. The average number of correct answers for human participants was 17.3 ± 2.3 . Subgroup analysis showed no statistically significant difference in the average number of correct answers between the residents and the specialists (88.0% vs 84.8%, $P = 0.202$) and between residents of the first- and second-year period (86.5% vs 90.5%, $P = 0.452$). Conversely, a statistically significant difference was found between university physicians (residents and specialists) and those working in public or private facilities (88.6% vs 81.6%, $P = 0.020$). The number of years of medical practice negatively correlated with the performance in the test (Pearson = -0.502 , $P = 0.004$). On January 9, 2023, ChatGPT performed 8 correct answers, while, on May 9, 2023, ChatGPT scored 13 correct answers. In both cases, the performances of ChatGPT were inferior to those of humans (40.0% vs 86.2%, $P < 0.001$; 65.0% vs 86.2%, $P < 0.001$; Figure 1 and Supplementary Table 3). Over a period of 122 days, ChatGPT showed an improvement of 62.5% in the accuracy of responses.

Discussion

ChatGPT is the latest frontier in AI systems. Its sophisticated design combined with the obvious ease of use makes it extremely appealing in the medical field. Current AI utilization in medicine has been rapidly expanding, although the potential utility in the therapeutic diagnostic process is still underexplored. Few studies have investigated the diagnostic skills of ChatGPT and, to our knowledge, this is the first study to explore its application in the gastroenterological field.

In our study, ChatGPT showed a significantly lower rate of correct answers compared with gastroenterologists. This result is in line with that of Huh et al. [7] who reported a poorer performance of ChatGPT compared with 77 medical students (60.8% vs 90.8%) in answering questions about parasitology. On the contrary, previous studies have reported better performances of AI [8, 9]. However, ChatGPT performance in our study is not to be considered negative at all. First of all, in just 122 days, ChatGPT improved its performance by 62.5%. Moreover, the submission of vignettes to a highly qualified cohort of specialists and residents is a possible contributing factor to the difference in performance between ChatGPT and humans. Finally, even in the case of an incorrect answer, ChatGPT often provided a proper assessment of the clinical case and useful suggestions for the subsequent diagnostic and therapeutic workup (Supplementary Materials 4 and 5).

In conclusion, this study highlights how ChatGPT, an easily available AI system, appears to be inferior to specialists and residents operating in tertiary centers in solving gastroenterological clinical problems. However, the software showed an impressive improvement in the accuracy of the answers over a 4-month period. If confirmed by prospective studies of direct application on patient management, these findings might soon justify the introduction of ChatGPT into daily clinical practice.

Supplementary Data

Supplementary data is available at *Gastroenterology Report* online.

Authors' Contributions

L.E. and F.B. conceived of the presented idea. F.B., T.P., M.C., M.M., N.N., L.S., and L.E. collected and interpreted the data. T.P. and F.B. developed the text. L.E., M.V., and P.L. supervised the findings of this work. All authors followed the development of the study, discussed the results, and contributed to the final manuscript. All authors have read and approved the final version of the manuscript.

Funding

This study was partially funded by Italian Ministry of Health, Current research IRCCS.

Acknowledgements

The authors acknowledge support from the University of Milan through the APC initiative.

Conflict of Interest

None declared.

References

1. Gupta R, Srivastava D, Sahu M et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers* 2021;**25**:1315–60.
2. Le Berre C, Sandborn WJ, Aridhi S et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology* 2020;**158**:76–94.e2.
3. Castelvechi D. Are ChatGPT and AlphaCode going to replace programmers? *Nature* 2022. <https://doi.org/10.1038/d41586-022-04383-z>.
4. Zhu J, Jiang J, Yang M et al. ChatGPT and environmental research. *Environ Sci Technol* 2023. <https://doi.org/10.1021/acs.est.3c01818>.
5. Cao R, Tang L, Fang M et al. Artificial intelligence in gastric cancer: applications and challenges. *Gastroenterol Rep (Oxf)* 2022;**10**: goac064.
6. Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. *Nat Rev Gastroenterol Hepatol* 2020;**17**:591–2.
7. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof* 2023;**20**:1.
8. Rao A, Pang M, Kim J et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J Med Internet Res* 2023;**25**:e48659.
9. Benoit JR. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. *medRxiv* 2023.