



## OPEN ACCESS

## EDITED BY

Weihua Yang,  
Southern Medical University, China

## REVIEWED BY

Shulin Liu,  
First Affiliated Hospital of Chongqing Medical  
University, China  
Hanyi Yu,  
South China University of Technology, China

## \*CORRESPONDENCE

Xiao Huang  
✉ sophiahx@foxmail.com  
Ruili Wei  
✉ ruiliwei@smmu.edu.cn  
Yuqing Chen  
✉ yuqingchan@126.com

<sup>†</sup>These authors have contributed equally to  
this work and share first authorship

RECEIVED 30 December 2024

ACCEPTED 27 January 2025

PUBLISHED 18 February 2025

## CITATION

Wang Y, Yang S, Zeng C, Xie Y, Shen Y, Li J,  
Huang X, Wei R and Chen Y (2025) Evaluating  
the performance of ChatGPT in patient  
consultation and image-based preliminary  
diagnosis in thyroid eye disease.  
*Front. Med.* 12:1546706.  
doi: 10.3389/fmed.2025.1546706

## COPYRIGHT

© 2025 Wang, Yang, Zeng, Xie, Shen, Li,  
Huang, Wei and Chen. This is an open-access  
article distributed under the terms of the  
[Creative Commons Attribution License  
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction  
in other forums is permitted, provided the  
original author(s) and the copyright owner(s)  
are credited and that the original publication  
in this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Evaluating the performance of ChatGPT in patient consultation and image-based preliminary diagnosis in thyroid eye disease

Yue Wang<sup>1†</sup>, Shuo Yang<sup>1†</sup>, Chengcheng Zeng<sup>1†</sup>, Yingwei Xie<sup>2</sup>,  
Ya Shen<sup>1</sup>, Jian Li<sup>1</sup>, Xiao Huang<sup>1\*</sup>, Ruili Wei<sup>1\*</sup> and Yuqing Chen<sup>1\*</sup>

<sup>1</sup>Department of Ophthalmology, Changzheng Hospital of Naval Medical University, Shanghai, China, <sup>2</sup>Department of Urology, Beijing Tongren Hospital of Capital Medical University, Beijing, China

**Background:** The emergence of Large Language Model (LLM) chatbots, such as ChatGPT, has great promise for enhancing healthcare practice. Online consultation, accurate pre-diagnosis, and clinical efforts are of fundamental importance for the patient-oriented management system.

**Objective:** This cross-sectional study aims to evaluate the performance of ChatGPT in inquiries across ophthalmic domains and to focus on Thyroid Eye Disease (TED) consultation and image-based preliminary diagnosis in a non-English language.

**Methods:** We obtained frequently consulted clinical inquiries from a published reference based on patient consultation data, titled *A Comprehensive Collection of Thyroid Eye Disease Knowledge*. Additionally, we collected facial and Computed Tomography (CT) images from 16 patients with a definitive diagnosis of TED. From 18 to 30 May 2024, inquiries about the TED consultation and preliminary diagnosis were posed to ChatGPT using a new chat for each question. Responses to questions from ChatGPT-4, 4o, and an experienced ocular professor were compiled into three questionnaires, which were evaluated by patients and ophthalmologists on four dimensions: accuracy, comprehensiveness, conciseness, and satisfaction. The preliminary diagnosis of TED was deemed accurate, and the differences in the accuracy rates were further calculated.

**Results:** For common TED consultation questions, ChatGPT-4o delivered more accurate information with logical consistency, adhering to a structured format of disease definition, detailed sections, and summarized conclusions. Notably, the answers generated by ChatGPT-4o were rated higher than those of ChatGPT-4 and the professor, with accuracy (4.33 [0.69]), comprehensiveness (4.17 [0.75]), conciseness (4.12 [0.77]), and satisfaction (4.28 [0.70]). The characteristics of the evaluators, the response variables, and other quality scores were all correlated with overall satisfaction levels. Based on several facial images, ChatGPT-4 twice failed to make diagnoses because of lacking characteristic symptoms or a complete medical history, whereas ChatGPT-4o accurately identified the pathologic conditions in 31.25% of cases (95% confidence interval, CI: 11.02–58.66%). Furthermore, in combination with CT images, ChatGPT-4o performed comparably to the professor in terms of diagnosis accuracy (87.5, 95% CI 61.65–98.45%).

**Conclusion:** ChatGPT-4o excelled in comprehensive and satisfactory patient consultation and imaging interpretation, indicating the potential to improve

clinical practice efficiency. However, limitations in disinformation management and legal permissions remain major concerns, which require further investigation in clinical practice.

#### KEYWORDS

thyroid eye disease, large language model, ChatGPT, virtual healthcare, clinical practice

## 1 Introduction

Equipped with extensive medical information and built-in mechanisms for self-checking (1), Large Language Model (LLM) chatbots represent notable advancements in artificial intelligence (AI), acting as one of the advanced technologies for practical tools in virtual healthcare (2). ChatGPT-4 (released on 14 March 2023) has demonstrated human-level proficiency across a range of professional and academic benchmarks, particularly in terms of factual accuracy and adaptability. A subsequent iteration, ChatGPT version 4o (“o” for “omni”), released on 13 May 2024, mainly pre-trained by publicly available data such as diverse web articles and videos, further improves the naturalness of human-computer interaction, particularly in non-English languages, enabling its broader global application (“GPT-4o System Card,” (3, 35).

Thyroid Eye Disease (TED), also known as Thyroid-associated Ophthalmopathy or Graves’ Ophthalmopathy, is a chronic and debilitating autoimmune disorder characterized by infiltrative and proliferative lesions in the retrobulbar and periorbital tissues (4, 5). Because some clinicians are unfamiliar with TED, the condition is frequently misdiagnosed, resulting in delayed or inappropriate treatment (6). Significant changes in appearance, such as severe eyelid retraction, pronounced exophthalmos, and corneal ulceration, have profound negative effects on patients’ work and daily lives, often leading to psychological distress, feelings of inferiority, and reduced social interactions (7, 8).

Importantly, with the aim of promoting public health, almost all hospital departments provide patients with educational materials, allowing them to access essential health information. However, patients tend to search for disease-related topics, focusing mainly on new or inappropriate therapies that can induce unrealistic expectations toward outcomes and potential disputes between clinicians and patients (9). Several studies have explored the role of ChatGPT in the medical field, with promising results in clinical domains. LLM chatbots can also assist in preliminary assessments based on patients’ reported symptoms in a multilingual global environment (10), particularly in remote regions where specialized ophthalmologists are scarce (2, 11).

A key gap in existing research is the lack of patient feedback on the quality of information generated by LLM chatbots. This study seeks to address that gap by involving patients with TED and subspecialty ophthalmologists to evaluate the quality of the responses generated in terms of accuracy, comprehensiveness, conciseness, and overall satisfaction. In this cross-sectional study, the objective of the study was to evaluate the performance of ChatGPT-4 and 4o in patient consultation from multiple dimensions and further explored whether ChatGPT could perform preliminary diagnoses based on images, serving as a valuable supplementary tool in clinical practice.

## 2 Methods

### 2.1 Questions and image source

The common consultation questions in this study were sourced from a printed patient education book, edited by R.L.W., titled *A Comprehensive Collection of Thyroid Eye Disease Knowledge*, which contains 141 commonly asked questions related to TED, focusing on disease treatment and rehabilitation (12). The questions were selected based on more than 40 years of clinical experience, relevance, and clarity from a popular online consultation platform, Haodafu Consultant App (Beijing Xinyi Qianguo Technology Co.). While excluding several typical prognosis reports, we only included the most straightforward questions for each type with similar meanings. After screening, 15 questions were finalized, covering a broad range of topics such as pathological mechanisms and treatment for TED. Answers to descriptive and binary questions contained elaborate explanations, so we identified the content categories (basic knowledge, diagnosis, mechanism, surgical, and non-surgical treatment, with three questions from each category) and difficulty levels (low, moderate, and high difficulty, with five questions from each category, in [Supplementary eTable 1](#)). In the image-based evaluation, 16 cases with a definitive diagnosis of TED were from the Shanghai Changzheng Hospital medical record system, each including several frontal and lateral facial appearance images and CT images from horizontal and coronal positions (detailed in [Supplementary eTable 2](#)).

### 2.2 Ethics approval

Ethical approval for this cross-sectional study was obtained from the Ethics Committee of the Changzheng Hospital Biomedical Research Ethics Committee (*Identifier: 2021SL044*). This research adhered to the principles outlined in the Declaration of Helsinki. Due to the utilization of de-identified images and personal information, the requirement for informed consent was waived.

### 2.3 Responses from ChatGPT

In response to the consultation questions, two versions of ChatGPT (Versions 4 and 4o, Open AI) were used from 18 May to 20 May 2024. Under inquiries, “Hi, ChatGPT. I am a patient with TED.” comment was posted. Following the common questions for patient consultation in Chinese, GPT-4 and GPT-4o generated responses for the 15 questions ([Supplementary eTable 3](#), translated by Grammarly software). All identifying elements were removed, including special characters, website links, greetings, etc. The objective characteristics,

the number of characters, paragraphs, segments, and photographs in the responses were collected.

When conducting the image-based evaluation, either facial personalized features below orbits or personal information had been mosaic processed, then ChatGPT-4 and 4o were used from 22 May to 30 May 2024. Under inquiries, “Hi, ChatGPT. Assuming as a certified clinician, please walk me through image-based diagnosis and list three in order of likelihood.” was posted in Chinese. Following images from patients with TED, ChatGPT (versions 4 and 4o) performed preliminary diagnoses for the 16 cases. The first section of the preliminary diagnosis was based on several facial images, followed by the section of several facial images combined with CT images (Supplementary eTable 2). In particular, every question was asked using a new chat to ensure consistency and avoid self-learning from the chat history in this study.

## 2.4 Response from the professor

Professor R.L.W.’s responses to 15 consultation questions were sourced from the publication he edited, *A Comprehensive Collection of Thyroid Eye Disease Knowledge* (Supplementary eTable 3, translated by Grammarly software). Objective characteristics, the number of characters, paragraphs, segments, and photographs in the answers were also collected. After observing the PowerPoint presentation that displayed facial images of a patient on one page and CT images of the same patients on the following page, the professor performed the preliminary diagnosis.

## 2.5 Quality evaluation of responses

The quality evaluation of responses to the consultation questions was performed as a questionnaire investigation. To enhance reliability, three questionnaires were designed to avoid simultaneously presenting the answers of three generators to the same question to the evaluators and following bias. The three questionnaires each contained the same questions about TED but with different corresponding answers, presented in a consistent and specific order. In other words, since the structure of the questionnaires remained uniform, the answers varied, featuring five responses from ChatGPT-4, five from ChatGPT-4o, and five from an ocular professor’s written publication, arranged randomly.

With the full informed consent of the patients and their accompanying persons, the paper questionnaires were distributed to patients visiting the consulting room and professional ophthalmologists (Department of Ophthalmology, Shanghai Changzheng Hospital). In addition to the instructions from the residents of the ophthalmologist face to face, there was a brief description of the essential matters at the beginning and some questions to collect the evaluator’s sex, age, disease course, duration, treatment experience, and level of knowledge of the disease in the patient questionnaires. In ophthalmologist questionnaires, we collected the evaluator’s sex, age, and professional titles using Likert scales. Despite 15 questions and their answers followed by four quality scores, there is an optional section to collect patient feedback (detailed in Supplementary eTable 4).

The voluntary patients in the questionnaire investigation were all literate, aged between 18 and 60 years old, with at least a preliminary

understanding of the TED and the courses of personal diseases. The professional participants were subspecialists in ophthalmology, with clinical experience assisting to diagnose and independently treating patients with TED for more than 2 years. Under specific guidance, all participants scored in a quiet environment without a time limit.

Concise and easy-to-understand materials are in great need for patients with poor disease knowledge. Therefore, evaluation tools were designed as predefined 5-point Likert scales, including common quality indicators (accuracy, comprehensiveness), conciseness, and overall satisfaction. The screening criteria were as follows: The questionnaires were considered invalid if at least one-third (five or more questions) scored the same on all four dimensions. Finally, 162 valid questionnaires from 12 ophthalmologists and 150 patients with TED were included (Figure 1A). When evaluating the performance of the image-based preliminary diagnosis, we excluded two non-definitive diagnoses from ChatGPT-4 and then calculated the accuracy rate of image interpretation to diagnose TED (Figure 1B).

## 2.6 Statistical analysis

The quality of responses to various difficulty and content questions about TED was compared in four dimensions: accuracy, comprehensiveness, conciseness, and satisfaction, using the Kruskal-Wallis test. We calculated the correlations between the characteristics of the evaluator, the quality evaluation scores, and the overall satisfaction scores using the Spearman’s correlation analysis. In addition, the independent samples *t*-test (Mann-Whitney test and Welch’s *t*-test) was used to compare quality scores between sex and disease courses using GraphPad Prism (version 10.2.0, GraphPad Software, Inc.). With numerous participants, the reliability or consistency of the questionnaire was verified using Cronbach’s alpha coefficient, and its validity was graded using the Kaiser-Meyer-Olkin and Bartlett sphere tests. The accuracy rate (95% confidence interval, CI) was calculated using Clopper-Pearson and the inter-group differences using the McNemar Chi-square test (SPSS Statistics, version 21.0, IBM). Descriptive analysis, such as median (interquartile range, IQR), mean (standard deviation, SD), and range, was used as appropriate in the manuscript.  $p < 0.05$  was considered statistically significant.

## 3 Results

### 3.1 Performance of ChatGPT in TED patient consultation

In combination with responses from ChatGPT-4 and 4o under inquiries to 15 consultation questions, responses generated by an ocular professor (R.L.W.) were collected (total responses for comparison:  $N = 45$ ). The quality variables ordered in three questionnaires were listed as 45 groups to evaluate reliability and validity. Cronbach’s alpha coefficient was 0.831, the Kaiser-Meyer-Olkin value was 0.778, the approximate chi-square value of Bartlett’s sphere test was 2649.992, the degree of freedom was 990, and the *p* value was 0.000, indicating good reliability of scales in questionnaires.

With joint efforts, the questionnaire investigation included 9,720 assessment items and 936 personal information items. The main

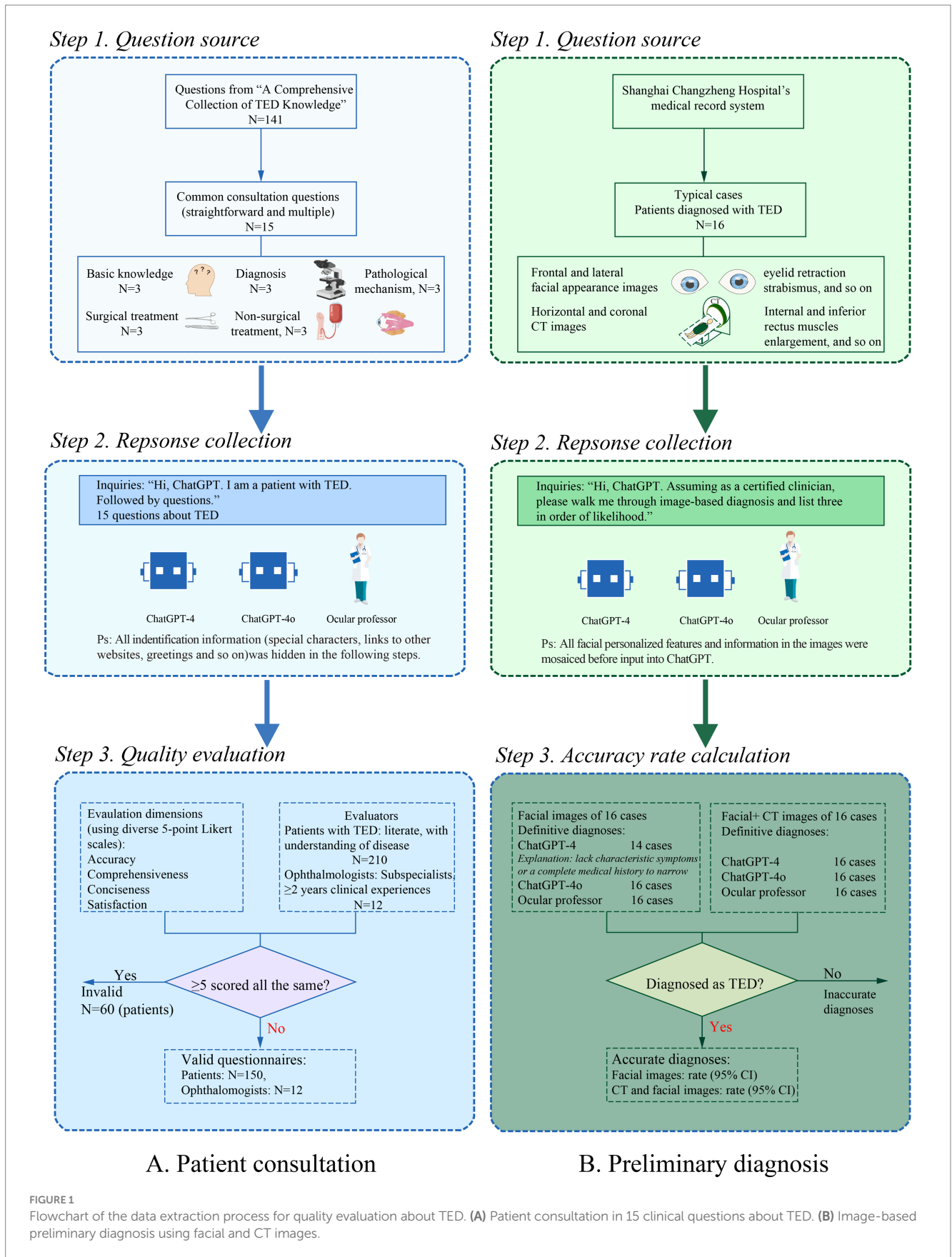


FIGURE 1

Flowchart of the data extraction process for quality evaluation about TED. (A) Patient consultation in 15 clinical questions about TED. (B) Image-based preliminary diagnosis using facial and CT images.

participants in the primary group were young and middle-aged women in the stable stage of TED who had moderate or higher knowledge levels and experiences of thyroid, intravenous glucocorticoid, or orbit surgeries (Table 1). All professional participants were junior to senior ophthalmologists at Shanghai Changzheng Hospital (Table 2).

The median (IQR) character count was 373 (274–4,449), while that of paragraphs, sections, and photographs was 5 (3–8), 3 (0–5), and 0 (0–0), respectively (Table 3). With hundreds of characters, separating the content into sections was more logical and beneficial to grasp key points. Since most of the answers generated by chatbots (version 4: 66.7%, 4o: 93.3%) remained a uniform structure, the mechanisms, advantages, and disadvantages of the clinical options were explained point by point. For example, in Question 6, interpreting the thyroid

TABLE 1 Personal information on patients participating in the questionnaire survey on consultation (N = 150).

Group	Number (Percentage)	
<b>Gender</b>		
	Male	38 (25.3%)
	Female	112 (74.7%)
<b>Ages (years)</b>		
	18–20	6 (4%)
	21–30	35 (23.3%)
	31–40	47 (31.3%)
	41–50	38 (25.3%)
	51–60	24 (16%)
<b>Disease stage</b>		
	Active	36 (24%)
	Stable	114 (76%)
<b>Disease course</b>		
	<6 months	6 (4%)
	Half-1 year	27 (18%)
	1–3 years	63 (42%)
	3–5 years	28 (18.7%)
	> 5 years	26 (17.3%)
<b>Treatment experiences<sup>a</sup></b>		
	Treatment on thyroid	110 (73.3%)
	Glucocorticoid	63 (42%)
	Radiotherapy	17 (11.3%)
	Monoclonal antibody	14 (9.3%)
	Orbital surgery	53 (35.3%)
<b>TED knowledge</b>		
	Poor	5 (3.3%)
	Inferior	24 (16%)
	Moderate	55 (36.7%)
	Basic	53 (35.3%)
	Excellent	13 (8.7%)

<sup>a</sup>Patients with TED may have experienced various treatments during the chronic disease course. Therefore, the total number of five different treatment experiences is more than 150, as the total percentage is more than 100%.

TABLE 2 Personal information on ophthalmologists participating in the questionnaire survey on consultation (N = 12).

Group	Number (Percentage)	
<b>Gender</b>		
	Male	6 (50%)
	Female	6 (50%)
<b>Age (years)</b>		
	18–20	0 (0%)
	21–30	8 (66.7%)
	31–40	3 (25%)
	41–50	1 (8.3%)
	51–60	0 (0%)
<b>Professional titles</b>		
	Residents	6 (50%)
	Attendings	5 (41.7%)
	Associate professor	1 (8.3%)

function checklist, ChatGPT-4o segmentally summarized three interpretation steps and gave examples of how indicators changed in hyperthyroidism and hypothyroidism, while the professor introduced their synthesis pathways and clinical meanings.

There were some inappropriate translations of terms in ChatGPT-4, which had been correctly updated in version 4o. In comparison with answers from ChatGPT-4, those of version 4o contained significantly more characters ( $p = 0.045$ ), and the same happened in the paragraphs and sections ( $p = 0.0034$  and  $p = 0.0005$ , respectively) between ChatGPT-4o and the professor (Table 4). In long text, visual impacts such as bold font and schematic diagrams are prominent to attract viewers' attention and improve understandability. Without special instruction, chatbots did not generate photographs in responses but highlighted key points in bold font at the front of each section. In the professor's answers, with several clear images, all paragraphs had first-line indentation without highlighting, in line with Chinese language habits.

### 3.2 Quality evaluation in different content and difficulty levels

From the perspective of patients, ChatGPT-4o performed the best in accuracy, comprehensiveness, conciseness, and satisfaction. In detail, the mean accuracy scores of ChatGPT-4, 4o, and professor showed no significant differences, all rated 4 on the 5-point scale (4.24 [0.74], 4.33 [0.69], 4.29 [0.72];  $p = 0.08$ ). In particular, the mean performance of ChatGPT-4o in comprehensiveness amounted to 4, surpassing version 4 (4.17 [0.75], 4.01 [0.78];  $p = 0.0001$ ). The same was observed between 4o and the professor in terms of satisfaction (4.28 [0.70], 4.18 [0.70];  $p = 0.0094$ ) and conciseness (4.12 [0.77], 3.95 [0.79];  $p < 0.0001$ , Table 2; Figure 2A). However, quality evaluation by ophthalmologists was not the same, showing no differences in all dimensions. An associate professor annotated the questionnaire in detail, highlighting some misinformation and briefly explaining several occasions (Table 5).

TABLE 3 The reactions of LLM chatbots and an experienced ocular professor to the 15 chosen thyroid eye disease inquiries.

Answer variables	ChatGPT-4	ChatGPT-4o	Ocular professor	Total	<i>p</i> value <sup>a</sup>
Characters Median (IQR)	313 (258–377)	404 (373–644)	360 (169–1,017)	373 (274–449)	0.049*
Paragraphs Median (IQR)	5 (3–6)	7 (5–10)	3 (1–9)	5 (3–8)	0.0045**
Sections Median (IQR)	3 (0–4)	5 (3–9)	0 (0–0)	3 (0–5)	0.0008***
Photographs Median (IQR)	0 (0–0)	0 (0–0)	0 (0–1)	0 (0–0)	0.0138*

<sup>a</sup>The *p* value was described as the Kruskal-Wallis test *p* value.

\* refers to the *p* < 0.05, \*\* refers to the *p* < 0.01, and \*\*\* refers to the *p* < 0.001.

TABLE 4 The performance of the LLM chatbots and an experienced ocular professor on different difficulty levels and content focuses.

Categories	Quality evaluation	ChatGPT-4	ChatGPT-4o	Professor	<i>p</i> value <sup>a</sup>
Low difficulty	Comprehensiveness	4.04 (0.82)	4.23 (0.73)	4.07 (0.79)	0.0126*
	Conciseness	4.13 (0.72)	4.18 (0.77)	4.05 (0.72)	0.0859
Moderate	Comprehensiveness	4.03 (0.72)	4.01 (0.78)	4.13 (0.86)	0.0387*
	Conciseness	4.19 (0.70)	4.03 (0.80)	4.01 (0.84)	0.0460*
High difficulty	Comprehensiveness	3.98 (0.78)	4.27 (0.71)	4.11 (0.70)	<0.0001****
	Conciseness	4.07 (0.78)	4.14 (0.74)	3.94 (0.79)	0.0160*
Diagnosis	Accuracy	4.25 (0.78)	4.49 (0.62)	4.35 (0.69)	0.0162*
	Comprehensiveness	4.02 (0.74)	4.27 (0.67)	4.26 (0.74)	0.0022**
	Conciseness	4.13 (0.74)	4.10 (0.82)	3.72 (0.81)	<0.0001****
	Satisfaction	4.15 (0.67)	4.36 (0.67)	4.17 (0.74)	0.0127*
Surgical treatment	Accuracy	4.25 (0.72)	4.35 (0.65)	4.46 (0.65)	0.0320*
	Comprehensiveness	4.06 (0.75)	4.22 (0.80)	4.25 (0.65)	0.0478*
	Conciseness	4.10 (0.79)	4.14 (0.69)	4.11 (0.74)	0.9739
	Satisfaction	4.17 (0.76)	4.28 (0.72)	4.33 (0.65)	0.1967
Total	Accuracy	4.24 (0.74)	4.33 (0.69)	4.28 (0.72)	0.0800
	Comprehensiveness	4.01 (0.78)	4.17 (0.75)	4.08 (0.75)	0.0002***
	Conciseness	4.13 (0.73)	4.12 (0.77)	3.95 (0.79)	<0.0001****
	Satisfaction	4.20 (0.72)	4.28 (0.70)	4.18 (0.70)	0.0102*

<sup>a</sup>All statistics were described as mean (SD), while the *p* value was referred to as the Kruskal-Wallis test *p* value.

\* refers to the *p* < 0.05, \*\* refers to the *p* < 0.01, \*\*\* refers to the *p* < 0.001, and \*\*\*\* refers to the *p* < 0.0001.

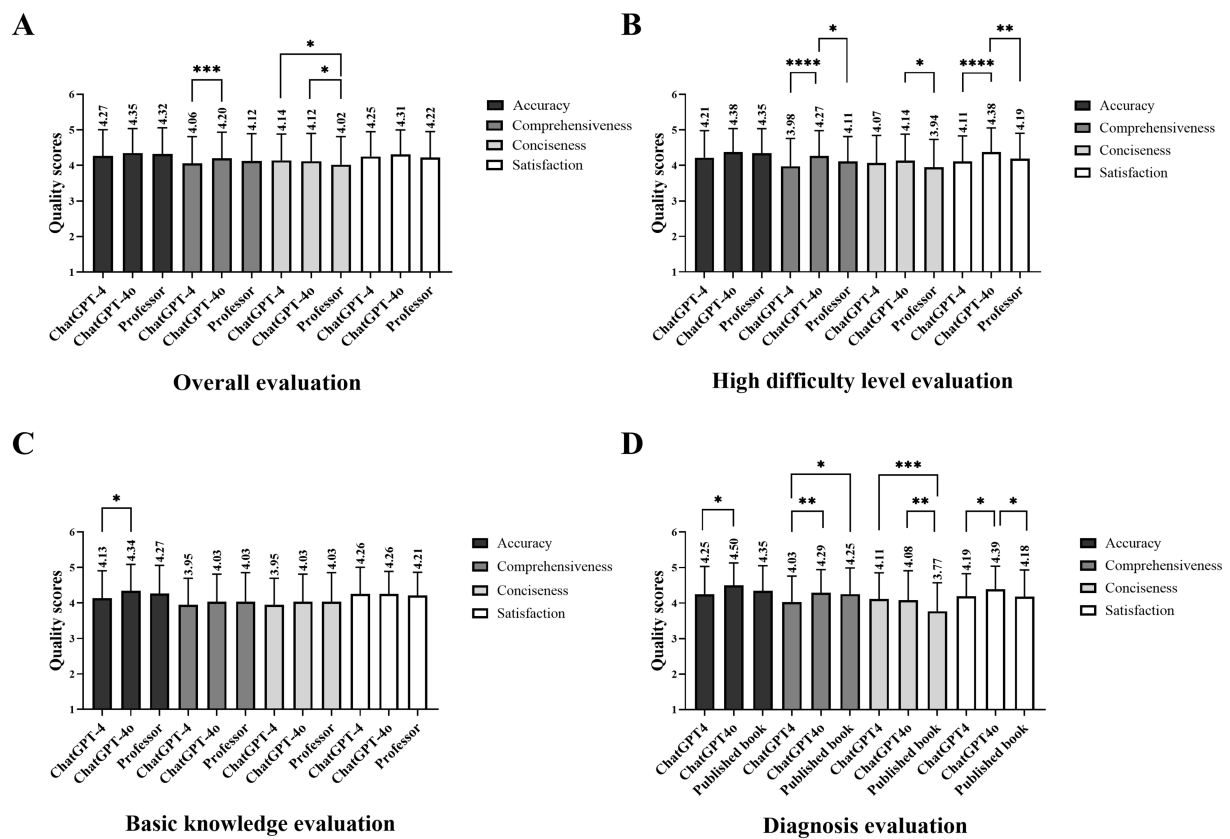
Regarding difficulty levels, ChatGPT-4o scored the highest and showed significant differences in comprehensiveness (low difficulty: *p* = 0.0126, moderate difficulty: *p* = 0.0387, high difficulty: *p* < 0.0001) and in high-difficulty level (accuracy: *p* = 0.0449, comprehensiveness: *p* < 0.0001, conciseness: *p* = 0.016, satisfaction: *p* < 0.0001). The accuracy scores of ChatGPT-4, 4o, and the ocular professor all amount to 4 (4.28 [0.74], 4.34 [0.71], and 4.34 [0.70], respectively). In terms of satisfaction and comprehensiveness, ChatGPT-4o's performance exceeded the professor for low (*p* = 0.2082 and *p* = 0.0478, respectively) and high-difficulty questions (*p* = 0.0043 and *p* = 0.0236, respectively, Table 2; Figure 2B).

There were inter-group differences in content focuses, including basic knowledge of the disease, diagnosis, and surgical treatment. The mean accuracy score of the answers generated by ChatGPT-4o was 4, surpassing ChatGPT-4 in the diagnosis section (4.49 [0.62] vs. 4.25 [0.78]; *p* = 0.0144, Figure 2C), behind the professor in surgical treatment (4.35 [0.65] vs. 4.46 [0.65]; *p* = 0.028, Figure 2D). The mean performance of ChatGPT-4 in the diagnosis section was unacceptable in terms of comprehensiveness (4.02 [0.74]), worse than version 4o

and professor (*p* = 0.008 and *p* = 0.0069, respectively). The mean conciseness scores of the professor lagged behind the chatbots (*p* both < 0.0001), while differences in satisfaction scores were found between ChatGPT-4 and 4o (4.15 [0.67] vs. 4.36 [0.67]; *p* = 0.0168, Figure 2C).

### 3.3 Performance of ChatGPT in image-based preliminary diagnosis

In specific inquiries on several facial appearance images, ChatGPT-4 failed to make possible diagnoses in two cases (2/16), complaining of the lack of characteristic symptoms or a complete medical history to narrow the general scope. For the three possible differential diagnoses, ChatGPT-4 demonstrated 4 correct diagnoses among the 14 cases with specific diagnoses (28.57%; 95% CI: 8.39–58.10%), while ChatGPT-4o correctly diagnosed 5 cases as TED (31.25%; 95% CI: 11.02–58.66%), showing no significant difference. Based on the facial appearance and CT images provided, both ChatGPT-4 and 4o achieved preliminary diagnoses in all 16 cases



**FIGURE 2** The performance of the LLM chatbot and the experienced ocular professor across quality dimensions. (A) Scores of overall quality evaluation. (B) Scores of quality evaluation on a high-difficulty level. (C) Scores of quality evaluation on diagnosis. (D) Scores of quality evaluation on surgical treatment (\* refers to the  $p < 0.05$ , \*\* refers to the  $p < 0.01$ , and \*\*\* refers to the  $p < 0.001$ ).

(16/16). When combined with CT images from the horizontal and coronal positions, ChatGPT-4 and 4o demonstrated 31.25% accuracy (95% CI 11.02–58.66%) and 87.5% accuracy (95% CI 61.65–98.45%,  $p = 0.004$ ), respectively.

### 3.4 Patients’ feedback on the questionnaire

Patients with TED have poor disease knowledge currently (13), partially because of deficiencies in high-quality patient consultation and education. In addition to the constant evolution of popularization, collecting patients’ evaluations and feedback is also necessary. The survey is a quality evaluation collection as well as a channel for acquiring knowledge and dispelling doubts about TED, and many patients take photographs for later learning. Furthermore, with expectations for disease popularization, 56 patients provided their opinions or advice, and the majority of them had positive comments. Astonishingly, a patient identified inappropriate information based on her great knowledge, while others concisely shared their personal experiences in diagnosis and treatment. However, several patients proposed a preference for treatment information, such as the approximate surgery cost, surgical indications, and complications, and a great need for professional explanation about eyelid retraction, the main reason for huge psychological pressure (14) (Supplementary eTable 5). These valuable

suggestions offer important directions for effective consultation and disease popularization.

### 3.5 Spearman’s correlation coefficients for the variables and satisfaction

Satisfaction, which is an overall reflection of the evaluator’s attitude, is worth determining whether it correlates with response variables and other quality scores. We discovered significant correlations between satisfaction, accuracy, comprehensiveness, and conciseness ( $r = 0.3179$ ,  $r = 0.3415$ ,  $r = 0.399$ , and  $p$  all  $< 0.0001$ , respectively). The sex and disease stages impacted the evaluation. In addition, in paragraphs and sections, the evaluators’ disease duration and knowledge levels were positively correlated with satisfaction scores ( $r = 0.0665$ ,  $p = 0.0016$ ;  $r = 0.0544$ ,  $p = 0.009$ ;  $r = 0.0502$ ,  $p = 0.0173$ ;  $r = 0.0550$ , and  $p = 0.0091$ , respectively, Table 6). From the perspective of ophthalmologists, satisfaction scores were positively correlated with accuracy, comprehensiveness, and conciseness ( $r = 0.5399$ ,  $p < 0.0001$ ;  $r = 0.6427$ ,  $p < 0.0001$ ;  $r = 0.2749$ , and  $p = 0.0002$ , respectively) and negatively correlated with the evaluator’s age and professional title ( $r = -0.2091$ ,  $p = 0.0048$ ;  $r = -0.2689$ , and  $p = 0.0003$ , respectively, Table 6). The dissimilar results revealed the potential biases between patients and ophthalmologists in the satisfaction evaluation.

TABLE 5 The earnest annotations of the questionnaires from an associate professor.

Question (content, difficulty)	Answers	Annotate with comments
Question 1. Do all patients with hyperthyroidism occur with Thyroid Eye Disease? (Basic knowledge, low difficulty)	<p>Ocular professor:</p> <p>Thyroid eye disease is often clinically manifested in two types: thyroid dysfunction ophthalmopathy (hyperthyroidism and hypothyroidism) and normal thyroid function ophthalmopathy. It is an autoimmune-related disease that is closely related to the immune response of the thyroid organ and is not directly related to hyperthyroidism.</p> <p>Thyroid eye disease occurs primarily in patients with Graves' disease, which accounts for about 80% of all patients and has symptoms of hyperthyroidism. Another 10% of patients with Hashimoto's thyroiditis (HT) have a more complex condition, showing hyperthyroidism, hypothyroidism, or normal thyroid function at different stages of their disease. In terms of time of onset, in about 10% of patients, thyroid-related eye disease occurs several years before thyroid disease. This condition is called euthyroid Graves' disease. These patients do not have any signs of thyroid disease, but a complete sequence of thyroid antibody detection (thyroid peroxidase antibodies, thyroglobulin antibodies, blocking, stimulating, and binding antibodies to TSH receptors, and thyroid growth-stimulating immunoglobulins) reveals in almost all of these patients autoimmune evidence of the thyroid gland. Therefore, not all thyroid eye disease is associated with hyperthyroidism.</p>	<p>Do all patients with TED occur with hyperthyroidism?</p> <p>The answer is inconsistent with the question.</p>
Question 4. What are the manifestations of Thyroid Eye Disease? Which one is the most characteristic and common? (Diagnosis, moderate difficulty)	<p>ChatGPT-4o:</p> <p>Protrusion (exophthalmos): This is the most characteristic and common manifestation of thyroid-associated ophthalmopathy. Exophthalmos causes the eye to look more prominent than usual.</p> <p>Eyelid swelling: Eyelid swelling may affect the appearance.</p> <p>Diplopia: Seeing double shadows when looking at objects, usually due to the involvement of extraocular muscles, resulting in uncoordinated eye movement.</p> <p>Dry eyes or tears: Eyes feel dry, uncomfortable, or excessive tears.</p> <p>Eye pain or tenderness: Pain or tenderness around the eyes, especially during activity.</p> <p>Light sensitivity: sensitive to light and uncomfortable.</p> <p>Visual acuity changes: severe cases may lead to blurred vision or loss of vision.</p> <p>Eyelid retraction: The upper eyelid retracts, and the white eye is exposed more, giving a "surprise" expression.</p> <p>Conjunctival congestion: Congestion or redness of the white part of the eye.</p> <p>Among these symptoms, protrusion (exophthalmos) is the most characteristic and common manifestation. Exophthalmos not only significantly affects the appearance but may also be accompanied by other symptoms, such as dry eyes, diplopia, and light sensitivity. Therefore, exophthalmos is often the first symptom that patients and doctors notice. If any of the above symptoms occur, especially with hyperthyroidism, it is recommended to consult an ophthalmologist or endocrinologist as early as possible for professional diagnosis and treatment.</p>	<p>Summary: Exophthalmos and eyelid retraction are the most characteristic and common symptoms for patients with TED.</p>
Question 14. I am a patient with Thyroid Eye Disease, but my thyroid function is not stable recently. Can I receive surgical treatment? (Surgical treatment, high difficulty)	<p>ChatGPT-4:</p> <p>The treatment of Thyroid Eye Disease usually requires a combination of multiple factors, including the stability of thyroid function.</p> <p>If thyroid function is unstable, surgical treatment may bring additional risks or complications. Usually, clinicians recommend stabilizing thyroid function first, which may be managed with drug therapy or other non-surgical treatments. After the stabilization of thyroid function, surgical treatment may be considered if eye symptoms are severe or progressive.</p> <p>It is recommended to communicate in detail with the medical team to understand all feasible treatment options and their potential pros and cons. The specific suitability for surgery is best determined by a joint assessment between endocrinologists and ophthalmologists, who provide personalized treatment recommendations, taking into account overall health, thyroid function test results, and the specific symptoms and severity of the eye disease.</p>	<p>Side effects: epilepsy, possible exacerbation of the Condition: weakened therapeutic effect.</p>



TABLE 6 Spearman's correlation analysis between response variables, other quality scores, and overall satisfaction.

Variables and analysis		Satisfaction from patients with TED	Satisfaction from professor
<b>Response variables</b>			
	Characters	$r = 0.0232$ , $p = 0.2704$	$r = 0.1348$ , $p = 0.0712$
	Paragraphs	$r = 0.0665$ , $p = 0.0016$	$r = 0.1379$ , $p = 0.0649$
	Sections	$r = 0.0544$ , $p = 0.009$	$r = 0.0882$ , $p = 0.2389$
	Images	$r = -0.0176$ , $p = 0.4052$	$r = 0.0893$ , $p = 0.2334$
<b>Quality scores</b>			
	Accuracy	$r = 0.3179$ , $p < 0.0001$	$r = 0.5399$ , $p < 0.0001$
	Comprehensiveness	$r = 0.3415$ , $p < 0.0001$	$r = 0.6427$ , $p < 0.0001$
	Conciseness	$r = 0.3990$ , $p < 0.0001$	$r = 0.2749$ , $p = 0.0002$

$r$  refers to Spearman's  $r$  values, while  $p$  refers to  $p$  (two-tailed).

## 4 Discussion

TED is an ocular disease related to endocrine disorders, and thousands of patients have consulted ophthalmologists and endocrinologists from nationwide hospitals but inevitably encountered difficulties during diagnosis and treatment. Due to the short inquiry time or inner anxiety emotions, sometimes they could not fully express their doubts and receive effective answers. In this case, almost every department delivers online educational health-disease materials to patients, enabling them to better understand their current conditions and assimilate treatment planning (15). In addition to the traditional Internet browsers and social media platforms, online AI-based tools like ChatGPT with encouraging performance on the simulated Medical Licensing Examination (11, 16, 17) and early detection and prediction of abnormalities (18–20) have gained increasing attention from patients seeking medical information.

This study serves as a foundation for assessing the practical application of LLM chatbots in a non-English language and promoting virtue healthcare innovation, especially in the ophthalmic field. Online, high-quality consultation and pre-diagnosis, in coping with clinical efforts, are fundamental in the effective patient-oriented management system. Abundant disease knowledge, identification of abnormal facial symptoms, especially eye appearances, and interpretation of imaging examination results are must-have skills for ophthalmologists as well as the key points in this study. We explored the capacity of the LLM chatbot (ChatGPT-4, 4o) to generate accurate, comprehensive, concise, and satisfactory responses to 15 frequently consulted questions and to conduct accurate preliminary diagnoses based on images in clinical practice. In

combination with the feedback collection, these responses, along with answers provided by an experienced ocular professor, were compiled into three questionnaires presented in random order. After exclusion, 162 valid questionnaires completed by 150 patients with TED and 12 verified ophthalmologists were included for further analysis.

The image interpretation, a new feature, has extended ChatGPT's capabilities beyond text (21, 22). The severity, course, and stage of the disease all have potential effects on the degree of exophthalmos and other symptoms in patients. In elderly patients, there were many wrinkles around the loose and swollen eyelids, which led to disease symptoms that were not clearly visible in images. The difficulty of image-based preliminary diagnosis varied significantly across different cases, which may cause selection bias. This should be discussed in further studies. For patients whose symptoms are not so obvious, the imaging examination results within the orbit are considered a supplement during TED diagnosis. Only based on several facial appearances and CT images, ChatGPT-4o diagnosed TED, displaying powerful capabilities of image information extraction and clinical matching, suppressing that of version 4. However, orbital cellulitis was the secondary preliminary diagnosis from ChatGPT, indicating a teaser deficiency in specialized information on the differential diagnosis of orbital lesions (23).

Every generator had its distinct features, and the LLM chatbot excelled in providing quantities of information, while the professor brought extensive clinical experiences into tailored advice. ChatGPT provided responses with a consistent structure: an introduction, point-by-point explanations, and a summary. There were several distinctive descriptions, including differences in indicator scope because of the equipment and the necessity to consult healthcare professionals for a personal plan. Compared with the cautious attitude in ChatGPT, the professor provided humanistic care in earnest.

With outstanding self-learning capacity, after some time, ChatGPT had significantly improved its responses to the same inquiries, especially on high-difficulty questions and image interpretation (24). We also observed notable advancements between versions (ChatGPT-4 and 4o), including amendments to translational errors inconsistent with language habits and improvements in comprehensiveness. Since long sentences and dense paragraphs make it difficult to maintain patience, it is necessary to separate into sections and use visual impacts such as bold fonts and website links, as in the logical and coherent responses from ChatGPT-4o. Therefore, unsatisfactory inquiries need to be repeated in further studies to verify self-learning and self-correct capabilities.

As seen in the feedback from patients, the majority of the patients expressed high satisfaction with the responses and proposed a preference for treatment information in further patient education, without identifying some of the answers generated by ChatGPT. Clinical decision-making recommendations are often considered weak spots of chatbots (16), classified as difficult-level questions in this study. ChatGPT-4o demonstrated prominence, much better than in previous research (25, 26), surpassing the professor in balancing accuracy, comprehensiveness, and conciseness. For example, in response to Question 11 (the timing of surgery), ChatGPT-4o pointed out that clinicians usually recommend surgery

under stable conditions as listed in the advantages and disadvantages. These results highlight the potential benefits of key points for evolving patient consultation and disease management.

The overall satisfaction among patients demonstrated a positive correlation with scores on accuracy, comprehensiveness, and conciseness, and the four dimensions should be guaranteed in consultation and education. However, it remains challenging for LLM chatbot to keep impeccable with the avoidance of inaccurate, incomprehensive information all the time (27). In this study, without specific instructions, ChatGPT generated answers including some abbreviations to refer to the disease in the whole answer only after a brief introduction. However, the appearance of multiple terms (GD, Graves' Disease; GO, Graves' Ophthalmopathy; TAO, Thyroid Eye Disease and TED, Thyroid Eye Disease) in a single questionnaire increased the complexity for patients, leading to confusion. The reason may be that the pre-trained data for each response varied, and ChatGPT failed to distinguish various terms. In conclusion, although the performance of ChatGPT-4o was satisfactory for most circumstances, there's still some distance from a perfect professor (11, 28).

Neither ChatGPT-4 nor 4o has the legal authority to provide diagnoses or treatment recommendations, under the guidelines of the American Academy of Ophthalmology (AAO) (29). To ensure the reliability and efficacy of eye care services assisted by chatbots, there's a desperate need for targeted domain-specific pretraining under reliable databases (30). It is necessary to carry out supervision (29) that harnesses ChatGPT's ability to tailor learning experiences (31) and ameliorate augmented learning (32). Healthcare professionals are supposed to address ethical concerns (33) and increase social acceptability (34) to enhance more beneficial integration of LLM chatbots into healthcare.

Due to the limited amount, the study included 15 common questions instead of specific questions about TED, whose bias may result in questions different from the essential topics to patients. Although some patients and ophthalmologists participated and left valuable feedback, including more TED experts, the latest disease guidelines, and widely recognized pathological views would have strengthened the quality comparison. Moreover, difficulty in distinguishing the exact quality was also a limitation, potentially leading to the reliance on self-criteria. Our findings on preliminary diagnosis were only based on several facial images and CT images from patients with TED, focusing on evaluating ChatGPT's capacity to interpret images. Lacking information on patients' medical histories, such as thyroid function indicators, specific symptoms, and underlying medical history, the occasion in this cross-sectional study may not be all the same in clinical practices, which should be considered in a more extensive study. The promising results of this cross-sectional study were specific to ChatGPT-4 and 4o and may not be generalized to other versions under special training. To apply Chatbot in clinical practices, its performance should be further assessed in the multicenter stimulation system.

In summary, ChatGPT-4o demonstrated satisfactory performance in generating responses to common clinical questions about TED and conducting accurate preliminary diagnoses based on images as a supplementary tool in the effective patient-oriented management system. Comprehensiveness and conciseness surpassing the ocular professor in decision-making questions. Although the performance of ChatGPT is promising in

specific clinical scenarios, suppressing or approaching the experienced ocular professor, its recommendations should act as a reference under supervision.

## Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding authors.

## Ethics statement

The studies involving humans were approved by Biomedical Research Ethics Committee of Shanghai Changzheng Hospital (Ethics approval number: 2021SL044). The studies were conducted in accordance with the local legislation and institutional requirements. Due to the utilization of de-identified images and personal information, written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin, in accordance with the national legislation and institutional requirements.

## Author contributions

YW: Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Conceptualization, Investigation, Methodology, Validation. SY: Writing-original draft, Writing – review & editing, Data curation, Formal analysis, Visualization. CZ: Writing – review & editing, Data curation, Methodology, Visualization, Formal analysis. YX: Writing – review & editing, Data curation, Methodology, Visualization. YS: Writing – review & editing, Data curation, Formal analysis, Methodology. JL: Writing – review & editing, Data curation, Formal analysis. XH: Writing – review & editing, Conceptualization, Supervision, Data curation. RW: Writing – review & editing, Conceptualization, Data curation, Funding acquisition, Supervision. YC: Conceptualization, Formal analysis, Visualization, Writing – review & editing, Supervision.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by grants from the National Natural Science Foundation of China (no. 81770959 and no. 82371100) and the Clinical Research Project of Shanghai Shenkang Hospital Development Center (Grant no. SHDC2020CR1043B).

## Acknowledgments

The authors appreciate the participation of 210 patients and 12 clinicians in the questionnaire survey and their valuable feedback. The authors also thank Evan Wang for his assistance in

network and information acquisition when using ChatGPT, OpenAI.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

## References

- Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*. (2023) 26:108163. doi: 10.1016/j.isci.2023.108163
- Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. *Ann Biomed Eng*. (2023) 51:2353–5. doi: 10.1007/s10439-023-03263-5
- Masalkhi M, Ong J, Waisberg E, Zaman N, Sarker P, Lee AG, et al. ChatGPT to document ocular infectious diseases. *Eye (Lond)*. (2023) 38:826–8. doi: 10.1038/s41433-023-02823-2
- Buonfiglio F, Ponto KA, Pfeiffer N, Kahaly GJ, Gericke A. Redox mechanisms in autoimmune thyroid eye disease. *Autoimmun Rev*. (2024) 23:103534. doi: 10.1016/j.autrev.2024.103534
- Smith TJ. Understanding pathogenesis intersects with effective treatment for thyroid eye disease. *J Clin Endocrinol Metab*. (2022) 107:S13–26. doi: 10.1210/clinem/dgac328
- Perros P, Žarković M, Pearce SH, Razvi S, Kolli H, Dickinson AJ. Inter-observer variability of clinical activity score: assessments in patients with thyroid eye disease. *Am J Ophthalmol*. (2023) 252:94–100. doi: 10.1016/j.ajo.2023.03.027
- Bartalena L, Tanda ML. Current concepts regarding graves' orbitopathy. *J Intern Med*. (2022) 292:692–716. doi: 10.1111/joim.13524
- Kahaly GJ, Douglas RS, Holt RJ, Sile S, Smith TJ. Teprotumumab for patients with active thyroid eye disease: a pooled data analysis, subgroup analyses, and off-treatment follow-up results from two randomised, double-masked, placebo-controlled, multicentre trials. *Lancet Diabetes Endocrinol*. (2021) 9:360–72. doi: 10.1016/s2213-8587(21)00056-5
- Yeung AWK, Tosevska A, Klager E, Eibensteiner F, Tsagkaris C, Parvanov ED, et al. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. *J Intern Med*. (2022) 24:e28152. doi: 10.2196/28152
- Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? *Int J Surg*. (2024) 110:6018–9. doi: 10.1097/j99.0000000000001754
- Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in Board Examinations for Specialists in the Japanese ophthalmology society. *Cureus*. (2023) 15:e49903. doi: 10.7759/cureus.49903
- Ruili W. A comprehensive collection of thyroid eye disease knowledge. Shanghai: Shanghai World Book Publishing Company (2021).
- Edmunds MR, Boelaert K. Knowledge of thyroid eye disease in Graves' disease patients with and without Orbitopathy. *Thyroid*. (2019) 29:557–62. doi: 10.1089/thy.2018.0665
- Sharma A, Stan MN, Rootman DB. Measuring health-related quality of life in thyroid eye disease. *J Clin Endocrinol Metab*. (2022) 107:S27–35. doi: 10.1210/clinem/dgac230
- Lee J, Koh J, Kim J-Y. Popularization of medical information. *Healthc Inform Res*. (2021) 27:110–5. doi: 10.4258/hir.2021.27.2.110
- Chow R, Hasan S, Zheng A, Gao C, Valdes G, Yu F, et al. The accuracy of artificial intelligence ChatGPT in oncology examination questions. *J Am Coll Radiol*. (2024) 21:1800–4. doi: 10.1016/j.jacr.2024.07.011
- Ebel S, Ehrengut C, Denecke T, Gößmann H, Beeskow AB. GPT-4o's competency in answering the simulated written European Board of Interventional Radiology exam compared to a medical student and experts in Germany and its ability to generate exam items on interventional radiology: a descriptive study. *J Educ Eval Health Prof*. (2024) 21:21. doi: 10.3352/jeehp.2024.21.21
- Huang X, Raja H, Madadi Y, Delsoz M, Poursoroush A, Kahook MY, et al. Predicting Glaucoma before onset using a large language model Chatbot. *Am J Ophthalmol*. (2024) 266:289–99. doi: 10.1016/j.ajo.2024.05.022

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2025.1546706/full#supplementary-material>

- Peng Z, Ma R, Zhang Y, Yan M, Lu J, Cheng Q, et al. Development and evaluation of multimodal AI for diagnosis and triage of ophthalmic diseases using ChatGPT and anterior segment images: protocol for a two-stage cross-sectional study. *Front Artif Intell*. (2023) 6:1323924. doi: 10.3389/frai.2023.1323924
- Santucci D, Ragone R, Vergantino E, Vaccarino F, Esperto F, Prata F, et al. Comparison between three Radiomics models and clinical nomograms for prediction of lymph node involvement in PCa patients combining clinical and Radiomic features. *Cancers*. (2024) 16:2731. doi: 10.3390/cancers16152731
- Dehdab R, Brendlin A, Werner S, Almansour H, Gassenmaier S, Brendel JM, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn J Radiol*. (2024) 42:1168–77. doi: 10.1007/s11604-024-01606-3
- Hu Y, Hu Z, Liu W, Gao A, Wen S, Liu S, et al. Exploring the potential of ChatGPT as an adjunct for generating diagnosis based on chief complaint and cone beam CT radiologic findings. *BMC Med Inform Decis Mak*. (2024) 24:55. doi: 10.1186/s12911-024-02445-y
- Gerrie SK, Rajani H, Navarro OM, Lyons CJ, Marie E, Frayn CS, et al. Pediatric orbital lesions: non-neoplastic extraocular soft-tissue lesions. *Pediatr Radiol*. (2024) 54:910–21. doi: 10.1007/s00247-024-05892-x
- Guo S, Li R, Li G, Chen W, Huang J, He L, et al. Comparing ChatGPT's and Surgeon's responses to thyroid-related questions from patients. *J Clin Endocrinol Metab*. (2024) 10:dgae235. doi: 10.1210/clinem/dgae235
- Marshall RF, Mallem K, Xu H, Thorne J, Burkholder B, Chaon B, et al. Investigating the accuracy and completeness of an artificial intelligence large language model about uveitis: an evaluation of ChatGPT. *Ocul Immunol Inflamm*. (2024) 32:2052–5. doi: 10.1080/09273948.2024.2317417
- Rojas-Carabali W, Cifuentes-González C, Wei X, Putera I, Sen A, Thng ZX, et al. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in uveitis. *Ocul Immunol Inflamm*. (2024) 32:1526–31. doi: 10.1080/09273948.2023.2253471
- Benitez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, van L, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. *J Am Med Inform Assoc*. (2024) 31:776–83. doi: 10.1093/jamia/ocad252
- Mihalache A, Huang RS, Popovic MM, Patil NS, Pandya BU, Shor R, et al. Accuracy of an artificial intelligence Chatbot's interpretation of clinical ophthalmic images. *JAMA Ophthalmol*. (2024) 142:321–6. doi: 10.1001/jamaophthalmol.2024.0017
- Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. *Eye (Lond)*. (2024) 38:1368–73. doi: 10.1038/s41433-023-02906-0
- Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology. *Ophthalmol Sci*. (2023) 3:100324. doi: 10.1016/j.xops.2023.100324
- Kaur K, Gurnani B. Enhancing ophthalmic education with ChatGPT: addressing concerns and future directions. *Eur J Ophthalmol*. (2024) 34:NP91–NP92. doi: 10.1177/11206721241258426
- Gurnani B, Kaur K. Leveraging ChatGPT for ophthalmic education: a critical appraisal. *Eur J Ophthalmol*. (2023) 34:323–7. doi: 10.1177/11206721231215862
- Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Intern Med*. (2023) 25:e48009. doi: 10.2196/48009
- Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol*. (2024) 85:146–53. doi: 10.1016/j.eururo.2023.10.014
- OpenAI. (2024). GPT-4o System Card. Available at: <https://openai.com/index/gpt-4o-system-card/> (Accessed August 8, 2024).