Large Language Models (LLMs) vs. Specialist Doctors: A Comparative Study on Health Information in specific medical domains.

Zelin Yan, Jingwen Liu, Shiyuan Lu, Dingting Xu, Yun Yang, Honggang Wang, Jie Mao, Hou-Chiang Tseng, Tao-Hsing Chang, Yan Chen, Yihong Fan

Submitted to: Journal of Medical Internet Research
on: June 03, 2024

Table of Contents

Large Language Models (LLMs) vs. Specialist Doctors: A Comparative Study on Health Information in specific medical domains.

Zelin Yan[1, 2, 3*] ms; Jingwen Liu[3*] phd; Shiyuan Lu[3] phd; Dingting Xu[3] md, phd; Yun Yang[4] bs; Honggang Wang[5] md, phd; Jie Mao[6] ms; Hou-Chiang Tseng[7] phd; Tao-Hsing Chang[8] phd; Yan Chen[3] md, phd; Yihong Fan[1] md, phd

[1]Department of Gastroenterology The First Affiliated Hospital of Zhejiang Chinese Medical University Zhejiang Provincial Key Laboratory of Gastrointestinal Diseases Pathophysiology Hangzhou CN
[2]Secretariat The China Crohn's & Colitis Foundation Hangzhou CN
[3]Center for Inflammatory Bowel Diseases, Department of Gastroenterology The Second Affiliated Hospital, Zhejiang University School of Medicine Hangzhou CN
[4]The Clinical Medical College Zhejiang University School of Medicine Hangzhou CN
[5]Department of Gastroenterology The Affiliated Huaian No.1 People's Hospital of Nanjing Medical University Huai'an CN
[6]The Second Clinical Medical College Zhejiang Chinese Medical University Hangzhou CN
[7]Graduate Institute of Digital Learning and Education National Taiwan University of Science and Technology Taipei TW
[8]Department of Computer Science and Information Engineering National Kaohsiung University of Science and Technology Kaohsiung TW
[*]these authors contributed equally

Corresponding Author:
Yan Chen md, phd
Center for Inflammatory Bowel Diseases, Department of Gastroenterology
The Second Affiliated Hospital, Zhejiang University School of Medicine
88 Jiefang Road,Shangcheng District.
Hangzhou
CN

Abstract

Background: Although Large Language Models (LLMs) such as ChatGPT show promise in providing specialized information, their quality require further evaluation, especially considering that these models are trained on internet text and the quality of health-related information available online varies widely.

Objective: The aim of this study was to evaluate the performance of ChatGPT in the context of patient education for individuals with chronic diseases, comparing it with that of industry experts to elucidate its strengths and limitations.

Methods: This evaluation was conducted by analyzing the responses of ChatGPT and specialist doctors to questions posed by patients with Inflammatory Bowel Disease (IBD), comparing their performance in terms of subjective accuracy, empathy, completeness, and overall quality, as well as readability to support objective analysis.

Results: In a series of 1578 binary choice assessments, ChatGPT was preferred in 48.4% (95% CI, 45.9%-50.9%) of instances. There were 12 instances where ChatGPT's responses were unanimously preferred by all evaluators, compared to 17 instances for specialist doctors. In terms of overall quality, there was no significant difference between the responses of ChatGPT (3.98; 95% CI, 3.93-4.02) and those of specialist doctors (3.95; 95% CI, 3.90-4.00) (t=0.95, p=0.34), both being considered "good". Although differences in accuracy (t=0.48, p=0.63) and empathy (t=2.19, p=0.03) lacked statistical significance, the completeness of textual output (t=9.27, p=0.00) was a distinct advantage of the Large Language Model (ChatGPT). In the sections of the question where patients and doctors responded together (Q223-Q242), ChatGPT demonstrated superior performance (p=0.006). Regarding readability, no statistical difference was found between the responses of specialist doctors (median: 7th grade, Q1: 4th grade; Q3: 8th grade) and those of ChatGPT (median: 7th grade, Q1: 7th grade; Q3: 8th grade) according to the Mann-Whitney U test (p=0.09). The overall quality of ChatGPT's output exhibited strong correlations with other sub-dimensions (with empathy: r=0.842; with accuracy: r=0.839; with completeness: r=0.795), and there was also a high correlation between the sub-dimensions of accuracy and completeness (r=0.762).

Conclusions: ChatGPT demonstrated more stable performance across various dimensions. Its output of health information content is more structurally sound, addressing the issue of variability in individual specialist doctors' information output. ChatGPT's performance highlights its potential as an auxiliary tool for health information, despite limitations such as AI

hallucinations. It is recommended that patients be involved in the creation and evaluation of health information to enhance the quality and relevance of the information.

(JMIR Preprints 03/06/2024:62857)
DOI: https://doi.org/10.2196/preprints.62857

Preprint Settings

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

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

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

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

Original Manuscript

Large Language Models (LLMs) vs. Specialist Doctors: A Comparative Study on Health Information in specific medical domains.

Zelin Yan[1,2,8*], MS; Jingwen Liu[2*], PhD; Shiyuan Lu[2], PhD; Dingting Xu[2], MD, PhD; Yun Yang[3], BS; Honggang Wang[4], MD, PhD; Jie Mao[5], MS; Hou-Chiang Tseng[6], PhD; Tao-Hsing Chang[7], PhD; **Yihong Fan[1], MD, PhD; Yan Chen[2,8], MD, PhD.**

[1]Department of Gastroenterology, The First Affiliated Hospital of Zhejiang Chinese Medical University, Zhejiang Provincial Key Laboratory of Gastrointestinal Diseases Pathophysiology, Hangzhou, China.

[2]Center for Inflammatory Bowel Diseases, Department of Gastroenterology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China.

[3]The Clinical Medical College, Zhejiang University School of Medicine, Hangzhou, China.

[4]Department of Gastroenterology, The Affiliated Huai'an No.1 People's Hospital of Nanjing Medical University, Huai'an, China.

[5]The Second Clinical Medical College, Zhejiang Chinese Medical University, Hangzhou, China.

[6]Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taipei, Taiwan.

[7]Department of Computer Science and Information Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan.

[8]The China Crohn's and Colitis Foundation, Hangzhou, China

*: These authors should be considered joint first author.

Correspondence to: **Yan Chen,** MD & PhD, Professor. Department of Gastroenterology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, No 88, Jiefang Road, Hangzhou, 310009, China. Tel: +86-571-87783777, Fax: +86-571-87783936, E-mail: chenyan72_72@zju.edu.cn

Co-corresponding: Yihong Fan, MD & PhD, Professor. Email: yhfsjr@163.com

ABSTRACT

Background: Although Large Language Models (LLMs) such as ChatGPT show promise in providing specialized information, their quality require further evaluation, especially considering that these models are trained on internet text and the quality of health-related information available online varies widely.

Objective: The aim of this study was to evaluate the performance of ChatGPT in the context of patient education for individuals with chronic diseases, comparing it with that of industry experts to elucidate its strengths and limitations.

Methods: This evaluation was conducted by analyzing the responses of ChatGPT and specialist doctors to questions posed by patients with Inflammatory Bowel Disease (IBD), comparing their performance in terms of subjective accuracy, empathy, completeness, and overall quality, as well as readability to support objective analysis.

Results: In a series of 1578 binary choice assessments, ChatGPT was preferred in 48.4% (95% CI, 45.9%-50.9%) of instances. There were 12 instances where ChatGPT's responses were unanimously preferred by all evaluators, compared to 17 instances for specialist doctors. In terms of overall quality, there was no significant difference between the responses of ChatGPT (3.98; 95% CI, 3.93-4.02) and those of specialist doctors (3.95; 95% CI, 3.90-4.00) (t=0.95, p=0.34), both being considered "good". Although differences in accuracy (t=0.48, p=0.63) and empathy (t=2.19, p=0.03) lacked statistical significance, the completeness of textual output (t=9.27, p=0.00) was a distinct advantage of the Large Language Model (ChatGPT). In the sections of the question where patients and doctors responded together (Q223-Q242), ChatGPT demonstrated superior performance (p=0.006). Regarding readability, no statistical difference was found between the responses of specialist doctors (median: 7th grade, Q1: 4th grade; Q3: 8th grade) and those of ChatGPT (median: 7th grade, Q1: 7th grade; Q3: 8th grade) according to the Mann-Whitney U test (p=0.09). The overall quality of ChatGPT's output exhibited strong correlations with other sub-dimensions (with empathy: r=0.842; with accuracy: r=0.839; with completeness: r=0.795), and there was also a high correlation between the sub-dimensions of accuracy and completeness (r=0.762).

Conclusion: ChatGPT demonstrated more stable performance across various dimensions. Its output of health

information content is more structurally sound, addressing the issue of variability in individual specialist doctors' information output. ChatGPT's performance highlights its potential as an auxiliary tool for health information, despite limitations such as AI hallucinations. It is recommended that patients be involved in the creation and evaluation of health information to enhance the quality and relevance of the information.

BACKROUND

In the medical field, Large Language Models (LLMs), represented by ChatGPT, have shown significant application potential: not only passing medical licensing exams in the United States and China but also demonstrating their value in interpreting medical records, assisting in medical decision-making, and improving patient follow-up compliance.[1-6] LLMs have showcased their "rich medical knowledge" and the ability to extract disease information from various languages and contexts. Their method of providing information in a "human-like" tone is considered more effective than traditional search engines.[7] Despite a lack of evidence, these tools are being adopted by patients and clinical doctors.[6,8] The reason behind their excellent performance is that their text training set comes from a vast amount of publicly available internet information, making the quality of medical information provided by LLMs comparable to existing internet information.[9,10]

However, the quality of health information on the internet about chronic diseases, such as Inflammatory Bowel Disease (IBD), varies widely and has been considered a challenge for patient self-management and education.[11-17] Inflammatory Bowel Disease, including Crohn's disease and ulcerative colitis, is an increasingly prevalent chronic intestinal disease in China, characterized by primary invasion of the digestive system and cumulative multi-system involvement of autoimmune diseases, with no cure currently available. Patients have a strong need to learn and reinforce self-care abilities, among which the WeChat public account of the China Crohn’s and Colitis Foundation (CCCF) is the most popular with IBD patients.[18] We use it as a representative to study the patient education ecosystem for chronic diseases.

This study aims to evaluate ChatGPT's ability to provide specialized vertical domain information, especially in the education of some chronic disease patients, and compare it with industry experts. Through this comparison, we can identify the strengths and limitations of LLMs in medical information services, providing a basis for further improvement and application of the technology. Additionally, this study also aims to enhance the public and medical professionals' awareness and acceptance of using artificial intelligence tools in medical information acquisition and education.

METHODS

Question Collection

The mode of one-on-one question-and-answer dialogue stands as a prevalent form of interaction within the healthcare domain. Across various medical applications, online forums, and instant messaging groups, a substantial portion of queries manifests as repetitive and amenable to categorization.

In earlier epochs, we undertook the aggregation of high-frequency, prototypical questions posed by patients afflicted with Inflammatory Bowel Disease (IBD) through online platforms and outpatient settings. We extended invitations to industry peers to collectively address these patient queries, culminating in the publication of a didactic tome titled "Q&A on Ulcerative Colitis and Crohn's Disease" tailored for the self-learning of individuals grappling with inflammatory bowel ailments. This publication reflects the cumulative outcomes of doctor-patient interactions over seven years at the CCCF and the Second Affiliated Hospital- Zhejiang University School of Medicine (SAHZU) IBD Center encompassing 9,000 cases. The compendium was predominantly curated by eight seasoned IBD specialist doctors, with contributions from 55 IBD practitioners and five experienced patients with a high level of cultural acumen. Upon its publication, the book garnered commendations and accolades from numerous esteemed figures within the IBD community in China and the United States. The content delves into various aspects of IBD, including etiology, symptoms, diagnosis, treatment, follow-up protocols, and emotional support. The questions encapsulated within are highly representative and encompass a broad spectrum (supplementary 1). Many analogous questions have surfaced on pertinent social media platforms, with the content of this tome serving as a primary source of representative patient inquiries. Presently, the book has undergone nine printings, with a distribution nearing 20,000 copies.

The thematic essence of the book comprises 263 distinct questions matched with corresponding responses from medical professionals. Apart from a minor subset of emotional support content provided by patients, all responses are underpinned by evidence-based rationales. This sample size is anticipated to afford us a statistical power of

90% to discern a 10% differential between responses generated by ChatGPT and those proffered by medical practitioners (55% vs 45%) (Figure 1).

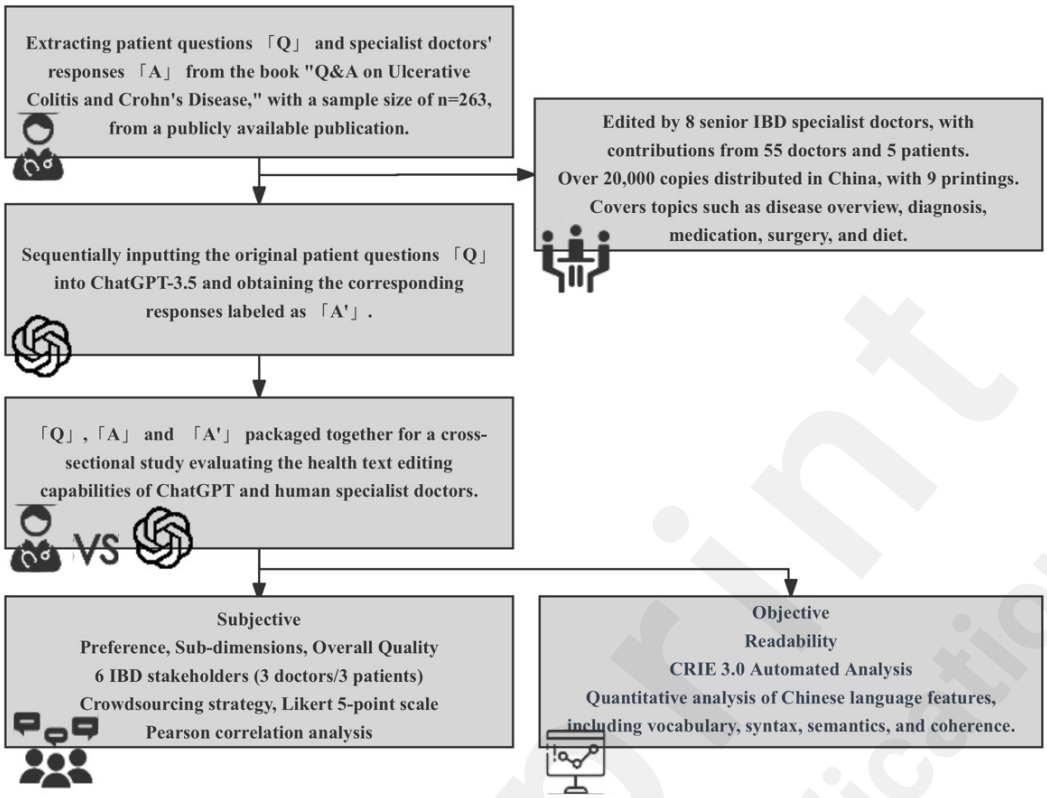

Figure 1: Schematic of Operational Workflow

To enhance reproducibility, we have not utilized anonymized online medical consultation text for doctor-patient interactions. The content in the book is derived from authentic doctor-patient interactions, pre-authorized and publicly disclosed, with many excerpts being republished on social media platforms in electronic format.[19] The data used in this study are publicly available and do not contain any identifiable personal information. Despite the data being publicly accessible, we still submitted our research plan to the ethics committee for evaluation to determine whether a formal ethical review was necessary or if it could be exempted. Based on the assessment and guidance from the ethics committee, given the nature and use of the research materials, it was determined that this study did not involve direct research on human participants, thus deeming informed consent unnecessary. The content is used under authorization and license from Zhejiang University Press.

Collecting ChatGPT Responses

In the period from September 8 to 22, 2023, ChatGPT responses were collected by inputting the original question texts into a new chatbot session (GPT-3.5 version, Open AI, August 3 version, 2023) and saving the chatbot replies (https://chat.openai.com/share/4be965cf-a024-4f8e-a438-a9064a1e0657□,Differing from some other experimental methodologies, we adopted a sequential prompting of all questions listed in the directory within the same bot link.[4,20] The rationale behind this approach includes:

1.The original questions in the book contained terminology descriptions presumed to be familiar and comprehensible to healthcare professionals; for instance, in Chinese, the term□□□□ "hormone" in the book and in IBD doctor-patient communication scenarios often specifically refers to □□□□□□□"glucocorticoids." In a typical context, using □□□□"hormone" in communication may commonly lead individuals to think of "chemical messengers between cells," and according to LLMs principles, ChatGPT would respond to the latter in the absence of contextual elucidation. This measure was taken to mitigate the potential for the bot to provide accurate "incorrect responses" due to a lack of contextual background.

2.Upon encountering the instances of the bot misinterpreting language contexts, we continued to supplement vocabulary prompts to guide ChatGPT in understanding the true intent of the questioner, thereby eliciting a response that aligns with it. However, prompts were limited to no more than three times, drawing from the routine search habits of patients on the web and previous experiments.[4,21]

3.To emulate the habitual reading practices and contexts of normal situations, we posed questions to ChatGPT in the same sequential order as presented in the book.

Quality Control

1. The final analytic sample encompassed 263 questions and their corresponding responses from doctors and ChatGPT, as featured in the ninth edition of "Q&A on Ulcerative Colitis and Crohn's Disease " printed in April 2022. Responses from doctors were designated as the benchmark.
2. Some original responses in the book were provided by patients, and we retained this text as it had undergone professional medical review before the book's publication. It can be understood that while drafted by patient volunteers, the expressions were approved by doctors and deemed suitable for new patients to view, primarily addressing psychological issues (supplementary2).
3. Original illustrations from the book were not excerpted, whereas tables were permitted. This decision was made because when using ChatGPT, the model itself could generate tables, thus remaining unaffected.
4. At that time, the version of ChatGPT would randomly present two response options for user selection when prompted, with the first option being the default choice.
5. Due to network issues, in the event of a crash or incomplete display, we would click "regenerate" once to select a complete text answer for material completion.

Text Content Evaluation

Subjective Assessment

Assessment was conducted by six evaluators (three licensed IBD doctors and three IBD patients). The doctors were experienced IBD physicians in patient education (YC, DTX, HGW): with over 10 years of clinical experience, having treated more than 500 IBD patients, and engaged in patient education for over 5 years. The patient characteristics required (as stated in the acknowledgments) were individuals aged between 20 and 60, with at least an undergraduate education level, diagnosed with IBD for more than a year, and who had not read the "Questions and Answers" book. To ensure evaluators were as unable as possible to distinguish the source of the text, we employed a blind method when presenting the materials to evaluators, concealing explanatory language such as "as an artificial intelligence," etc. The doctor responses and ChatGPT responses for the same question were anonymized and randomly labeled as Response 1 and Response 2. Evaluators were required to first read the question along with the corresponding doctor response and ChatGPT response, followed by a two-step evaluation process. 1. Selecting the preferred answer version. 2. Subjectively rating the two answers on a Likert 5-point scale for overall quality and dimensional evaluation, referencing dimensions from previous health information research[14,15,20,21]: including accuracy, empathy, and completeness. A higher score indicates greater evaluator approval of the response text's performance in that dimension. (Table 1 for details)

Dimension	Definition
Accuracy	Whether the response scientifically and impartially explains the issue, such as providing explanations on medication usage and dosage, and clarifying surgical timing limitations.
Completeness	Whether there are any omissions of important information or concepts in the explanation.
Empathy	Whether the response demonstrates an understanding of the question from the perspective of the "patient" or the inquirer.
Overall Quality	Subjective perception of the overall quality of the text.

Table 1. Definitions of Each Dimension, Pre-training Required for Evaluators.

Objective Evaluation

Simplified Chinese Readability Analysis

The Chinese Readability Index Explorer (CRIE, version 3.0, http://www.chine-sereadability.net/CRIE) was utilized. In addition to evaluators' subjective assessments, we introduced a quantitative Chinese readability tool,

CRIE. It consists of four subsystems, comprising 82 multi-level language features.[22] CRIE employs multi-level language features for text analysis, including vocabulary, syntax, semantics, and cohesion. This tool aids in analyzing various types of texts, such as Chinese textbooks,[23] foreign language learning materials,[24] and domain-specific knowledge texts.[25] Numerous studies have validated its reliability and practicality in the Chinese health domain.[26,27] Results can be interpreted using the Flesch-Kincaid English readability assessment method: the higher the grade, the greater the text complexity. Quantitative natural language processing and text mining tools serve as valuable supplements to subjective human evaluations.[4]

Data Statistics and Analysis

Aligned with the research objectives, we employed a crowdsourced scoring strategy for data collection, a method that aggregates data across a collective of evaluators. Primarily applied in the field of linguistics, where language usage is a fundamental domain for the general populace, the central idea is to harness the collective expertise of both experts and the public to pioneer new concepts through crowd annotations. This method is well-suited for subjective evaluations, such as scoring by singing judges or the exploration of novel concepts. Calculating average scores for each dimension reflects the consistency variances among evaluators, encapsulating individual uncertainties and subjective biases within the variance of the scores.[28] In the context of health text evaluation, the involvement of judges and the assessment method, involving direct quantification by both IBD healthcare providers and consumers, represents a feasible, efficient, cost-effective, and relatively accessible evaluation strategy.

Primary outcomes: We conducted descriptive analysis and assessed evaluators' preference ratios for ChatGPT using a chi-square goodness-of-fit test. A two-tailed Welch's t-test was utilized to compare the mean values of the two responses. We defined a threshold score of 3 (acceptable) and calculated the proportion exceeding or falling below this threshold score (3), comparing them using prevalence ratios (PR). Furthermore, we evaluated the Pearson correlation coefficients between the various sub-dimensions of quality to observe or predict correlations between different dimensions. Given that the readability of each response text is a calculated ordinal variable, non-parametric tests were employed for comparison.

Secondary outcomes: Subgroup t-test analyses were conducted to assess the impact of evaluator identity (medical professional/patient) and the original response creator's source (solely medical professional/healthcare provider-patient collaboration) on mean scores.

A significance level of $P < 0.05$ was set, and Bonferroni correction was applied for multiple tests. All statistical analyses were performed using R software (version 4.3.1 GUI 1.79 Big Sur ARM build) and RStudio (version 2023.09.1+494). Data visualization was created based on code references from the open-source platform Hiplot.

RESULTS

Preferred Response Ratio

Out of 1578 evaluations, evaluators showed a preference for ChatGPT responses at a rate of 48.4% (95% CI, 45.9%-50.9%, p=0.00). Among these, 6 evaluators exclusively favored ChatGPT responses for a total of 12 questions: 22, 28, 42, 56, 73, 120, 121, 124, 127, 161, 174, 195. Evaluators exclusively favored doctor responses for 17 questions: 5, 27, 41, 83, 85, 92, 156, 175, 180, 198, 205, 210, 219, 234, 237, 251, 263. The questions and corresponding responses from both doctors and the AI model are detailed in (supplementary3).

Comparison of Mean Scores and Prevalence Ratios of Threshold Scores

Overall, the proportion of responses rated below acceptable quality (<3) was 1.26 times higher for doctors responses compared to ChatGPT responses (doctors: 3.3%; 95% CI, 2.5%-4.4%; ChatGPT: 2.7%; 95% CI, 1.9%-3.5%).Simultaneously, the proportion of responses rated as good or very good quality was 1.10 times higher for ChatGPT compared to doctors (doctors: 69.4%; 95% CI, 67.1%-71.7%; ChatGPT: 76.0%; 95% CI, 73.7%-78.0%). While ChatGPT had a slight advantage in the overall quality distribution, there was no significant difference between ChatGPT and doctor responses (t=0.95, p=0.34), with doctors (3.95; 95% CI, 3.90-4.00) and ChatGPT responses (3.98; 95% CI, 3.93-4.02) both rated at a "good" level.

In terms of the completeness dimension, ChatGPT responses significantly outperformed doctor responses (t=9.27, p=0.00), although both doctor responses (3.88; 95% CI, 3.83-3.94) and ChatGPT responses (4.21; 95% CI, 4.17-4.26) were rated at a "good" level. The proportion of low completeness responses was 3.66 times higher for doctor responses (6.7%; 95%, 5.53%-8.97%) compared to ChatGPT responses (1.8%; 95% CI, 1.2%-2.6%); the proportion of high completeness responses was 1.20 times higher for ChatGPT responses (81.1%; 95% CI, 79.2%-83.1%) compared to doctor responses (67.4%; 95% CI, 65.1%-69.7%). Pre-trained models and structured outputs contributed to ChatGPT receiving more favor in this dimension.

In the empathy dimension, ChatGPT responses were inferior to doctor responses (t=2.19, p=0.03). Due to the significance correction for multiple tests, we conservatively state that there is no significant difference between doctor responses (3.99; 95% CI, 3.95-4.03) and ChatGPT responses (4.06; 95%, 4.01-4.11). The proportion of low empathy responses was 1.12 times higher for ChatGPT responses (2.8%; 95% CI, 2.0%-3.7%) compared to doctor responses (2.5%; 95%, 1.8%-3.4%), while the proportion of high empathy responses was 1.04 times higher for ChatGPT responses (75.0%; 95% CI, 72.8%-77.1%) compared to doctor responses (72.0%; 95% CI, 69.6%-74.1%).

In the accuracy dimension, there was no significant difference between ChatGPT and doctor responses (t=0.48, p=0.63). Doctor responses (4.11; 95% CI, 4.07-4.15) and ChatGPT responses (4.12; 95% CI, 4.08-4.17) were comparable. The proportion of low accuracy responses was 2.2 times higher for ChatGPT responses (2.4%; 95% CI, 1.7%-3.3%) compared to doctor responses (1.1%; 95% CI, 0.6%-1.7%), while the proportion of high accuracy responses was 1.05 times higher for ChatGPT responses (81.1%; 95% CI, 79.0%-83.0%) compared to doctor responses (76.9%; 95% CI, 74.7%-78.9%).

Subgroup Comparisons

Discrepancies in Evaluator Perspectives (Figure 2)

In terms of overall quality, physicians (p=0.09) and patients (p=0.88) perceived no difference between ChatGPT and doctors. Regarding completeness, physicians (p=0.00) and patients (p=0.00) unanimously agreed that ChatGPT outperformed doctors. On the empathy dimension, while physicians (p=0.70) did not perceive a difference between the two, patients believed that doctors' responses exhibited more emotional depth compared to ChatGPT (p=0.00). In terms of accuracy, physicians (p=0.02) considered doctors’ responses to be more accurate, while patients believed that ChatGPT responses held a slight edge in accuracy (p=0.00).

Differences in Responders' Performance (Figure 3)

We selected original questions (Q223-Q242) where patients assisted doctors in crafting responses and found that in this subset, human performance significantly surpassed that of the ChatGPT AI model (p=0.001).in terms of overall quality and across various dimensions (Figure 4), ChatGPT's ability to curate health information in the specialized field of medicine is now on par with professional doctors, reaching a level of excellence.

Subdimension Correlation Analysis

Using Pearson's test, a correlation analysis was conducted on the scores of ChatGPT responses across different text dimensions, revealing strong correlations between overall quality and other subdimensions (with empathy: r=0.842; with accuracy: r=0.839; with completeness: r=0.795). Additionally, there was a high correlation between accuracy and completeness among subdimensions (r=0.762). Similar patterns were observed in text responses from doctors, where overall quality exhibited correlations with completeness: r=0.857; with empathy: 0.849; with accuracy: 0.828, and a correlation between accuracy and completeness: r=0.785. (Figure 5)

Readability Analysis

A Mann-Whitney U test on the readability of responses from doctors (median: 7th grade, Q1: 4th grade; Q3: 8th grade) and ChatGPT (median: 7th grade, Q1: 7th grade; Q3: 8th grade) revealed no significant difference (p=0.09). (Figure 6)

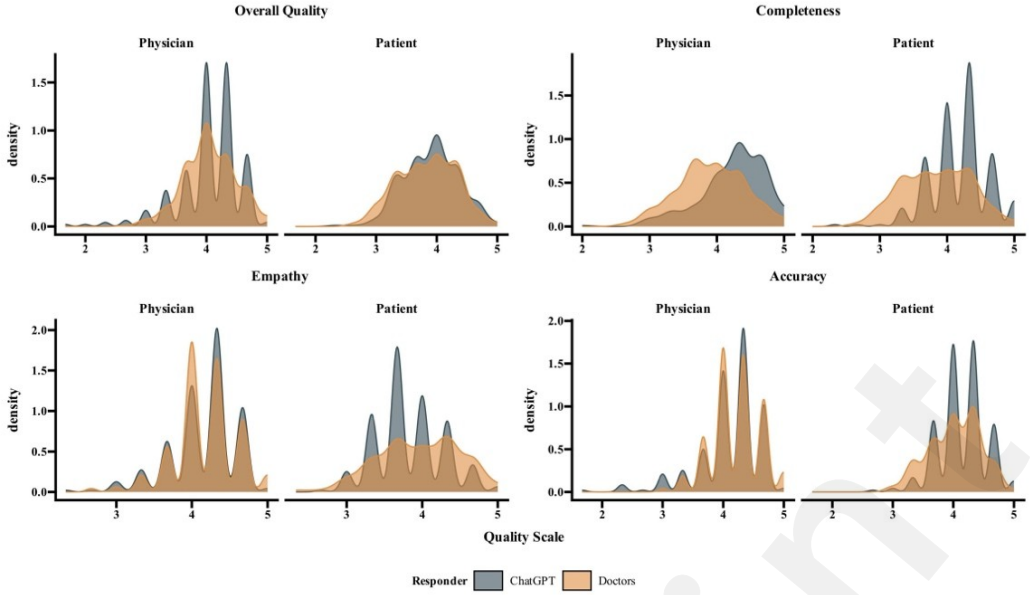

Figure 2. Kernel Density Plot illustrating quality assessment based on two identities of evaluators.

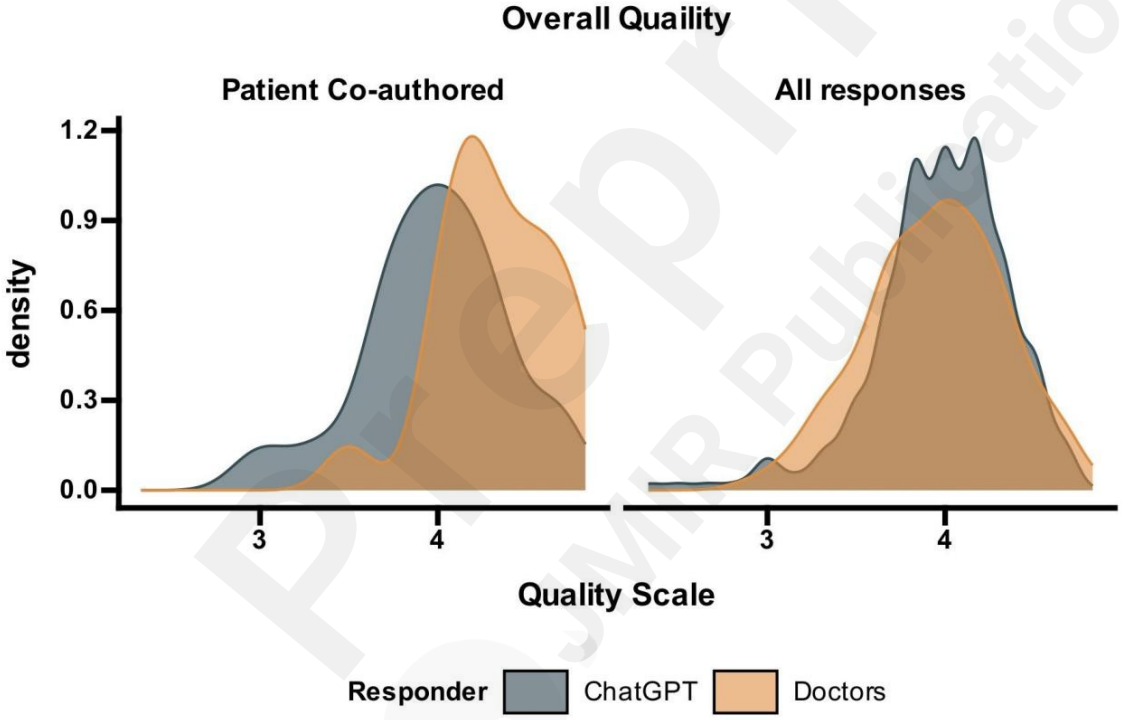

Figure 3. Kernel Density Plot. In the original responses, there was no significant difference between the responses of general doctors and ChatGPT (p=0.34). However, for questions assisted by patients (Q223-Q242), the performance was significantly better than ChatGPT (p=0.006).

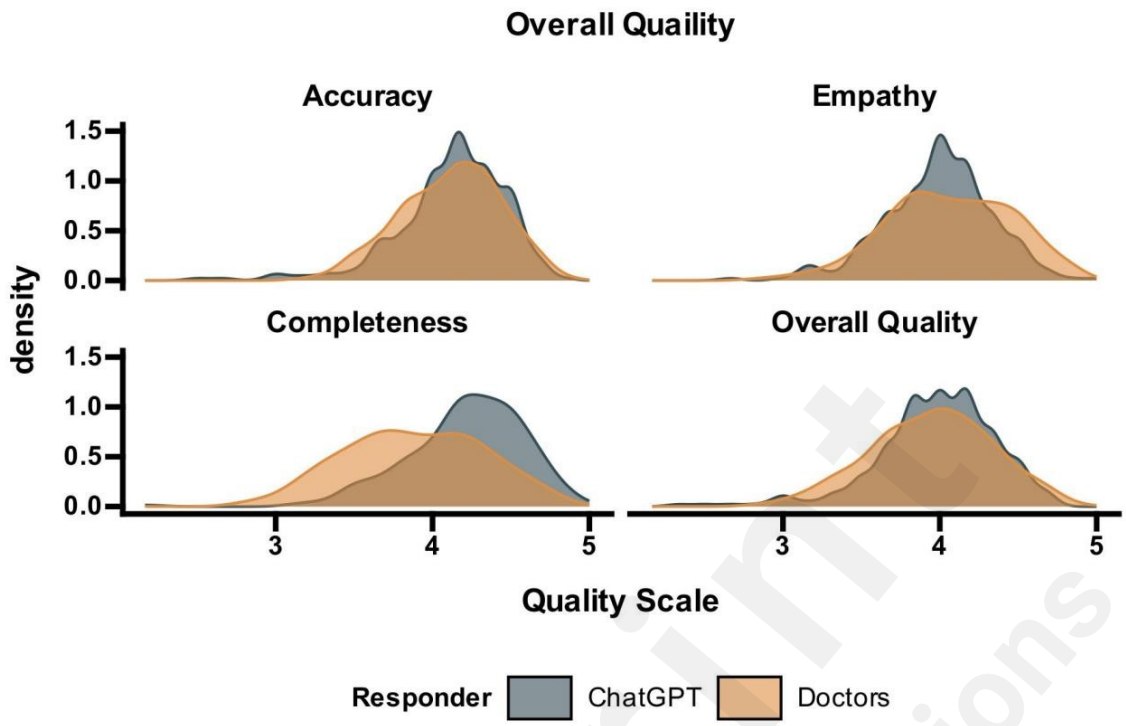

Figure 4. The kernel density plot illustrates the distribution of assessors' evaluations on overall quality and sub-dimensions, highlighting ChatGPT's significantly superior capability in completeness compared to doctors.

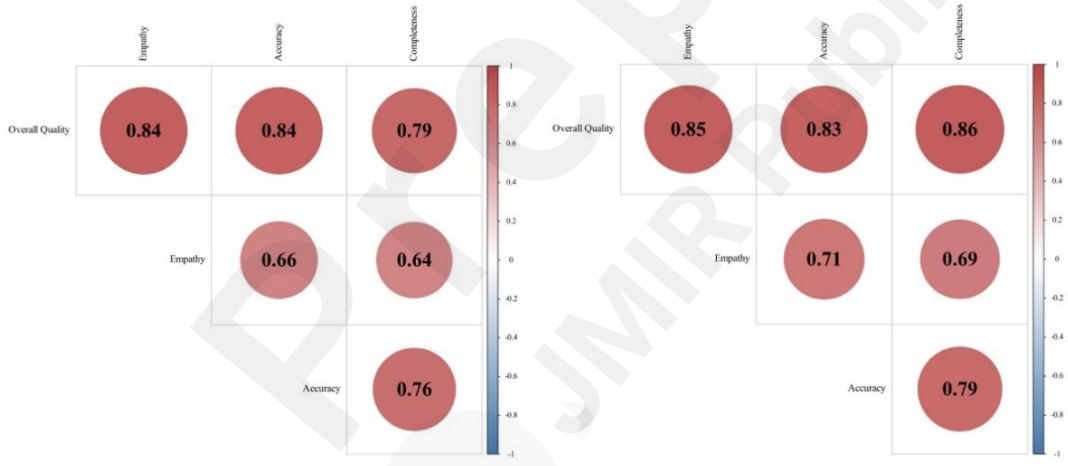

Figure 5. A. Responses from ChatGPT, B. Responses from doctors, indicating a high degree of correlation between overall quality and various dimensions. The correlation between empathy and accuracy, completeness falls within a moderate range (0.3-0.70).

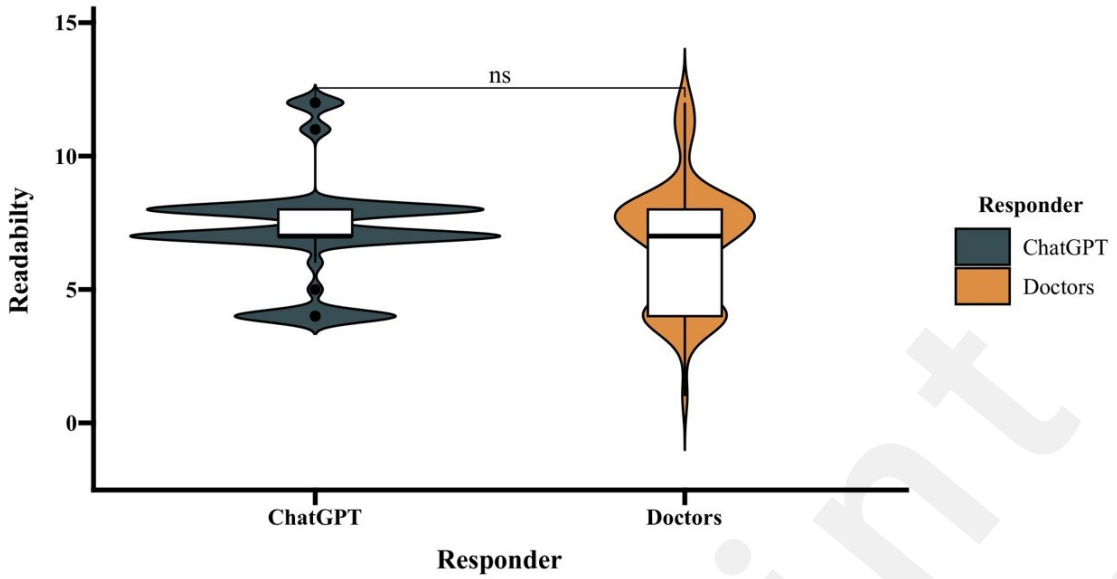

Figure 6. An accordion plot illustrating significant fluctuations in the text editing abilities of original doctor responses (median: 7th grade, Q1: 4th grade, Q3: 8th grade). ChatGPT, as an AI, demonstrates greater stability in its output compared to humans, with a narrower distribution of readability in the generated text.

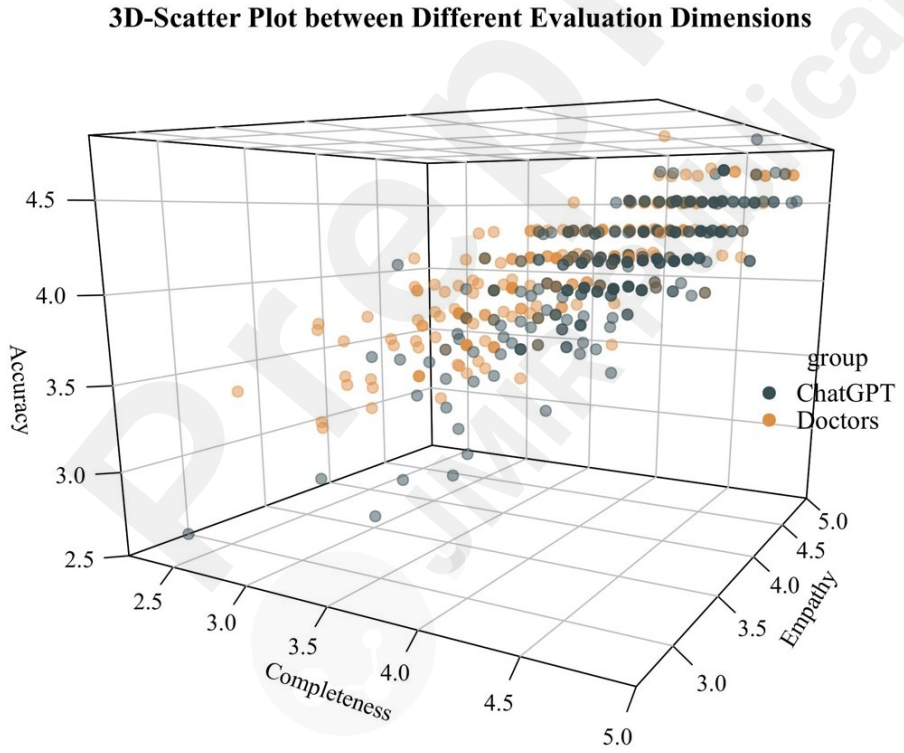

Figure 7. The 3D scatter plot intuitively reflects that 1. On the left side of the plot, ChatGPT exhibited a few instances of extremely low responses; 2. ChatGPT responses are densely distributed on the high completeness dimension, indicating ChatGPT's enhanced ability to provide comprehensive information.

DISCUSSIONS

Stable and Comprehensive: ChatGPT's Health Information Output Capability.

High-quality health information contributes to favorable medical outcomes, especially for patients with chronic

conditions.[12] Conversely, erroneous, incomplete, and unregulated information may mislead patients into making detrimental choices.[29] Exploring the application of Large Language Models (LLMs) in health information retrieval and chronic disease patient education holds significant practical relevance.

Although the selection ratio did not exceed that of doctors to a statistically significant level, it is important to note that the latter are highly specialized professionals in the field of Chinese IBD and are endorsed by the expert association (Chinese IBD board) that follows them.

However, its responses are more comprehensive. From a selection of 12 fully optimized ChatGPT responses (partially referenced in Supplementary 3), the advantages of ChatGPT are evident. Moreover, in the Likert-5 scale evaluation, it significantly outperformed human experts. This stems from GPT's structured approach based on pre-set models, featuring a brief introductory paragraph, followed by a list of answers with bullet points or numbering, and a standard concluding paragraph. In terms of accuracy and empathy, there were no significant differences between the two in comparison to completeness. Previous studies have found that GPT provides appropriate and easily understandable answers to questions regarding diagnosis and treatment choices but falls short when it comes to explaining diagnostic tests and recommending complex management strategies.[7] While its responses are structurally sound, they often lack critical insights into decision thresholds and treatment timing.[20] Our IBD specialist physician (YC) provided a sharp critique: its answers are superficial and lack sensitivity and understanding of medication efficacy and monitoring timeframes. Interestingly, despite our blind randomized process, two evaluators admitted towards the end of the experiment that they could discern distinctly different styles between the two groups. Our study does not seem to support ChatGPT's tendency to offer technically correct but insufficient textual conclusions.[20]

From various visual distribution charts (Figures 2, 3, 4, 6), it is evident that ChatGPT exhibits lower score variance, indicating a more stable performance in this dimension. If you have ever coordinated a large group of people, compiling group publications and information without being able to control the format, you would deeply appreciate the commendable ability of LLMs in this regard. With an equal amount of learning material provided, machines produce more consistent outputs compared to humans.

How to make others understand (whether they are LLMs or patients)

In its official description, ChatGPT is merely a language model, pre-trained for general cognitive tasks.[10] Its performance may decrease when faced with tasks that require specialized and highly professional skills. Subsequent strategies include:1. Utilizing secondary LLMs tailored for various professional scenarios, such as Med-PalM;[30] 2. Application of prompt engineering.

Prompt engineering, a concept that combines artistry and science,[10] led to a sudden realization in our experimental design, respecting the working principles of ChatGPT. LLMs are based on large-scale learning, reflecting the collective knowledge level of most learnable materials, implying that its understanding of information is based on widely applicable domains. However, our questions were based on the IBD community, sourcing information from a vertically specialized field. Specific contexts give rise to “slang” and a plethora of "terminology." Broadly speaking, ChatGPT interpreting □□□□"hormones" as "chemical messengers between cells" is the most accurate, as outside of clinical contexts, such abbreviations are rarely used. When faced with unsatisfactory responses or doubts about AI Hallucination, consider first whether the prompts (terminology) you provide as the output party have been broken down for laypersons to understand.In previous evaluations of English and Chinese IBD information,[14,15,31] the readability levels of web health information were generally too high, unsuitable for public dissemination.[13]Simultaneously, there are also articles indicating that the English output generated by ChatGPT is at a university level.[9] This aligns with the common complaint heard by the author in work settings from patients: doctors chatter on, but I can't understand a word they're saying. Explaining one term with another is not a joke but a satirical reality.

In English health information research, readability analysis is commonplace. Common assessment tools include the Flesch-Reading Ease score, Flesch-Kincaid Grade Level.[14,15,21,32] However, the analysis and application of Chinese readability are still in their infancy. We hope to see more industry experts, not just medical professionals, engage in such research, valuing information, and its underlying power. The popularity of " Q&A on Ulcerative Colitis and Crohn's Disease " in the IBD community remains inexplicable, but we speculate that its readability matches the general educational levels of the Chinese population and the recommended grade level for popular science publications.[33] Additionally, it is pleasantly surprising that the Chinese readability of ChatGPT's response information is also very good, showing no significant difference from the level of professional doctors and exhibiting greater stability (narrower kernel density variance).

In the cross-sectional comparison of subdimensions, we observed a strong correlation between the overall quality

of health information and completeness, accuracy, and empathy. Furthermore, there is a high predictive function between completeness and accuracy, as depicted in Figures 5 and 6, with a "more words, more reason" phenomenon. This same trend is confirmed in sensitivity analysis.[28] While each aspect can enhance the persuasiveness of textual information, empathy as an emotional dimension is not strongly correlated with rational dimensions such as accuracy and completeness.

Disparities in Cognitive Understanding Between Patients and Healthcare Providers.

In the overall comprehension of quality and completeness, it is evident that there is no disagreement among assessors in the roles of healthcare providers and patients. Both parties unanimously consider ChatGPT and medical experts to perform similarly, with the former providing more comprehensive information. However, upon conducting subgroup analysis, we discovered that healthcare providers have a delayed grasp on empathy and are more sensitive to accuracy. Healthcare providers can discern more accurately sourced information from their peers, while patients may not. These disparities lay the foundation for the communication conflicts between healthcare providers and patients in real-life scenarios. Patients may not perceive ChatGPT's information to be more erroneous than that of medical professionals, possibly due to their lack of professional knowledge to comprehend the underlying facts. This mirrors the headache-inducing situation for healthcare providers when patients prefer to believe exaggerated television advertisements for health products rather than opting for industry-reviewed experts and guidelines. This serves as a reminder that medical and health information must be developed and tested with patients (consumers) at the center.[34]

AI Hallucination

Errors in responses from LLMs are referred to as "AI hallucination," and chatbots typically present themselves in a convincing manner, leading the inquirer to potentially believe in their authenticity. [6,10]We believe this is also a key reason why patient assessors cannot differentiate between the accuracy of ChatGPT and medical experts.

Despite emphasizing the importance of prompt engineering, we are still amazed by ChatGPT's ability to identify spelling errors, ambiguities, and highly condensed issues, based on our experimental responses structured as progressive inquiries following textbook content. As feedback, ChatGPT even comprehends outdated drug translations (e.g., the new official translation□□□□□□□ for "infliximab" and the old term □□□□□□□□.It also gave us a few "AI hallucinations"(correspondingly, numerous poorly performing outliers are evident in Figure 7), where commonly used drug names in clinical practice are interpreted as the scientific names of mosquitoes and, when questioned further, ChatGPT refuses to acknowledge the error (supplementary 4). We attribute the causes of these AI delusions to a lack of background knowledge and insufficient prompts.

In the realm of medicine, a discipline that relentlessly pursues zero errors as a necessity of natural science and ethics, allowing AI to engage in self-expression is inappropriate.Therefore, we also agree that it is imperative for professionals to verify the output of ChatGPT,[10,35,36] despite our observations indicating that it often performs at a level comparable to that of experts.

Evaluating the achievements of AI should first be based on how humans assess their own accomplishments.

While ChatGPT may provide outdated or incorrect information, the level at which an LLM operates is a key consideration. Care should be taken when comparing ChatGPT to various experts/professional guidelines. Additionally, we must also consider whether our human experts are capable of effectively dissecting and conveying complex, obscure, and uncommon terms and concepts to laypersons.[37] Criticisms and warnings about LLMs are prevalent, reminding us of the need to contemplate the baseline definition of medical practice. Questions arise as to whether outputs need to strictly adhere to guidelines and if the discrepancies in guidelines among different countries, regions, and medical associations have been fully addressed. If not, evaluating LLMs or AI outputs will always involve subjective differences.

Based on our findings, we cautiously endorse the view that ChatGPT has the potential to improve patients' access to disease information in healthcare settings.[7,20] Its performance may be even better when assessed by non-specialist doctors or young medical students. If we were to compare humans to AI in the context of online community doctors, we speculate that the positive outcomes would be significantly pronounced!

Next Steps in Exploration.

To our knowledge, this study represents the first invitation for IBD health information consumers and providers to participate in a crowdsourced evaluation. It is also an exploration of the readability of Simplified Chinese characters in the context of IBD.

Introducing tools like ChatGPT in a timely manner into patient communities and basic patient education settings seems feasible: initiating the use of ChatGPT to draft medical information for healthcare providers (health self-media practitioners, healthcare professionals, medical institution promoters), followed by expert review and refinement, appears to be a viable and convenient production pathway. Undoubtedly, the quality of ChatGPT's responses will gradually improve with version updates and over time, making the tool even more promising.[38]

While not the primary hypothesis of our experiment, we also observed variations in text quality between different disease types generated by ChatGPT.[35] Furthermore, ChatGPT has an overwhelming advantage over human experts in terms of speed of content creation. Many participating doctors acknowledge that crafting understandable content for patients in health education efforts requires significant dedication and effort.[39]

Previously, on social media platforms and in online medical consultation scenarios, ChatGPT's response capabilities have surpassed those of ordinary doctors in addressing common disease symptoms.[28] However, in this study's specialized vertical field (specifically referring to IBD specialization), professional doctors still demonstrate superior judgment and threshold control in information decision-making. We can speculate that LLMs have critical threshold points in disseminating information in specialized vertical fields. It is essential for us to identify these thresholds rationally: disseminating information to laypersons below the threshold and utilizing tools to assist professionals above the threshold.

We envision a brighter future in healthcare, advocating for outstanding organizations (such as national cancer research centers or high-quality industry databases) to promote the dissemination of untainted high-quality data through independent reviews and exploration. Subsequently, leveraging digital tools like LLMs to share this data freely or affordably with patients, their families, and young doctors in need of accessing such information.[4,37]

LIMITATIONS

Tool and Method Selection

In order to achieve a sufficient sample size for significant effects, we temporarily set aside the assessment of evaluator consistency and well-validated information tools such as PEMAT and DISCERN (not disregarding them).[40,41]Compared to subjective crowdsourced rating strategies, these questionnaires or systems have relatively higher thresholds and specific use cases. Some researchers have suggested that certain health information assessment tools may not be universally suitable[32].If resources permit and the context is appropriate, we also recommend considering the simultaneous use of the aforementioned tools in the future, and when necessary, conducting accuracy assessments based on medical guidelines for evidence-based evaluation.[20]

Reply Randomness and Answer Reproducibility

Many researchers argue that a key limitation of the application and reproducibility of large-scale language models lies in the inherent randomness of their generated responses.[6]This inherent randomness refers to the unpredictability of these models, primarily because they are trained on various text data and use probabilistic algorithms to generate answers. Even with multiple inputs of the same or similar content, this inherent randomness can lead to variations in the quality and accuracy of the outputs.[5,7] However, some experiments suggest that repeating questions to ChatGPT multiple times results in excellent consistency of answers, reaching 90.48%-100%.[9,20] Our preliminary findings also indicate that if the prompts are the same, although not identical in every aspect, the structure and substantive content of the responses from ChatGPT are generally similar. Despite the good reproducibility of ChatGPT mentioned above, we have not yet fully overcome this limitation in our experiment.

Model Version Changes

The utilization of LLMs in patient education represents an interdisciplinary field at the intersection of medicine and technology. Large-scale AI language models possess the capability for improvement and learning, rendering similar research findings potentially outdated in a short span of time.[9,10,38] Just as we completed the compilation of responses in the second week, Open AI released a new version. This is why we adhere to an open science approach, utilizing publicly available and traceable materials as the textual sources for this comparison. Given improved conditions, we suggest that peers could build upon this foundation to conduct more comprehensive experiments or expand them into randomized controlled trials.

Network Latency and Restrictions

The blocking of ChatGPT in specific regions' IP addresses (such as mainland China and Hong Kong) has added

additional challenges to the use of this technology. High latency and instances of crashes have made the entire response process lengthy. While these limitations indeed exist, they were not reflected in our results. Exceptional technology not only guides outstanding experimental outcomes but also relies on the accessibility and low barriers to entry of that technology. It is hoped that in the future, all individuals, especially those in underdeveloped regions, can benefit from this technology. Gratitude is extended to OpenAI and ChatGPT 3.5 for their free and open-source demonstration of the allure of large language models and the exploration of their application scenarios.

Current Limitations in Usage Scenarios

Countries worldwide have enacted citizen health information privacy protection measures,[33] such as China's Personal Information Protection Law (PIPL) and the United States' Health Insurance Portability and Accountability Act (HIPAA). Under current circumstances, we cannot and do not recommend researchers to directly extract patient questions from patient communities, communication social media platforms (such as WeChat, WhatsApp), or outpatient settings to pose queries to ChatGPT. This is why, when considering the adoption of patient question sources, we collect authorized publications with management oversight.

Conclusions

In all dimensions, regardless of subjective or objective evaluation, ChatGPT demonstrates greater stability compared to human experts. When it comes to responses to specialized medical questions, ChatGPT's overall performance is on par with that of human specialist doctors. Its output of health information exhibits a better structural coherence, addressing the differentiation in outputs caused by cognitive and knowledge variations among individual specialist doctors. Utilizing ChatGPT-3.5 for drafting patient education materials, with doctors refining, supplementing, and proofreading the information, is acceptable and worth promoting. However, direct patient consultations and health education using ChatGPT are not feasible due to the presence of AI Hallucination. Differences in empathy and accuracy may exist between healthcare providers and patients. As primary consumers of health information, patients need be involved in the creation and evaluation of health information. Before extensively applying LLMs in medical practice, more clinical trials and case studies are needed to assess their effectiveness and potential side effects. Ethical and privacy concerns, user training and education, as well as ongoing monitoring and evaluation, are all issues that we need to consider and carefully deliberate on.

Author Contributions:

Zelin Yan had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and Design: Zelin Yan, Yan Chen, Yihong Fan.
Data Collection and Assembly of Facts: Zelin Yan, Yun Yang, Dingting Xu, Honggang Wang, Yun Yang, Yan Chen.
Data Transformation or Technical Support: Jie Mao, Hou-Chiang Tseng, Tao-Hsing Chang, Yun Yang.
Drafting of the Manuscript: Zelin Yan, Yan Chen.
Critical Revision of the Manuscript for Important Intellectual Content: Shiyuan Lu, Yan Chen, Yihong Fan.
Statistical Analysis: Zelin Yan, Jingwen Liu.
Funding Acquisition: Yan Chen.
Administrative, Technical, or Material Support: Yan Chen, Yihong Fan.
Supervision: Yan Chen.

Conflict of Interest Disclosure:

Zelin Yan previously acted as an Operations Coordinator at CCCF, executing the public awareness campaign on IBD and playing a key role in organizing the knowledge base content for IBD patients.
Yan Chen was the Chairman of CCCF.
Yan Chen and Yihong Fan have been one of authors of the " Q&A on Ulcerative Colitis and Crohn's Disease ".

Funding/Support: This research was supported by the Qingfeng Scientific Research Fund of The China Crohn’s & Colitis Foundation (CCCF) under Grant No. CCCF-QF-2022A60-1.

Acknowledgements:

We extend our sincere gratitude for the authorization provided by CCCF and Zhejiang University Press. We deeply appreciate the generous support from the CRIE 3.0 team and the invaluable assistance of Tseng.
Our heartfelt thanks go to the CCCF patient community and the three IBD patient volunteers (HYC, XXQ, JMC) for their selfless dedication in participating in the assessment.
The translation and editing of this article were facilitated by the translation and refinement features of ChatGPT.
We thank openbiox community and Hiplot team (https://hiplot.org) for providing technical assistance and valuable tools for data analysis and visualization.

1. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol* 2023.
2. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-English-speaking outpatient clinic settings. *Digit Health* 2023; **9**: 20552076231184091.
3. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. *Int J Med Inform* 2023; **177**: 105173.
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; **2**(2): e0000198.
5. Gorelik Y, Ghersin I, Maza I, Klein A. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. *Gastrointest Endosc* 2023; **98**(4): 639-41 e4.
6. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol* 2023.
7. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA. Evaluation of the Potential Utility of an Artificial Intelligence Chatbot in Gastroesophageal Reflux Disease Management. *Am J Gastroenterol* 2023.
8. Kim J, Cai ZR, Chen ML, Simard JF, Linos E. Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes. *JAMA Netw Open* 2023; **6**(10): e2338050.
9. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res* 2023; **25**: e47479.
10. 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**(13): 1233-9.
11. Madrigal L, Escoffery C. Electronic Health Behaviors Among US Adults With Chronic Disease: Cross-Sectional Survey. *J Med Internet Res* 2019; **21**(3): e11240.
12. Zhao J, Han H, Zhong B, Xie W, Chen Y, Zhi M. Health information on social media helps mitigate Crohn's disease

symptoms and improves patients' clinical course. *Computers in Human Behavior* 2021; **115**.
13. Bai XY, Zhang YW, Li J, Li Y, Qian JM. Online information on Crohn's disease in Chinese: an evaluation of its quality and readability. *J Dig Dis* 2019; **20**(11): 596-601.
14. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol* 2007; **102**(9): 2070-7.
15. Langille M, Bernard A, Rodgers C, Hughes S, Leddin D, van Zanten SV. Systematic review of the quality of patient information on the internet regarding inflammatory bowel disease treatments. *Clin Gastroenterol Hepatol* 2010; **8**(4): 322-8.
16. Mukewar S, Mani P, Wu X, Lopez R, Shen B. YouTube and inflammatory bowel disease. *J Crohns Colitis* 2013; **7**(5): 392-402.
17. He Z, Wang Z, Song Y, et al. The Reliability and Quality of Short Videos as a Source of Dietary Guidance for Inflammatory Bowel Disease: Cross-sectional Study. *J Med Internet Res* 2023; **25**: e41518.
18. Yu Q, Xu L, Li L, et al. Internet and WeChat used by patients with Crohn's disease in China: a multi-center questionnaire survey. *BMC Gastroenterol* 2019; **19**(1): 97.
19. Zhi M, Yan Q. [Q&A Section] I am a patient with inflammatory bowel disease. Can I have a normal pregnancy?
20. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023; **29**(3): 721-32.
21. van der Marel S, Duijvestein M, Hardwick JC, et al. Quality of web-based information on inflammatory bowel diseases. *Inflamm Bowel Dis* 2009; **15**(12): 1891-6.
22. Sung YT, Chang TH, Lin WC, Hsieh KS, Chang KE. CRIE: An automated analyzer for Chinese texts. *Behav Res Methods* 2016; **48**(4): 1238-51.
23. Sung YT, Chen JL, Cha JH, Tseng HC, Chang TH, Chang KE. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behav Res Methods* 2015; **47**(2): 340-54.
24. Sung YT, Lin WC, Dyson SB, Chang KE, Chen YC. Leveling L2 Texts Through Readability. *The Modern Language Journal* 2015; **99**(2): 371-91.
25. Tseng HC, Chin B, Chang TH, Sung YT. Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering* 2019; **25**(PT.3): 1-31.
26. Li Y, Zhou X, Zhou Y, et al. Evaluation of the quality and readability of online information about breast cancer in China. *Patient Educ Couns* 2021; **104**(4): 858-64.
27. Zheng Y, Tang Y, Tseng HC, et al. Evaluation of quality and readability of over-the-counter medication package inserts. *Res Social Adm Pharm* 2022; **18**(9): 3560-7.
28. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023; **183**(6): 589-96.
29. Fortinsky KJ, Fournier MR, Benchimol EI. Internet and electronic resources for inflammatory bowel disease: a primer for providers and patients. *Inflamm Bowel Dis* 2012; **18**(6): 1156-63.
30. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; **620**(7972): 172-80.
31. Baker DM, Marshall JH, Lee MJ, Jones GL, Brown SR, Lobo AJ. A Systematic Review of Internet Decision-Making Resources for Patients Considering Surgery for Ulcerative Colitis. *Inflamm Bowel Dis* 2017; **23**(8): 1293-300.
32. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: A comparative analysis between plastic surgeons and laypersons. *Int J Med Inform* 2023; **179**: 105219.
33. Personal Information Protection Law of the People's Republic of China. https://www.gov.cn/xinwen/2021-08/20/content_5632486.htm.
34. Coulter A, Entwistle V, Gilbert D. Sharing decisions with patients: is the information good enough? *Bmj* 1999; **318**(7179): 318-22.
35. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023; **5**(4): e179-e81.
36. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc* 2023; **30**(7): 1237-45.
37. Butte AJ. Artificial Intelligence-From Starting Pilots to Scalable Privilege. *JAMA Oncol* 2023.
38. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq* 2023.
39. Ayre J, Mac O, McCaffery K, et al. New Frontiers in Health Literacy: Using ChatGPT to Simplify Health Information for People in the Community. *J Gen Intern Med* 2023.
40. Deborah Charnock SS, Gill Needham, Robert Gann. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of epidemiology and community health* 1999; **53**(2): 105-11.

41. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns* 2014; **96**(3): 395-403.

Supplementary Files