Assessing ChatGPT as a Medical Consulting Assistant for Chronic Hepatitis B: A Cross-Language Study of English and Chinese

Yijie Wang, Yining Chen, Jifang Sheng

Submitted to: JMIR Medical Informatics
on: January 16, 2024

Table of Contents

Assessing ChatGPT as a Medical Consulting Assistant for Chronic Hepatitis B: A Cross-Language Study of English and Chinese

Yijie Wang[1] MD; Yining Chen[2] MD; Jifang Sheng[1] MD

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases Collaborative Innovation Center for Diagnosis and Treatment of Infectious Disease The First Affiliated Hospital, Zhejiang University School of Medicine Hangzhou CN
[2]Department of Urology Sir Run Run Shaw Hospital Zhejiang University School of Medicine Hangzhou CN

Corresponding Author:
Jifang Sheng MD
State Key Laboratory for Diagnosis and Treatment of Infectious Diseases
Collaborative Innovation Center for Diagnosis and Treatment of Infectious Disease
The First Affiliated Hospital, Zhejiang University School of Medicine
79 Qingchun Road
Hangzhou
CN

Abstract

Background: Chronic hepatitis B imposes substantial economic and social burdens globally. Managing CHB involves intricate monitoring and adherence challenges, particularly in regions like China, where a high prevalence intersects with healthcare resource limitations. This study explores the potential of ChatGPT-3.5, an emerging AI assistant, to address these complexities. With notable capabilities in medical education and practice, ChatGPT-3.5's role is examined in managing CHB, particularly in regions with distinct healthcare landscapes.

Objective: This study aims to uncover insights into ChatGPT-3.5’s potential and limitations in delivering personalized medical consulting assistance for chronic hepatitis B patients across diverse linguistic contexts.

Methods: Questions sourced from published guidelines, online chronic hepatitis B communities, and search engines in English and Chinese were refined, translated, and compiled into 96 inquiries. These questions were independently presented to ChatGPT-3.5 in dialogues. Responses underwent evaluation by senior physicians, focusing on informativeness, emotional management, consistency across repeated inquiries and cautionary statements regarding medical advice.

Results: Over half of the responses from ChatGPT-3.5 were deemed comprehensive. Superior performance was observed in English, particularly in informativeness and consistency across repeated queries. However, deficiencies were noted in emotional management guidance.

Conclusions: In this study, ChatGPT demonstrates potential as a medical consulting assistant for chronic hepatitis B management. The choice of working language by ChatGPT is identified as a potential factor influencing its performance, particularly concerning the utilization of terms and jargon, which may impact the applicability of ChatGPT within specific target populations. This study highlights the significance of providing language-specific training and incorporating emotional management strategies when deploying ChatGPT for medical purposes similar to those investigated.

(JMIR Preprints 16/01/2024:56426)
DOI: https://doi.org/10.2196/preprints.56426

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

Original Paper

Yijie Wang[1], Yining Chen[2], Jifang Sheng[1]

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Disease, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China
[2]Department of Urology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 Qingchun Road, Hangzhou, 310016, China.

Assessing ChatGPT as a Medical Consulting Assistant for Chronic Hepatitis B: A Cross-Language Study of English and Chinese

Abstract

Background: Chronic hepatitis B imposes substantial economic and social burdens globally. Managing CHB (Chronic hepatitis B) involves intricate monitoring and adherence challenges, particularly in regions like China, where a high prevalence intersects with healthcare resource limitations. This study explores the potential of ChatGPT-3.5, an emerging AI assistant, to address these complexities. With notable capabilities in medical education and practice, ChatGPT-3.5's role is examined in managing CHB, particularly in regions with distinct healthcare landscapes.
Objective: This study aims to uncover insights into ChatGPT-3.5's potential and limitations in delivering personalized medical consulting assistance for chronic hepatitis B patients across diverse linguistic contexts.
Methods: Questions sourced from published guidelines, online chronic hepatitis B communities, and search engines in English and Chinese were refined, translated, and compiled into 96 inquiries. Subsequently, these questions were independently presented to both ChatGPT-3.5 and ChatGPT-4.0 in independent dialogues. The responses were then evaluation by senior physicians, focusing on informativeness, emotional management, consistency across repeated inquiries and cautionary statements regarding medical advice. Additionally, a true-or-false questionnaire was employed to further discern the variance in information accuracy for closed questions between ChatGPT-3.5 and ChatGPT-4.0.
Results: Over half of the responses from ChatGPT-3.5 were considered comprehensive (61.62%). In contrast, ChatGPT-4.0 exhibited a higher percentage at 74.48% ($P < 0.0001$). Notably, superior performance was evident in English, particularly in terms of informativeness and consistency across repeated queries. However, deficiencies were identified in emotional management guidance, with only 3.23% in ChatGPT-3.5, and 9.74% in ChatGPT-4.0 ($P = 0.0432$). ChatGPT-3.5 tend to include disclaimer in 10.81% responses, while ChatGPT-4.0 used disclaimers in 13.06% responses ($P = 0.4642$). When responding to true-or-false questions, ChatGPT-4.0 achieved an accuracy rate of up to 93.33%, significantly surpassing ChatGPT-3.5's 65.00% ($P < 0.0001$).
Conclusions: In this study, ChatGPT demonstrates basic capabilities as a medical consultation assistant for chronic hepatitis B management. The choice of working language for ChatGPT-3.5 was considered a potential factor influencing its performance, particularly in the use of terminology and colloquial language, potentially affecting its applicability within specific target populations. However, as an updated model, ChatGPT-4.0 exhibits improved information processing capabilities, overcoming the language impact on information accuracy. This suggests that the implications of

model advancements on application need to be considered when selecting large language models (LLMs) for medical consultation assistants. Given that both models performed inadequately in emotional guidance management, this study also highlights the importance of providing specific language training and emotional management strategies when deploying ChatGPT for medical purposes. Furthermore, the tendency of these models to use disclaimers in conversations should be further investigated to understand the impact on patients' experience in practical applications.

Key words: Chronic hepatitis B; Artificial intelligence; Large language models; Chatbots; Medical Consultation; AI in healthcare; Cross-linguistic study

Introduction

Chronic Hepatitis B: A Dual Burden on Patients and Society

Chronic hepatitis B (CHB) imposes significant economic and social burdens. In 2019, approximately 296 million people were affected, resulting in an estimated 820 thousand deaths [1]. The World Health Organization (WHO) noted that among those chronically infected with hepatitis B and C, about 20% or more would develop end-stage chronic liver disease, such as cirrhosis and hepatocellular carcinoma [2].

Hepatitis B virus (HBV) primarily spreads through blood contact, unprotected sexual intercourse, and mother-to-infant transmission. Effective management of chronic infection necessitates daily monitoring and self-care [3]. Nevertheless, the intricacy of regular monitoring, encompassing multiple tests such as HBsAg, HBeAg, HBV-DNA, ALT, and fibrosis assessment, as endorsed by authoritative bodies in hepatitis B diagnosis and treatment, including the European Association for the Study of the Liver (EASL), presents hurdles to patient compliance[4]. Additionally, the prolonged, often lifelong, administration of antiviral medications contributes to further adherence issues [4, 5]. Unique considerations for pregnant individuals and children add another layer of complexity, demanding targeted counseling and specialized management. This intricate management landscape not only burden s patients with emotional stress but also jeopardizes adherence to treatment regimens [6, 7]. The complexity of CHB management requires personalized healthcare strategies, easing individual and societal burdens, emphasizing the importance of diverse health approaches.

ChatGPT as a Prospective Medical Assistant

Currently, artificial intelligence (AI) has become integral in the medical domain, particularly in medical research and clinical practice. Notably, according to Shahid Ud Din Wani et al., traditional machine learning methodologies required the supervision of skilled individuals and structured input data, resulting in considerable resource-intensive processes [8]. Recognizing the limitations of traditional approaches, Charlotte J Haug and a colleague proposed chatbots for capabilities in medical practice assistance [9].

Released in June 2020, GPT-3.5 underpins ChatGPT's emergence in AI-assisted medical applications. As a Large Language Model (LLM), it shows potential for medical assistance [10], though challenges and concerns persist in its application within the field [11]. The functioning of large language models involves predicting and generating a coherent and contextually relevant response based on materials pre-input, necessitating training on massive amounts of diverse textual data. Various studies have explored ChatGPT's capacity to act as a virtual doctor or medical tutor on diagnosis or treatment [12].

The study by Aidan Gilson and colleagues revealed that ChatGPT performed well in medical knowledge assessments, demonstrating potential as a virtual medical tutor [13]. Yee Hui Yeo, etc. evaluated ChatGPT's performance in answering questions regarding cirrhosis and hepatocellular

carcinoma [14]. Most studies have compared its performance to that of real doctors or medical students, aiming to determine whether AI assistant could surpass human medical service providers. However, there were challenges and risks of employing ChatGPT in clinical practice, including the potential generation of plausible yet inaccurate information and ethical considerations [15]. According to these studies, LLMs could potentially assist on medical consulting and auxiliary diagnosis even if in traditional medical research, treatment and education, but there are still unidentified risks and problems.

ChatGPT-4.0, released on March 14, 2023, represents an updated version of the ChatGPT model. Plenty of researches have compared the application of ChatGPT-3.5 and ChatGPT-4.0 in medical practice [16-18]. In this research, we involved ChatGPT-4.0 as the comparative model to further assess the application problem of this model.

Medical Assistance in Hepatitis B Management with Chinese as the Primary Language

Bearing the highest global burden of hepatitis B, China recorded over 90 million people were living with chronic hepatitis B in 2017[1, 2]. Research reveals troubling trends in treatment non-compliance for HBV in China, including challenges in preventing vertical transmission [19-21]. Beyond China, studies in various regions highlight the impact of factors like family income, employment, and patient gender on medical treatment compliance for chronic hepatitis B [22].

Physician encouragement proves crucial for patient compliance with medication regimens [23]. Despite a rising number of medical doctors in China, there is a shortage of medical practitioners including licensed physicians and physician assistants, who face high workloads and burnout rates [24-26]. While research indicates Chinese physicians generally adhere to hepatitis B guidelines [27], medical errors due to workload demands could undermine intended impact on patient compliance [26]. Amidst these challenges, exploring medical assistance using Chinese as the primary working language becomes crucial. This inquiry is vital for enhancing patient compliance in hepatitis B management and alleviating strain on healthcare professionals amid work-related stress. However, a study involving ChatGPT's performance in medical exam in Chinese emphasized the significance of exploring cross-language difference of ChatGPT's performance in the future study [28].

In brief, a dialogue-based medical assistant is increasingly recognized as essential in clinical practice. Exploring the application of ChatGPT in this domain shows promise for medical research and clinical usage. This study assesses ChatGPT-3.5 and ChatGPT-4.0 in tasks such as diagnosis, providing management advice, and addressing counseling needs for patients with chronic hepatitis B. Given that English data accounts for the largest proportion (approximately 92%) in the original training language of this model [29], it is reasonable to assume that among all the languages included in the pre-training resource, this model performs best in English. However, the Chinese language serves the world's largest group of chronic hepatitis B patients, underscoring the irreplaceable role of Chinese in studies regarding medical AI assistant. Therefore, the research involves both English and Chinese as working languages and compares ChatGPT's performance in both languages. Additionally, the study includes a comparison between ChatGPT-3.5 and ChatGPT-4.0 to investigate the improvements from the former to the latter. Through this investigation, we aim to uncover the potential and limitations of this application in medical practice.

Methodology

Questionnaire Development Process

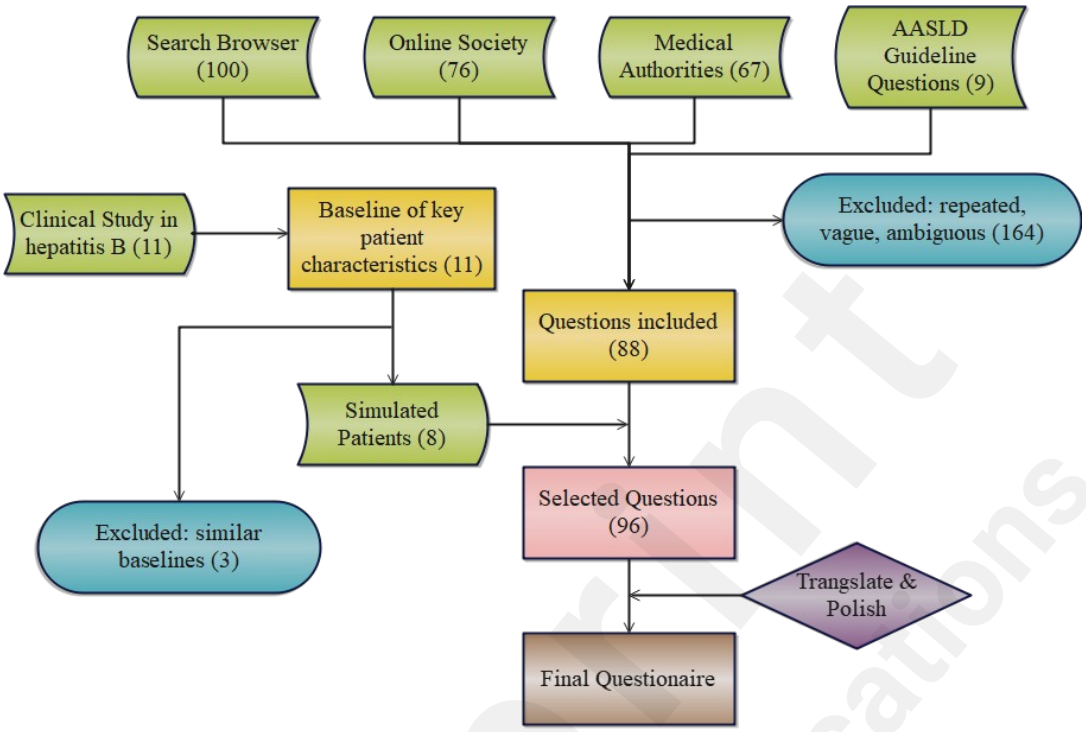


Figure 1. Workflow of questionnaire design process. The figure shows specific information of each stage of the questionnaire compiling process.

Following the workflow shown in Figure 1, we systematically compiled a set of questions relevant to both patients and physicians in clinical practice. This compilation process involved:

1. Questions sourced from esteemed professional associations and institutions such as the World Health Organization (WHO), American Centers for Disease Control and Prevention (CDC), American Association for the Study of Liver Disease (AASLD), European Association for the Study of the Liver (EASL), and Asian Pacific Association for the Study of the Liver (APASL).

2. Queries about hepatitis B found on online social media platforms, particularly in patient support groups and disease-specific forums. Inclusion criteria prioritized relevance to diagnosis, treatment, daily monitoring, lifestyle, and other hepatitis B-related concerns. Inclusion and exclusion criteria: questions with precise wording, minimal grammatical errors, and clarity were included, while non-medical inquiries, ambiguously framed questions, and those related to non-medical issues were excluded. Questions with significant updates after September 2021 were also omitted.

3. We conducted an exploration of associative keywords following the entry of "hepatitis B" or "HBV" into widely utilized search engines, such as Google, Bing, and Baidu, in both Chinese and English languages.

4. Based on diverse published hepatitis B clinical studies, we systematically extracted key patient characteristics, including age, gender, hepatitis B serum markers, HBV-DNA levels, ALT levels, and concomitant diseases. We developed profiles for eight simulated patients using these data with a random number function. ChatGPT was tasked with providing advice to these simulated individuals on various aspects, involving treatment/examination recommendations, treatment strategies, daily monitoring practices, lifestyle adjustments, and more.

Among all the questions gathered, multiple questions were separated into single entities, while repeated questions were excluded. To avoid ambiguity and misunderstanding resulting from language vagueness, which could potentially impact the assessment of the model's information

accuracy, we carefully polished all the collected questions, refining their grammar and phrasing, and performed localized translations between Chinese and English. Examples of revised or excluded questions are shown in Supplementary Table 1. In total, we gathered 96 questions about hepatitis B. Among the questions, there was 2 with only English version and 5 with only Chinese version. These language-specific questions focused on issues specific to the country or region where the questioner was located.

Section Allocation

We systematically categorized all questions into five distinct sections: Term Explanation Questions, Short Answer Questions, Clinical Problem Questions, AASLD Guideline Questions, Simulated Patient Questions.

The "Term Explanation Questions" section featured 17 terms associated with hepatitis B, including one term exclusively for Chinese responses. In the "Short Answer Questions" section, there were 22 questions, with one specifically designed for Chinese responses. Questions within the "Clinical Problem Question" section were primarily sourced from online hepatitis B societies, totaling 40 questions. Within this section, there was one question intended solely for English responses, and two exclusively for Chinese responses. The questions in "AASLD Guideline Questions" section were derived from the AASLD guidelines for hepatitis B in 2016 and 2018 (updated version), including nine questions which were all translated into Chinese. The "Simulated Patient Questions" section consisted of eight questions related to simulated patient information, as previously constructed. These questions were provided in both English and Chinese versions.

Gathering Responses

The questions were submitted to ChatGPT-3.5 during 1st to 30th in April, 2023. with each question forming a separate dialogue. Each question was sent twice for Chinese and English separately to ensure a comprehensive evaluation, and responses were collected. In the case of a system error preventing ChatGPT-3.5 from responding, the question was resubmitted in a new dialogue. All responses were compiled into a table for further assessment.

Assessment of Responses

Two senior physicians independently evaluated all responses. In case when discrepancies occurred in information accuracy, consistency of repeated responses, and emotional management guidance assessments, a third senior physician with over 30 years of experience in hepatitis B diagnosis and treatment would conduct a final review for the ultimate assessment, and give the final scores. The criterion of assessment was discussed and voted by a committee of 5 senior physicians in hepatitis B diagnosis and treatment. The assessment process referred to the research of Yee Hui Yeo, etc [14].

Information Accuracy Assessment

The assessment of information accuracy assessment was mainly focused on the correctness and comprehensiveness. Four assessment grades (1-4) were assigned: 1, correct and comprehensive;2, correct, but with missing information; 3, a mix of correct and incorrect details; 4, wholly incorrect or irrelevant information.

Categorization of the Types of Mistakes

Mistakes in responses assessed as “a mix of correct and incorrect details” and “wholly incorrect or irrelevant information” were analyzed. and categorized the types of mistakes. The mistakes were classified into five categories: A. misunderstanding of medical terms or jargon; B. incorrect usage of medical terms; C. mistakes in diagnosis/treatment/management without mistakes in terms or jargon;

D. total irrelevant information; E. a mixture of two or more kinds of mistakes among A-C.

Content Consistency of Repeated Responses Assessment

A binary assessment (“Yes” or “No”) was employed to indicate the consistency of the two responses for each question. This evaluation was independent of the information accuracy assessment, solely focusing on the consistency of response content.

Emotional Management Guidance Assessment

For all responses in the “Clinical Problem Questions” and “Simulated Patient Questions” sections (48 in total, 1 with only English version and 2 with only Chinese version), an emotional management guidance assessment was conducted. The assessment comprised three levels: 1, sufficient emotional and psychological management guidance; 2, respectful but lacking or inadequate emotional or psychological management guidance; 3, disrespectful or negative emotional guidance.

Analysis of ChatGPT's Cautionary Statements Regarding Medical Advice

We quantified the instances where ChatGPT recommended consulting a genuine healthcare provider or doctor. Meanwhile, we counted the frequency of ChatGPT explicitly stating disclaimers such as "I am not a doctor" or "I cannot give diagnosis or treatment" among all questions involving clinical practice (including the section of Clinical Problem Questions, AASLD Guideline Questions and Simulated Patient Questions).

Parallel Assessment of ChatGPT-4.0’s Performance

We replicated the above assessment process for ChatGPT-4.0. Considering that ChatGPT-4.0 is the updated version of the model, we omitted sections involving only the basic medical knowledge in the questionnaire. As a more intuitive alternative, we chose closed questions to evaluate the fundamental knowledge differences between the two model versions (see the following section in Methodology). The assessment of ChatGPT-4.0 included questions from “Clinical Problem Questions”, “AASLD Guideline Questions” and “Simulated Patient Questions” sections of the questionnaires used in the previous assessment. However, mistake analysis was omitted as there were no responses from ChatGPT-4.0 that were assessed as incorrect.

Compare of ChatGPT-3.5 and ChatGPT-4.0 Using Closed Questions (True-or-false Statements)

In this assessment, we formulated 30 statements based on AASLD Guidelines for Treatment of Chronic Hepatitis B including all its updates up to September in 2021. These statements were input into the models in separate dialogues. We utilized prompts to ask the models to judge whether the statements were correct and to provide a judgment with “Yes” or “No”. The prompts are detailed in Table 1. Each statement was input into the model three times, and the response for each iteration was recorded. All responses of the models were collected, and their accuracy and stability (the consistency of 3 responses to a repeated statement) were assessed.

Table 1. An example of the prompts used in closed questions.

	Prompts for ChatGPT-4.0	Prompts for ChatGPT-3.5
English	Now, I would like you to act as a hepatologist in the upcoming conversation and determine whether the statements are true and answer with only “Yes” or “No”. Here are the statements: []ᵃ	Now, I would like you to act as a hepatologist in the upcoming conversation and determine whether the statements are true and answer with only “Yes” or “No”, and do not add any explanation. Here are the

		statements: []
Chinese	□□“□”□“□”□□□□□□[a]	□□“□”□“□”□□□□□□□□□□□□□□□□□□□

[a]The statements were added in the square brackets.

Statistical Analysis

All statistical analyses were performed with the SPSS 26.0 statistical package (IBM, Armonk, NY, USA). Cohen's kappa coefficients were used to determine interobserver reliabilities. Assessment grades were calculated and reported as percentages. Comparative analysis of ranked data employed Mann-Whitney test. Categorical data were compared using Chi-square tests. Wilcoxon Signed Rank Test was applied to compare the grades of response1 and response2 to each question. Statistical significance was set at $P < 0.05$.

Result

Information Accuracy Assessment of ChatGPT-3.5

The interobserver reliability κ was 0.6020, $P < 0.0001$ for information accuracy assessment. The results of this assessment are shown in Figure 2. Across all the questions, 90.81% of responses from ChatGPT-3.5 contained no incorrect information (including comprehensive responses and correct but incomprehensive responses). The likelihood of ChatGPT giving correct and comprehensive responses was 61.62%, while there was a 29.19% probability of responses being correct with missing information (see Supplementary Table 2). Responses with mixture of correct and incorrect information accounted for 7.30%. There were 1.89% of the responses wholly incorrect or irrelevant to the questions.

Performance of ChatGPT-3.5 varied across the sections, and the differences were statistically significant ($P < 0.0001$, see Supplementary Table 1). In the section "Term Explanation Questions", the highest percentage of responses assessed as complete and comprehensive was observed (92.86% in English and 88.24% in Chinese, see Figure 2A) while in the section "AASLD Guideline Questions", the highest percentage of responses totally wrong or irrelevant, or mixed with incorrect information was noticed (22.22% in English and 55.56% in Chinese).

The language environment in which ChatGPT-3.5 operated also influenced its performance (see Figure 2B). In Chinese, ChatGPT demonstrated poorer performance compared to English ($P = 0.0013$), particularly in the section "Clinical Problem Questions" ($P = 0.0337$) and "AASLD Guideline questions" ($P = 0.0022$). However, performance in the sections Term Explanation Questions ($P = 0.5434$), Short Answer Questions ($P = 0.6235$), and Simulated Patient Questions ($P = 0.3268$) showed no significant difference between the two working languages. The evaluation table is in Supplementary Table 3.

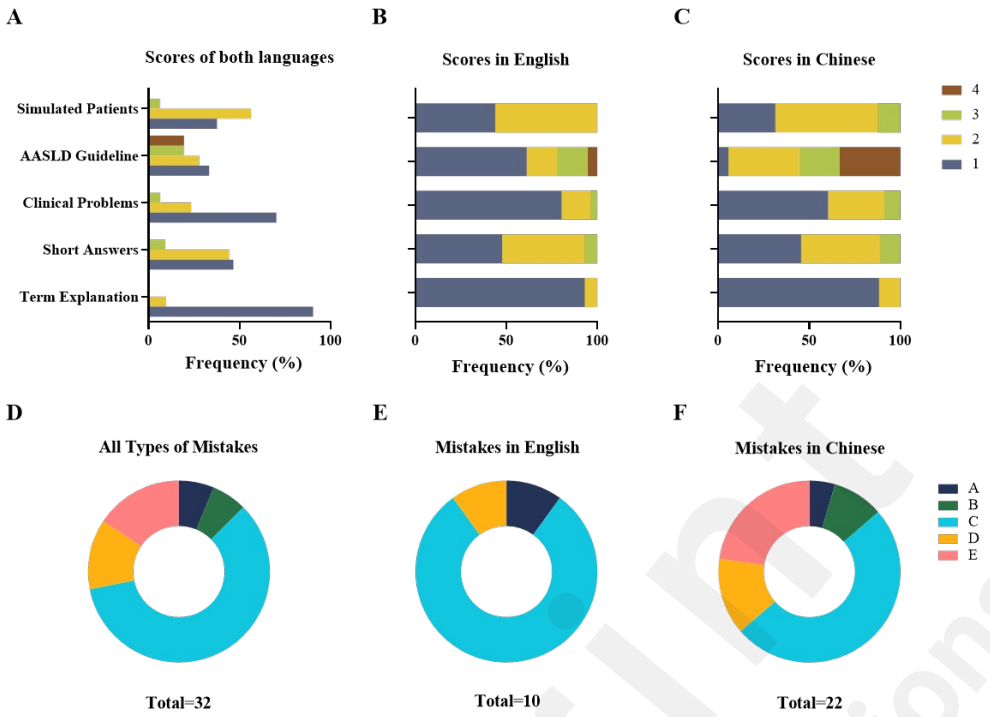


Figure 2. Results of information accuracy assessment and mistake analysis of ChatGPT-3.5. **A.** Comparation of percentages for each grade across all responses in distinct question sections. **B.** Percentages of each grade of responses in English in separated question sections. **C.** Percentages of each grade of responses in Chinese in separated question sections. **D.** Overview of mistake types across all responses. **E.** Breakdown of mistake types specifically among responses in English. **G.** Breakdown of mistake types specifically among responses in Chinese.

Categorization of the Types of Mistakes of ChatGPT-3.5

Figure 2D-F summarizes the types of mistakes in the responses. In both languages, the most common error pertained to diagnosis, treatment or disease management (see Figure 2D). Notably, in Chinese, 10 out of 32 mistakes involved incorrect usage or misunderstanding of technical terms (31.25%, see Figure 2F), while in English, there was no such mistakes (see Figure 2E). The evaluation tables are provided in Supplementary Table 4.

Content Consistency of Repeated Responses Assessment of ChatGPT-3.5

The interobserver reliability κ was 0.6532, $P < 0.0001$ for content consistency of repeated responses assessment of ChatGPT-3.5. Figure 3 shows the content consistency of repeated responses. For all questions, the probability of content consistency between two responses was 54.05%. In English, the consistency was 62.22%, while in Chinese, it was 46.32%, showing a significant difference ($P = 0.0387$, see Figure 3A and Supplementary 2). This disparity was also significant in the section of clinical problem questions ($P = 0.0375$, see Figure 3B and Supplementary 1). The section with the highest consistency was the Term Explanation Questions section (93.55% consistent), while the Short Answer Questions section had the lowest (27.91% consistent). Despite poor consistency in content, the two responses exhibited similarity in grades ($P = 0.6535$). The evaluation tables are provided in Supplementary Table 5.

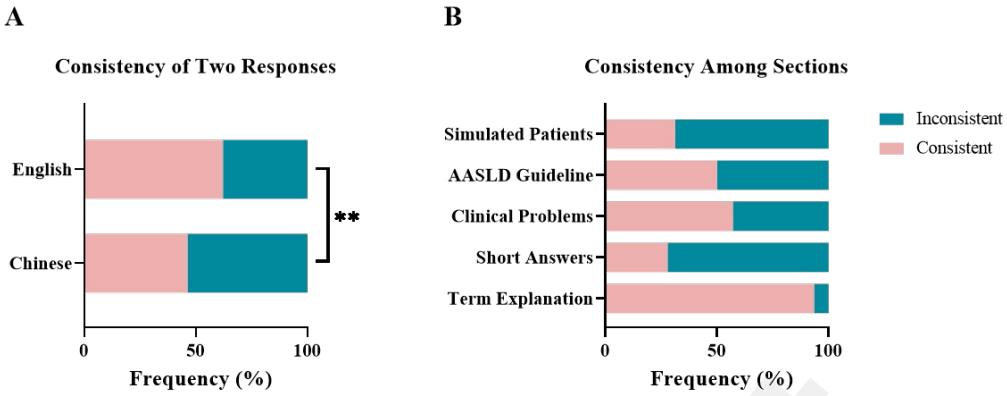


Figure 3. Assessment of content consistency of responses to repeated questions. **A.** Comparison of content consistency between responses in different working languages. **B.** Examination of content consistency in different sections of questions.

Emotional Management Guidance Assessment of ChatGPT-3.5

Among responses to questions within the “Clinical Problem Questions” and “Simulated Patient Questions” sections, only 3.23% were deemed to provide sufficient emotional management support (see Table 2). Related responses were listed in Supplementary Table 5. Most responses were assessed as “respectful but lacking or inadequate emotional or psychological management guidance” (96.77%). No response was assessed as “disrespectful or negative emotional guidance”. ChatGPT-3.5 exhibited comparable performance in both languages (P = 0.3928).

Table 2. Results of emotional management guidance assessment.

	Clinical Problem Questions			Simulated Patient Questions			Total			P value
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	
Chinese	1	77	0	1	15	0	2 (2.13%)	92 (97.87%)	0 (0.00%)	
English	4	72	0	0	16	0	4 (4.35%)	88 (95.65%)	0 (0.00%)	
P value	0.4751[a]									
Total	5 (3.25%)	149 (96.75%)	0 (0.00%)	1 (3.13%)	31 (96.88%)	0 (0.00%)	6 (3.23%)	180 (96.77%)	0 (0.00%)	0.3928[b]

[a]P value across the grades of each section.
[b]P value between the grades of different working languages.

Analysis of ChatGPT-3.5's Cautionary Statements Regarding Medical Advice

Figure 4 shows the results of this part. ChatGPT-3.5 exhibits distinct characteristics as a medical assistant. In most responses, ChatGPT-3.5 tends to remind patients to consult a healthcare provider or a physician (62.02% mentioned). This percentage is consistent in both English (65.56% mentioned, see Figure 4A) and Chinese (58.60% mentioned), with no significant difference (P = 0.1963). These responses were listed in Supplementary Table 7.

Among all questions involving clinical practice, the probability of ChatGPT-3.5 using the phrase “I am not a doctor” or “as a language model” was 10.81% (24 in 222 responses). In Chinese the probability was 11.61%, while in English the probability was 10.00% (see Figure 4B). No significant difference was observed between the two languages (p > 0.9999). These responses were listed in Supplementary Table 8.

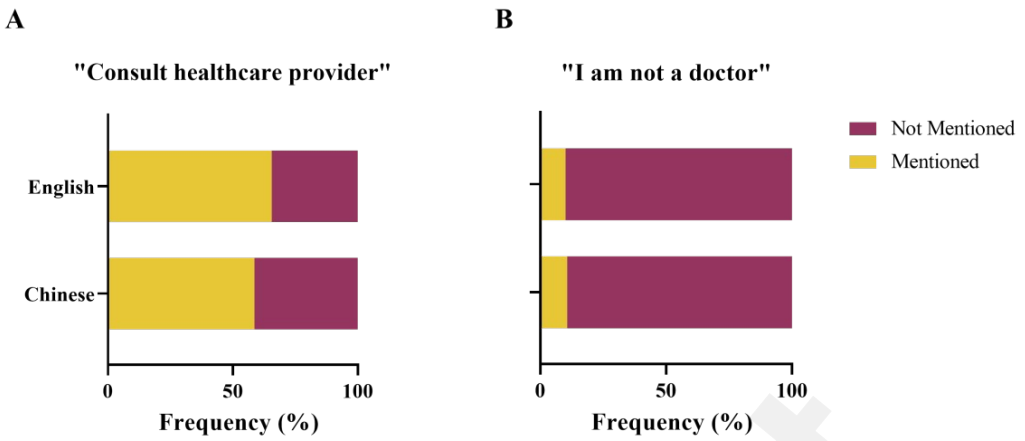


Figure 4. Percentages of ChatGPT cautionary statements regarding medical advice and disclaimers. **A.** Percentages of responses that include the recommendation to "consult healthcare providers or doctors". **B.** Percentages of responses containing the disclaimer phrases "I am not a doctor" or "I cannot give diagnosis or treatment".

Parallel Assessment of ChatGPT-4.0 in Sections Involving Clinical Practice

The interobserver reliability κ was 0.6896, $P < 0.0001$ for information accuracy assessment. Notably, ChatGPT-4.0 demonstrated distinct performance compared to ChatGPT-3.5. The scores of ChatGPT-4.0 are presented in Figure 5A-C. Across the three sections including “Clinical Problem Questions”, “AASLD Guideline Questions” and “Simulated Patient Questions”, the percentage of responses assessed as complete and comprehensive, as well as “Grade 1,” was higher for ChatGPT-4.0 compared to ChatGPT-3.5 (ChatGPT-4.0: 77.48%, ChatGPT-3.5: 59.46%), with a significant difference ($P < 0.0001$). Furthermore, variations in grades were observed across the sections ($P < 0.0001$). The “Clinical Problem Questions” section exhibited the highest percentage of responses assessed as complete and comprehensive (86.36%), surpassing ChatGPT-3.5 (70.13%, $P = 0.0006$). Importantly, no responses from ChatGPT-4.0 were assessed as “a mix of correct and incorrect details” and “wholly incorrect or irrelevant information”. In general, ChatGPT-4.0 demonstrated superior information accuracy compared to ChatGPT-3.5. Moreover, ChatGPT-4.0 showed improved performance in responding to Chinese questions. Although there was a slightly lower percentage of responses assessed as “Grade 1” for Chinese (73.21%) compared to English (81.82%), the difference in performance between the languages was not significant ($P = 0.1249$). See evaluation tables in Supplementary Tables 9, 10.

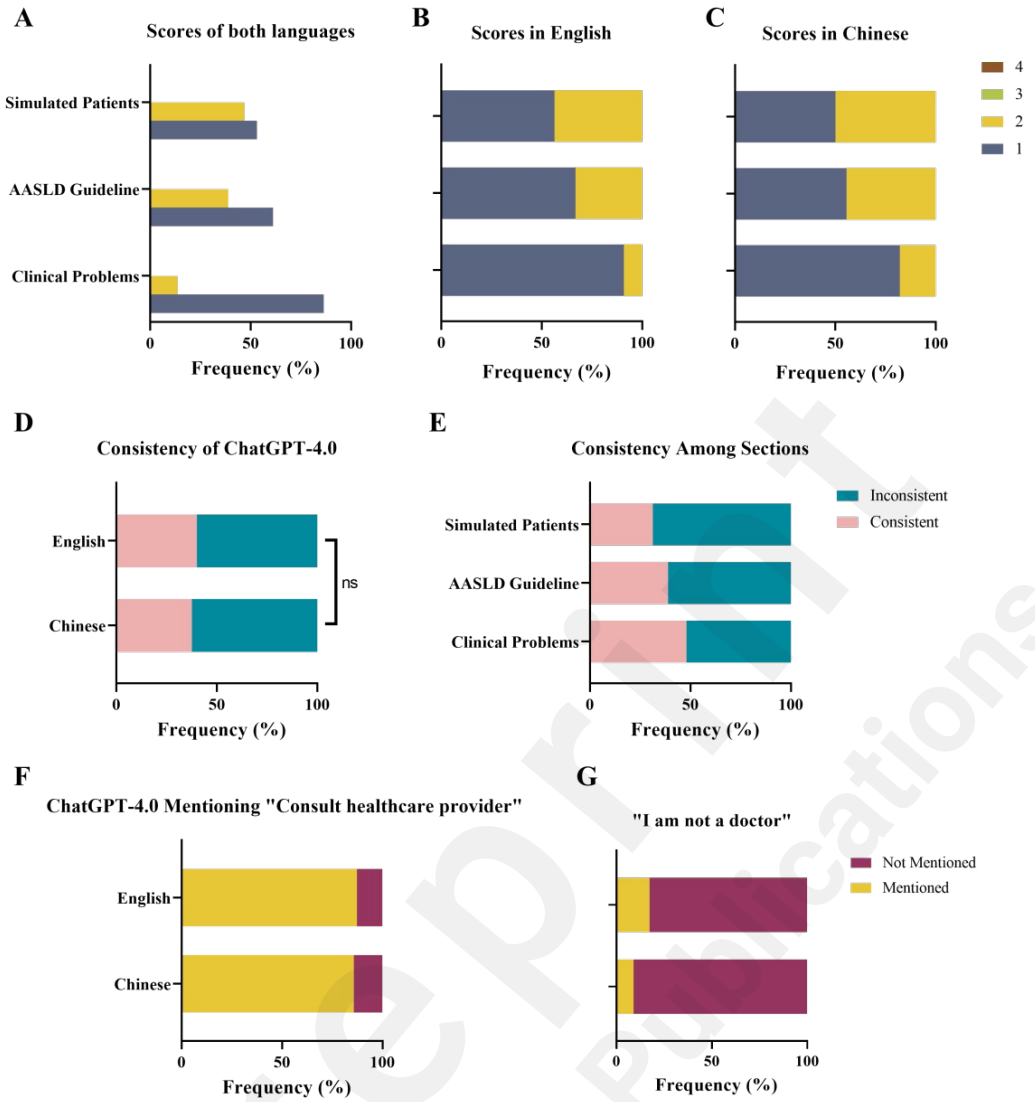


Figure 5. Results of parallel assessment of ChatGPT-4.0 in sections involving clinical practice. **A.** Comparation of percentages for each grade across all responses in distinct question sections. **B.** Percentages of each grade of responses in English in separated question sections. **C.** Percentages of each grade of responses in Chinese in separated question sections. **D.** Comparison of content consistency between responses in different working languages. **E.** Examination of content consistency in different sections of questions. **F.** Percentages of responses that include the recommendation to "consult healthcare providers or doctors". **G.** Percentages of responses containing the disclaimer phrases "I am not a doctor" or "I cannot give diagnosis or treatment".

The interobserver reliability κ was 0.6052, $P < 0.0001$ for content consistency of repeated responses assessment. ChatGPT-4.0 showed poorer consistency in responses to repeated questions. Across all questions, ChatGPT-4.0 provided 44.14% stable repeated responses, lower than ChatGPT-3.5's 52.25%. However, this difference was not significant ($P = 0.2267$). Specifically, in Chinese, ChatGPT-4.0's stability percentage was 37.5%, and in English, it was 50.91% (Figure 5D), with no significant difference ($P = 0.1549$). Among all the sections, responses in "Clinical Problem Questions" exhibited the highest rate of consistency at 48.05% (Figure 5E). The difference in consistency across sections was not significant ($P = 0.4153$). Detailed evaluation tables are provided in Supplementary Table 9 and 11.

In responses to questions within the "Clinical Problem Questions" and "Simulated Patient Questions" sections, ChatGPT-4.0's responses were assessed to provide sufficient emotional management support 9.74% of the time (see Table 3). This performance differed significantly from

that of ChatGPT-3.5 (P = 0.0432). The percentage was similar between Chinese and English (7.45% in Chinese and 91.30% in English, P = 0.7545). No response was assessed as “disrespectful or negative emotional guidance”. ChatGPT-4.0 showed similar performance between the two sections (6.49% in Clinical Problem Questions and 15.63% in Simulated Patient Questions assessed as Grade 1, P = 0.0843). However, among all responses assessed as “unstable”, there was no significant difference between the scores of response 1 and response 2 (P = 0.0593). All the responses assess as “sufficient emotional and psychological management guidance” are listed in Supplementary Table 12.

Table 3. Results of emotional management guidance assessment of ChatGPT-4.0.

	Clinical Problem Questions			Simulated Patient Questions			Total			P value
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	
Chinese	5	73	0	2	14	0	7(7.45%)	87 (92.55%)	0 (0.00%)	
English	5	71	0	3	13	0	8 (8.70%)	84 (91.30%)	0 (0.00%)	
P value			0.0843[a]							
Total	10 (6.49%)	144 (93.51%)	0 (0.00%)	5 (15.63%)	27 (84.38%)	0 (0.00%)	15 (9.74%)	171 (91.94%)	0 (0.00%)	0.7545[b]

[a]P value across the grades of each section.
[b]P value between the grades of different working languages.

As shown in Figure 5F, ChatGPT-4.0 demonstrated comparable performance to ChatGPT-3.5 across all responses, with 86.49% of responses emphasizing the importance of seeking medical assistance. In Chinese, 96 out of 112 responses (85.71%) stressed this need, while in English, 96 out of 110 responses (87.27%) did the same. Notably, no significant difference was observed between the languages (P = 0.7342). In responses from ChatGPT-3.5 of the sections “Clinical Problem Questions”, “AASLD Guideline Questions” and “Simulated Patient Questions”, the in total percentage of responses with medical service recommendation mentioned was 81.53% (181 out of 222 responses), which was not different from ChatGPT-4.0 (P = 0.1544). All the responses emphasizing the necessity of seeking for medical service are listed in Supplementary Table 13.

Figure 5G illustrates that among all responses to questions involving clinical practice, ChatGPT-4.0 used the phrase “I am not a doctor” or “as a language model” with a probability of 13.06% (29 out of 222 responses). This percentage did not significantly differ from that of ChatGPT-3.5 (P = 0.4642). In Chinese, the probability was 8.93% (10 out of 112), while in English, the probability was 17.27% (19 out of 110, see Figure 5B). No significant difference was observed between the two languages (P = 0.0651). These responses are detailed in Supplementary Table 14.

Assessment of Responses to Closed Questions Across ChatGPT-4.0 and ChatGPT-3.5

When assessing the accuracy of statements derived from AASLD Guideline of Treatment of Chronic Hepatitis B, ChatGPT-4.0 exhibited significantly superior performance compared to ChatGPT-3.5 (see Figure 6A-B). ChatGPT-4.0 achieved a correctness percentage of 93.33%, with the same percentage accuracy in both Chinese and English (93.33% for each language). Conversely, ChatGPT-3.5 yielded an overall accuracy of 65.00% (117 out of 180 responses), with a split of 50.00% in Chinese (45 out of 90 responses) and 80.00% in English (72 out of 90 responses).

Furthermore, ChatGPT-4.0 displayed enhanced consistency in repeated responses (Figure 6C). Stable responses accounted for 98.33% (59 out of 60 questions) in ChatGPT-4.0, whereas ChatGPT-3.5 provided only 66.67% stable responses (40 out of 60 questions). The difference in response

stability between the models was statistically significant (P < 0.0001).

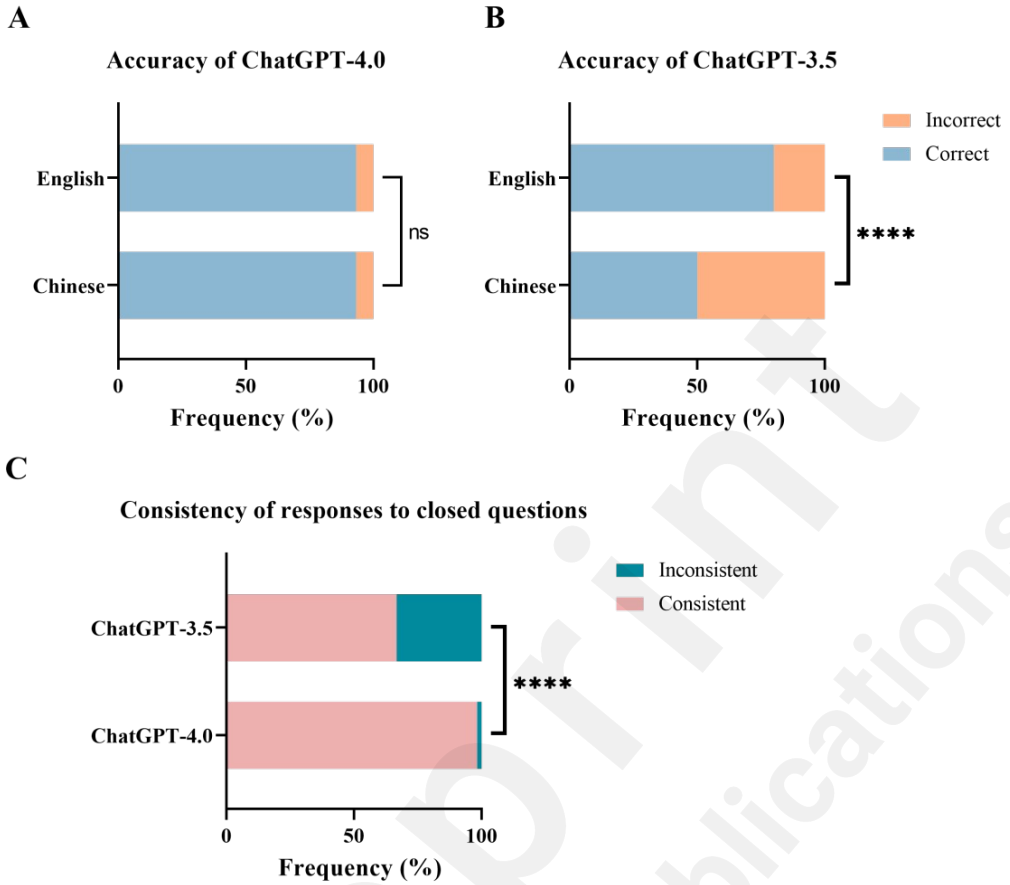


Figure 6. Results of responses to true-or-false questions. A. Rate of Accuracy of ChatGPT-4.0. B. Rate of Accuracy of ChatGPT-3.5. C. Compare of the consistency of responses to true-or-false questions between ChatGPT-4.0 and ChatGPT-3.5.

Discussion

ChatGPT-3.5 Working as a Medical Consulting Assistant

Our evaluation highlighted the proficiency of ChatGPT-3.5 as a medical consulting assistant. ChatGPT-3.5 provided predominantly accurate information, but there was a notable limitation in the comprehensiveness of the responses, indicating a need for targeted medical professional input. Continuous enhancement of LLMs may contribute to more specific and reliable guidance. Despite its strengths, ChatGPT-3.5 displayed limitations in emotional management support, a crucial aspect of chronic disease management [30]. Facilitating emotional modulation is integral to fostering patient willingness for self-management and treatment compliance [7, 30].

Therefore, it is imperative to consider emotional cognition and regulation in medical diagnosis and treatment. Our study suggested that the potential for ChatGPT to serve as an emotional management assistant for chronic patients warrants further study, with related localized training considered if LLMs are to be employed in in clinical practice as health consulting assistants.

Impact of Working Language on Performance

By revealing ChatGPT's inferior performance in Chinese compared to English, the study emphasized the influence of the choice of working language on stability and correctness. ChatGPT-3.5 showed worse performance on information accuracy in Chinses, implying the insufficient input

of knowledgeable materials in Chinese. Reflected in lower consistency rate of responses to the same questions, both ChatGPT-3.5 and ChatGPT-4.0 showed less stability in Chinese. Such challenge stemmed from variations in language resources during the model's original training, primarily centered around English-based medical guidelines. Though there are Chinese translation version of these guideline, there timeliness and accuracy of Chinese materials are limited. To enhance ChatGPT's efficacy in diverse language environment, the model should undergo additional training based on data sourced from specific language resources. This targeted training should focus on potential misunderstandings related to terms and phrasings in local languages, thereby addressing language-specific nuances and enhangcing overall performance. Notably, ChatGPT-3.5 exhibited language-specific mistakes, with Chinese responses showing errors related to misunderstanding or incorrect usage of terms. This underscores the importance of targeted language training for large language models to minimize inaccuracies, especially in medical contexts.

Cautionary Statements and Patient-Oriented Usage

In discussions related to diagnosis and therapy, ChatGPT-3.5 consistently emphasizes the importance of consulting a healthcare provider, indicating a cautious approach. Owing to constraints in both timeliness and accuracy inherent in language models, ChatGPT-3.5 occasionally emphasized its non-doctor status, thus refraining from providing direct diagnosis or therapy in the conversation. However, such statements may imply the unreliability of the medical judgment, especially in Chinese culture context. Thus, further inquiries are warranted to evaluate the potential risks and benefits of this response mode, considering its impact on patient trust and compliance challenges.

Implications for Future Development in Clinical Medicine

As artificial intelligence, including large language models, progressively integrates into clinical medicine, understanding the advantages and disadvantages is paramount. While ChatGPT demonstrates promise as a medical consulting assistant for chronic hepatitis B patients, future research and development should prioritize targeted language input and emotional management training. Besides, establishing and updating prompts, which are specific order and templates based on which LLMs could provide responses in a standardized format would significantly enhance ChatGPT's performance. Overcoming language barriers and addressing emotional support deficiencies will be crucial for maximizing the potential benefits of large language models in medical assistance.

ChatGPT-4.0 compared to ChatGPT-3.5

ChatGPT-4.0 demonstrated superior performance than ChatGPT-3.5 in terms of information accuracy. This improvement aligns with the expected advancements in ChatGPT-4.0 as a more advanced iteration. However, ChatGPT-4.0 did not exhibit better response stability in open-ended questions. This could be attributed to a reduced ability to follow chain-of-thought prompting [31]. Despite this inconsistency, it did not affect the accuracy of information, suggesting that LLMs tend to employ diverse language patterns and content combinations.

In responses to closed questions (30 true-or-false statements based on the AASLD Guideline of Treatment of Hepatitis B), ChatGPT-4.0 demonstrated a higher rate of accuracy and stability, indicating substantial improvement in the model's understanding of hepatitis B medical knowledge as the model progressed.

The improvement of ChatGPT-4.0 in terms of information accuracy suggested the tremendous benefit of model update, but the deficiency in emotional management maintained. Therefore, additional training related to emotion management guidance and humanistic care is essential for the preparation of the model before application.

Notably, in responses to open questions, ChatGPT-4.0 displayed interesting changes compared to

ChatGPT-3.5. ChatGPT-4.0 included reference information in 5 of the responses, all of which were verified to be accurate. This suggests an enhancement in the format and reliability of ChatGPT-4.0. However, the impact of such changes on the patient experience warrants further exploration. Additionally, ChatGPT-4.0 was more likely to use direct disclaimer like “I am an AI model...” or “I’m not a doctor...”, and even “Disclaimer: I’m not a doctor...”, indicating a more stringent approach. However, the increase of possibility was too subtle to be considered as significant.

Comparison to Prior Work

Numerous studies have explored the potential application of ChatGPT in clinical practice. John W Ayers and colleagues observed that ChatGPT tends to deliver longer, more empathetic responses of higher quality compared to real doctors [32]. In a study by Marco Cascella and team, ChatGPT demonstrated proficiency in composing medical notes for ICU patients and scientific writing, despite lacking medical expertise. The researchers highlighted the model's effectiveness in providing medical advice and its potential in patient communication [12]. Several studies have emerged, evaluating ChatGPT's responses in various medical specialties [14, 33-35]. In contrast, our study uniquely focuses on ChatGPT's cross-language performance in clinical counseling, revealing language choice impacts accuracy and answer stability. This emphasizes the importance of language selection for practical applications of LLMs.

Limitations

It's important to acknowledge certain limitations in our study. The evaluation did not comprehensively assess ChatGPT's knowledge and ability in guiding emotional management for patients due to the questionnaire resource constraints. The lack of standardized questionnaire also limited the reliability of the questionnaire we use for lack of related inter-rater reliability measure. Meanwhile, as the first work in hepatitis B medical consulting AI assessment, it was difficult to estimate the possible affect of vagueness, ambiguity and other understanding due to the grammar mistake or vagueness of the questions. The researchers revised the questions to address such concerns, which created new concerns about discrepancy between these “standard” questions and practical application scenario. These problems should be fixed in the future research. Additionally, while cautionary statements promote responsible usage, the potential risks and benefits of this approach require further exploration. Future studies should address these limitations for a more comprehensive understanding of ChatGPT-'s application in medical assistance.

Conclusion

In summary, ChatGPT-3.5 exhibits promising capabilities as a medical consulting assistant, providing accurate yet occasionally less comprehensive information. As the improved version of the model, ChatGPT-4.0 showed stronger application potential than ChatGPT-3.5. Recognizing their limitations in emotional support and language-specific performance, future developments should prioritize targeted language training and enhanced emotional management features. While cautionary statements underscore responsible usage, the model's potential in aiding chronic hepatitis B patients is evident. As artificial intelligence continues shaping medical practices, refining LLMs for nuanced healthcare contexts is imperative. Striking a balance between linguistic accuracy, emotional sensitivity, and ethical patient engagement remains key for successful integration into clinical settings.

Acknowledgments

Yijie Wang and Yining Chen conceived the study, collected and filtered questions and developed assessment criteria. Yijie Wang translated and polished the questions, conducted questionnaire assessment process and wrote the manuscript. Yining Chen revised the manuscript. Jifang Sheng

directed the entire research process, invited two independent reviewers and acted as the final senior specialist. This study was supported by the Young Scientists Fund of the National Natural Science Foundation of China (no.82100640). The study was also assisted by Ruihong Zhao, Yu Shi and Hong Zhao.

Conflict of interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Abbreviations

CHB: Chronic hepatitis B
WHO: World Health Organization
HBV: Hepatitis B virus
EASL: European Association for the Study of the Liver
LLM: Large Language Model
CDC: Centers for Disease Control and Prevention
AASLD: American Association for the Study of Liver Disease
APASL: Asian Pacific Association for the Study of the Liver

Reference

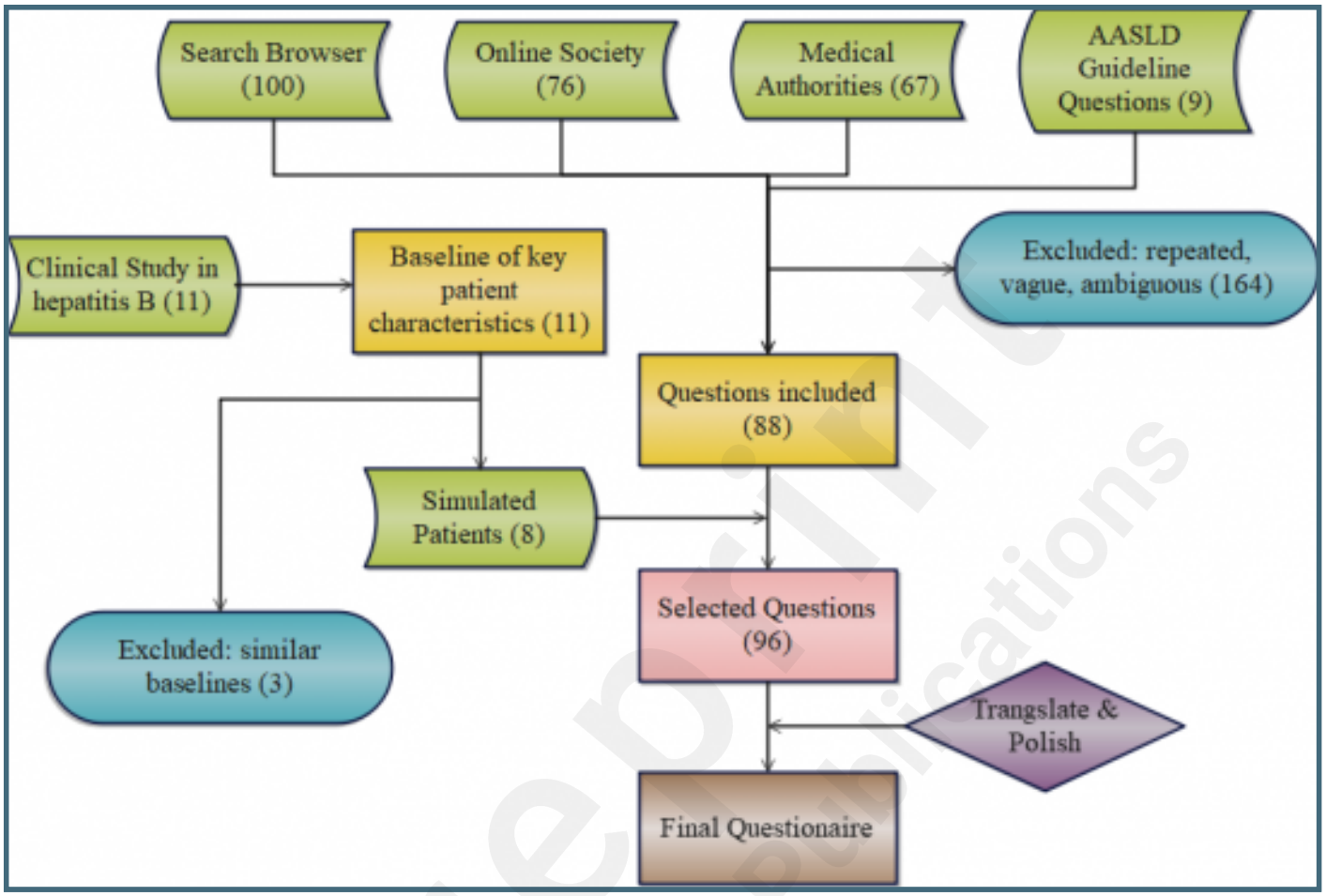
1. *WHO Key Facts about Hepatitis B*. 2023/07/18; Available from: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b.
2. *Global hepatitis report, 2017*. 19 April 2017; Available from: https://www.who.int/publications/i/item/9789241565455.
3. Han, S.H. and T.T. Tran, *Management of Chronic Hepatitis B: An Overview of Practice Guidelines for Primary Care Providers.* J Am Board Fam Med, 2015. **28**(6): p. 822-37.DOI: 10.3122/jabfm.2015.06.140331.
4. *EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection.* J Hepatol, 2017. **67**(2): p. 370-398.DOI: 10.1016/j.jhep.2017.03.021.
5. Degasperi, E., M.P. Anolli, and P. Lampertico, *Towards a Functional Cure for Hepatitis B Virus: A 2022 Update on New Antiviral Strategies.* Viruses, 2022. **14**(11).DOI: 10.3390/v14112404.
6. Appleton, A.A., et al., *Divergent associations of adaptive and maladaptive emotion regulation strategies with inflammation.* Health Psychol, 2013. **32**(7): p. 748-56.DOI: 10.1037/a0030068.
7. de Ridder, D., et al., *Psychological adjustment to chronic disease.* Lancet, 2008. **372**(9634): p. 246-55.DOI: 10.1016/S0140-6736(08)61078-8.
8. Wani, S.U.D., et al., *Utilization of Artificial Intelligence in Disease Prevention: Diagnosis, Treatment, and Implications for the Healthcare Workforce.* Healthcare (Basel), 2022. **10**(4).DOI: 10.3390/healthcare10040608.
9. Haug, C.J. and J.M. Drazen, *Artificial Intelligence and Machine Learning in Clinical Medicine, 2023.* N Engl J Med, 2023. **388**(13): p. 1201-1208.DOI: 10.1056/NEJMra2302038.
10. Au, K. and W. Yang, *Auxiliary use of ChatGPT in surgical diagnosis and treatment.* Int J Surg, 2023. **109**(12): p. 3940-3.DOI: 10.1097/js9.0000000000000686.
11. van Dis, E.A.M., et al., *ChatGPT: five priorities for research.* Nature, 2023. **614**(7947): p. 224-226.DOI: 10.1038/d41586-023-00288-7.
12. Cascella, M., et al., *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.* J Med Syst, 2023. **47**(1): p. 33.DOI: 10.1007/s10916-023-01925-4.
13. Gilson, A., et al., *How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment.* JMIR Med Educ, 2023. **9**: p. e45312.DOI: 10.2196/45312.
14. Yeo, Y.H., et al., *Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma.* Clin Mol Hepatol, 2023. **29**(3): p. 721-732.DOI: 10.3350/cmh.2023.0089.
15. Liu, J., C. Wang, and S. Liu, *Utility of ChatGPT in Clinical Practice.* J Med Internet Res, 2023. **25**: p. e48568.DOI: 10.2196/48568.
16. Deng, L., et al., Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg, 2024. 110(4): p. 1941-1950.DOI: 10.1097/JS9.0000000000001066.
17. Frosolini, A., et al., Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol, 2023. 280(11): p. 5129-5133. DOI: 10.1007/s00405-023-08205-4.
18. Lee, T.J., et al., Evaluating ChatGPT-3.5 and ChatGPT-4.0 Responses on Hyperlipidemia for Patient Education. Cureus, 2024. 16(5): p. e61067. DOI: 10.7759/cureus.61067.
19. Zheng, H., et al., *Compliance among infants exposed to hepatitis B virus in a post-vaccination serological testing program in four provinces in China.* Infect Dis Poverty, 2019. **8**(1): p. 57.DOI: 10.1186/s40249-019-0568-y.
20. Wang, M.L. and E.Q. Chen, *[Impact of treatment compliance in chronic hepatitis B].* Zhonghua Gan Zang Bing Za Zhi, 2022. **30**(11): p. 1266-1269.DOI: 10.3760/cma.j.cn501113-20201201-00635.
21. Zhou, X., et al., *Diagnosis experiences from 50 hepatitis B patients in Chongqing, China: a qualitative study.* BMC Public Health, 2021. **21**(1): p. 2195.DOI: 10.1186/s12889-021-11929-9.
22. Tutuncu, E.E., et al., *Adherence to Nucleoside/Nucleotide Analogue Treatment in Patients with Chronic Hepatitis B.* Balkan Med J, 2017. **34**(6): p. 540-545.DOI: 10.4274/balkanmedj.2016.1461.
23. Ozyigitoglu, D., et al., *Adherence to Treatment with Oral Nucleoside/Nucleotide Analogs in Patients with Chronic Hepatitis B.* Sisli Etfal Hastan Tip Bul, 2022. **56**(4): p. 543-551.DOI: 10.14744/SEMB.2022.82608.
24. Jiang, H., et al., *The main features of physician assistants/associates and insights for the development of similar professions in China.* J Evid Based Med, 2022. **15**(4): p. 398-407.DOI: 10.1111/jebm.12504.
25. Yu, Q., et al., *Trend and equity of general practitioners' allocation in China based on the data from 2012-2017.* Hum Resour Health, 2021. **19**(1): p. 20.DOI: 10.1186/s12960-021-00561-8.
26. Wen, J., et al., *Workload, burnout, and medical mistakes among physicians in China: A cross-sectional study.* Biosci Trends, 2016. **10**(1): p. 27-33.DOI: 10.5582/bst.2015.01175.
27. Wei, L., et al., *Treating chronic hepatitis B virus: Chinese physicians' awareness of the 2010 guidelines.* World J Hepatol, 2016. **8**(18): p. 762-9.DOI: 10.4254/wjh.v8.i18.762.

28. Yu, P., et al., *Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study.* JMIR Med Educ, 2024. **10**: p. e48514.DOI: 10.2196/48514.
29. Brown, T., et al., Language models are few-shot learners. Advances in neural information processing systems, 2020. 33: p. 1877-1901.DOI: 10.48550/arXiv.2005.14165.
30. Wierenga, K.L., R.H. Lehto, and B. Given, *Emotion Regulation in Chronic Disease Populations: An Integrative Review.* Res Theory Nurs Pract, 2017. **31**(3): p. 247-271.DOI: 10.1891/1541-6577.31.3.247.
31. Chen, L., M. Zaharia, and J.J.a.p.a. Zou, How is ChatGPT's behavior changing over time? 2023.DOI: 10.48550/arXiv.2307.09009.
32. Ayers, J.W., 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): p. 589-596.DOI: 10.1001/jamainternmed.2023.1838.
33. Grünebaum, A., et al., *The exciting potential for ChatGPT in obstetrics and gynecology.* Am J Obstet Gynecol, 2023. **228**(6): p. 696-705.DOI: 10.1016/j.ajog.2023.03.009.
34. Fournier, A., et al., *Assessing the applicability and appropriateness of ChatGPT in answering clinical pharmacy questions.* Ann Pharm Fr, 2023.DOI: 10.1016/j.pharma.2023.11.001.
35. Antaki, F., et al., *Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings.* Ophthalmol Sci, 2023. **3**(4): p. 100324.DOI: 10.1016/j.xops.2023.100324.

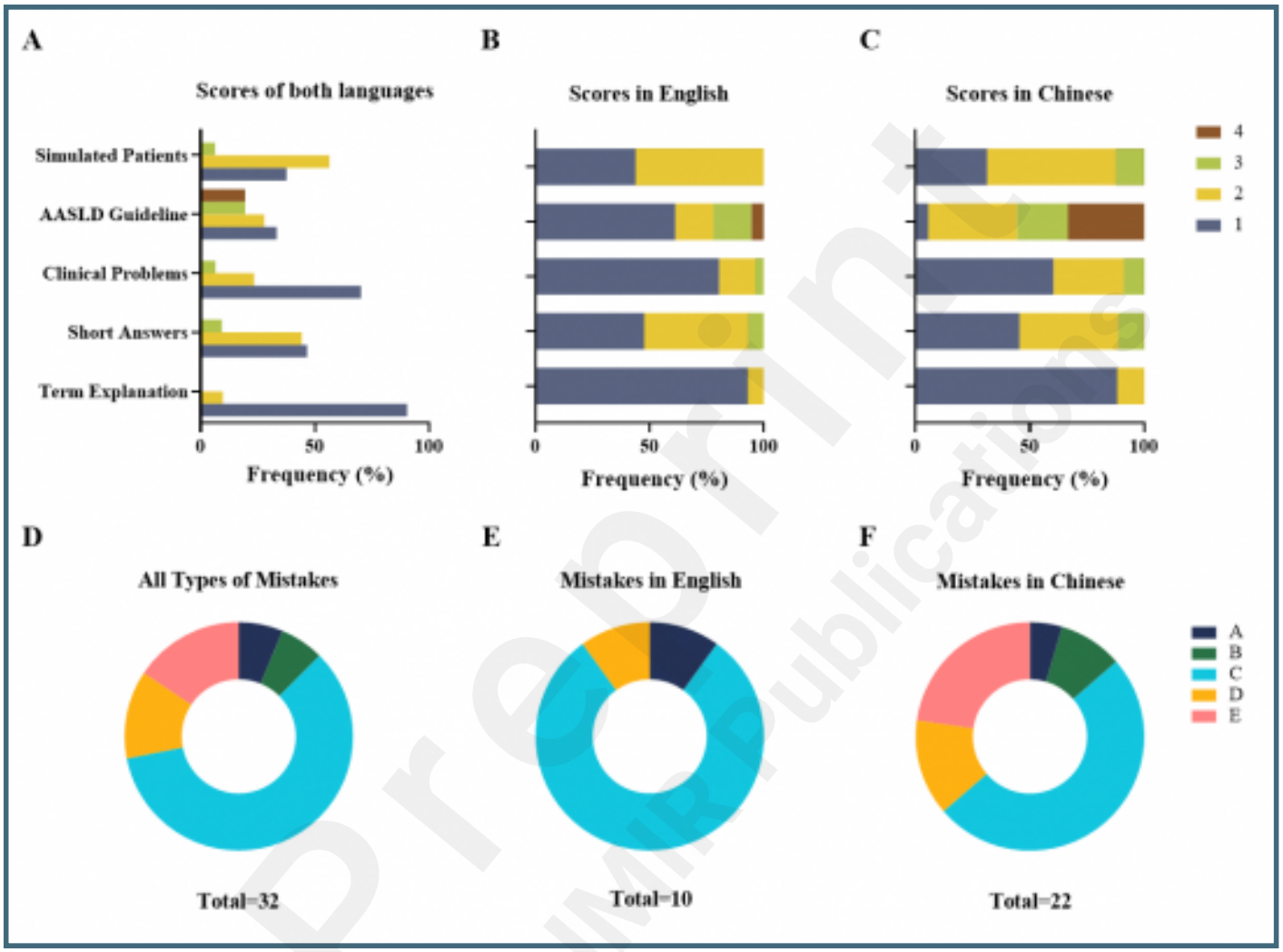
Supplementary Files

Figures

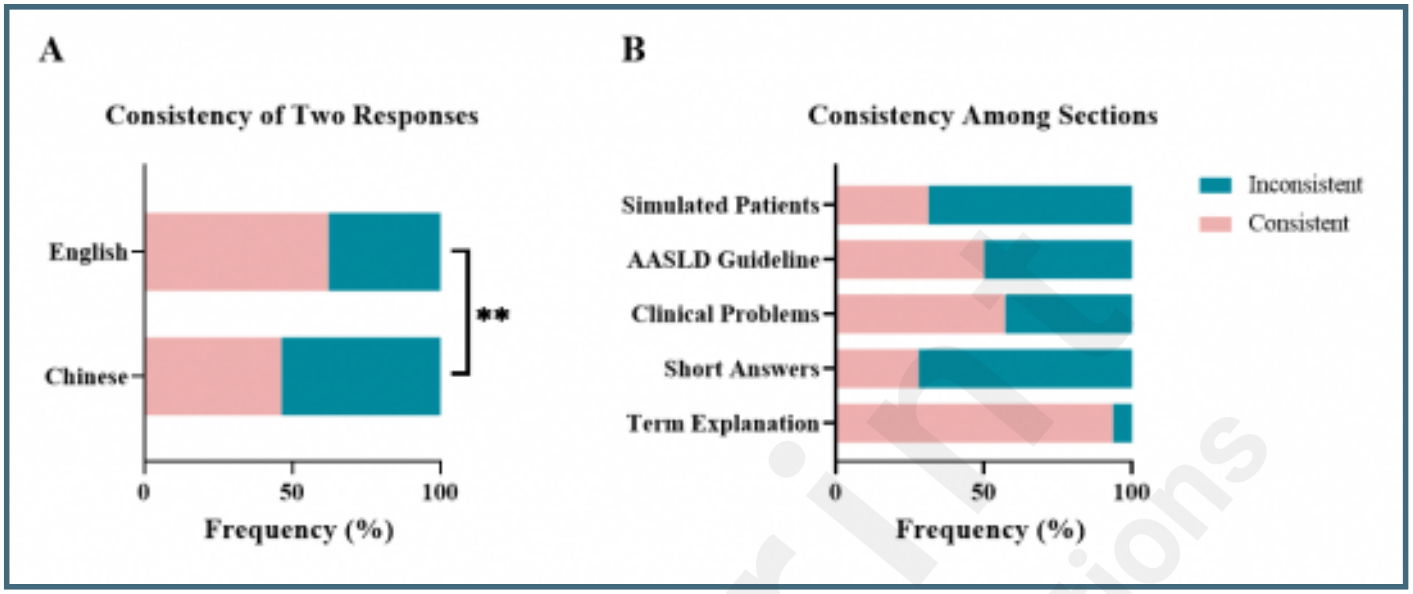
Workflow of questionnaire design process. The figure shows specific information of each stage of the questionnaire compiling process.



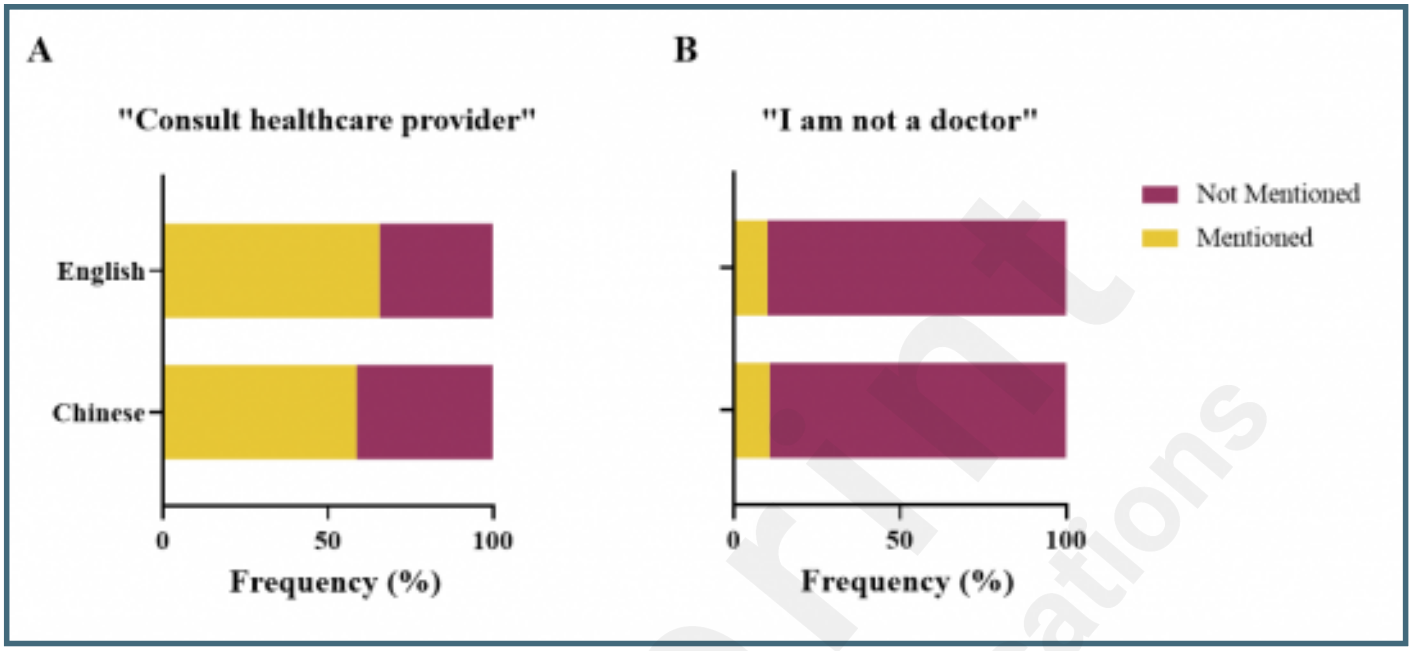
Results of information accuracy assessment and mistake analysis of ChatGPT-3.5. A. Comparation of percentages for each grade across all responses in distinct question sections. B. Percentages of each grade of responses in English in separated question sections. C. Percentages of each grade of responses in Chinese in separated question sections. D. Overview of mistake types across all responses. E. Breakdown of mistake types specifically among responses in English. G. Breakdown of mistake types specifically among responses in Chinese.



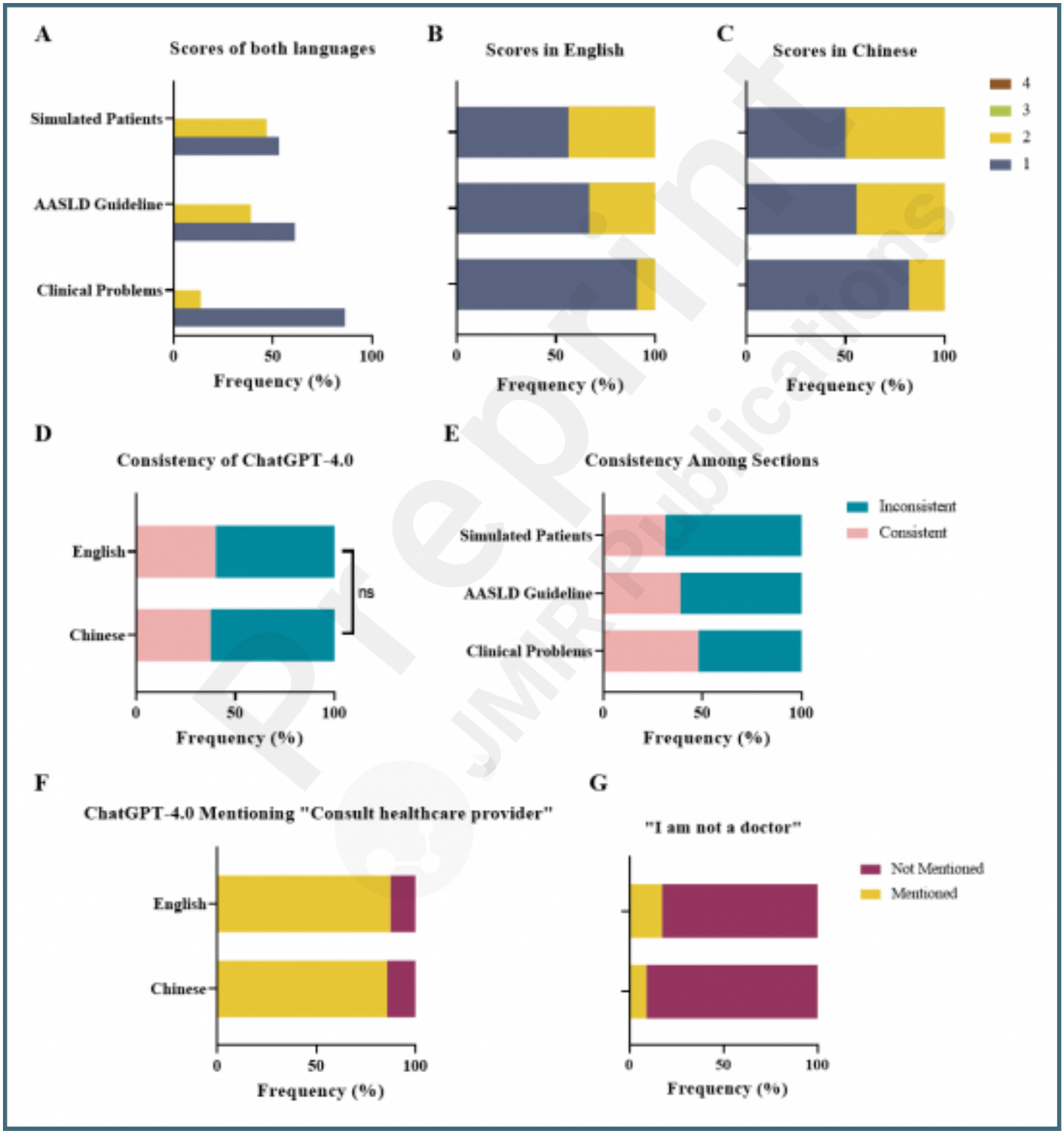
Assessment of content consistency of responses to repeated questions. A. Comparison of content consistency between responses in different working languages. B. Examination of content consistency in different sections of questions.



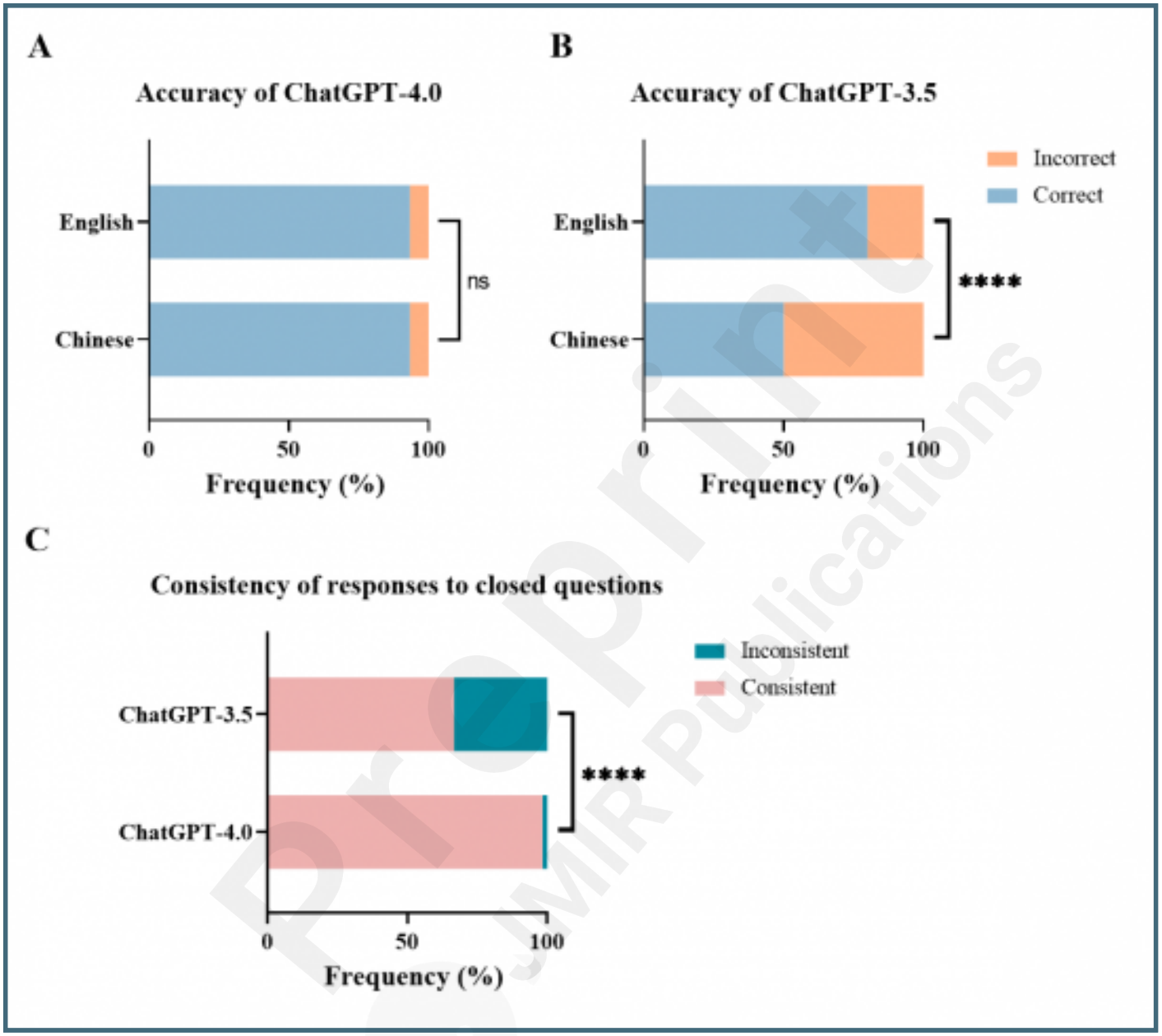
Percentages of ChatGPT cautionary statements regarding medical advice and disclaimers. A. Percentages of responses that include the recommendation to "consult healthcare providers or doctors". B. Percentages of responses containing the disclaimer phrases "I am not a doctor" or "I cannot give diagnosis or treatment".



Results of parallel assessment of ChatGPT-4.0 in sections involving clinical practice. A. Comparation of percentages for each grade across all responses in distinct question sections. B. Percentages of each grade of responses in English in separated question sections. C. Percentages of each grade of responses in Chinese in separated question sections. D. Comparison of content consistency between responses in different working languages. E. Examination of content consistency in different sections of questions. F. Percentages of responses that include the recommendation to "consult healthcare providers or doctors". G. Percentages of responses containing the disclaimer phrases "I am not a doctor" or "I cannot give diagnosis or treatment".



Results of responses to true-or-false questions. A. Rate of Accuracy of ChatGPT-4.0. B. Rate of Accuracy of ChatGPT-3.5. C. Compare of the consistency of responses to true-or-false questions between ChatGPT-4.0 and ChatGPT-3.5.



Multimedia Appendixes

Examples of questions revised or eliminated.
URL: http://asset.jmir.pub/assets/51feffb8bc5bc6238f624133bf2b61a6.docx

Summary of information accuracy grades and consistency of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/1cb9a0fe308a5e026f48532a1542a925.docx

Assessment of information accuracy of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/4a30e2b0c7d3c56a957df4495d9f7be9.docx

Results of mistake type evaluation of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/169b1345873253895c24575575816681.docx

Content consistency assessment of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/c48fad2b7247dc95f9631c1f0ad0e21d.docx

Responses provided with sufficient emotional management guidance of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/7b8693d22092114de782687cc0741a36.docx

Evaluation on whether responses mentioned about consulting healthcare providers or doctors of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/cd27152b4574f4865e65b8235feb794e.docx

Responses mentioned “I am not a doctor” or “I cannot give diagnosis or treatment” of ChatGPT-3.5.
URL: http://asset.jmir.pub/assets/919eea58a5accf32dba8ffc893d21e03.docx

Summary of information accuracy grades and consistency of ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/b61e6034a62b08c28d468217fa0c59ec.docx

Assessment of information accuracy of ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/b821cb7eea6d6b1da529a7464efc64fe.docx

Content consistency assessment of ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/51923ffe8c1ce121733e09d9aa72f618.docx

Responses provided with sufficient emotional management guidance of ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/0200e34dc3d83567cd0862149e55866e.docx

Evaluation on whether responses mentioned about consulting healthcare providers or doctors of ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/93bcfa402315d9e881b682912d1da24d.docx

Responses mentioned “I am not a doctor” or “I cannot give diagnosis or treatment” of ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/169c2f817091802633dda95c09bfee51.docx

Results of responses to closed questions (true-or-false) of ChatGPT-3.5 and ChatGPT-4.0.
URL: http://asset.jmir.pub/assets/9108ba721f34ed2ea6f80ba020cc0019.docx