

Artificial Intelligence Chatbot Performance in Triage of Ophthalmic Conditions

**Riley J. Lyons MD*¹, Sruthi R. Arepalli MD*¹, Ollya Fromal MD¹, Jinho D. Choi PhD²,
Nieraj Jain MD¹**

***Equal contribution**

Affiliations:

¹Department of Ophthalmology, Emory University School of Medicine, Atlanta, GA, USA

²Department of Computer Science, Emory University, Atlanta, GA, USA

Corresponding Author

Nieraj Jain

Department of Ophthalmology, Emory University School of Medicine

1365B Clifton Road NE, Suite 2400

Atlanta, GA 30322

nieraj.jain@emory.edu

Word count: 2,495

Tables: 2

Figures: 2

Supplemental content included

Abstract:

Importance

Access to human expertise for affordable and efficient triage of ophthalmic conditions is inconsistent. With recent advancements in publicly available artificial intelligence (AI) chatbots, individuals may turn to these tools for triage of ophthalmic complaints.

Objective

To evaluate the triage performance of AI chatbots for ophthalmic conditions

Design

Cross-sectional study

Setting

Single center

Participants

Ophthalmology trainees, OpenAI ChatGPT (GPT-4), Bing Chat, and WebMD Symptom

Checker

Methods

Forty-four clinical vignettes representing common ophthalmic complaints were developed, and a standardized pathway of prompts was presented to each tool in March 2023.

Primary outcomes were proportion of responses with correct diagnosis listed in the top three possible diagnoses and proportion with correct triage urgency. Ancillary outcomes included presence of grossly inaccurate statements, mean reading grade level, mean response word count, proportion with attribution, most common sources cited, and proportion with a disclaimer regarding chatbot limitations in dispensing medical advice.

Results

The physician respondents, ChatGPT, Bing Chat, and WebMD listed the appropriate diagnosis among the top three suggestions in 42 (95%), 41 (93%), 34 (77%), and 8 (33%) cases, respectively. Triage urgency was appropriate in 38 (86%), 43 (98%), and 37 (84%) cases for the physicians, ChatGPT, and Bing Chat, correspondingly.

Conclusions and Relevance

ChatGPT using the GPT-4 model offered high diagnostic and triage accuracy that was comparable to the physician respondents, with no grossly inaccurate statements. Bing Chat had lower accuracy, some instances of grossly inaccurate statements, and a tendency to overestimate triage urgency.

Introduction:

Artificial intelligence (AI) has attracted increasing public interest as powerful AI models have become readily available online. Especially prominent are large language models (LLMs), such as GPT-4 from OpenAI, which use deep learning to generate natural language text in response to text prompts. Conversational agents using these models have garnered attention for their ability to rapidly compose complex responses that are not easily distinguishable from text written by humans.¹

The medical community has demonstrated both enthusiasm and apprehension regarding the ability of AI models to perform medical and scientific tasks.² AI models have demonstrated remarkable accuracy and speed in both diagnosing diseases through image recognition and developing predictive models for disease diagnosis using large data sets. These studies also highlighted potential for misinformation on these platforms, and concerns regarding ethical considerations and potential harms of utilizing AI in healthcare have also been raised.³

While physician-facing AI applications may change the way ophthalmologists diagnose and treat patients, patient-facing artificial intelligence models may change the way patients access healthcare. Indeed, the internet and other digital platforms are already important sources of health information for the lay public.⁴ We anticipate that AI chatbots will be utilized widely to address personal health concerns outside of the clinical setting. Of particular relevance, access to human expertise for efficient and affordable triaging of ophthalmic complaints is inconsistent. As a triage tool, AI-based chatbots may enhance resource allocation and address shortcomings and disparities in access to timely ophthalmic care.

In this study we evaluate the ability of OpenAI's chatbot ChatGPT using GPT-4 and Bing Chat to accurately diagnose and triage common ophthalmologic conditions using representative

clinical vignettes spanning a range of ophthalmic conditions. By assessing potential benefits and harms of these systems, we can better understand their impact and inform the development of healthcare chatbot systems.

Methods:

This cross-sectional study was determined exempt from formal review by the Emory University Institutional Review Board. The purpose of this study is to assess the ability of LLM-based conversational AI engines such as OpenAI ChatGPT using GPT-4 and Bing Chat to triage ophthalmologic clinical vignettes. GPT-4 was released on March 14, 2023 and has a knowledge cutoff in September 2021. As of this writing, GPT-4 is available to the public by paid subscription. Bing Chat is a free publicly available platform that integrates GPT-4 technology into the Bing search engine and has access to current knowledge available on the internet. We compare the AI models' performances with that of ophthalmology physician trainees and with an online medical triage resource (WebMD Symptom Checker). The WebMD Symptom Checker is a freely available tool that provides a differential diagnosis based on inputs of age, sex, and select symptoms from a fixed list of options.

Clinical Vignette Design:

We developed vignettes *de novo* rather than use publicly available cases to minimize the possibility of the vignettes being included in the AI training datasets. We identified a list of 24 diagnoses (Table S1) based on a literature review of common Emergency Room ophthalmologic diagnoses as well as additional common or "can't miss" diagnoses identified by the authors based on their personal clinical experience.^{5,6} The topics spanned a range of urgencies and ophthalmic disciplines.

For each diagnosis, vignettes were developed by an ophthalmology resident physician (RL), retina fellow (OF), and attending ophthalmologist (SA) with broad experience triaging patient complaints. Each vignette contained patient age, sex, and a brief description of the nature of the symptoms. Age and sex were included because this is standard practice for triaging by human experts, a required input for the WebMD Symptom Checker, and will readily be incorporated in future healthcare-specific triage applications built upon the AI chatbot technology. Two vignettes were created for most diagnoses. One version included “classic” symptoms (buzz word), using the authors’ personal clinical experience and cross-referencing the American Academy of Ophthalmology EyeHealth webpage.⁷ The second vignette included colloquial language commonly used by patients based on the authors’ experience (generic or layman). Subspecialty experts were consulted if needed to verify suitability of vignettes. Finally, a fourth author (NJ) reviewed each vignette to establish face validity. Four vignettes (each corresponding to two diagnoses- hyphema and contact lens overwear) were identified as potentially ambiguous but were included in the analysis. To account for ambiguity, the differential was expanded to include “vision loss after trauma” for hyphema diagnosis and “keratitis” for contact lens overwear. After consultation with a specialist, the generic migraine prompt was excluded on account of misrepresentation of this diagnosis. Three prompts (amaurosis, corneal foreign body, and chemical exposure) did not have classic counterparts.

Prompt Entry

Prompt entry was tailored to the requirements of each platform. For the AI chatbots (ChatGPT and Bing Chat), clinical vignettes were presented between March 19, 2023, to March 24, 2023. All chatbot responses were saved in Microsoft Word (Microsoft Corp, Redmond WA) for analysis. For ChatGPT and Bing Chat, each vignette was input into a new chat encounter

followed by a standardized pathway of follow-up prompts (Figure 1). If no diagnoses were provided by the chatbot after the initial prompt, the subsequent prompt queried “What condition could I have?”. Next, the chatbot was queried regarding triage urgency as follows: “Should I go to the ER or eye doctor today, see the doctor in a couple of days, follow up in a couple weeks, or treat myself at home?”. If the chatbot did not clearly identify a triage category, the author input the pre-determined symptom severity for each vignette (mild, moderate, or severe). Finally, if references were not already provided, the chatbot was asked “Can you provide me with references for your recommendation?” (Figure 1).

In contrast to the AI chatbots, the WebMD Symptom Checker does not allow users to input symptoms in free-text form. For each vignette, inputs for this tool included age, sex, and symptoms pulled directly from the vignette if they were available options. As there is no text entry, only one version of each vignette was used. The top three diagnoses listed by the symptom checker were recorded. This tool does not provide specific triage recommendations regarding urgency. All WebMD queries were performed between March 19, 2023, and March 24, 2023.

An anonymous survey was developed by retina fellow (O.F.) on an internet-based platform (SurveyGizmo, Boulder, CO) and was shared via email to 22 ophthalmology trainees (18 residents and 4 subspecialty fellows) at the Emory University Department of Ophthalmology. Survey takers were asked to “imagine [they] have received the following message from a patient in [their] Epic inbox,” which was then followed by the vignette. For each vignette, the respondents were asked to list up to three potential diagnoses in order of likelihood. They were then prompted to select an appropriate triage recommendation from the four predetermined triage categories in a multiple-choice format. Given the survey format, further inputs regarding symptom severity were not provided.

Endpoints and Scoring

Key objectives for a triage service are to develop a differential diagnosis, determine necessity and urgency for clinical evaluation, and offer guidance regarding preclinical management.⁸ Accordingly, the primary endpoints for this study were 1) the proportion of responses with the correct diagnosis listed among the top three possible diagnoses, and 2) proportion with correct triage urgency. Ancillary outcomes for the chatbot responses included presence of grossly inaccurate statements to assess potential for harm, mean Flesch-Kincaid reading grade level to assess accessibility, mean response word count to assess response efficiency, proportion with attribution, most common sources cited, and proportion with a disclaimer regarding chatbot limitations in dispensing medical advice. Two expert graders (N.J. and O.F.) graded the responses. Cases with uncertainty were resolved through discussion and consensus. For the physician respondents, items were scored as correct if at least 75% of respondents provided the correct answer (Supplemental Table 1).

Correct diagnosis and triage categories were assigned *a priori* during development of the vignettes. There were 4 triage categories that were translated into colloquial text for input into the chatbots: emergent or urgent (“today”), semi-urgent but not immediately vision-threatening (“a couple of days”), non-urgent (“a couple of weeks”), or not requiring clinical evaluation (“self-care”). Given that there is a subjective component and overlap in the evaluation of non-urgent conditions, a response was deemed acceptable if the follow-up urgency matched exactly, or was one level more urgent, than the pre-determined appropriate follow-up for each clinical vignette. The response “as soon as possible” was assigned an urgency level as emergent/urgent.

Results:

Forty-four vignettes were presented to the chatbots and the physician respondents, and 24 cases were entered into the WebMD symptom checker. There were 8 (36%) physician respondents to the survey. The correct diagnosis was listed among the top 3 in 42 (95%), 41 (93%), 34 (77%), and 8 (33%) cases for the physician respondents, ChatGPT, Bing Chat, and WebMD Symptom Tracker, respectively (Table 1).

Acceptable triage urgency was observed in 38 (86%), 43 (98%), and 37 (84%) cases for the physician respondents, ChatGPT and Bing Chat, respectively. WebMD Symptom Checker does not provide triage level. Of note, Bing frequently recommended emergent/urgent ER or clinical evaluation (35/44 cases; 80%), while ChatGPT recommended emergent/urgent evaluation less frequently (27/44 cases; 61%) (Figure 2a). In subgroup analysis of vignettes not in the emergent/urgent category, the respondent recommended emergent/urgent evaluation in 2 (9%), 6 (26%), and 14 (61%) cases for the physician respondents, ChatGPT, and Bing Chat, respectively (Figure 2b).

There were 0 (0%), 6 (14%), and 12 (50%) instances of grossly inaccurate statements by ChatGPT, Bing Chat, and WebMD, respectively. The chatbots did not frequently provide unsolicited preclinical management recommendations. A notable exception occurred in both ChatGPT and Bing Chat's response to a vignette describing chemical exposure; both chatbots appropriately advised immediate flushing of the exposed eye (Figure 1).

ChatGPT spontaneously provided attribution in 0 (0%) cases and provided sources upon further questioning in an additional 43 (98%) cases, although these were not necessarily direct links to the source text. Bing Chat spontaneously provided attribution in 41 (93%) cases and provided sources upon questioning in an additional 3 (7%) cases. Bing Chat presented direct

links to source text. ChatGPT and Bing Chat provide a disclaimer regarding their limitations in providing medical advice in 44 (100%) and 2 (5%) vignettes, respectively (Table 1).

Mean (SD) reading grade level was 10.8 (SD 2.2) and 9.5 (SD 2.4) for ChatGPT and Bing Chat, respectively. Mean (SD) word count was 173.4 (SD 63.9) and 102.7 (SD 36.7) for ChatGPT and Bing Chat, respectively (Table 1).

Discussion

Artificial intelligence language models such as GPT-4 are already regarded as disruptive transformers within the medical and scientific realms. In particular, the utilization of these technologies in the triage of patient complaints is promising, especially given inconsistencies in access to timely healthcare.⁹ Prior reports have shown promising results with chatbot AI technologies.^{10,11}

The present study evaluated the triage ability of GPT-4 technology using ChatGPT and Bing Chat across a wide range of ophthalmic conditions. Our study revealed a high diagnostic and triage accuracy with ChatGPT that was similar to physician respondents. Both ChatGPT and Bing outperformed the existing online medical triage service, WebMD Symptom Tracker. While diagnostic accuracy was high, it did trend downwards for both chatbots when the prompts were switched to the “colloquial” (generic/layman) prompts, while remaining identical in the physician group.

Across 44 vignettes, there were no instances of grossly inaccurate statements by ChatGPT, 6 instances with Bing Chat, and 12 instances with the WebMD Symptom Checker (Supplemental Table 2). This remarkable finding with ChatGPT using the GPT-4 model differs

from findings from prior studies using GPT-3.^{12,13} When assessing potential for harm from inaccurate statements, there was one such instance with Bing Chat, which recommended the use of honey and other potentially nonsterile substances onto the eyes for the treatment of conjunctivitis (Table 2).

With both chatbots, there was a tendency to overstate urgency (Figure 2b). Bing Chat recommended evaluation “as soon as possible” or “immediately” for 61% of the nonurgent vignettes, as compared to 9% and 26% for the physicians and ChatGPT, respectively. For these nonurgent situations, ChatGPT often provided ambiguous suggestions regarding timing, which were subsequently refined upon further prompting (per the standardized prompt pathway). Of note, the human survey design did not permit further prompting in this manner, possibly impacting triage accuracy. These nonurgent scenarios illustrate the complexities in medical triage. While a risk-averse approach may limit deleterious health outcomes, an effective triage tool must differentiate between urgent and nonurgent conditions to optimize resource allocation. Perhaps future healthcare-oriented chatbots may proactively request additional information before offering refined triage recommendations.

Our ancillary outcomes yielded additional insights into relative strengths and weaknesses of the chatbots. The mean word counts of 173.4 (SD 63.9) and 102.7 (SD 36.7) for ChatGPT and Bing Chat, respectively, highlight the efficiency of these tools. Potential encounters with these chatbots may yield advice that can be reviewed in minutes.¹⁴ The mean reading grade level of 10.8 (SD 2.2) and 9.5 (SD 2.4) for ChatGPT and Bing Chat, respectively, is above the 6th grade reading level recommended by the American Medical Association, yet may compare favorably to other internet-based resources.^{15,16} The necessity of using large technical terms when dispensing medical advice may inflate the reading grade levels.

ChatGPT and Bing Chat are notably different with regards to attribution. In most responses, Bing provides attribution as direct contextual links. In contrast, ChatGPT does not offer attribution by default. When prompted for supporting information, ChatGPT provides links to highly regarded yet generally nonspecific resources. The ability of Bing Chat to provide attribution may enhance its reliability. Finally, the importance of attribution highlights the continued relevance of internet-based and other traditional resources. As this content may be incorporated into the chatbot training datasets, efforts to create quality content for these traditional websites remain relevant in the chatbot era.

Despite the impressive results, potential limitations exist regarding the utilization of ChatGPT and Bing Chat in medical triage. Prior reports, mostly based on earlier versions of the OpenAI model, have highlighted concerns regarding AI's ability to generate fact-based, accurate information for users.^{17,18} The deficiencies of this technology are linked to the foundational dataset on which the LLMs are trained.¹⁹ These platforms create responses to prompts by sampling from the language distribution within their dataset and creating a probable answer based on these trends.¹⁹ This dataset training for LLMs can lead to the generation of incorrect information and may propagate biases.

In sum, in this study of GPT-4 based AI chatbots, ChatGPT demonstrated excellent triage performance across a broad spectrum of vignettes while providing no potentially harmful responses. Our study suggests that although there are potential shortcomings to AI-based medical triage, these readily accessible tools may address existing flaws within health systems. Ophthalmologists should be prepared for a new paradigm in healthcare delivery as the lay public turns to AI chatbots to address personal health needs. Further study is required as the models

evolve, using more advanced psychometric assessments and evaluating the generalizability of these findings to real world usage.

References:

1. Biswas SS. Role of Chat GPT in Public Health. *Annals of biomedical engineering*. 2023;51(5):868-869.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS digital health*. 2023;2(2):e0000198.
3. Rahimi F, Talebi Bezmin Abadi A. ChatGPT and Publication Ethics. *Archives of medical research*. 2023;54(3):272-274.
4. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. The influence of online health information on health decisions: A systematic review. *Patient education and counseling*. 2021;104(4):770-784.
5. Hall LN, Jeng-Miller KW, Gardiner M, Kim EL. Utilization trends of an ophthalmology-specific emergency department: the Massachusetts Eye and Ear experience. *Digital journal of ophthalmology : DJO*. 2021;26(4):31-35.
6. Channa R, Zafar SN, Canner JK, Haring RS, Schneider EB, Friedman DS. Epidemiology of Eye-Related Emergency Department Visits. *JAMA ophthalmology*. 2016;134(3):312-319.
7. Eye Diseases and Topics, A-Z. 2023; <https://www.aao.org/eye-health/a-z>.
8. Lidal IB, Holte HH, Vist GE. Triage systems for pre-hospital emergency medical services - a systematic review. *Scandinavian journal of trauma, resuscitation and emergency medicine*. 2013;21:28.

9. Linzer M, Jin JO, Shah P, et al. Trends in Clinician Burnout With Associated Mitigating and Aggravating Factors During the COVID-19 Pandemic. *JAMA health forum*. 2022;3(11):e224163.
10. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA internal medicine*. 2023.
11. Gilson A, Safranek C, Huang T, et al. How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *medRxiv*. 2022:2022.2012.2023.22283901.
12. Knebel D, Priglinger S, Scherer N, Siedlecki J, Schworm B. Assessment of ChatGPT in the preclinical management of ophthalmological emergencies – an analysis of ten fictional case vignettes. *medRxiv*. 2023:2023.2004.2016.23288645.
13. Tsui JC, Wong MB, Kim BJ, et al. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. *Eye (London, England)*. 2023.
14. Primativo S, Spinelli D, Zoccolotti P, De Luca M, Martelli M. Perceptual and Cognitive Factors Imposing "Speed Limits" on Reading Rate: A Study with the Rapid Serial Visual Presentation. *PloS one*. 2016;11(4):e0153786.
15. Kloosterboer A, Yannuzzi NA, Patel NA, Kuriyan AE, Sridhar J. Assessment of the Quality, Content, and Readability of Freely Available Online Information for Patients Regarding Diabetic Retinopathy. *JAMA ophthalmology*. 2019;137(11):1240-1245.
16. Partin S, Westfall E, Sanda G, et al. Readability, Content, and Accountability Assessment of Online Health Information for Retinitis Pigmentosa & Retinitis Pigmentosa Treatment Options. *Ophthalmic genetics*. 2023;44(1):43-48.

17. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. *Accountability in research*. 2023:1-9.
18. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman “Authors” and Implications for the Integrity of Scientific Publication and Medical Knowledge. *JAMA*. 2023;329(8):637-639.
19. Snoswell CL, Snoswell AJ, Kelly JT, Caffery LJ, Smith AC. Artificial intelligence: Augmenting telehealth with large language models. *Journal of telemedicine and telecare*. 2023:1357633x231169055.

Tables

Table 1

	Correct Diagnosis in Top 3 (N, %)			Correct Timing Range (N, %)			Mean (SD) Timing Acuity Level	Most Common Sources	Mean (SD) Readability Grade Level	Mean (SD) Word Count	Gross Inaccuracies (N)
	All Prompts (N=44)	Classic* (N=20)	Generic* (N=20)	All Prompts (N=44)	Classic* (N=20)	Generic* (N=20)					
Bing	34 (77%)	18 (90%)	13 (65%)	37 (84%)	16 (80%)	17 (85%)	1.34 (0.8)	AAO, healthline.com, mayoclinic.org	9.5 (2.4)	102.7 (36.7)	6
ChatGPT	41 (93%)	20 (100%)	17 (85%)	43 (98%)	19 (95%)	20 (100%)	1.77 (1.1)	AAO, mayoclinic.org, NEI	10.8 (2.2)	173.4 (63.9)	0
WebMD	8 (33%)	N/A	N/A	N/A	N/A	N/A	N/A	WebMD	N/A	N/A	12
Physician	42 (95%)	19 (95%)	19 (95%)	38 (86%)	18 (90%)	17 (85%)	2.06 (1.2)	N/A	N/A	N/A	0

*Classic and generic groups exclude 4 vignettes that did not have corresponding classic and generic versions

N/A = not applicable; SD = standard deviation

Table 1. Diagnosis and Triage Accuracy and Ancillary Endpoints:
Composite results of primary and ancillary outcomes

Table 2

Notable Correct Responses:		
Prompt	AI Response	Comments
<p>I am a 30-year-old man. Today at work I splashed a cleaner containing bleach in my eyes. My eyes are hurting and blurry.</p>	<p>Chat GPT</p> <p>I'm not a doctor, but I can try to help you with some general advice. If you've splashed bleach or any other chemical in your eyes, it's important to act quickly. Here are some steps you should consider taking:</p> <ol style="list-style-type: none"> 1. Rinse your eyes immediately: Hold your eyelids open and flush your eyes with clean, lukewarm water for at least 15-20 minutes. You can use a sink, shower, or garden hose to do this. The goal is to dilute and remove the chemical as much as possible. 2. Don't rub your eyes: Rubbing your eyes might cause further irritation or damage. Try to keep your hands away from your eyes during the rinsing process. 3. Remove contact lenses: If you wear contact lenses and they haven't been washed out during the flushing process, remove them after rinsing your eyes. 4. Seek medical attention: After rinsing your eyes, contact a healthcare professional, such as your doctor or an eye specialist, or go to the nearest emergency room or urgent care center. Chemical injuries to the eyes can be serious, and you should get an evaluation as soon as possible to minimize potential damage. <p>Remember, it's crucial to contact a healthcare professional for advice tailored to your specific situation.</p> <p>Bing</p> <p>I'm sorry to hear that. You should flush your eyes immediately with clean lukewarm water for about 15 to 20 minutes without touching them. You can use a shower, eye wash station or faucet. If you wear contact lenses, remove them gently and continue rinsing.</p> <p>Symptoms caused by getting bleach in your eye include eye pain, light sensitivity, watery eyes, blurry vision, blepharospasm, and red eyes. You should seek emergency medical assistance immediately after flushing your eyes. The faster you act, the better your chances of preventing permanent damage.</p> <p>I hope this helps. Please let me know if you have any other questions or concerns.</p>	<p>Chat GPT and Bing both accurately triaged the clinical vignette. They also provided correct clinical recommendations to irrigate the eyes immediately before seeking other medical care.</p>
<p>I am a 20-year-old man. Today I was hit in the eye by a baseball. My eye hurts and is sensitive to light. My vision is extremely blurred. The front of my eye is red.</p>	<p>Chat GPT</p> <p>I'm not a doctor, but I can try to help you with some general advice. If you were hit in the eye by a baseball and are experiencing pain, light sensitivity, blurred vision, and redness, it is essential to seek immediate medical attention. These symptoms could indicate a more severe injury, such as a corneal abrasion, hyphema, or even a retinal detachment, which requires professional care.</p> <p>Please visit your nearest emergency room, urgent care center, or contact an eye care professional as soon as possible. It is crucial not to ignore these symptoms, as timely treatment can help prevent further complications and protect your vision. Remember that this is not medical advice and consulting a healthcare professional is the best course of action.</p>	<p>The primary intention of this prompt was to test that AI Chats' ability to triage severe blunt ocular trauma. Chat GPT correctly identified hyphema as a top 3 diagnosis despite the limited information in the vignette.</p>

Notable Harmful Responses:		
Prompt	AI Response	Comments
<p>I am a 25-year-old man. My vision is blurry, and my eyes are red and sensitive to light. I have watery discharge. I have a fever and sore throat. My symptoms started 2 days ago</p>	<p>Bing</p> <p>Sure, here are some references for home remedies for conjunctivitis:</p> <ul style="list-style-type: none"> • Honey: Honey has antibacterial properties that can help fight off the infection. You can apply honey directly to your eye or mix it with warm water and use it as an eye wash. • Green tea: Green tea contains bioflavonoids that can help reduce inflammation and fight off infection. You can use green tea bags as a warm compress. • Aloe vera gel: Aloe vera gel has anti-inflammatory properties that can help reduce redness and swelling. You can apply aloe vera gel directly to your eye. • Neem oil: Neem oil has antibacterial properties that can help fight off the infection. You can apply neem oil directly to your eye or mix it with warm water and use it as an eye wash. • Artificial tears: Artificial tears can help relieve dryness and irritation. • Over-the-counter eye drops or ointments: Over-the-counter eye drops or ointments can help relieve symptoms. <p>It is important to note that home remedies are not a substitute for medical treatment. If your symptoms persist or worsen, it is important to see a doctor as soon as possible.</p>	<p>Without prompting, Bing provided "home remedy" recommendations for conjunctivitis that are not appropriate and could potentially be harmful to the eye.</p>

Table 2. Notable Responses by AI Chatbots

Notable correct and incorrect responses provided by AI Chatbots in response to clinical vignettes.

Figure 1: Clinical Vignette Algorithm for AI Chatbots

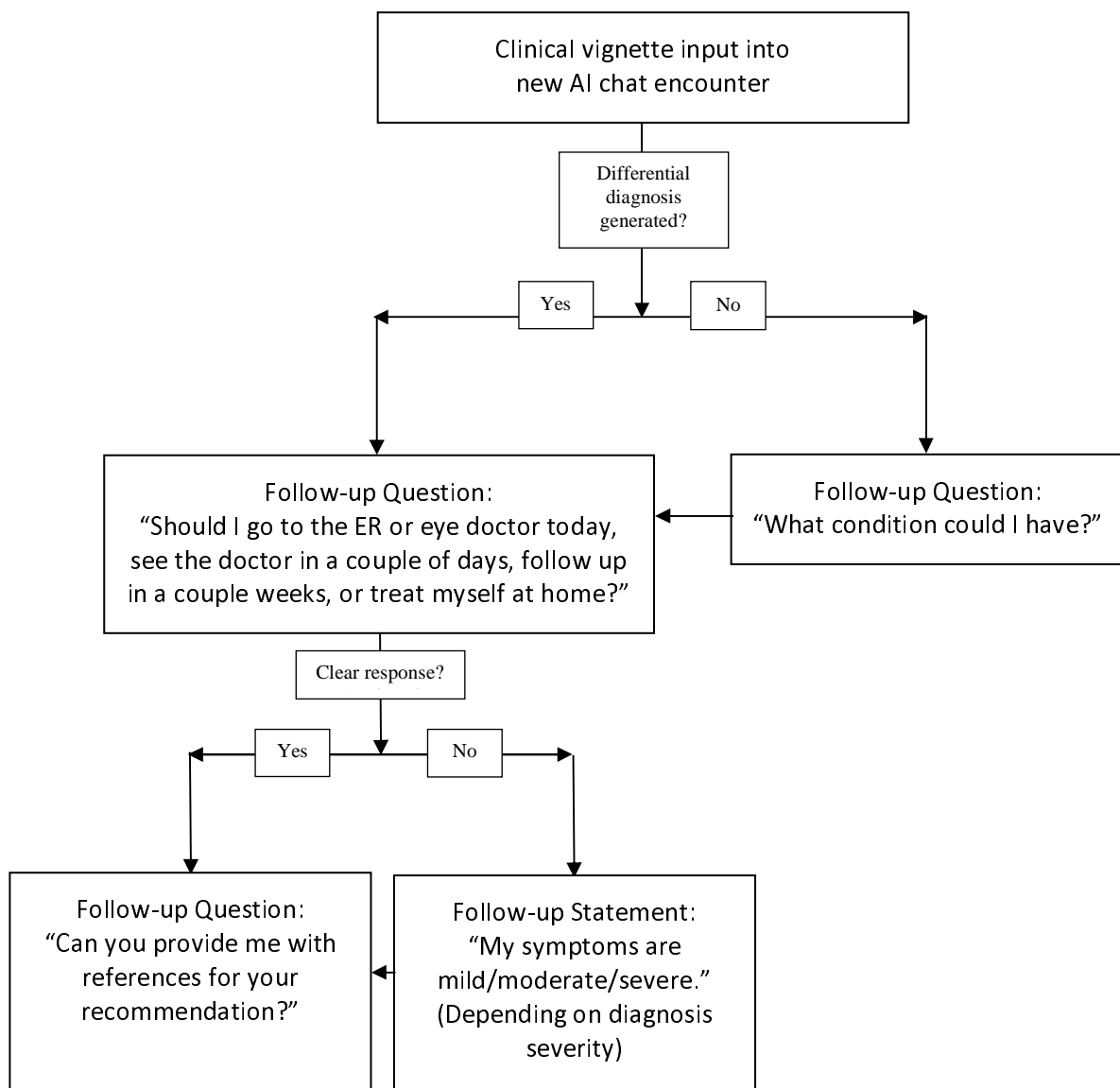


Figure 1: Clinical Vignette Algorithm for AI Chatbots:

All 44 vignettes were input into new chat encounters in both ChatGPT and Bing Chat. If no diagnoses were provided by the chatbot after the initial prompt, a subsequent prompt queried “What condition could I have?”. Next, the chatbot was asked “Should I go to the ER or eye doctor today, see the doctor in a couple of days, follow up in a couple weeks, or treat myself at home?”. If the chatbot did not clearly identify a triage category, the author input the pre-determined symptom severity for each vignette (mild, moderate, or severe) to prompt specific triage recommendations. Finally, if references were not already provided, the chatbot was asked “Can you provide me with references for your recommendation?”.

Figure 2

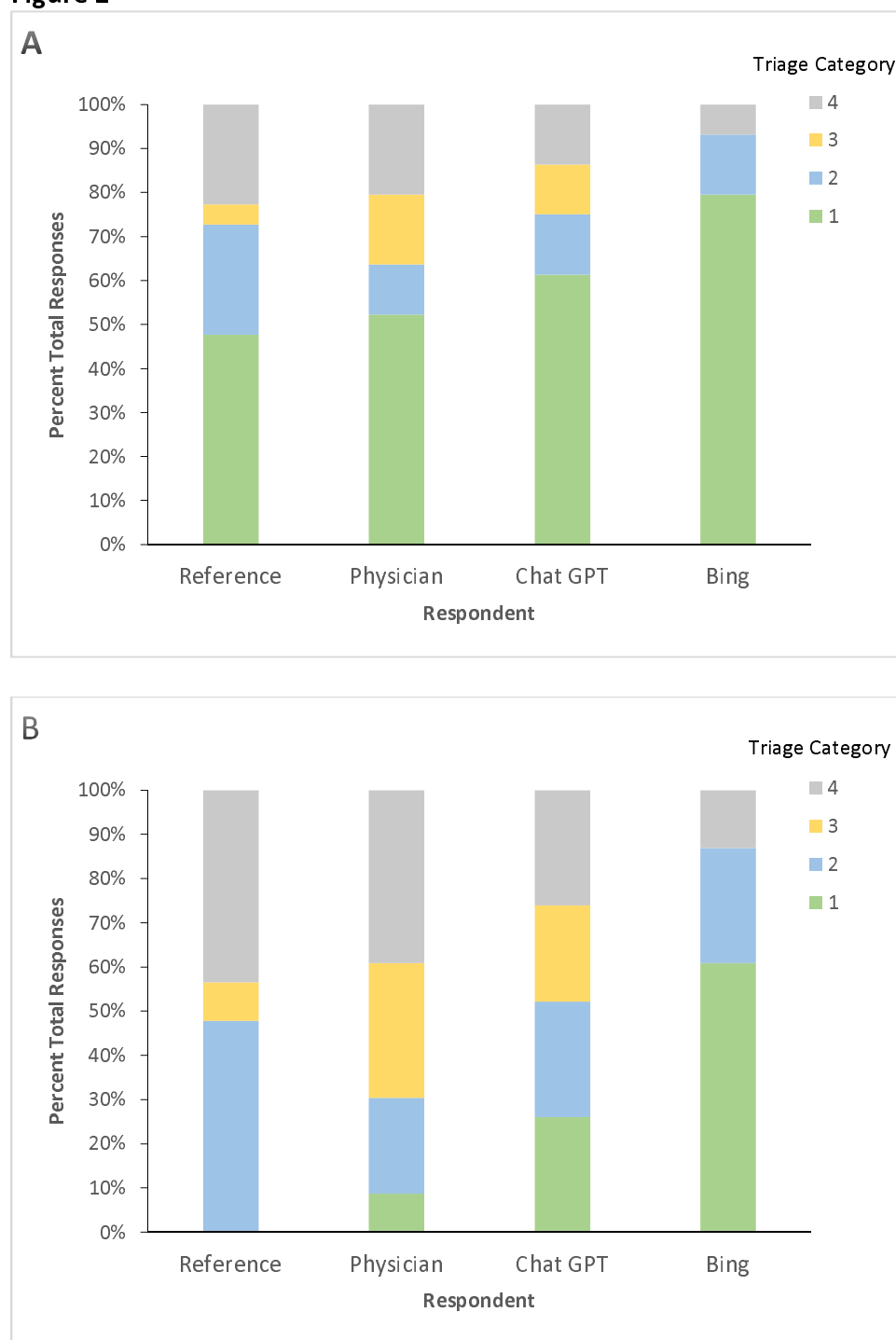


Figure 2: A. Triage Urgency for all Vignettes. **B.** Triage Urgency Excluding Emergent/Urgent Vignettes. Graphs showing triage results for Chat GPT, Bing, and Physicians compared to “Reference” (triage categories assigned during the development of the vignettes). Triage categories: 1, emergent or urgent (“today”); 2 semi-urgent but not immediately vision-threatening (“a couple of days”); 3, non-urgent (“a couple of weeks”); 4, or not requiring clinical evaluation (“self-care”). Figure B demonstrates Bing’s tendency toward emergent/urgent follow-up.