Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Assessing the ChatGPT aptitude: A competent and effective Dermatology doctor?

Chengxiang Lian[a,1], Xin Yuan[b,1], Santosh Chokkakula[c,1], Guanqing Wang[d], Biao Song[e,f], Zhe Wang[e], Ge Fan[g], Chengliang Yin[h,*]

[a] *Department of Dermatology and Venereology, The First Affiliated Hospital of Guang-xi Medical University, Nanning, 530021, China*
[b] *Department of Dermatology, GuiZhou Provincial People's Hospital, Guiyang, 550000, China*
[c] *Department of Microbiology, Chungbuk National University College of Medicine and Medical Research Institute, Cheongju, Chungbuk, 28644, South Korea*
[d] *Department of Dermatology, Shanghai General Hospital (South), Shanghai Jiao Tong University, No. 650, New Songjiang Road, Shanghai, 200000, China*
[e] *Zhihui Big Data Research Institute of Inner Mongolia, Inner Mongolia, 010020, China*
[f] *Collaborative Innovation Center of Big Data Application Research of Inner Mongolia University of Finance and Economics, Inner Mongolia, 010020, China*
[g] *Lightspeed & Quantum Studios, Tencent Inc., Shenzhen, 693388, China*
[h] *Faculty of Medicine, Macau University of Science and Technology, Macau, 999078, China*

ARTICLE INFO

Keywords:
ChatGPT
Atopic dermatitis
Evidence-based medicine
Autoimmune blistering skin diseases

ABSTRACT

Background: The efficacy and adeptness of ChatGPT 3.5 and ChatGPT 4.0 in the precise diagnosis and management of conditions like atopic dermatitis and Autoimmune blistering skin diseases (AIBD) remain to be elucidated. So this study examined the accuracy and effectiveness of the ChatGPT responses related to understanding, therapies, and specific cases of these two conditions.
Method: Firstly, the responses provided by ChatGPTs to a set of 50 questionnaires underwent evaluation by five distinct dermatologists, with complete adjudication of the third-party reviewer. The comparative analysis included the evaluative efficacy of both ChatGPT3.5 and ChatGPT4.0 against the diagnostic abilities exhibited by three distinct cohorts of qualified clinical professionals. And then, an examination was conducted to assess the diagnostic proficiency of ChatGPT3.5 and ChatGPT4.0 in the context of diagnosing specific instances of skin blistering autoimmune diseases.
Results: In assessing the proficiency of ChatGPTs in generating responses related to fundamental knowledge about AD it is noteworthy that both versions of ChatGPTs, despite their lack of specialized training on medical databases, exhibited a commendable capacity to yield solutions that exhibited a substantial degree of concurrence with evidence-based medical information. Accordingly we observed that the performance of ChatGPT-4.0 beyond that of the ChatGPT-3.5. However, it it crucial to emphasize that ChatGPT-4.0 did not show the ability to offer answers surpassing those provided by associate senior, and senior medical professionals. In the assessment designed to determine the proficiency of ChatGPTs in recognizing particular type of AIBD, it is evident that both ChatGPT-4 and ChatGPT-3.5 demonstrated inadequacy in providing responses that are both precise and accurate for each individual occurrence of this skin condition.

[*] Corresponding author. Faculty of Medicine, Macau University of Science and Technology, Taipa, 999078, China.
E-mail address: chengliangyin@163.com (C. Yin).

[1] These authors contributed equally to this article and share first authorship.

https://doi.org/10.1016/j.heliyon.2024.e37220
Received 13 July 2023; Received in revised form 29 August 2024; Accepted 29 August 2024
Available online 30 August 2024

Conclusion: Both ChatGPT-3.5 and ChatGPT-4.0 satisfactory for addressing fundamental inquiries related to atopic dermatitis, however they prove insufficient for diagnosing AIBD. The progress of ChatGPT in achieving utility within the professional medical domain remains a considerable journey ahead.

1. Introduction

ChatGPT is a conversational system that utilizes a large pre-trained language model called Generative Pre-trained Transformer (GPT), which was developed by OpenAI [1]. ChatGPT was built upon the foundation of the GPT-3.5 architecture bolstered by an expansive training dataset exceeding 570 GB in size and comprising a staggering 175 billion model parameters [2,3]. On November 30, 2022, version 3.5 of GPT was launched, followed by an even more advanced GPT-4.0 on March 14, 2023 [4]. Within just two months of its release, by end of January 2023, the number of monthly active users had reached 100 millions, establishing it as the fastest-growing consumer app in history [5,6]. ChatGPT-4.0 represents a substantial enhancement over ChatGPT-3.5, demonstrating superior performance across a wide range of tasks owing to its enhanced generation capabilities [7,8].

ChatGPT known for its comprehensive understanding across the wide spectrum of topics, has garnered significant attention from experts in diverse fields, including medicine [3,4]. To streamline operations within the medical realm, it is imperative to consider utilizing ChatGPT as a substitute for medical practitioners. This approach ensures that patients receive prompt and accurate responses to fundamental inquiries [9]. However, it is worth noting that substantial portion of the data used to train both ChatGPT-3.5 and ChatGPT-4.0 is sourced from public outlets rather than professional medical databases. This raises pertinent questions regarding their ability to generate responses consistent with evidence-based medicine (EBM) [10].

Atopic dermatitis (AD), also known as eczema, is a common, chronic skin disease. Acute lesions appear as redness, swelling, and oozing, while chronic ones result in dryness, thickening, and skin color changes [11]. It often starts in infancy but can occur at any age. There are different subtypes based on age groups: infantile (<2 years), childhood (2–12 years), adolescent (12–18 years), and adult-onset (after 18 years). There is also a type that occurs after 60, known as elderly-onset AD [12,13]. Diagnosis considers symptoms like itching, thickened skin, and a history of atopic conditions [14]. AD has two types, IgE-allergic and non-IgE-allergic. Pathophysiology of AD is complex which involves epidermal barrier issues, immune dysfunction, and changes in skin microbiota [15]**.** AD treatment involves a multifaceted approach, comprising patient and caregiver education, proper skin care, anti-inflammatory drugs (corticosteroids and TCIs), Phosphodiesterase 4-inhibitors, JAK-STAT inhibitors and artificial UV radiation exposure [16,17].

Differentiating AD from certain skin disorders such as seborrheic dermatitis, psoriasis, nummular dermatitis and contact dermatitis is often difficult; nevertheless, in certain instances, a familial history of atopy and lesions scattering is beneficial in deter-mining the diagnosis. Seborrheic dermatitis (SD) is a common inflammatory skin condition on sebum-rich areas like the scalp and face. It begins in infancy due to hormone-induced sebum, decreases in childhood, and surges in adolescence and adulthood. Both SD and AD are common in infants and can co-occur, causing confusion [18]. Psoriasis in infants and children can be mistaken for AD due to less prominent scaling and lesions on the face. Psoriasis is common in the diaper area, unlike AD. Nail pits can help distinguish psoriasis from AD [19]. Nummular dermatitis (ND) features round or oval, well-defined lesions that can be itchy, often resembling atopic dermatitis [20]. Contact dermatitis (CD) is the most prevalent form, characterized by skin redness and swelling. It can be acute, chronic, persistent, or recurrent, and is traditionally divided into irritant and allergic types, which can overlap and often complicate other skin conditions, including AD [21]. AD patients, with *S. aureus* on affected and unaffected skin, pose challenges in distinguishing colonization from infections like Impetigo and Secondary Syphilis through skin cultures [22,23]. Along with these infections, immunodeficiency disorders such as Netherton Syndrome and HIV/AIDS-Related Skin Changes are often giving challlinging to clinicians in diagnosis of the AD [24,25]. Given the persistent and unresolved nature of AD, a significant number of patients necessitate frequent medical consultations. Consequently, clinicians often seek advanced tools like ChatGPT to assist in the diagnosis and treatment of AD.

AIBD is a prevalent dermatological illness, and its etiology involves immunological, genetic, and other variables [26,27]. The precise and timely diagnosis of the numerous AIBD is highly crucial as it affects the subsequent choice of therapy and the prognosis of patients [26]. Similar to AD, the management of AIBD presents clinicians with considerable challenges. To aid in creating a medical ChatGPT version for clinical use, we empirically evaluated diagnostic capabilities of ChatGPT- 3.5 and enhanced ChatGPT-4.0. For the testing and validation accuracy of these two models, we performed a small-scale investigation involving 50 questions along with standardized responses concerning Atopic Dermatitis (AD), and 39 questions with standardized responses related to a AIBD. The aim was to evaluate the effectiveness and capacities of both ChatGPT-3.5 and ChatGPT-4.0, as validated by medical experts.

2. Methods

2.1. Questions generations and strategies for assessing ChatGPT proficiency in responding to knowledge about AD

Five clinical experts thoroughly reviewed the guidelines (Supplementary Table 1) and generate 50 questions and standard answers related to AD. To ensure a diverse range of prospective, three sets of doctors were included, each with a junior, medium, and an associate and senior grade titles. Each set comprised three doctors, and each doctor within their respective set was provided with a questionnaire containing the aforementioned 50 questions. Simultaneously, both ChatGPT-3.5(OpenAI Inc., San Francisco, CA, USA) and ChatGPT-4.0(OpenAI Inc., San Francisco, CA, USA) were subjected to three rounds of evaluation, each involving the same set of 50 questions in English. Some questions were consisting of a serial of step-by-step issues. If ChatGPT-3.5 or ChatGPT-4.0 could not answer the first issue correctly, then the asking for follow-up issues ceased. The result of each round was recorded. The process of questions generations and asking strategies is illustrated in Fig. 1.

2.2. *Assessing process for the answers*

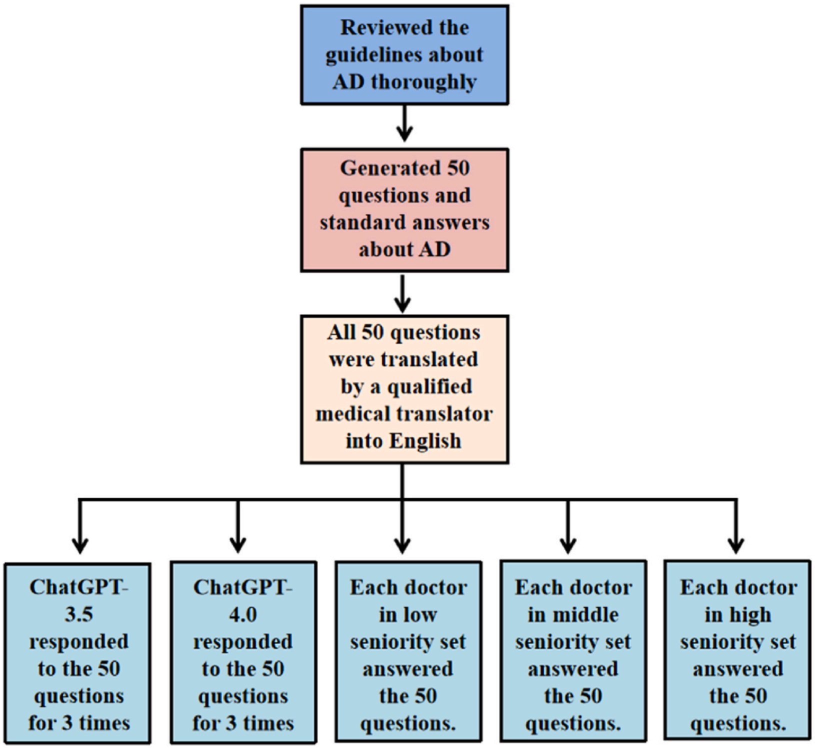
Three reviewers sequentially assessed the responses provided by ChatGPTs, determining the consistency of each answer by comparing it to the standardized answers. The responses were labeled as either “consistent” or “inconsistent” based on their evaluation. The cumulative count of “consistent” responses for each question generated the score, ranging from 0 to 3. If the answer is far-fetch to the question, it will be regarded as 0 points. Throughout the evaluation process, solid agreement was maintained among the three reviewers, ensuring a consistent and unbiased assessment of the responses.

2.3. *Strategy for validating the ability of ChatGPTs to identify particular type of AIBD*

Three clinical experts extracted a variety of reported pemphigus cases from the China Medical Journal Full Text Database (website: https://www.yiigle.com/index). They formulated a set of 39 cases consisting detailed information about the main complaint, medical history, and pathology slide findings for each case. These questions facilitated the accurate diagnosis of the specific type of pemphigus associated with each case. Throughout the process of generating questions, solid agreement was maintained among the three dermatology doctors, ensuring generating accurate and unbiased questions. Both ChatGPT-3.5 and ChatGPT-4.0 were subjected to three rounds of evaluation, each involving the same set of 39 cases in English. Some questions were consisting of a serial of step-by-step issues. If ChatGPT-3.5 or ChatGPT-4.0 could not answer the first issue correctly, then the asking for follow-up issues ceased. The result of each round was recorded. The process of questions generations and asking strategies is illustrated in Fig. 2.

2.4. *Data collection process for validating the ability of ChatGPTs to identify particular type of AIBD*

Three dermatology clinicians scored the answers provided by ChatGPTs, by assessing the consistency of each answer to the standardized diagnosis given in the indicated initial articles. In instances where ChatGPTs offer multiple diagnoses for a given case, only the initial or final diagnosis supplied by ChatGPTs is acknowledged as the response. The cumulative count of coherent responses for each query establishes the score, which is ranging from 0 to 3. If ChatGPTs accurately determine a correct answer without specifying a particular subtype, a score of 0.5 is allotted. When the correct diagnosis given by ChatGPTs is only part of the standard answer,


Reviewed the guidelines about AD thoroughly

Generated 50 questions and standard answers about AD

All 50 questions were translated by a qualified medical translator into English

ChatGPT-3.5 responded to the 50 questions for 3 times

ChatGPT-4.0 responded to the 50 questions for 3 times

Each doctor in low seniority set answered the 50 questions.

Each doctor in middle seniority set answered the 50 questions.

Each doctor in high seniority set answered the 50 questions.

Fig. 1. The summary of questions generations and the research strategy.

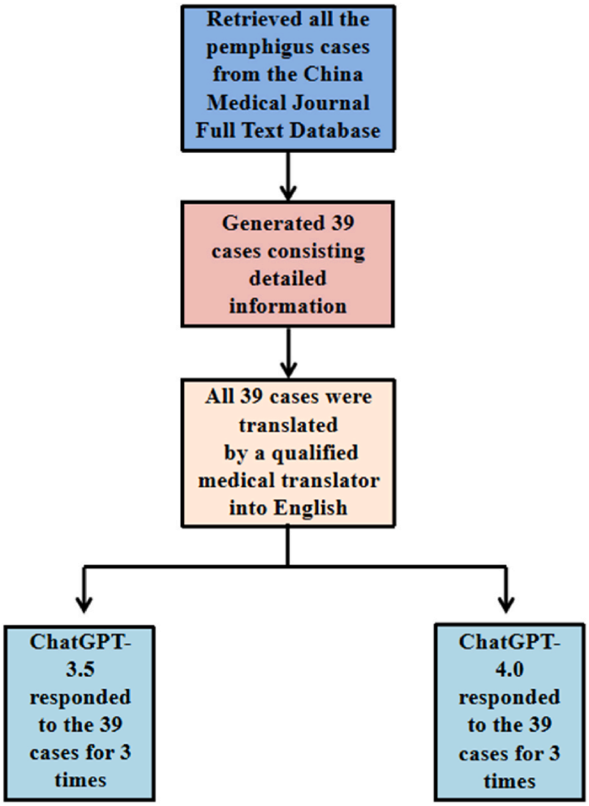

Retrieved all the pemphigus cases from the China Medical Journal Full Text Database
Generated 39 cases consisting detailed information
All 39 cases were translated by a qualified medical translator into English
ChatGPT-3.5 responded to the 39 cases for 3 times
ChatGPT-4.0 responded to the 39 cases for 3 times

Fig. 2. The summary of questions generations and the research strategy for validating the ability of ChatGPTs to identify particular type of AIBD.

the diagnosis is given 0.5 points. Responses from ChatGPTs that are incorrect or absent are valued at 0 points. In cases where responses of ChatGPTs are predominantly incorrect partially consistent with the standard answer, 0 points are granted. Throughout the evaluation process, solid agreement was maintained among the three dermatology doctors, ensuring a consistent and unbiased assessment of the responses.

2.5. Statistical analysis

All statistical analyses were conduct by GraphPad Prism software version 6.01 (GraphPad Prism Software Inc., La Jolla, CA, USA). The quantified data were expressed as the mean ± SD. The mean values between two groups were compared by utilizing student t-test and $P < 0.05$ were considered to be statistically significant.

3. Results

3.1. Ability of ChatGPTs to respond to fundamental inquiries about AD

The cumulative scores achieved by ChatGPT-3.5, ChatGPT-4.0, junior, medium, and associate senior and senior doctors set were determined as 106, 122, 111, 117 and 131 respectively. After accomplishing three different tasks, ChatGPT-3.5 provided accurate responses to 44 % of the AD-related inquiries (22/50). For during three separate instances, ChatGPT-3.5 generated entirely disparate responses from the standard answer on all four occasions. Among the incorrect responses, the open-ended inquiries pertaining to AD constituted 50 % of the overall count. Out of the total of 50 queries, 70 % that (35/50) were in accordance with the guidelines. Among the entirely incorrect answers, the TCI use in the treatment for AD accounted for 67 % of the total (2/3). The respective rates of test results aligning with the correct answers categorized by medical professional were as follows, junior −46 % (23/50), medium −46 % (23/50), and associate senior and senior −68 % (34/50) (Table 1).

Among the all 50 questions，there are 12 questions related to basic knowledge about AD, 4 questions related to diagnosis about AD and 34 questions related to treatment about AD. The mean scores of AD diagnosis associated answers given by doctors from high seniority group were 3.000 which is significantly higher than those of ChatGPT-3.5($p < 0.05$) and ChatGPT-4.0($p < 0.05$). The mean scores of answers given by doctors from high seniority group about the AD treatment questions were 2.559 which is significantly higher than that of ChatGPT-3.5($p < 0.05$). The details of answers given by each set about the basic knowledge, diagnosis and treatment about AD were shown in Table 2 and Fig. 3A–C.

Table 1
Assessment of the responses from ChatGPT-3.5, ChatGPT-4, Junior professional title , Medium-grade professional title, Associate senior title, and Senior title for the clinical knowledge of AD.

Classification	Serial number	Questions	ChatGPT −3.5	ChatGPT −4.0	J[a]	M[b]	A[c]
Basic knowledge of AD	1	Is the risk of psycho-nervous system disease and cardiovascular disease increased or decreased in chronic AD patients?	3	3	3	2	3
	2	Is Th2 inflammation a basic feature of AD? Please answer “yes” or “No".	0	3	1	2	3
	3	Which cytokines produced by Th2 cells, basophils, and innate lymphoid cells are important for the onset of AD?	0	2	1	1	2
	4	Can the ecological imbalance of skin microflora considerably enhance the risk of AD onset, according to current studies on microflora? Please answer “yes” or “No".	3	3	3	2	3
	5	Are increased colonization of *Staphylococcus aureus* and decreased bacterial diversity the main manifestations of skin microflora disturbance in AD patients? Please answer “yes” or “No".	3	3	3	3	3
	6	According to current clinical practice, does AD exist in the old? If yes, is it more likely to occur in men or women?	0	2	1	2	3
	7	Is the pattern of skin inflammation in AD patients at different ages the same or different?	0	3	3	2	2
	8	Can a cold or an infection cause or worsen AD symptoms? Please answer “yes” or “No".	3	3	2	3	2
	9	Over the past 30 years, has the prevalence of AD grown globally? Please answer “yes” or “No".	3	3	2	3	3
	10	Does the incidence of extra-cutaneous allergy disorders rise or fall with age in patients with chronic AD?	2	1	1	1	2
	11	Do increased *Staphylococcus aureus* colonization and decreased bacterial diversity contribute to the progression of skin inflammation in AD patients? Please answer “yes” or “No".	3	3	2	3	3
	12	Which age group has the highest risk of AD onset?	3	3	3	2	3
Diagnosis of AD	13	Is a history of allergic disease in a family member, such as a parent, the strongest risk factor for a patient with AD? Please answer “yes” or “No".	1	1	3	2	3
	14	Do the majority of children with AD have severe symptoms or not? Please answer “yes” or “No".	2	3	2	2	3
	15	Does the difference in blood total IgE levels determine whether AD patients are endogenous type or exogenous type? Please answer “yes” or “No".	2	1	1	3	3
	16	Which symptom, itching or dry skin, is main for AD patients at each stage?	3	2	1	3	3
Treatment of AD	17	Can TCI be used to treat AD patients with skin lesions located in the facial-neck, fold site, breast, genital areas? Please answer “yes” or “No".	2	3	3	3	3
	18	Can patients with light AD be treated with UV light? Please answer “yes” or “No".	3	2	2	2	1
	19	When a patient is diagnosed with AD, is it advisable to use glucocorticoids systematically for treat at very first beginning? Please answer “yes” or “No".	3	3	3	3	3
	20	Should we employ TCS of a sufficient intensity immediately to treat patients of acute AD, or should we gradually raise the amount of TCS from a minimal dose?	3	3	3	3	3
	21	Can a patient with moderate AD be treated with immunosuppressive medications?	1	0	2	2	1
	22	Can whole-body UV treatment be used to treat AD in children less than 6 years old? Please answer “yes” or “No".	2	1	3	3	2
	23	A patient was diagnosed with chronic AD. Is more bathing or less bathing advised as a basic treatment method for the patient?	2	2	3	3	2
	24	Can AD sufferers continually scrape the skin lesion? Why or why not? Please answer “yes” or “No".	3	3	3	3	3
	25	A patient with AD was in the acute phase. Is it proper to recommend short-term use of a mild topical glucocorticoid (TCS) in the face, neck and folds to the patient? Please answer “yes” or “No".	2	3	3	2	3
	26	For the majority of patients with AD, can the adverse reactions caused by long-term use of TCI gradually disappear with the extension of medication time? Please answer “yes” or “No".	3	2	2	3	2
	27	Is narrow spectrum ultraviolet (NB-UVB) and large dose UVA1 the priority selection for the treatment of chronic and lichtenized lesions in moderate-severe adult AD patients? Please answer “yes” or “No".	3	3	2	3	2
	28	A patient was diagnosed with AD.Is it necessary to stop eating foods he/she has been allergic to before? Does food allergy play an important role in the onset of AD? Please answer “yes” or “No".	3	3	1	2	2
	29	For chronic and lichenized skin lesions in moderate-severe adult AD patients, can UV be used to treat the itching symptoms of patients? Please answer “yes” or “No".	3	3	3	3	3

(*continued on next page*)

Table 1 (*continued*)

Classification	Serial number	Questions	ChatGPT −3.5	ChatGPT −4.0	J[a]	M[b]	A[c]
	30	When a patient is diagnosed with AD, is it important to educate the patient in aspects such as “food, wearing, housing, transportation and bath”? Please answer “yes” or “No".	3	3	3	3	3
	31	Can topical emollients reduce the number and severity of AD attacks in patients? Please answer “yes” or “No".	2	3	3	2	3
	32	Should a patient with AD be counseled to avoid eating a large number and various of highly allergic foods? Please answer “yes” or “No".	0	3	2	2	3
	33	Can sodium hypochlorite be added to the bath (0.0005 % bleach powder bath) when it is discovered that the skin lesions of an AD patient are susceptible to infection? Please answer “yes” or “No".	2	3	1	2	3
	34	Can Dupilumab be used for the treatment of patients with light AD? Please answer “yes” or “No".	2	0	2	2	3
	35	Do AD patients require long-term treatment given the chronic and relapsing nature of the disease?	1	3	0	3	3
	36	Are TCS the first-line topical medications for treating AD patients? Please answer “yes” or “No".	1	3	3	3	3
	37	Can a powerful TCS be applied for a long-term use to the face, neck, and folds to cure acute AD symptoms? Please answer “yes” or “No".	3	3	3	3	3
	38	Can acute bouts of AD be treated with NB-UVB? Why ? Please answer “yes” or “No".	1	3	2	2	3
	39	Can topical calcineurin inhibitors (TCI) be used to treat AD externally? Please answer “yes” or “No".	2	3	3	2	3
	40	Will a patient with persistent AD experience local burning and irritability from long-term TCI use? Please answer “yes” or “No".	1	0	3	1	2
	41	Are Janus kinase (JAK) inhibitors and Dupilumab effective treatment approaches for severe AD patients? Please answer “yes” or “No".	3	3	3	3	3
	42	Can UV be used as “maintenance treatment” for treating persistent and lichtenized skin lesions in moderate-severe adult AD patients? Please answer “yes” or “No".	3	3	3	3	3
	43	Can symptoms be mitigated with UVA1 in an AD patient at acute phage? Please answer “yes” or “No".	2	3	2	1	1
	44	Can AD patients receive UV therapy while taking glucocorticoids and humectants at the same time? Please answer “yes” or “No".	3	3	3	2	3
	45	Can UV therapy be administered when an AD patient’s symptoms are aggravated by sunlight exposure? Please answer “yes” or “No".	1	3	2	2	3
	46	Can patients with AD be treated with a combination of TCI and UV? Please answer “yes” or “No".	1	1	1	2	2
	47	Should TCS or TCI be used initially to treat acute symptoms in some AD patients who cannot tolerate medication response (particularly in the acute phase)?	2	3	2	3	3
	48	Will long-term use of TCI result in adverse reactions such skin barrier deterioration and skin atrophy in a patient with chronic AD? Please answer “yes” or “No".	2	3	2	1	2
	49	Does TCS/TCI work well for treating patients with severe AD? Please answer “yes” or “No".	2	0	1	1	2
	50	After the skin symptoms subside, should patients with moderate-severe AD or prone to AD recurrence switch to long-term “active maintenance therapy”? Please answer “yes” or “No".	3	3	2	3	3
Total score			**106**	**122**	**111**	**117**	**131**

[a] L: doctors with junior professional title.

[b] M: doctors with medium-grade professional title.

[c] H: doctors with associate senior title or senior titled.

The mean scores of correct answers given by ChatGPT-3.5, ChatGPT-4.0 about the whole 50 questions were 1.917 and 2.667, respectively. The mean score of correct answers given by doctors from low seniority set about the 50 questions was 2.083 which was significantly lower than that of ChatGPT-4.0($p < 0.05$). The mean score of correct answers given by doctors from high seniority group about the 50 questions was 2.667 which was higher than that of ChatGPT-3.5($p < 0.05$), significantly. The details of correct and incorrect answers given by ChatGPT-3.5, ChatGPT-4.0, low seniority, middle seniority and high seniority were shown in Table 3 and Fig. 4.

3.2. *Performance ability of ChatGPTs on diagnosis particular types of AIBD*

ChatGPT-3.5 has answered questions 21, 3, 6, 6, 1, and 2 times, corresponding to a total score of 0, 0.5, 1, 1.5, 2, and 3, respectively. Similarly, ChatGPT-4.0 has responded to questions 11, 6, 4, 14, 2, and 2 times, resulting in a total score of 0, 0.5, 1, 1.5, 2, and 3. The cumulative scores attained by the ChatGPT-3.5 and ChatGPT-4.0 groups were calculated as 24.5 and 38 consecutively. After conducting three distinct investigations, ChatGPT-3.5 exhibited an inability to supply accurate answers for 21 questions, comprising

Table 2
Comparison of mean scores according to question type.

Characteristic	ChatGPT-3.5	ChatGPT-4.0	low seniority			middle seniority			high seniority		
	mean ± SD	mean ± SD	mean ± SD	P value (with GPT-3.5)	P value (with GPT-4.0)	mean ± SD	P value (with GPT-3.5)	P value (with GPT-4.0)	mean ± SD	P value (with GPT-3.5)	P value (with GPT-4.0)
Basic knowledge of AD	1.917 ± 1.443	2.667 ± 0.6513	2.083 ± 0.9003	0.7375	0.0826	2.167 ± 0.7177	0.5965	0.0877	2.667 ± 0.4924	0.1025	>0.9999
Diagnosis of AD	2.000 ± 0.8165	1.750 ± 0.9574	1.750 ± 0.9574	0.7049	>0.9999	2.500 ± 0.5774	0.3559	0.2283	3.000 ± 0	0.0498	0.0401
Treatment of AD	2.147 ± 0.8575	2.441 ± 1.050	2.324 ± 0.8061	0.3851	0.6060	2.382 ± 0.6970	0.2188	0.7863	2.559 ± 0.6602	0.0300	0.5821

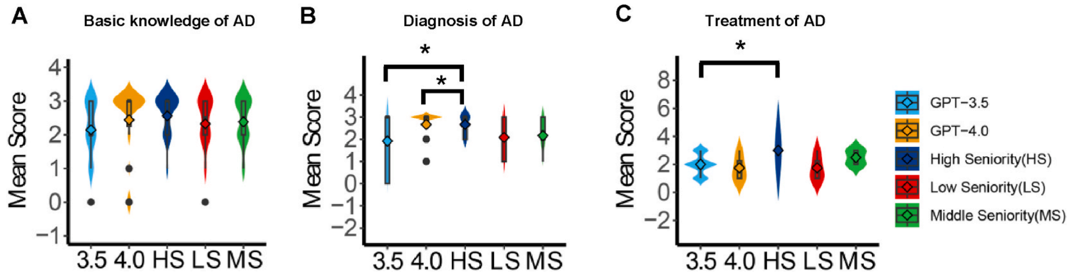


Fig. 3. The mean scores of answers given by ChatGPT-3.5, ChatGPT-4.0, low seniority, middle seniority and high seniority group depicted by violin plots.(a) Violin plots depicting the mean scores of each group replying for questions of basic knowledge of AD.(b)Violin plots depicting the mean scores of each group replying for questions of diagnosis of AD.(c)Violin plots depicting the mean scores of each group replying for questions of AD treatment. * represents p-value <0.05. **Abbreviations** 3.5, GPT-3.5; 4.0, GPT-4.0; HS, high seniority; LS, low seniority; MS, middle seniority.

Table 3
Comparison of mean scores according to answers' classification.

Characteristic	ChatGPT-3.5	ChatGPT-4.0	low seniority			middle seniority			high seniority		
	mean ± SD	mean ± SD	mean ± SD	P value (with GPT-3.5)	P value (with GPT-4.0)	mean ± SD	P value (with GPT-3.5)	P value (with GPT-4.0)	mean ± SD	P value (with GPT-3.5)	P value (with GPT-4.0)
Correct	2.120 ± 0.9613	2.440 ± 0.9723	2.020 ± 0.7690	0.5670	0.0185	2.360 ± 0.6312	0.1432	0.6266	2.740 ± 0.4870	<0.0001	0.0539
Incorrect	0.8800 ± 0.9613	0.5600 ± 0.9723	0.9800 ± 0.7690	0.5670	0.0185	0.6400 ± 0.6312	0.1432	0.6266	0.2600 ± 0.487	<0.0001	0.0539

roughly 54 % of the overall inquiries. Notably, within the category of entirely inaccurate responses, instances of pemphigus represented a significant proportion at 33 % (7/21). Similarly, cases of BP constituted 29 % of the total (6/21). Approximately 28 % of ChatGPT-4 responses (11/39) exhibited complete divergence. Of these inaccurate answers, pemphigus cases constituted 18 % (2/11), while BP cases represented an equivalent 18 % (2/11). During three separate instances for each case, only two responses from ChatGPT-3.5 and ChatGPT-4.0 were graded as completely correct, respectively. ChatGPT-3.5 diagnosed the "BP caused by sitagliptin phosphate" and "Secondary acquired hemophilia A(AHA) after BP" completely correct and ChatGPT-4.0 diagnosed the "IgA pemphigus" and "BP" completely correct (Table 4).

Among the 3 times chances of answering each question by ChatGPT-3.5 and ChatGPT-4.0, a result could be correct, partial correct or incorrect. The mean scores of correct answers given by ChatGPT-3.5, ChatGPT-4.0 about the whole 39 questions were 0.2308 and 0.2564, respectively. The mean score of partial correct answers given by ChatGPT-4.0 was 1.436 which was significantly higher compared with those of ChatGPT-3.5($p < 0.05$). The mean score of incorrect answers given by ChatGPT-4.0 was 1.256 which was lower than that of ChatGPT-3.5($p < 0.05$), significantly. The more details about the answers given by ChatGPT-3.5 and ChatGPT-4.0 were illustrated in Table 5 and Fig. 5.

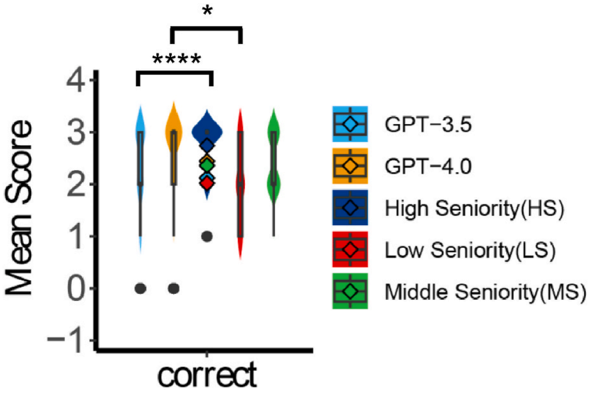


Fig. 4. The mean scores of correct answers given by ChatGPT-3.5, ChatGPT-4.0, low seniority, middle seniority and high seniority group depicted by violin plots. * represents p-value <0.05; **** represents p-value <0.0001.

Table 4
Assessment of the responses from ChatGPT-3.5, ChatGPT-4 for the diagnosis of diversity Pemphigus diseases.

	Diagnosis	ChatGPT-3.5 [a]	ChatGPT-4.0 [a]
1	Pemphigus foliaceus	1. Herpetic urticarial (0) 2. Psoriasis (0) 3. Psoriasis (0) **0**	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**
2	Castleman disease with PNP	1 Erythema multiforme (0) 2. T-cell non-Hodgkin's lymphoma (0) 3. Erythema multiforme (0) **0**	1. Non-hodgkin's lymphoma (NHL) (0) 2. Castleman disease (0.5) 3. Castleman disease (0.5) **1**
3	Pemphigoid nodularis (PN , severe)	1. Linear IgG deposition (0) 2. Linear IgG deposition (0) 3. Linear IgG deposition (0) **0**	1.Dermatitis (0) 2.SLE (0) 3. Pigmented skin disease (0) **0**
4	Cheilitis-like Pemphigus vulgaris	1. BP (0) 2. BP (0) 3. BP (0) **0**	1. pemphigus (0.5) 2. pemphigus (0.5) 3. pemphigus (0.5) **1.5**
5	Mucosal BP induced laryngeal stenosis	1. Laryngeal edema (0.5) 2. Laryngeal edema (0.5) 3. Laryngeal edema (0.5) **1.5**	1. BP (0.5) 2. BP (0.5) 3. BP (0.5) **1.5**
6	Chronic Lymphocytic leukemia/Small lymphocytic Lymphoma (CLL/SLL) with PNP	1. Lymphocytic leukemia/small B-cell lymphoma (0.5) 2. Lymphocytic leukemia/small B-cell lymphoma (0.5) 3. None (0) **1**	1. pemphigus (0.5) 2. pemphigus (0.5) 3. Chronic lymphocytic leukemia with pemphigus (0.5) **1.5**
7	Secondary acquired hemophilia A (AHA) after BP	1 AHA with BP (1) 2. AHA with BP (1) 3. AHA with BP (1) **3**	1. BP (0.5) 2. BP (0.5) 3. BP (0.5) **1.5**
8	BP induced by long-term use of rifampicin	1. Behcet's disease (0) 2. BP and TB (0.5) 3. TB and BP (0.5) **1**	1. BP (0.5) 2. BP (0.5) 3. BP (0.5) **1.5**
9	Pemphigus vulgaris secondary to scalp scarring	1. Mucosal-associated ocular disease (0) 2. Skin cancer complicated by Mucosal-associated ocular disease (0) 3. Skin cancer complicated by Mucosal-associated ocular disease (0) **0**	1. BP (0) 2. BP (0) 3. BP (0) **0**
10	Pemphigus vegetans (Neumann type)	1. Pemphigus (0.5) 2. Lesions of the oral mucosa (0) 3. Lesions of the oral mucosa (0) **0.5**	1. Pemphigus (0.5) 2. BP (0) 3. Pemphigus vulgaris (0) **0.5**
11	Severe BP of the ocular mucosa	1. Acute keratoconjunctivitis in the left eye (0) 2. Purulent keratitis in the left eye and chronic conjunctivitis in the right eye (0) 3. Purulent keratitis in the left eye and chronic conjunctivitis in the right eye (0) **0**	1. Acute bacterial keratoconjunctivitis of the left eye (0) 2. Acute bacterial keratoconjunctivitis of the left eye (0) 3. Acute keratoconjunctivitis in the left eye (0) **0**
12	PNP presented as toxic epidermal necrolysis	1. Bullous oral lichen planus (0) 2. A tumor-related vasculitis induced by Castleman's disease (0) 3. Urticaria caused by an allergic reaction or infection (0) **0**	1. Severe acute idiopathic erythema multiforme (0) 2. Severe acute idiopathic erythema multiforme (0) 3. BP (0) **0**
13	Pemphigoid nodularis	1. Linear IgA disease (0) 2. Linear IgA disease (0) 3. Linear IgA disease (0) **0**	1. BP (0.5) 2. Eosinophilic cellulitis with BP (0) 3. Eosinophilic dermatosis (0) **0.5**
14	BP caused by sitagliptin phosphate	1. BP caused by sitagliptin phosphate (1) 2. BP caused by sitagliptin phosphate (1) 3. BP caused by sitagliptin phosphate (1) **3**	1. BP (0.5) 2. BP (0.5) 3. Drug-induced dermatitis (possibly related to sitagliptin phosphate use) (0) **1**
15	Mucosal BP	1. Mucosal pemphigus (0) 2. Mucosal BP (1)	1. BP (0.5) 2. Pustular Epidermolysis (0)

(*continued on next page*)

Table 4 (*continued*)

	Diagnosis	ChatGPT-3.5 [a]	ChatGPT-4.0 [a]
		3. Mucosal BP (1) **2**	3. Pustular Epidermolysis (0) **0.5**
16	Pemphigus vulgaris of the oral cavity and vulva	1. Inflammation in oral mucosa and genital (0) 2. Inflammation in oral mucosa and genital (0) 3. Inflammation in genital (0) **0**	1. Pemphigus (0.5) 2. Oral mucosal BP (0) 3. Mucous erosive dermatitis (0) **0.5**
17	PNP secondary to Castleman disease	1. No skin lesions were found (0) 2. No skin lesions were found (0) 3. Acanthocytosis (0) **0**	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**
18	IgA pemphigus	1. Urticaria-like linear IgA bullous dermatosis (0) 2. Urticaria-like LABD (0) 3. Urticaria-like LABD (0) **0**	1. IgA pemphigus (1) 2. IgA pemphigus (1) 3. IgA pemphigus (1) **3**
19	PNP	1. Epidermolysis disease (0) 2. Pemphigus (0.5) 3. LABD (0) **0.5**	1. BP (0) 2. BP (0) 3. BP (0) **0**
20	Non-hodgkin lymphoma with PNP onset	1. PNP (0.5) 2. PNP (0.5) 3. PNP (0.5) **1.5**	1. BP (0) 2. BP (0) 3. BP (0) **0**
21	Cephalic and facial BP	1. Oral mucosal granulomatosis (0) 2. Oral mucosal granulomatosis (0) 3. BP (0.5) **0.5**	1.BP (0.5) 2.Local BP (0.5) 3.BP (0.5) **1.5**
22	Cephalic and facial BP	1. BP (0.5) 2. BP (0.5) 3. BP (0.5) **1.5**	1. BP (0.5) 2. BP (0.5) 3. BP (0.5) **1.5**
23	Pemphigus vulgaris	1. Herpetic contact dermatitis (0) 2. PEMP (0) 3. PEMP (0) **0**	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**
24	Local BP	1. Mucous Membrane Pemphigoid(MMP) (0) 2. MMP (0) 3. MMP (0) **0**	1. BP (0.5) 2. BP (0.5) 3. BP (0.5) **1.5**
25	Severe Pemphigus vulgaris	1. Pemphigus (0.5) 2. Early pemphigus (0.5) 3. Dermatitis herpetiformis after vaccination (0) **1**	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**
26	Pemphigus foliaceus	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**
27	Familial benign chronic pemphigus (FBCP)	1. B-cell lymphoma (Skin type) (0) 2. Verrucous growth of vulva (0) 3. Condyloma acuminatum (0) **0**	1. Chronic skin fungal infection (0) 2. Tonic and Clonic disease (0) 3. Chronic recurrent infectious eczema (0) **0**
28	Cicatricial BP in both eyes	1. Generalized choroiditis (0) 2. Multiple oral ulcer (0) 3. Multiple oral ulcer (0) **0**	1. Mucosal BP (0.5) 2. Mucous dermatitis (0) 3. No diagnosis (0) **0.5**
29	Congenital Pemphigus vulgaris	1. Congenital pemphigus (0.5) 2. Congenital pemphigus (0.5) 3. Congenital pemphigus (0.5) **1.5**	1. Congenital Pemphigus vulgaris(1) 2. Congenital pemphigus(0.5) 3. Congenital pemphigus(0.5) **2**
30	Psoriasis and Pemphigus foliaceus	1. Pemphigus vulgaris (0) 2. Pemphigus vulgaris (0) 3. Pemphigus vulgaris (0) **0**	1. Acute generalized pustular Psoriasis (0) 2. Pemphigus combined with *Staphylococcus aureus* infection and Psoriasis (0) 3. Pemphigus coinfection (0) **0**

(*continued on next page*)

Table 4 (*continued*)

	Diagnosis	ChatGPT-3.5 [a]	ChatGPT-4.0 [a]
31	BP	1. Drug eruption (0) 2. Pemphigus vulgaris (0) 3. Pemphigus vulgaris (0) **0**	1. BP (1) 2. BP (1) 3. BP (1) **3**
32	Pemphigus vulgaris	1. Pemphigus vulgaris (1) 2. Recurrent oral, genital, and anal ulcers (0) 3. Behcet's disease (0) **1**	1. Pemphigus vulgaris (1) 2. Behcet's disease (0) 3. Behcet's disease (0) **1**
33	Bullous lichen planus	1. Lichen Planus (0.5) 2. Dermatosis associated with xerosis (0) 3. Lichen Planus (0.5) **1**	1. Lichen Planus (0.5) 2. Pemphigus vulgaris (0) 3. Pemphigus vulgaris (0) **0.5**
34	Bullous lichen planus	1. PNP (0) 2. PNP (0) 3. PNP (0) **0**	1. PNP (0) 2. PNP (0) 3.PNP (0) **0**
35	Milium	1. Papular seborrheic keratosis (0) 2. Epidermal cyst (0) 3. Acne (0) **0**	1. Keratosis pilaris (0) 2. Keratosis pilaris (0) 3. Keratosis pilaris (0) **0**
36	BP	1. LABD (0) 2. LABD (0) 3. Herpetic dermatitis (0) **0**	1. BP (1) 2.BP (1) 3. Epidermolysis bullosa (0) **2**
37	Ocular cicatricial BP	1. Corneal ulcer in left eye with hyperemia and edema (0) 2. Corneal ulcer in the left eye with severe inflammation (0) 3. Corneal ulcer in the left eye with severe inflammation (0) **0**	1. Mild corneal opacity in the right eye (0) 2. Mild corneal opacity in the right eye (0) 3. Right eye subcutaneous bullosa (0) **0**
38	IgA pemphigus	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Linear IgA disease (0) **1**	1. Pemphigus (0.5) 2. BP (0) 3. Pemphigus (0.5) **1**
39	Pemphigus foliaceus	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**	1. Pemphigus (0.5) 2. Pemphigus (0.5) 3. Pemphigus (0.5) **1.5**
Total score		24.5	38

[a] The number in the last row of each box is the total score of the three responses.

4. Discussion

At the time of composing this text, the present study stands as the primary descriptive analysis of ChatGPT capability to furnish insights pertinent to the fundamental comprehension, diagnosis, and treatment of AD. AD is a prevalent chronic inflammatory skin disease that significantly impacts on quality of life [28,29]. To enable early diagnosis and effective treatment of AD, it is imperative to employ reliable and accurate evaluation methods [30]. The evidence-based medicine underscores the importance of healthcare professionals delivering personalized care by leveraging robust and current medical research data [31]. In order to assess the adherence of ChatGPT-3.5 and ChatGPT-4.0 to evidence-based medicine, we conducted a comparative analysis of their performance against three distinct groups of clinicians with varying qualifications. Surprisingly, both ChatGPT versions, despite not being specifically trained on medical databases, delivered solutions that were largely congruent with evidence-based medicine. Accordingly, we observed that ChatGPT-4.0 exhibited superior performance compared to ChatGPT-3.5, scoring higher than doctors with junior professional titles (111) and medium-grade professional titles (117). It achieved an impressive total score of 122 out of 50 questions,

Table 5
Comparison of mean scores according to answers' classification.

Characteristic	ChatGPT-3.5	ChatGPT-4.0	
	mean ± SD	mean ± SD	P value with GPT-3.5
Correct	0.2308 ± 0.7420	0.2564 ± 0.7511	0.8799
Incorrect	1.974 ± 1.246	1.256 ± 1.312	0.0154
Partial	0.7949 ± 1.174	1.436 ± 1.334	0.0271

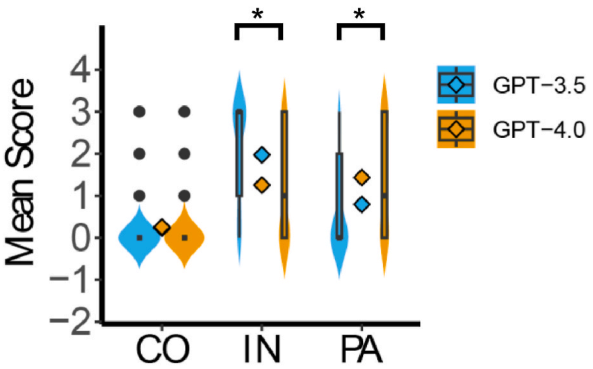


Fig. 5. The mean scores of correct, incorrect and partial correct answers given by ChatGPT-3.5 and ChatGPT-4.0 group depicted by violin plots. * represents p-value <0.05. **Abbreviations** CO, correct; IN, incorrect; PA, partial correct.

surpassing the score of ChatGPT-3.5 (106). ChatGPT-4.0(35) demonstrated a higher frequency of providing complete correct answers after three rounds of evaluation compared to doctors with medium-grade professional titles (23), associate senior titles, and senior titles (34). Despite ChatGPT-4.0 performing better than ChatGPT-3.5, it did not exceed the quality of answers given by doctors with associate senior and senior title.

Multiple published studies have indicated that the incorporation of higher-order problem-solving skills were linked to lower accuracy for ChatGPT-3.5, but not for ChatGPT-4.0 [32–34]. Likewise, we have noticed a similar situation when it comes to responding to questions in both ChatGPT-3.5 and ChatGPT-4.0. In contract, we discovered instances where neither version of ChatGPT provided accurate responses to certain queries. Both versions of ChatGPT failed to provide satisfactory answers to questions regarding, family history in relation to the emergence of AD, the use of immunosuppressive medications, the likelihood of local recurrence due to long-term TCI usage, and the effectiveness of combination therapy involving TCI and UV. Therefore, relying on the current version of the Chatbots for disseminating medical knowledge is not fully recommended. Recognizing the need to acknowledge its capacity for producing erroneous conclusions with undue certainty is crucial, as sometimes this could lead to misleading outcomes. Yeo et al. and Javaid et al. have also reported the ChatGPT can only response to the question precisely and well when the question is clear and specify [9,35].

Considering the mean score, the scores attained by the ChatGPT-3.5 for the basic knowledge about AD, diagnosis and treatment is low which is almost equal to low seniority levels health professionals. The ChatGPT-4.0 showed a significant improvement over professionals at lower and intermediate seniority levels and same was reported by the other studies [36]. The notable improvement observed can be credited to the ongoing refinement of algorithms and the iterative incorporation of user feedback facilitated by reinforcement learning [37]. Nevertheless, the high seniority professionals consistently achieved significantly higher median scores across all three types of questionnaires than ChatGPT-4.0 which signifies the importance of the experienced health professionals in any filed including the precision medicine. This could be attributed to the extensive and varied training data set, which allows it to more effectively understand the subtleties and intricacies of medical terminology and concepts [38,39]. These discoveries underscore the significance of consistently updating and fine-tuning AI models. While ChatGPT-4.0 outperforms lower and middle-tier medical professionals, it is crucial to recognize the effectiveness of senior medical practitioners in consistently providing high-quality answers. This underscores the irreplaceable value of their experience and expertise in the field. Their expertise and experience continue to be invaluable in the field of AD and should not be underestimated. This study highlights the potential for AI, particularly ChatGPT-4.0, to complement and enhance medical knowledge, but emphasizes that it should be viewed as a supportive tool rather than a replacement for the expertise and clinical judgment of seasoned healthcare professionals.

In the responses provided by ChatGPT regarding knowledge about AD, both versions of ChatGPT exhibited reduced accuracy when addressing open-ended queries in contrast to "yes or no" inquiries. To illustrate, considering the example of the query "which cytokines produced by Th2 cells, basophils, and innate lymphoid cells are important for the onset of AD?", it was observed that ChatGPT-3.5 failed to yield any accurate response, whereas ChatGPT-4.0 managed to provide a correct answer. ChatGPT 4.0 also provided precise information regarding treatment options for AD, addressing queries related to diagnosis. This enhancement strengthens its role in diagnostic and treatment decision-making, minimizing errors and bolstering effectiveness. This can be particularly beneficial for junior and mid-level clinicians. Despite this promising advancement, it is important to view ChatGPT as a supplementary source of information and guidance in the diagnostic process. The ultimate decision should still rest with the physician, and ChatGPT should not be seen as a replacement for expert medical advice [40,41]. Initial concerns about misinformation and harm were addressed, considering user influence, stereotypes, and incredulity about broader applicability of ChatGPT. However, only given the precise conditions of "please answer yes or no", ChatGPT was able to answer these 50 questions well, which indicates the quality of replies by both ChatGPT versions relies on the quality of input questions. These observations clearly suggested that unlike numerous other domains, the field of medicine is not reliant on specific tools. While occurrences of this nature are infrequent, they can occasionally yield incorrect responses. Overcoming inaccuracies in Chatbot provided medical information presents several challenges. The model faces difficulty in determining a definitive source of truth and evaluating the reliability of its training data, which may lead it to prioritize less credible sources. Supervised training can introduce biases or misinformation based on the human supervisor knowledge limitations. Minor variations in input phrasing can lead to significantly different responses. The model is unable to seek clarification for ambiguous queries, often resorting to guesswork instead. Additionally, there are transparency concerns as the Chatbot may provide inaccurate citations upon request for sources.

AIBDs are a set of severe skin diseases with clinical prevalence [42,43]. The diagnosis of various type of AIBDs requires dermatologists comprehensively examine the patients complaints, medical history, pathological slides and other factors [44,45]. Moreover, the evaluation of therapy and prognosis is inherently linked to the expertise of dermatologists [46]. Our findings suggest that substituting doctors with AI is unfeasible. This conclusion arises from the inconsistent quality of responses provided by ChatGPT regarding the AIBD. In certain instances, the responses are inadequate, which can be attributed to the limitations in the quality of the data set employed for training ChatGPT models [47]. As of the current moment, the data utilized by the two versions of ChatGPTs is still derived from the two preceding years and has not undergone recent updates. This has resulted in limited effectiveness in terms of accuracy and processing capability for the purpose of diagnosing intricate clinical cases [48,49]. Both results of the AD and AIBD query indicate that ChatGPT-4.0 is more accurate in giving answers than ChatGPT-3.5 because ChatGPT-4.0 employs a bigger training data set.

5. Conclusion

While the data collection and analytical capabilities of ChatGPT superior than humans, it remains insufficient for practical clinical diagnosis and treatment. Significant advancements are needed before ChatGPT can find effective utilization within the professional medical field.

Ethical approval

This study does not include any individual-level data and thus does not require any ethical approval.

Data availability statement

Data will be found in https://pan.baidu.com/s/1eosdy7iyUcMeauTManpFLg, the access code is:8ed8.

CRediT authorship contribution statement

Chengxiang Lian: Writing – original draft, Project administration, Data curation. **Xin Yuan:** Software, Resources. **Santosh Chokkakula:** Writing – review & editing, Resources, Methodology, Investigation. **Guanqing Wang:** Resources, Investigation. **Biao Song:** Supervision, Project administration. **Zhe Wang:** Formal analysis, Data curation. **Ge Fan:** Software. **Chengliang Yin:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

Chengliang yin, declare that he have no conflicts of interest related to the research, authors, or funding sources of the articles. He is handling as an AE of this journal. He maintain the highest standards of objectivity, fairness, and integrity in his evaluations and decisions. It is his responsibility to ensure that the articles the handle are evaluated and published based solely on their scientific merit and relevance to the journal's scope, without any undue influence or bias. He am committed to upholding the journal's standards of excellence and ethical conduct in all my work as an AE.

Acknowledgements

We would like to express our sincere gratitude to colleagues and peers for their valuable insights and suggestions during the preparation of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e37220.

References

[1] A. Gilson, 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 Medical Education 9 (1) (2023) e45312.

[2] A.J. Thirunavukarasu, et al., Large language models in medicine, Nat. Med. (2023) 1–11.

[3] Y.K. Dwivedi, et al., "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, Int. J. Inf. Manag. 71 (2023) 102642.

[4] J.C. de Winter, Can ChatGPT Pass High School Exams on English Language Comprehension, Researchgate. Preprint, 2023.

[5] T. Wu, et al., A brief overview of ChatGPT: the history, status quo and potential future development, IEEE/CAA Journal of Automatica Sinica 10 (5) (2023) 1122–1136.
[6] J. Whalen, C. Mouza, ChatGPT: challenges, opportunities, and implications for teacher education, Contemp. Issues Technol. Teach. Educ. 23 (1) (2023) 1–23.
[7] P.P. Ray, ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems. 3 (2023) 121–154.
[8] Y. Liu, et al., Summary of chatgpt/gpt-4 research and perspective towards the future of large language models, arXiv preprint arXiv:2304.01852, 2023 n.pag.
[9] M. Javaid, A. Haleem, R.P. Singh, ChatGPT for healthcare services: an emerging stage for an innovative perspective, BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (1) (2023) 100105.
[10] M.S. Rahaman, et al., From ChatGPT-3 to GPT-4: a significant advancement in ai-driven NLP tools, Journal of Engineering and Emerging Technologies 2 (1) (2023) 1–11.
[11] C. Oliveira, T. Torres, More than skin deep: the systemic nature of atopic dermatitis, Eur. J. Dermatol. 29 (2019) 250–258.
[12] T. Tsakok, et al., Does atopic dermatitis cause food allergy? A systematic review, J. Allergy Clin. Immunol. 137 (4) (2016) 1071–1078.
[13] D. Gustafsson, O. Sjöberg, T. Foucard, Development of allergies and asthma in infants and young children with atopic dermatitis–a prospective follow-up to 7 years of age, Allergy 55 (3) (2000) 240–245.
[14] H. Gu, et al., Evaluation of diagnostic criteria for atopic dermatitis: validity of the criteria of Williams et al. in a hospital-based setting, Br. J. Dermatol. 145 (3) (2001) 428–433.
[15] H. Enomoto, et al., Filaggrin null mutations are associated with atopic dermatitis and elevated levels of IgE in the Japanese population: a family and case–control study, J. Hum. Genet. 53 (7) (2008) 615–621.
[16] A.C. Krakowski, L.F. Eichenfield, M.A. Dohil, Management of atopic dermatitis in the pediatric population, Pediatrics 122 (4) (2008) 812–824.
[17] D. Becker, et al., Clinical efficacy of blue light full body irradiation as treatment option for severe atopic dermatitis, PLoS One 6 (6) (2011) e20566.
[18] A. Alexopoulos, et al., Retrospective analysis of the relationship between infantile seborrheic dermatitis and atopic dermatitis, Pediatr. Dermatol. 31 (2) (2014) 125–130.
[19] D. Wilsmann-Theis, et al., Facing psoriasis and atopic dermatitis: are there more similarities or more differences? Eur. J. Dermatol. 18 (2) (2008) 172–180.
[20] M. Napolitano, et al., Children atopic dermatitis: diagnosis, mimics, overlaps, and therapeutic implication, Dermatol. Ther. 35 (12) (2022) e15901.
[21] M. Aquino, L. Fonacier, The role of contact dermatitis in patients with atopic dermatitis, J. Allergy Clin. Immunol. Pract. 2 (4) (2014) 382–387.
[22] D.J. Hetem, M.J. Bonten, Clinical relevance of mupirocin resistance in Staphylococcus aureus, J. Hosp. Infect. 85 (4) (2013) 249–256.
[23] D.L. Brown, J.E. Frank, Diagnosis and management of syphilis, Am. Fam. Physician 68 (2) (2003) 283–290.
[24] A.D. Breithaupt, A. Alio, S.F. Friedlander, A comparative trial comparing the efficacy of tacrolimus 0.1% ointment with Aquaphor ointment for the treatment of keratosis pilaris, Pediatr. Dermatol. 28 (4) (2011) 459–460.
[25] A. Allen, et al., Significant absorption of topical tacrolimus in 3 patients with Netherton syndrome, Arch. Dermatol. 137 (6) (2001) 747–750.
[26] K.T. Amber, et al., Autoimmune subepidermal bullous diseases of the skin and mucosae: clinical features, diagnosis, and management, Clin. Rev. Allergy Immunol. 54 (2018) 26–51.
[27] V. Ruocco, et al., Pemphigus: etiology, pathogenesis, and inducing or triggering factors: facts and controversies, Clin. Dermatol. 31 (4) (2013) 374–381.
[28] C.H. Na, J. Chung, E.L. Simpson, Quality of life and disease impact of atopic dermatitis and psoriasis on children and their families, Children 6 (12) (2019) 133.
[29] A. Faraz, V. Jui, A.Y. Finlay, Counting the burden: atopic dermatitis and health-related quality of life, Acta Derm. Venereol. 100 (12) (2020).
[30] J.D. Johansen, et al., European Society of Contact Dermatitis guideline for diagnostic patch testing–recommendations on best practice, Contact Dermatitis 73 (4) (2015) 195–221.
[31] B. Kulhanek, K. Mandato, Healthcare Technology Training: an Evidence-Based Guide for Improved Quality, Springer Nature, 2022.
[32] R.S. Goodman, et al., Accuracy and reliability of Chatbot responses to physician questions, JAMA Netw. Open 6 (10) (2023).
[33] I. Levkovich, Z. Elyoseph, Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study, JMIR mental health 10 (2023) e51232.
[34] R. Ali, et al., Performance of ChatGPT and GPT-4 on neurosurgery written board examinations, medRxiv (2023) 2023, 03. 25.23287743.
[35] Y.H. Yeo, et al., Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma, medRxiv (2023) 2023, 02. 06.23285449.
[36] M. Lewandowski, et al., An original study of ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the dermatology specialty certificate examinations, Clin. Exp. Dermatol. (2023) llad255.
[37] I. Ayub, et al., Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis, Cureus 15 (8) (2023).
[38] M. Joly-Chevrier, et al., Performance of ChatGPT on a practice dermatology board certification examination, J. Cutan. Med. Surg. 27 (4) (2023) 407–409.
[39] J. Haemmerli, et al., ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health & Care Informatics 30 (1) (2023).
[40] A. Rao, et al., Evaluating ChatGPT as an adjunct for radiologic decision-making, medRxiv (2023) 2023, 02. 02.23285399.
[41] F.C. Kitamura, ChatGPT Is Shaping the Future of Medical Writing but Still Requires Human Judgment, Radiological Society of North America, 2023 e230171.
[42] M. Kurzeja, et al., Ocular involvement in autoimmune bullous diseases, Clin. Dermatol 41 (4) (2023) 481–490.
[43] A. Ajayi, S. Sathi, V. Petronic-Rosic, Autoimmune bullous diseases in skin of color, Clin. Dermatol. 40 (6) (2022) 676–685.
[44] K. Balighi, et al., Retrospective study of gingival involvement in pemphigus: a difficult to treat phenomenon, Dermatol. Ther. 35 (6) (2022) e15475.
[45] X. Li, et al., Autoimmune blistering diseases, volume II, Front. Immunol. 14 (2023) 1175962.
[46] L. Borradori, et al., Updated S2 K guidelines for the management of bullous pemphigoid initiated by the European Academy of Dermatology and Venereology (EADV), J. Eur. Acad. Dermatol. Venereol. 36 (10) (2022) 1689–1704.
[47] Y.H. Yeo, et al., GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis, medRxiv (2023) 2023, 05. 04.23289482.
[48] R. Vaishya, A. Misra, A. Vaish, ChatGPT: is this version good for healthcare and research? Diabetes Metabol. Syndr.: Clin. Res. Rev. 17 (4) (2023) 102744.
[49] Y. Shen, et al., ChatGPT and Other Large Language Models Are Double-Edged Swords, Radiological Society of North America, 2023 e230163.