frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 20 August 2025
DOI 10.3389/frai.2025.1579375

OPEN ACCESS

EDITED BY
Rashid Ibrahim Mehmood,
Islamic University of Madinah, Saudi Arabia

REVIEWED BY
Thiago Gonçalves dos Santos Martins,
Federal University of São Paulo, Brazil
Carlos Ruiz-Nuñez,
University of Malaga, Spain
Gülcan Gencer,
Afyonkarahisar Health Sciences University, Türkiye

*CORRESPONDENCE
Muhammad Hasnain
✉ drhasnain.it@leads.edu.pk
Khursheed Aurangzeb
✉ kaurangzeb@ksu.edu.sa

RECEIVED 19 February 2025
ACCEPTED 17 July 2025
PUBLISHED 20 August 2025

CITATION
Hasnain M, Aurangzeb K, Alhussein M, Ghani I and Mahmood MH (2025) AI in conjunctivitis research: assessing ChatGPT and DeepSeek for etiology, intervention, and citation integrity via hallucination rate analysis.
Front. Artif. Intell. 8:1579375.
doi: 10.3389/frai.2025.1579375

AI in conjunctivitis research: assessing ChatGPT and DeepSeek for etiology, intervention, and citation integrity via hallucination rate analysis

Muhammad Hasnain[1]*, Khursheed Aurangzeb[2]*, Musaed Alhussein[2], Imran Ghani[3] and Muhammad Hamza Mahmood[1]

[1]Department of Computer Science, Lahore Leads University, Lahore, Pakistan, [2]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, [3]Department of Computer and Information Sciences, Virginia Military Institute, Lexington, KY, United States

Introduction: The advent of large language models and their applications have gained significant attention due to their strengths in natural language processing.
Methods: In this study, ChatGPT and DeepSeek are utilized as AI models to assist in diagnosis based on the responses generated to clinical questions. Furthermore, ChatGPT, Claude, and DeepSeek are used to analyze images to assess their potential diagnostic capabilities, applying the various sensitivity analyses described. We employ prompt engineering techniques and evaluate their abilities to generate high quality responses. We propose several prompts and use them to answer important information on conjunctivitis.
Results: Our findings show that DeepSeek excels in offering precise and comprehensive information on specific topics related to conjunctivitis. DeepSeek provides detailed explanations and in depth medical insights. In contrast, the ChatGPT model provides generalized public information on the infection, which makes it more suitable for broader and less technical discussions. In this study, DeepSeek achieved a better performance with a 7% hallucination rate compared to ChatGPT's 13%. Claude demonstrated perfect 100% accuracy in binary classification, significantly outperforming ChatGPT's 62.5% accuracy.
Discussion: DeepSeek showed limited performance in understanding images dataset on conjunctivitis. This comparative analysis serves as an insightful reference for scholars and health professionals applying these models in varying medical contexts.

KEYWORDS

ChatGPT, comprehensiveness, DeepSeek, eye infection, prompts

1 Introduction

Conjunctivitis, also known as "pink eye," is an ocular condition that can affect individuals of all ages. It has adverse impacts on public and private economies and their productivity (Gunay et al., 2015). Various antibiotics were prescribed for bacterial conjunctivitis, which are now not frequently used due to bacterial resistance and safety concerns (Karpecki et al., 2010). Healthcare practitioners cannot discriminate between the viral and bacterial infections of conjunctivitis, resulting in over 80% of patients receiving antibiotics. Recent reports indicate that over 900,000 antibiotic prescriptions were dispensed in the Netherlands, incurring a cost of $10.9 million (Rietveld et al., 2004).

Recently, Pakistan faced a big health issue of conjunctivitis. We witnessed an alarming surge in infection in different areas of Pakistan. More than 86,133 cases were confirmed in Pakistan (Ahmed et al., 2024). This infection rapidly spread to other parts of the country. This issue demanded immediate attention from healthcare professionals, government officials, the research community, policymakers, and pharmaceuticals. Medical and research communities could play a big role in a coordinated and strategic way to overcome the prevalence and pressing issue of conjunctivitis infection (Alves et al., 2023). Figure 1 is the illustration of pink eye conjunctivitis.

Several factors contribute to the escalating cases of conjunctivitis. Notably, the lack of clean water and sanitation facilities can facilitate the spread of this highly contagious infection (Jasper et al., 2012). Moreover, insufficient education and public awareness regarding eye hygiene practices may contribute to its transmission. Recent research does not definitively establish conjunctivitis as the sole sign of a sexually transmitted infection (Abedifar et al., 2023). A recent study stated that the conjunctivitis rate was highest among children under 7 years of age. Viral conjunctivitis could be caused by contact (Hashmi et al., 2024). This study supports the observation that conjunctivitis is common in people of all ages. A child with no previous history of conjunctivitis could catch this infection from family members. Previous studies evidenced that bacterial conjunctivitis was more common in children than adults.

FIGURE 1
Pink eye conjunctivitis.

Large language models (LLMs) have several applications in clinical settings. LLMs show potential in image analysis when integrated with the computer-aided diagnosis (CAD). LLMs summarize information in natural language that help in diagnosis and generating health reports (Wang et al., 2024). In a recent research, researchers revealed that ChatGPT-4 and PaLM2 LLMs achieved better performance in answering questions on ocular surface diseases (Ling et al., 2025). However, ChatGPT-4 offered potential in improving the accuracy of question-answering and may serve as a valuable tool for medical professionals in future. Previous studies have not compared ChatGPT and emerging AI application such as DeepSeek on health topics. Most recent research employed variants of ChatGPT, such as ChatGPT 3.5 and ChatGPT 4.0, and compared their abilities to diagnose glaucoma in a ocular hypertension treatment study (OHTS) dataset (Raja et al., 2025). ChatGPT-4 achieved better accuracy in diagnosing glaucoma compared to ChatGPT 3.5. ChatGPT serves as a beneficial tool in exploring ocular hypertensive eye when specific information is available. However, none of the studies examined the role of ChatGPT-4 compared to DeepSeek, an emerging LLM, on conjunctivitis topics. Based on the available data on conjunctivitis, our study fills up this research gap by comparing the performance of ChatGPT 4.0 and DeepSeek on health topics related to conjunctivitis.

This study has the following contributions:

- This study aims to reveal how AI models may assist in recalling or identifying relevant factors that support the clinical diagnosis of conjunctivitis.
- It compares the performance of ChatGPT and DeepSeek on topics related to conjunctivitis infection and evaluates how the two models are powerful in giving precise and comprehensive information.
- We use the hallucination rate metric to test factual accuracies and compare the performance of ChatGPT and DeepSeek models.
- This study compares the capabilities of ChatGPT, Claude, and DeepSeek models in classifying healthy and conjunctivitis-affected eyes from a real-world dataset.

The layout of this study is as follows:

We detail research methods in Section 2. The results and their discussion are given in Section 3. The main points of this study are concluded in Section 4.

2 Literature review

Ophthalmology has experienced significant improvements from artificial intelligence (AI) in recent years through its enhanced management of eye infections. The Knowledge-Guided Diagnosis Model (KGDM) emerged from research (Chen et al., 2020) as an interpretable AI system to support infectious keratitis diagnosis. This model combines expert clinical understanding with data intelligence to generate visual reasoning systems that use AI-based biomarkers and comparable case searches. The KGDM exhibited excellent diagnostic capabilities across various datasets. At the

same time, it improved medical worker performance by working alongside human experts, which presents significant potential for AI use in eye infection diagnosis.

The study by Rajarajeshwari and Selvi (2024) conducted a detailed review of AI usage in major retinal conditions. The research emphasized early detection capabilities while assessing multiple diagnostic and structural anomaly detection algorithms for retinal diseases. According to the authors, future research should concentrate on boosting computational complexity while using existing public datasets to increase the reliability of AI-based detection methods.

The DATUM alpha study by Odaibo et al. (2019) evaluated the cloud-based mobile AI technology for retinal disease detection within mobile health technology. The mobile application named Fluid Intelligence underwent testing to detect subretinal fluid and macular edema from optical coherence tomography (OCT) scans. The AI system demonstrated a weighted average sensitivity of 89.3% along with a specificity of 81.23%, indicating its value as a screening tool, particularly for underserved areas.

In a study (Yoo et al., 2019), authors analyzed current developments in central serous retinopathy (CSR) detection, which combines imaging with artificial intelligence methods. The researchers analyzed traditional imaging approaches such as OCT and fundus imaging alongside machine and deep learning methods. The study findings show that deep learning classifiers perform CSR detection with high accuracy while also achieving fast results and reliable outcomes. However, more research is necessary to optimize the computational performance and establish their effectiveness on open-source datasets.

ChatGPT supports graduates to score better than their peers from medical schools. ChatGPT and other similar applications provide interactive support, personalized learning, motivation boosting, and self-assessment (Gencer and Gencer, 2024). Consequently, most recent study presents a bibliographic analysis to show current research trends. The results show that ChatGPT, clinical management, natural language processing (NLP), Chatbots, and virtual reality indicate a shift toward addressing implications at a higher level for LLMs and their applications in healthcare (Gencer and Gencer, 2025). ChatGPT applications are widely discussed in the literature. One of the studies highlights applications for improving the effectiveness and efficiency of educational, clinical, and research work in medicine (Thirunavukarasu et al., 2023).

3 Research methods

3.1 ChatGPT 4.0

ChatGPT application adds to medical education clinical decision-making. A recent study examined the role of ChatGPT and its performance in assessing ophthalmology-related queries (Alexander et al., 2024). ChatGPT and its variants demonstrate a strong capability in handling medical knowledge. Even ChatGPT and Bard as large language models (LLMs) show remarkable performance in domains such as intelligence diagnostics (Caruccio et al., 2024a).

During times of crisis, we rely on invaluable information and support from generative artificial intelligence applications such as ChatGPT, which can promptly respond to inquiries posed to the application. In the context provided above, we posed several questions to ChatGPT for public awareness and education regarding conjunctivitis infection.

3.2 DeepSeek

DeepSeek was developed by a Chinese AI company, which can process users' queries at a low cost. Open source and cost-efficient DeepSeek is disrupting the conventional AI approaches and their applications in various fields (Krause, 2025). The basic architecture of DeepSeek is based on the transformer framework.

3.3 Claude

Anthropic released its AI model, Claude 3.5 Sonnet, on 21 June 2024. As per official statements, Claude 3.5 Sonnet is superior in performance compared to GPT-4, Gemini, and other predecessors. The Claude model provides rapid response to prompts for generating discharge summaries for kidney patients (Anthropic, 2024). It can handle complex medical tasks and effectively integrate with medical field (Jin et al., 2024). Research on the Claude model showed an improved performance using imaging data and medical history. It showed improved tumor, node, metastasis (TNM) classification results using the radiology reports of pancreatic cancer patients (Suzuki, 2025). Hence, we use the Claude model for classification task and compare its performance with ChatGPT and DeepSeek models.

3.4 Prompt engineering

The method of designing optimized input prompts functions such as prompt engineering to guide GPT models toward generating specific output content. The process of prompt engineering requires the creation of questions or instructions that match both model abilities and necessary task functions. Zero-shot prompt engineering enables the creation of prompts that allow the system to execute untrained functions without any supplemental fine-tuning operations. The generalization capabilities of the model allow this approach to work. When GPT receives text summarization or problem-solving requests before any specific training occurs, it operates through zero-shot prompting. The majority of studies used a zero-shot prompt engineering technique, followed by a few-shot technique. The few-shot technique outperforms a zero-shot technique in addressing multiple-choice questions, while its performance remains inconsistent in NLP tasks. A mixed prompt engineering technique has been reported to lead to a better performance than using any prompt engineering technique alone (Zaghir et al., 2024). Hence, we used a mix-up of zero-shot and few-shot prompt engineering techniques in this study. Prompt engineering may enhance the learning capability and generalization of an LLM on a small dataset and fully exploit the potential performance of a model.

3.5 Hallucination rate

The hallucination rate is used to quantify the proportion of Generative models' produced references, which are irrelevant, incorrect, and unsupportive by the available literature (Chelli et al., 2024). When LLMs are confronted with factual questions, hallucinations may occur, resulting in information that is factually incorrect but syntactically and grammatically correct. The hallucination rate is judged by using a systematic verification process based on PubMed and Google Scholar to ensure accuracy, legitimacy, and transparency. Each reference is verified by searching for its title in PubMed and Google Scholar databases (Aljamaan et al., 2024). In case of failing to reach a reference in the PubMed database, Google Scholar was the second choice for verification using a paper's title. Hallucination rate can be expressed as given in the Equation 1.

$$\text{Hallucination rate} = \frac{\text{No of fake or inaccurate references}}{\text{Total references generated}} \times 100 \quad (1)$$

The hallucination rate is measured by conducting human evaluations. Annotators compare the model's generated output with ground truth references, thereby flagging fabricated, incorrect, or irrelevant instances of responses. Next, we calculate the hallucination rate by using the formula in Equation 1. We preferred using human evaluations over automated metrics, such as BLEU or ROUGE, because humans can detect contextual hallucination, which is overlooked by automated metrics.

3.6 Data synthesis

The process of data synthesis creates artificial datasets that replicate actual data to overcome data shortages and unbalanced distributions and protect sensitive information. The creation of realistic dataset succeeds through the application of algorithms alongside simulations and models, including Generative Adversarial Networks (GANs), and no synthetic dataset was employed in this study. Data synthesis enables important tasks such as model training and sensitive domain research while protecting actual data from disclosure.

3.7 Contour count

Eyelid contour abnormalities are widely reported in the literature. In a study, researchers reported that postoperative and preoperative digital images were used for eyelid analysis (Choudhary et al., 2016). A contour represents a curve that joins continuous points along a boundary with the same intensity or color. Thus, contour count is the number of distinct boundaries, which are detected in an eye image. The rationale behind using contour count as a feature of conjunctivitis relates to the iris, pupil, and sclera being smooth and distinct for a healthy eye. For conjunctivitis eye, inflammation or redness creates more edges or irregular shapes. When blood vessels become more prominent, it could increase the contour detected. Alongside contour count, the mean intensity feature is employed in this study to increase accuracy and reliability in results. Contour count and mean intensity features could help us differentiate between healthy eyes and those affected by conjunctivitis.

Computer vision directly interprets images, faces, and scenes and detects objects using convolutional neural networks (CNNs) and transformers. On the other hand, natural language processing (NLP) processes texts extracted from images using the optical character recognition (OCR) technique. NLP techniques are used to analyze words but not visuals. Computer vision understands pixels, while NLP relies on the pre-trained text. NLP allows computers to understand human languages such as Spanish and English, by converting letters into numbers using some algorithms (Manovich, 2021).

3.8 Mean intensity

Mean intensity as a feature to distinguish between healthy eye and eyes affected by conjunctivitis is used in this study. Mean intensity refers to the average pixel brightness within the eye region (Singh et al., 2021). A healthy eye provides moderate values of the mean intensity due to consistent illumination and reflectance. In conjunctivitis, vascular changes and inflammation alter reflectance, resulting in irregular or elevated intensity patterns. This feature is highly supportive for early detection and differentiation and provides the quantitative examination of tissue health.

3.9 Pre-processing and standardization

For LLMs used in this study, image classification involves converting visual information into textual prompts because of the text-based nature of these models. Pre-processing image data include resizing images and converting them to the textual description. Moreover, we standardize input information and ensure a uniform prompt structure. The input lines of a model are carefully constructed prompts that embed the clinical context and image-driven characteristics and allow a model to infer a reliable diagnostic classification from structured textual prompts.

3.10 Classification accuracy metric

3.10.1 Accuracy

Accuracy is a measure that shows the proportion of accurate predictions relative to all predictions. The formula of accuracy measure is expressed in the Equation 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

3.10.2 Precision

The precision measure is the ratio of correctly classified positive values to the estimated total number of positive samples measured. The precision is expressed in the following Equation 3.

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (3)$$

3.10.3 Recall

The recall measure is about determining the proportion of positive values, which are correctly classified. The recall formula is given in the Equation 4.

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (4)$$

3.10.4 F1-score

The F1-score can be described as a harmonic mean of precision and recall measures. Precision and recall equally contribute to the score and ensure the reliability of the classifier. This measure can be expressed in the following Equation 5.

$$F1 - Score = 2\left(\frac{Precison^{*}Recall}{Precision + Recall}\right) \quad (5)$$

3.11 Ethical considerations

AI ethical concerns aim to establish three major standards related to design approaches and system usage, which consist of promoting fairness and transparency and maintaining accounting responsibilities. Three main ethical challenges stem from correcting data bias, protecting user privacy, and managing consent within sensitive application frameworks. Medical AI systems operating in eye infection management should first prioritize concentrating on patient protection combined with model transparency alongside equal technology access for all patients. The development of responsible innovation depends heavily on ethical guidelines together with regulatory compliance to build trust between stakeholders.

4 Results

4.1 Dataset description

This study employed a real-world Kaggle conjunctivitis dataset. The dataset contains 181 healthy eye's images and 177 infected images with conjunctivitis. This dataset can be accessed from the Kaggle website (Sofiya, 2025).

4.2 Results description

We asked targeted questions to evaluate the ChatGPT's capability in processing and analyzing diverse clinical information. Moreover, we assessed the ChatGPT's utility to manage the disease with multiple etiologies.

The first prompt was, "Is there any multifaceted approach to mitigate conjunctivitis infection?" In response to this prompt, ChatGPT's response was as follows:

A multifaceted approach can be employed to mitigate conjunctivitis. This includes ensuring access to clean water, promoting hygiene, providing education, enabling early diagnosis, establishing treatment protocols, implementing surveillance, administering vaccinations, fostering research collaboration, and advocating for policies. DeepSeek provides the following response for the same prompt. In addition to the response from DeepSeek, we compared the responses generated by DeepSeek to ChatGPT 4.0 in the following Table 1.

TABLE 1 Comparison of ChatGPT and DeepSeek on various aspects of conjunctivitis.

Aspect	DeepSeek	ChatGPT 4.0
Scope	Focused on conjunctivitis-specific prevention and management	Broader, addressing infectious diseases at a systemic level
Clean water	Not mentioned	Ensuring access to clean water for hygiene and sanitation
Hygiene	Emphasizes personal hygiene practices	Promotes hygiene at community and systemic levels
Education	Limited to public awareness about conjunctivitis	Includes broader education on infectious diseases and prevention
Early diagnosis	Focused on prompt medical consultation for conjunctivitis	Enabling early diagnosis for a wide range of infectious diseases
Treatment protocols	Specific to conjunctivitis medications	Establishes protocols for treating various infectious diseases
Surveillance	Not mentioned	Implements disease surveillance systems
Vaccinations	Not mentioned	Includes administering vaccinations for preventable diseases
Research collaboration	Not mentioned	Fosters collaboration for research on infectious diseases
Policy advocacy	Not mentioned	Advocates for policies to support public health initiatives

Based on the data in Table 1, we can make a difference between two AI models with the same prompt. DeepSeek provided us with specific information on conjunctivitis, while ChatGPT's reply was focused on a comprehensive framework for controlling infectious diseases. DeepSeek does not fully capture some systemic focus, such as access to clean water, surveillance, and policy advocacy. ChatGPT response incorporates vaccinations and research collaboration strategies, which are absent in a response from DeepSeek. Response from DeepSeek is primarily focused on individual and community-level actions to overcome the conjunctivitis infection.

On the other hand, ChatGPT, in contrast to DeepSeek, provides broader systematic health measures that are common to other infections. Therefore, DeepSeek provides more precise information about health issues. Each measure given by DeepSeek is focused on either preventing or managing the health issue. DeepSeek recommends antihistamines for allergic conjunctivitis; however, it missed the mention of clean water in its response and instead responds with good hygiene, which depends on the clean water. DeepSeek analyzes thoroughly while answering the prompt and significantly checks the context of the prompt before responding

to the users. Therefore, DeepSeek is superior to ChatGPT in terms of contextual values and provides precise, practical, and directly applicable information. For public health, ChatGPT is superior since it reveals information on addressing systematic gaps, scalability, and multi-disease resilience. Next, we have prompt 2 and its response from ChatGPT 4.0 in the Figure 2.

While AI models are showing success in healthcare, it is necessary to use them appropriately. As conjunctivitis symptoms vary from person to person, AI is helpful but cannot replace the medical professional. Sometimes, AI models trained on incomplete data may provide incorrect diagnoses. To ensure accurate diagnosis, comprehensive data on patients' medical history, lifestyle, and allergies are essential. In response to prompt 2, DeepSeek provides a roadmap for how AI tools accomplish the tasks of differentiating various types of conjunctivitis. The roadmap given by DeepSeek is very precise and accurate, including data collection using symptom duration, symptom severity, and additional features of conjunctivitis. The next phase is feature extraction, in which three prominent types of conjunctivitis, such as viral, bacterial, and allergic conjunctivitis, are used. The pattern recognition and diagnostic models follow this phase. In the following, we can see the response from DeepSeek about pattern recognition and diagnostic algorithms (Figure 3).

In addition to the abovementioned phases, DeepSeek, in its response, further explores additional data integration to improve accuracy. For example, data on patient history, environmental factors, and lab results can be used to improve the accuracy of the results. Depending on the information and processing of AI models, types of conjunctivitis are identified.

In the third prompt, we asked, "What is the best treatment for conjunctivitis infection?". In response, we have learned about five types of conjunctivitis infections and their best treatments by ChatGPT, as given in the following Table 2.

Table 2 presents some key types of conjunctivitis infection along with treatment and remarks. Based on the conjunctivitis type, better treatment can be suggested to patients. However, we did not know how to spread these types, so we asked ChatGPT to reveal the causes.

In comparison with the ChatGPT, DeepSeek, in response to the prompt "What is the best treatment of conjunctivitis infection?", revealed useful information. However, we found some common information from the two models. For instance, both DeepDeek and ChatGPT models recommend supportive care for viral conjunctivitis. However, DeepSeek recommends cold compresses for comfort, while ChatGPT favors warm compresses and reveals a resolution of 1–2 weeks. The information given by the two models for bacterial and allergic conjunctivitis aligns with each other. In summary, both models present complementary management strategies.

The fourth prompt was, "Is conjunctivitis infection dangerous and can it cause eye impairment?" The response of ChatGPT was something as follows: conjunctivitis infection is typically not dangerous but can cause discomfort and vision issues in patients. In severe cases, complications may arise, but permanent eye impairment is rare. However, practicing good hygiene and seeking prompt treatment can effectively address the issue and aid in a swift recovery. In a recent study (Jasim and Al-Jubory, 2024), the authors stated that conjunctivitis is a common type of eye condition that can cause discomfort and inconvenience. However, its effects could be minimized using appropriate prevention and management techniques.

In Figure 4, we provide prompt five and its corresponding response regarding diagnostic support for the ChatGPT application.

ChatGPT and other AI tools can serve as decision-supporting tools to analyze complex clinical data. Based on the analysis and response from the ChatGPT application, some diagnostic tests can be suggested. Based on variations in symptoms of infection, AI models can identify the patterns indicating viral, bacterial, or allergic conjunctivitis. AI models are capable of accurately

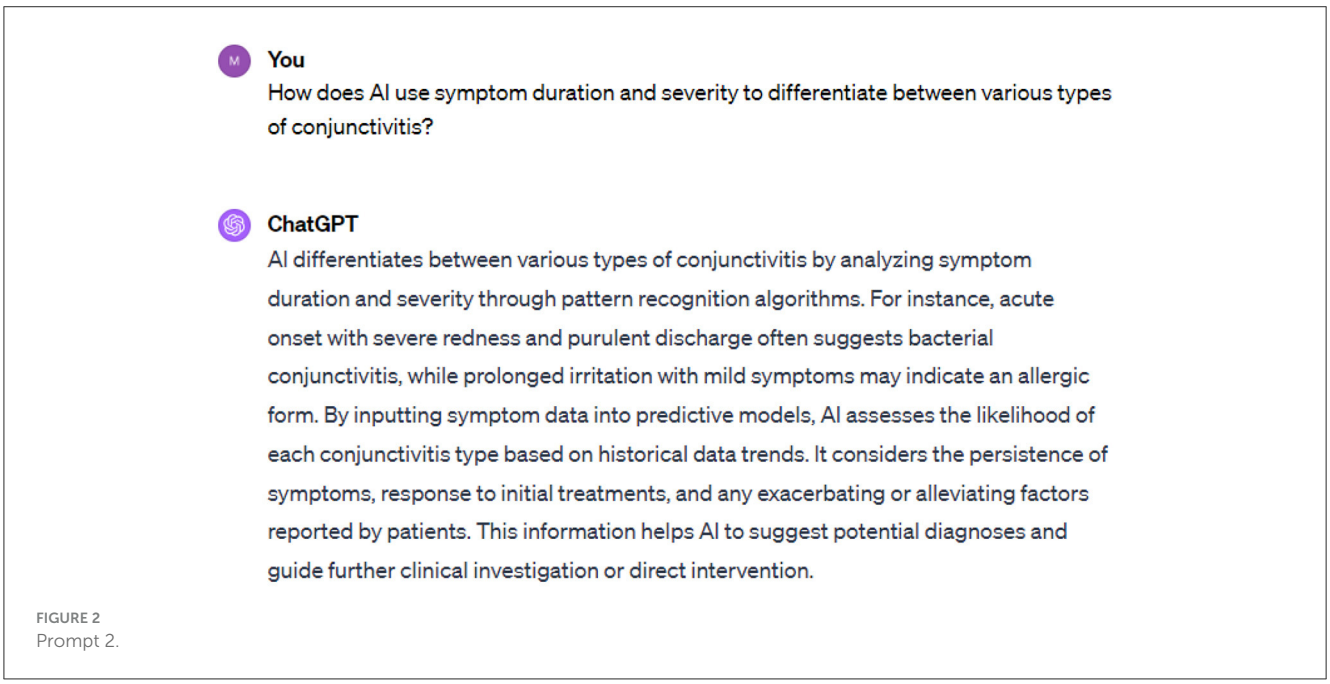

FIGURE 2
Prompt 2.

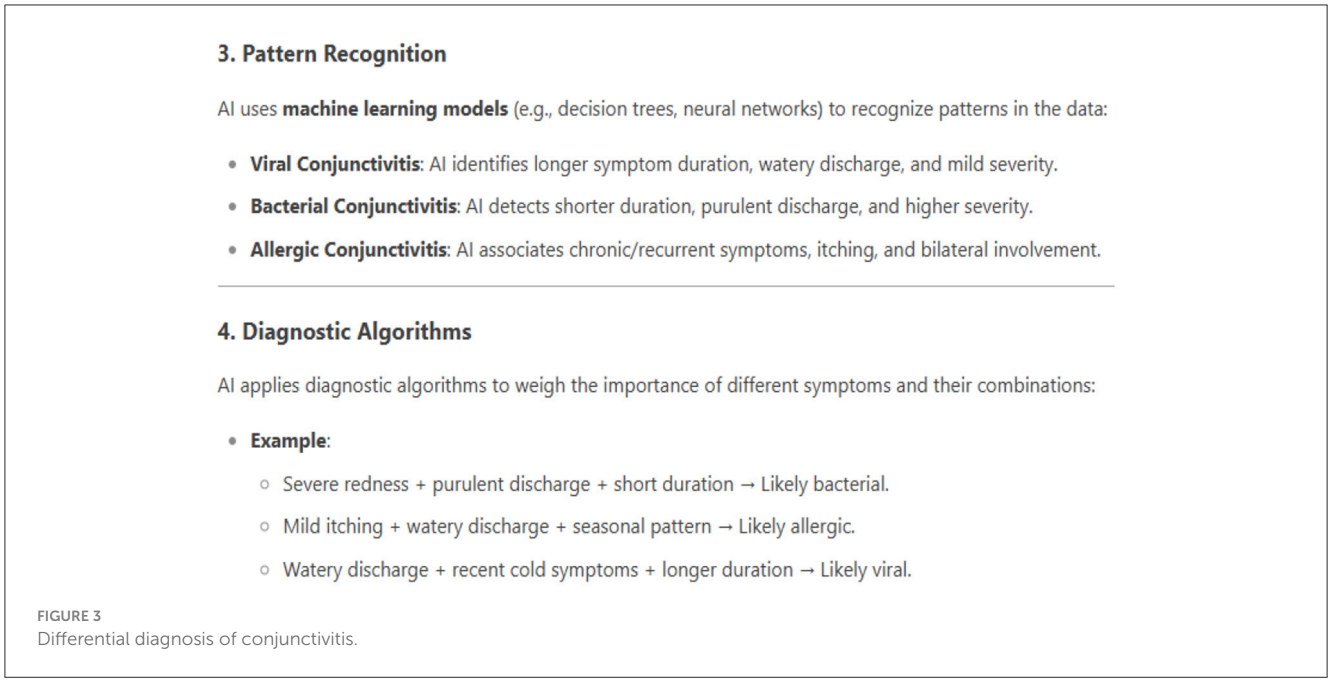

FIGURE 3
Differential diagnosis of conjunctivitis.

TABLE 2 Conjunctivitis infection types and treatment.

Sr. No.	Conjunctivitis infection type	Treatment	Remarks
1.	Viral conjunctivitis	Warm compresses and artificial tears can provide relief	It takes 1–2 weeks to resolve the infection
2.	Bacterial conjunctivitis	Antibiotics, eye drops, and ointment are prescribed	A complete course of antibiotics is important
3.	Allergic conjunctivitis	Antihistamine eye drops and oral medicine can relieve the patients	No remarks
4.	Irritant conjunctivitis	Flushing the eye with clean water and	Physicians can prescribe eye drops
5.	Giant papillary conjunctivitis (GPC)	Avoid using contact lenses and eye drops	Physicians can prescribe effective medicines

diagnosing the disease by integrating imaging information with electronic health records. This could result in reducing the diagnostic time and enhancing the care of patients.

The ChatGPT application helps healthcare professionals in identifying the etiology of conjunctivitis without suggesting any laboratory test (Figure 5). For instance, an allergic type is best recognized in the context of allergy season and viral etiology in the presence of upper respiratory symptoms.

In response to a prompt "*How PubMed library is used for Conjunctivitis discussion. Cite this discussion with 30 references and provide a complete list of references*," we received a response to this prompt, we obtained 15 and 30 references from ChatGPT and DeepSeek models, respectively. We cross-checked all references in PubMed and Google databases, from which we received the following hallucination score, as shown in Table 3.

Table 3 shows us the hallucination rate of two models in the context of references related to conjunctivitis. ChatGPT model provided us with a list of 15 references. After manually checking the list, we found some additional errors in the references. Some references did not contain a complete list of authors, while others showed incorrect issue and volume numbers. We were interested in searching for articles using their titles. Based on the titles, DeepSeek provided 30 correct references. However, we found duplicate titles in the list generated by DeepSeek.

In addition, the DeepSeek model only provided the list of titles from the PubMed database. The hallucination rate for DeepSeek was better than that of ChatGPT; notably, DeepSeek produced 30 references as were asked in the prompt. The contextual understanding of DeepSeek is superior to that of ChatGPT.

Compared to ChatGPT (13%), DeepSeek appears to be a more reliable model for generating academic references on conjunctivitis topics because of its low hallucination rate (7%). On the other hand, ChatGPT produced fewer references and a low proportion of references, which could be a concern and challenging to various stakeholders. DeepSeek is better in this performance scenario, and it becomes worrying for ChatGPT users to rely on the limited and fabricated information. Therefore, it is necessary for researchers to manually check all references to ensure academic rigor. These findings could better serve as a basis for future research or for refining prompt techniques to minimize hallucination in generating references. Prompt engineering is highly essential in highly accurate scenarios, such as medical and healthcare applications (Guo et al., 2024). However, inherent bias in training can give rise to discriminatory viewpoints, which, as a result, may produce hallucinations stemming from the overfitting of training data or lack of contextual understanding. The high rate of hallucination indicates that an AI model identifies some keywords in the prompt and aims to search for relevant references on some topics. However, it fails to find relevant and correct references. Therefore, depending only on the reference title is risky, and we

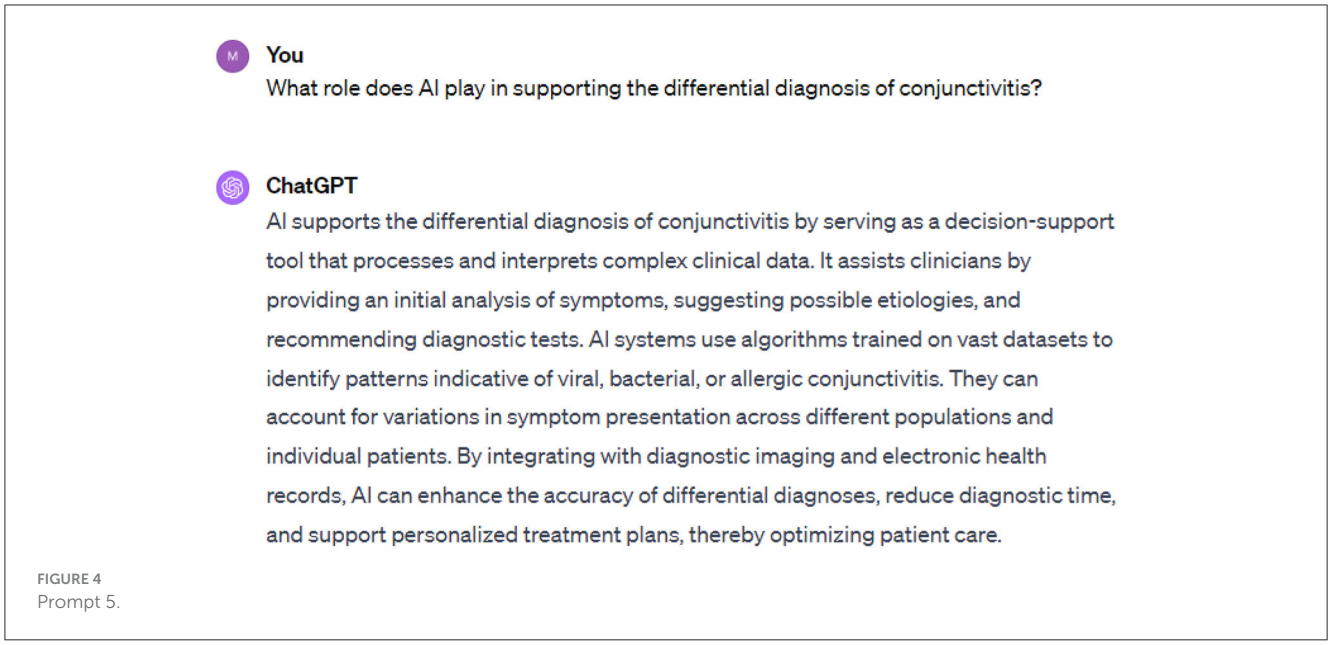

FIGURE 4
Prompt 5.

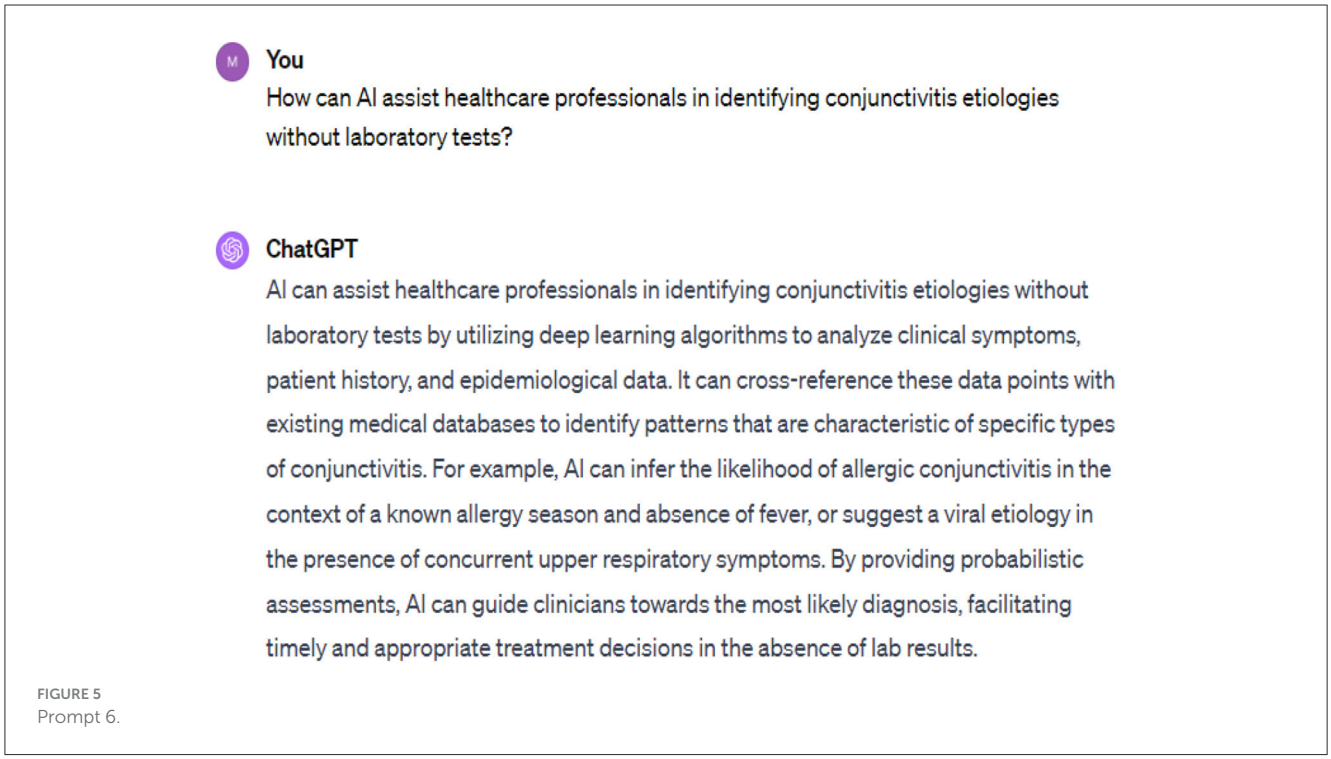

FIGURE 5
Prompt 6.

need to use identifier hallucination, such as journal names or author names. The findings of our study align with a previous study, which showed a high hallucination rate for ChatGPT 3.5 and ChatGPT 4.0 in various aspects of bibliography items. Despite ChatGPT 4.0 showing an improved hallucination rate over ChatGPT 3.5, fabricated references persist.

4.3 Classification results

ChatGPT works on principles using contour count and mean intensity for the classification of healthy eyes and those affected by conjunctivitis in this research. The number of contour indicates the structural complexity of the eye. In conjunctivitis, the increased inflammation and vascular activity lead to a higher contour count. Mean intensity is applied to measure the overall brightness of images and closely relates with the redness of eyes. Both factors evidence the conjunctivitis.

Contour number and mean intensity factors, as shown in Table 4, help in classifying ocular images. Compared to conjunctivitis, a healthy eye exhibits fewer contours with the smooth transition. ChatGPT's classification of conjunctivitis and health images is based on the proposal of classification criteria.

10.3389/frai.2025.1579375

Contour count and mean intensity were used as factors for the classification of conjunctivitis.

Figure 6 presents confusion matrix results obtained from the ChatGPT model. We have a total of 10 instances on conjunctivitis. Out of the 10 instances, five were correctly predicted and rest of the instances were incorrectly identified. On the other hand, the ChatGPT model correctly identified only one out of six instances.

Table 5 shows us classification results of conjunctivitis from the Claude model. These results show a perfect accuracy, as images of all six healthy eyes and 10 conjunctivitis-affected eyes were correctly classified into their respective categories. Data in Table 4 and confusion matrix results in Figure 7 show no

TABLE 3 Hallucination rate of ChatGPT and DeepSeek models.

Model	Total references generated	Fake references	Hallucination rate (%)
ChatGPT	15	2	13%
DeepSeek	30	2	7%

TABLE 4 Classification results from ChatGPT model.

Image	Contour count	Mean intensity	Classification
10.jpg	98	113.17	Healthy
11.jpg	181	143.82	Conjunctivitis
12.jpg	430	139.59	Conjunctivitis
13.jpg	66	163.46	Conjunctivitis
15.jpg	279	152.60	Conjunctivitis
17.jpg	114	158.07	Conjunctivitis

TABLE 5 Claude classification results.

Image	Contour count	Mean intensity	Classification
1	5–8	95–105	Healthy
2	6–9	100–110	Healthy
3	7–10	105–115	Healthy
4	5–8	90–100	Healthy
5	8–11	100–110	Healthy
6	6–9	95–105	Healthy
7	35–40	165–175	Severe conjunctivitis
8	25–30	155–165	Moderate conjunctivitis
9	40–45	170–180	Severe conjunctivitis
10	30–35	160–170	Moderate–severe conjunctivitis
11	25–30	155–165	Moderate conjunctivitis
12	35–40	165–175	Severe conjunctivitis
13	30–35	160–170	Moderate–severe conjunctivitis
14	25–30	155–165	Moderate conjunctivitis
15	30–35	160–170	Moderate–severe conjunctivitis
16	35–40	165–175	Severe conjunctivitis

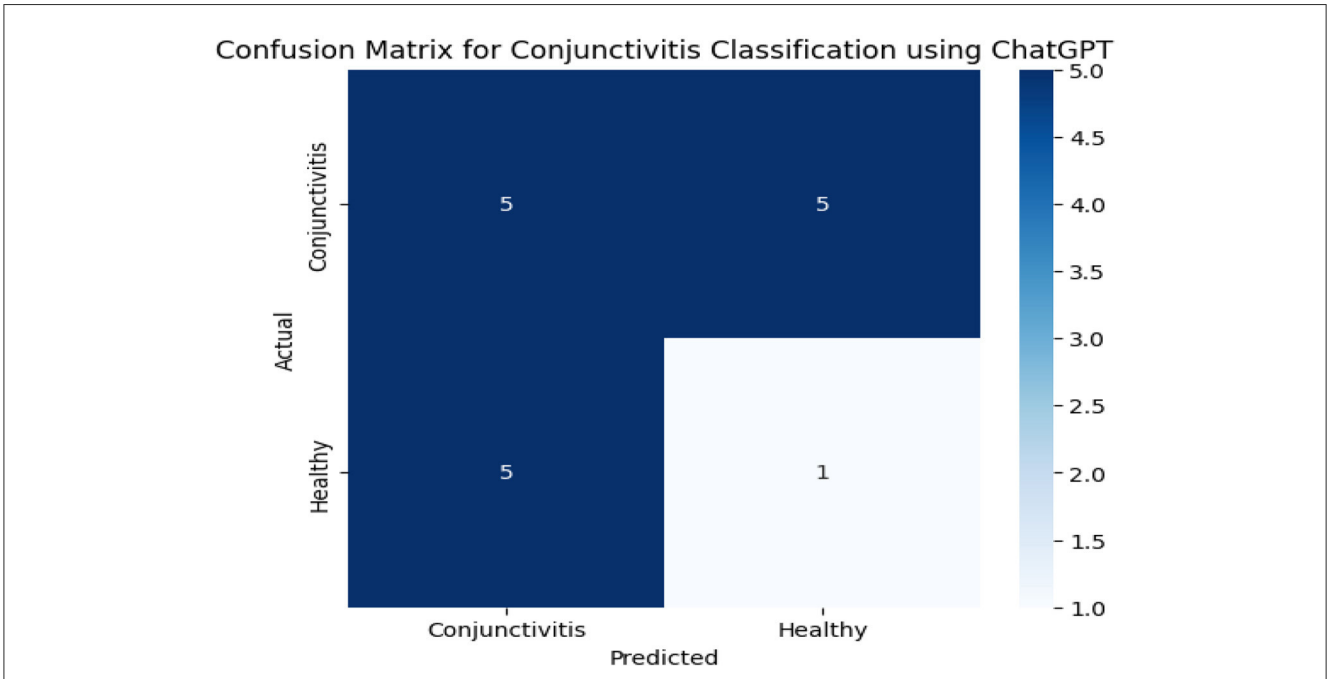

FIGURE 6
Confusion matrix results of the ChatGPT model.

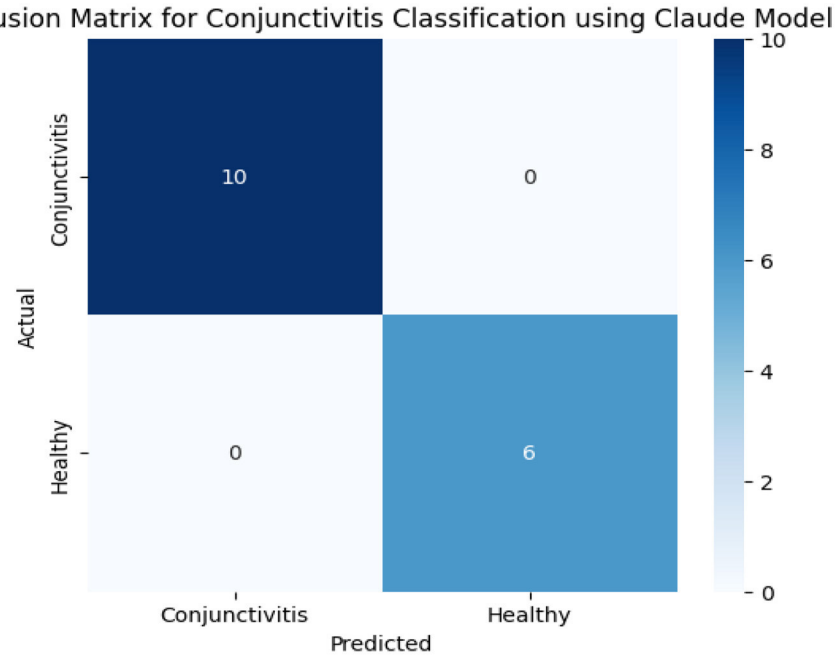

FIGURE 7
Confusion matrix results of the Claude model.

TABLE 6 Performance comparison results.

Model	Accuracy	Precision	Recall	F1-score	Specificity
Claude	100%	100%	100%	100%	100%
ChatGPT	62.5	83.33	50.00	63.50	83.33
DeepSeek	–	–	–	–	–

misclassification and ensure the high precision, recall, and F1-score. This reveals the effectiveness of the Claude model in distinguishing between healthy and conjunctivitis cases. Based on the contour count and mean intensity values, Claude further divided conjunctivitis cases into its sub-categories such as severe conjunctivitis, moderate conjunctivitis, and moderate–severe conjunctivitis.

Using classification results from ChatGPT, Claude, and DeepSeek models, we obtained performance accuracy measures, as shown in Table 6.

The ChatGPT and Claude models show classification of dataset on conjunctivitis. DeepSeek does not favor to classify the image dataset and present some error messages when uploading images.

5 Discussion

We studied most recent literature on DeepSeek and other LLMs and did not find evidence on the success of DeepSeek for image classification. Several preprints of the research articles were accessed from the Google Scholar database. We even searched for articles from other databases such as ScienceDirect and Springer for related papers.

This study used only 16 images from real-world dataset containing a total of 359 images (healthy/infected eyes), but each selected case represented key conjunctivitis types, which ensured clinically relevant analysis despite the limited sample size. Each image was carefully selected by experts and represented the etiological subtypes. A small but curated dataset provided us with more meaningful performance metrics compared to a larger noisier collection. Our methodology aligns with FDA guidelines for early-stage machine learning or AI medical device validation, where small but high-quality datasets are accepted for initial capability demonstration. This is supported by a recent study that reported the sample size used for a task using LLMs can be small yet effective (Majdik et al., 2024). Furthermore, this study's results indicated that only small samples were required for substantial improvement.

In this study, we found some discrepancies in contour count and mean intensity values for yielding classification results. The inconsistencies in contour count metrics between ChatGPT and Claude arise due to differences in their underlying image processing pipelines and model-specific designs. Factors such as post-processing technique, sensitivity thresholds, and distinct edge-detection features of algorithms inherently influence contour quantification. For example, post-processing logic in noise reduction thresholds further contributes to discrepancies. ChatGPT prioritizes generalizability across various image types, while Claude's training emphasizes precision in medical imaging. However, we need to calibrate both models against controlled datasets with ground truth annotations to isolate the biases in the two models from true performance gaps.

ChatGPT and Claude models present approximate values for mean intensity while distinguishing between healthy and conjunctivitis eyes. However, two models showed strong differences in contour count values. For example, ChatGPT marked an image healthy with 98 contour count while Claude took an image as healthy with 5–11 contour count. This discrepancy may arise as each model has its own algorithm and image preprocessing phases. The Claude model might use various thresholds or edge-detection techniques to count contours and results in lower or higher counts compared to the ChatGPT's simulated metrics. In essence, unique configuration and preprocessing techniques of ChatGPT and Claude models led to different numerical outputs, and even both models were used to analyze similar features.

The most recent version of ChatGPT (ver. 4.0) has shown higher accuracy (85%) in diagnosing corneal eye disease compared to the previous version (ver. 3.5), which achieved an accuracy of 60% (Delsoz et al., 2024). Performance optimization of ChatGPT might be due to its training on more eye information records. Another study presented a comparison of ChatGPT with Isabel on the ophthalmic dataset. The role of ChatGPT as a diagnostic tool has been examined. Isabel only correctly identified one out of 10 cases, and ChatGPT identified nine cases correctly (Balas and Ing, 2023). This helps us to improve the diagnostic accuracy of health issues by learning from past cases of ophthalmic diagnosis.

A pilot study was performed to assess the capability of ChatGPT-4 for diagnosing the rare eye disease. A set of 10 treatable rare ophthalmic case studies were used in this study. The results indicated that GPT-4 could serve as a consultation tool for patients and families to obtain referral suggestions and assistance (Hu et al., 2023). However, despite these strengths, ChatGPT-4 shows limitations such as concern for patients' privacy. Moreover, ChatGPT may provide misconceptions as it was originally developed for general purposes. Later, researchers used them to make decisions in clinical diagnosis.

Claude model in conjunctivitis classification. The ChatGPT model struggles with the differentiating features, while the Claude model achieved perfect classification and correctly classified all cases with no false positives or negatives. These results are aligned with the findings of a previous study that achieved better classification accuracy for the Claude model compared to the ChatGPT model (Caruccio et al., 2024b). Our study's classification performance is overall better in comparison with the previous study. Claude 3.5 Sonnet has stronger suitability and discriminative abilities for sensitive diagnostic tasks where precision and recall are critical. However, ChatGPT requires refined prompts or supplementary context to improve reliability.

Our study is better at leveraging LLMs for analyzing conjunctivitis with etiology-specific insights, intervention, and hallucination rate quantification. The scope of a previous study is limited as it only dealt with information extraction from electronic health records about antibiotics using NLP and machine learning (ML) algorithms (Sanz et al., 2022). In a previous study, an old dataset collected between 2014 and 2018 was used. Our study employs a recent dataset on conjunctivitis that makes it more clinically relevant and representative of current trends compared to an older dataset. To evaluate ocular surface pain, an earlier study used a multimodal approach and achieved 86% accuracy from a random forest model (Kundu et al., 2022). The accuracy was inherently limited by subjective pain due to variability in imaging. In contrast to this study, the Claude model achieved 100% accuracy for conjunctivitis diagnosis. In addition, our study ensures precision in outcomes through hallucination-free LLMs.

To determine the hallucination rate, we adopted this method from a previous study (Chelli et al., 2024). Although the hallucination rate as a performance metric is widely used in the literature, its relevancy to LLMs is new and is liable for future verifications. This study employed a limited number of prompts on conjunctivitis topics, so the generalization of results might be limited and could be enhanced by proposing more prompts on conjunctivitis topics overlooked by this study. Moreover, we can expand image diversity with demographic and pathological features and incorporate a multi-centered dataset to enhance the clinical applicability of findings. Advanced data augmentation techniques could be applied to expand training set while preserving important clinical features. The current study used a set of focused prompts, which were designed to evaluate core clinical competencies. In addition, we used a rigorous and curated dataset on conjunctivitis. The proposed approach using limited prompts ensures depth over breadth, which validates real-world diagnostic reliability. Future work can expand prompts, as our methodology provides actionable and clinically relevant insights, which surpass generic evaluation. In future research, we can improve image diversity using multi-institutional collaboration on AI-based approaches to strengthen diagnostic accuracy and preserve clinical relevance.

The medical prompts used for conjunctivitis have a limited scope, which does not cover the full spectrum of conjunctivitis topics and scenarios. The proposed prompts to assess the hallucination rate of referencing by AI Chatbots may require future refinements and the introduction of AI models that are specific to the medical field. To determine the hallucination rate, we used only two databases, namely, PubMed and Google, which may have their limitations in indexing or recognizing literature, and researchers might use other search engines or databases such as Scopus and Web of Science in the future in this regard. Standardized algorithms can be used to minimize indexing biases and ensure comprehensive coverage.

6 Conclusion

In conclusion, this article highlights the critical situation of conjunctivitis. It discusses the role of the ChatGPT application in addressing the disease, emphasizing the importance of distinguishing between different types of conjunctivitis for accurate identification. With the ongoing advancements in AI technology, Chatbots such as ChatGPT, Claude, and DeepSeek models have the potential to mitigate the situation by delivering precise information to their users.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.kaggle.com/datasets/alisofiya/conjunctivitis.

Author contributions

MH: Conceptualization, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing. KA: Conceptualization, Funding acquisition, Investigation, Writing – review & editing, Project administration, Resources. MA: Formal analysis, Software, Supervision, Writing – review & editing, Resources. IG: Data curation, Investigation, Project administration, Visualization, Writing – review & editing. MM: Conceptualization, Methodology, Visualization, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This Research is funded by Ongoing Research Funding Program (ORF-2025-947), King Saud University, Riyadh, Saudi Arabia.

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 author(s) declare that Gen AI was used in the creation of this manuscript. Some contents were generated by Generative AI models, which have completely edited.

Publisher's note

References

Abedifar, Z., Fallah, F., Asadiamoli, F., Bourrie, B., and Doustdar, F. (2023). *Chlamydia trachomatis* Serovar distribution in patients with follicular conjunctivitis in Iran. *Turk. J. Ophthalmol.* 53:218. doi: 10.4274/tjo.galenos.2022.12080

Ahmed, A., Irfan, H., and Islam, M. A. (2024). Unraveling the conjunctivitis crisis: understanding the spiking incidence in Karachi and Lahore-Pakistan. *Ann. Med. Surg.* 86, 920–922. doi: 10.1097/MS9.0000000000001623

Alexander, A. C., Raghupathy, S. S., and Surapaneni, K. M. (2024). An assessment of the capability of ChatGPT in solving clinical cases of ophthalmology using multiple choice and short answer questions. *Adv. Ophthalmol. Pract. Res.* 4, 95–97. doi: 10.1016/j.aopr.2024.01.005

Aljamaan, F., Temsah, M. H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., et al. (2024). Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med. Inform.* 12:e54345. doi: 10.2196/54345

Alves, M., Asbell, P., Dogru, M., Giannaccare, G., Grau, A., Gregory, D., et al. (2023). TFOS lifestyle report: impact of environmental conditions on the ocular surface. *Ocul. Surf.* 29, 1–52. doi: 10.1016/j.jtos.2023.04.007

Anthropic (2024). *Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet*. Available online at: https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf (Accessed April 28, 2025).

Balas, M., and Ing, E. B. (2023). Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro differential diagnosis generator. *JFO Open Ophthalmol.* 1:100005. doi: 10.1016/j.jfop.2023.100005

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., and Tortora, G. (2024a). Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst. Appl.* 235:121186. doi: 10.1016/j.eswa.2023.121186

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., and Tortora, G. (2024b). Claude 2.0 large language model: tackling a real-world classification problem with a new iterative prompt engineering approach. *Intell. Syst. App.* 21:200336. doi: 10.1016/j.iswa.2024.200336

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., et al. (2024). Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. *J. Med. Internet Res.* 26:e53164. doi: 10.2196/53164

Chen, T., Lin, L., Chen, R., Hui, X., and Wu, H. (2020). Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1371–1384. doi: 10.1109/TPAMI.2020.3025814

Choudhary, M. M., Chundury, R., McNutt, S. A., and Perry, J. D. (2016). Eyelid contour following conjunctival müllerectomy with or without tarsectomy blepharoptosis repair. *Ophthalmic Plast. Reconstr. Surg.* 32, 361–365. doi: 10.1097/IOP.0000000000000545

Delsoz, M., Madadi, Y., Raja, H., Munir, W. M., Tamm, B., Mehravaran, S., et al. (2024). Performance of ChatGPT in diagnosis of corneal eye diseases. *Cornea* 43, 664–670. doi: 10.1097/ICO.0000000000003492

Gencer, G., and Gencer, K. (2024). A comparative analysis of ChatGPT and medical faculty graduates in medical specialization exams: uncovering the potential of artificial intelligence in medical education. *Cureus* 16:e66517. doi: 10.7759/cureus.66517

Gencer, G., and Gencer, K. (2025). Large language models in healthcare: a bibliometric analysis and examination of research trends. *J. Multidiscip. Healthc.* 18, 223–238. doi: 10.2147/JMDH.S502351

Gunay, M., Goceri, E., and Danisman, T. (2015). "Automated detection of adenoviral conjunctivitis disease from facial images using machine learning," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (Miami, FL: IEEE), 1204–1209. doi: 10.1109/ICMLA.2015.232

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., and Li, K. (2024). Large language models for mental health applications: systematic review. *JMIR Ment. Health* 11:e57400. doi: 10.2196/57400

Hashmi, M. F., Gurnani, B., Benson, S., and Price, K. L. (2024). *Conjunctivitis* (*nursing*).

Hu, X., Ran, A. R., Nguyen, T. X., Szeto, S., Yam, J. C., Chan, C. K. M., et al. (2023). What can Gpt-4 do for diagnosing rare eye diseases? A pilot study. *Ophthalmol. Ther.* 12, 3395–3402. doi: 10.1007/s40123-023-00789-8

Jasim, H. Z., and Al-Jubory, M. (2024). Understanding conjunctivitis: symptoms, causes, and treatment. *Sch. J. App. Med. Sci.* 3, 255–258. doi: 10.36347/sjams.2024.v12i03.007

Jasper, C., Le, T.-T., and Bartram, J. (2012). Water and sanitation in schools: a systematic review of the health and educational outcomes. *Int. J. Environ. Res. Public Health* 9, 2772–2787. doi: 10.3390/ijerph9082772

Jin, H., Guo, J., Lin, Q., Wu, S., Hu, W., and Li, X. (2024). study of Claude 3.5-Sonnet and human physicians in generating discharge summaries for patients with renal insufficiency: assessment of efficiency, accuracy, and quality. *Front. Digit. Health* 6:1456911. doi: 10.3389/fdgth.2024.1456911

Karpecki, P., Paterno, M. R., and Comstock, T. L. (2010). Limitations of current antibiotics for the treatment of bacterial conjunctivitis. *Optom. Vis. Sci.* 87, 908–919. doi: 10.1097/OPX.0b013e3181f6fbb3

Krause, D. (2025). DeepSeek and FinTech: the democratization of AI and its global implications. *SSRN* 5116322. doi: 10.2139/ssrn.5116322

Kundu, G., Shetty, R., D'Souza, S., Khamar, P., Nuijts, R. M. M. A., Sethu, S., et al. (2022). A novel combination of corneal confocal microscopy, clinical features and

artificial intelligence for evaluation of ocular surface pain. *PLoS ONE* 17:e0277086. doi: 10.1371/journal.pone.0277086

Ling, Q., Xu, Z. S., Zeng, Y. M., Hong, Q., Qian, X. Z., Hu, J. Y., et al. (2025). Assessing the possibility of using large language models in ocular surface diseases. *Int. J. Ophthalmol.* 18:1. doi: 10.18240/ijo.2025.01.01

Majdik, Z. P., Graham, S. S., Shiva Edward, J. C., Rodriguez, S. N., Karnes, M. S., Jensen, J. T., et al. (2024). Sample size considerations for fine-tuning large language models for named entity recognition tasks: methodological study. *JMIR AI* 3:e52095. doi: 10.2196/52095

Manovich, L. (2021). Computer vision, human senses, and language of art. *AI Soc.* 36, 1145–1152. doi: 10.1007/s00146-020-01094-9

Odaibo, S. G., MomPremier, M., Hwang, R. Y., Yousuf, S. J., Williams, S. L., and Grant, J. (2019). Mobile artificial intelligence technology for detecting macula edema and subretinal fluid on OCT scans: initial results from the DATUM alpha Study. *arXiv* [Preprint]. arXiv:1902.02905. doi: 10.48550/arXiv.1902.02905

Raja, H., Huang, X., Delsoz, M., Madadi, Y., Poursoroush, A., Munawar, A., et al. (2025). Diagnosing glaucoma based on the ocular hypertension treatment study dataset using chat generative pre-trained transformer as a large language model. *Ophthalmol. Sci.* 5:100599. doi: 10.1016/j.xops.2024.100599

Rajarajeshwari, G., and Selvi, G. C. (2024). Application of artificial intelligence for classification, segmentation, early detection, early diagnosis, and grading of diabetic retinopathy from fundus retinal images: a comprehensive review. *IEEE Access* 12, 172499–172536. doi: 10.1109/ACCESS.2024.3494840

Rietveld, R. P., ter Riet, G., Bindels, P. J., Sloos, J. H., and van Weert, H. C. (2004). Predicting bacterial cause in infectious conjunctivitis: cohort study on informativeness of combinations of signs and symptoms. *BMJ* 329, 206–210. doi: 10.1136/bmj.38128.631319.AE

Sanz, N. V., García-Layana, A., Colas, T., Moriche, M., Moreno, M. M., and Ciprandi, G. (2022). Clinical characterization of inpatients with acute conjunctivitis: a retrospective analysis by natural language processing and machine learning. *Appl. Sci.* 12:12352. doi: 10.3390/app122312352

Singh, R. B., Liu, L., Anchouche, S., Yung, A., Mittal, S. K., Blanco, T., et al. (2021). Ocular redness–I: etiology, pathogenesis, and assessment of conjunctival hyperemia. *Ocul. Surface* 21, 134–144. doi: 10.1016/j.jtos.2021.05.003

Sofiya, A. (2025). *Conjubtivitis Dataset*. Available online at: https://www.kaggle.com/datasets/alisofiya/conjunctivitis/data (Accessed February 12, 2025).

Suzuki, K. (2025). Claude 3.5 Sonnet indicated improved TNM classification on radiology report of pancreatic cancer. *Jpn. J. Radiol.* 43, 56–57. doi: 10.1007/s11604-024-01681-6

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nat. Med.* 29, 1930–1940. doi: 10.1038/s41591-023-02448-8

Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., and Shen, D. (2024). Interactive computer-aided diagnosis on medical image using large language models. *Commun. Eng.* 3:133. doi: 10.1038/s44172-024-00271-8

Yoo, T. K., Choi, J. Y., Seo, J. G., Ramasubramanian, B., Selvaperumal, S., and Kim, D. W. (2019). The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *Med. Biol. Eng. Comput.* 57, 677–687. doi: 10.1007/s11517-018-1915-z

Zaghir, J., Naguib, M., Bjelogrlic, M., Névéol, A., Tannier, X., and Lovis, C. (2024). Prompt engineering paradigms for medical applications: scoping review. *J. Med. Internet Res.* 26:e60501. doi: 10.2196/60501