Article

Not peer-reviewed version

A Performance Evaluation of Large Language Model in Keratoconus: A Comparative Study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity

Ali Hakim REYHAN *, Çağrı MUTAF, İrfan UZUN, Funda YÜKSEKYAYLA

Posted Date: 5 September 2024

doi: 10.20944/preprints202409.0257.v1

Keywords: Keratoconus; Chatbots; Large Language Models

 doi:10.20944/preprints202409.0257.v1

Article

A Performance Evaluation of Large Language Model in Keratoconus: A Comparative Study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity

Ali Hakim Reyhan *, Çağrı Mutaf, İrfan Uzun and Funda Yüksekyayla

* Correspondence: alihakimreyhan@gmail.com

Abstract: Background: This study evaluates the ability of six popular chatbots; ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity to provide reliable answers to questions concerning keratoconus. **Methods:** Chatbots responses were assessed using mDISCERN and Global Quality Score (GQS) metrics. Readability was evaluated using nine validated readability assessments. We also addressed the quality and accountability of websites from which the questions originated. **Results:** We analyzed 20 websites, 65% "Private practice or independent user" and 35% "Official patient education materials." The mean JAMA Benchmark score was 1.5±0.68, indicating low accountability. Reliability, measured using mDISCERN, ranged from 42.9±3.16 (ChatGPT-3.5) to 46.95±3.53 (Copilot). The most frequent question was "What is Keratoconus?" with 70% of websites providing relevant information. This received the highest mDISCERN score (49.33±4.96) and a relatively high GQS score (3.50±0.55), with an Automated Readability Level Calculator score of 13.17±2.13. Moderate positive correlations were determined between the website numbers and both mDISCERN (r=0.265,p=0.25) and GQS (r=0.453,p=0.05) scores. The quality of information, assessed using the GQS, ranged from 3.01±0.51 (ChatGPT-3.5) to 3.3±0.65 (Gemini) (p=0.34). The differences between the texts were statistically significant. Gemini emerged as the easiest to read, while ChatGPT-3.5 and Perplexity were the most difficult. Based on mDISCERN scores, Gemini and Copilot exhibited the highest percentage of responses in the "Good" range (51-62 points). For the GQS, the Gemini model exhibited the highest percentage of responses in the "Good" quality range, with 40% of its responses scoring 4-5. **Conclusions:** While all chatbots performed well, Gemini and Copilot showed better reliability and quality. However, their readability often exceeded recommended levels. Continuous improvements are essential to match information with patients' health literacy for effective use in ophthalmology.

Keywords; keratoconus; chatbots; large language models

Introduction

The use of artificial intelligence (AI) in healthcare is expanding rapidly, with applications ranging from diagnostic support to patient education. However, the medical field demands a high level of accuracy and reliability, since misinformation can lead to adverse health outcomes. The advent of large language models (LLMs) has revolutionized the field of natural language processing (NLP), enabling machines to generate human-like and contextually appropriate responses. Models such as ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity have attracted significant attention for their potential applications across various domains, particularly in healthcare. However, the accuracy and reliability of LLMs in specific medical contexts remain underexplored.

Keratoconus is a progressive eye disease characterized by the thinning and bulging of the cornea, leading to visual impairment. Patients and caregivers frequently seek information regarding

 doi:10.20944/preprints202409.0257.v1

the symptoms, diagnosis, and therapeutic options for this condition that affects a significant part of the population [1]. Due to the complexity and specificity of medical information, it is crucially important to evaluate the performance of LLMs in providing accurate and reliable answers to questions related to keratoconus. Obtaining early and accurate information is essential for effective management and treatment. This trend underscores the importance of evaluating the quality of information provided by LLMs, which are increasingly being used to answer health-related queries.

This study was intended to assess the performance of six leading LLMs (ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity) in the context of keratoconus. We planned to determine the extent to which these models can be considered reliable sources of medical information by comparing their responses to frequently asked questions (FAQs) about keratoconus. The findings of this study will yield valuable insights into the capabilities and limitations of LLMs in the medical sphere, guiding future developments in and applications of this technology in the health care setting.

Materials and Methods

Ethics

Since the LLMs used in this study are public applications and no patients were involved, ethics committee approval was not required.

Data Collection and Search Strategy

All Google searches used in data collection were executed using a clean-installed Google Chrome (Menlo Park, CA) browser in Incognito Mode. In order to avoid bias from previous searches and targeted search results based on geography, we disabled all location filters, advertisements, and sponsored results. The search terms used were "Keratoconus FAQ," and the "People also ask" box was used to obtain FAQs generated by Google's machine learning algorithms.

Question Selection and Categorization

The first 20 websites were reviewed. The 20 most frequently asked questions concerning keratoconus were selected by two experienced ophthalmologists (AHR, ÇM). These subsequently transformed similar question patterns into a common question template. Websites used to answer each of the 20 FAQs in this study were first categorized according to the information source: (1) educational institution, including academic medical centers, (2) private practice or independent user, (3) crowd-sourced reference (such as Wikipedia), or (4) official patient education materials published by a national organization (such as the American Academy of Ophthalmology).

JAMA Accountability Analysis

All websites were evaluated for accountability (scores of 0–4) using JAMA benchmarks. According to JAMA guidelines, a website containing patient education materials should (1) include all authors and their relevant credentials, (2) list references, (3) provide disclosures, and (4) provide the date of the most recent update.

Large Language Model (LLM)

The LLM was trained on extensive bodies of text data, including books, scholarly articles, and web pages, covering a wide array of subjects including medicine, sports, and politics. The LLM models employed were ChatGPT-3.5, ChatGPT-4, Gemini, Copilot, Chatsonic, and Perplexity. These were asked 20 FAQs related to 'keratoconus,' and their responses were recorded.

Evaluation of LLM-Chatbot Responses

As shown in Table 3, DISCERN is a scoring system developed by Oxford University, consisting of three parts and 16 questions and used to evaluate the reliability and quality of online health information. The DISCERN scoring system result range is 15–75, and the results are classified as

 doi:10.20944/preprints202409.0257.v1

excellent (63–75 points), good (51–62), reasonable (39–50), poor (27–38), or very poor (15–26). The Global Quality Scale (GQS) was applied to assess the quality of LLM responses. Accordingly, 1 point indicates poor quality, and 5 points indicate excellent quality (Table 3). Additionally, this scale was also used for quality classification, 1-2 points representing low quality, 3 points moderate quality, and 4-5 points high quality.

Table 3. mDISCERN and GQS Content and Readability indexes.

DICERN scoring system	**Total score (15–75 points)**
1.Are the aims clear?	1-5 points
2.Does it achieve its aims?	1-5 points
3.Is it relevant?	1-5 points
4.Is it clear what sources of information were used to compile the publication (other than the author or producer)?	1-5 points
5.Is it clear when the information used or reported in the publication was produced?	1-5 points
6. Is it balanced and unbiased?	1-5 points
7. Does it provide details of additional sources of support and information?	1-5 points
8. Does it refer to areas of uncertainty?	1-5 points
9. Does it describe how each treatment works?	1-5 points
10. Does it describe the benefits of each treatment?	1-5 points
11. Does it describe the risks of each treatment?	1-5 points
12. Does it describe what would happen if no treatment is used?	1-5 points
13. Does it describe how the treatment choices affect overall quality of life?	1-5 points
14. Is it clear that there may be more than 1 possible treatment choice?	1-5 points
15. Does it provide support for shared decision making?	1-5 points
16. Based on the answers to all of these questions, rate the overall quality of the publication	1-5 points
Global Quality Score	**Score**
Poor quality, very unlikely to be of any use to patients	0-1 Points
Poor quality but some information present, of very limited use to patients	0-1 Points
Suboptimal flow, some information covered but important topics missing, somewhat useful	0-1 Points
Good quality and flow, most important topics covered, useful to patients	0-1 Points
Excellent quality and flow, highly useful to patients	0-1 Points
Readability indexes	
Flesch reading ease score (FRE)	206.835 - (1.015 (W/S)) - (84.6 * (S/W)

 doi:10.20944/preprints202409.0257.v1

Flesch–Kincaid grade level (FKGL)	0.39 * (W/S) + 11.8 * (B/W)−15.59
Gunning FOG Index (GFI)	0.4 × [(W/S) + 100 × (C*/W)]
Coleman-Lıau Readabılıty Index (CLI)	(0.0588 × L)−(0.296 × S*)−15.8
Automated Readabılıty Index (ARI)	(4.71 * (C/W)) + (0.5 * (W/S)) - 21.43
Simple measure of Gobbledygook (SMOG)	1.0430 × √C + 3.1291
Linsear Write Readability Formula (LW)	(ASL + (2 * HDW)) / SL
Forcast Readability Formula (FORCAST)	20 - (# of Single Syllable Words x 150 / # of Words x 10)
Average Readıng Level Consensus Calc (ARLC)	Based on (8) above popular readability formulas, your text yielded a final result

B Number of syllables, W Number of words, S Number of sentences, C Complex words (≥3 syllables), C*: Complex words with exceptions including, LW = Linsear Write Readability Formula result, ASL = Average sentence length (i.e., the number of words divided by the number of sentences), HDW = Number of "hard words" (words with more than two syllables), SL = Number of sentences, **L** = Average number of characters per 100 words, **S** = Average number of sentences per 100 words.

The LLM-Chatbots responses were evaluated and scored in a double-blinded manner by two experienced ophthalmologists. The LLM-Chatbot responses represented the average scores given by two experienced ophthalmologists using DISCERN (15-75 points) and GQS (1-5 points) (AHR, ÇM). A consensus score was then determined.

Readability Analysis

Each of the 20 websites that provided answers to the 20 FAQs examined in this study was evaluated for readability using nine validated readability assessments: Flesch Reading Ease (FRE), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG), Coleman-Liau Index (CLI), Automated Readability Index (ARI), Linsear Write Formula (LINSEAR), FORCAST Readability Formula, and the Automated Readability Level Calculator (ARLC).

Statistical Analysis

Statistical analyses were conducted using R software (Version 4.1.1, R Foundation, Vienna, Austria). Descriptive statistics were used to categorize the sources of online information regarding keratoconus. Categorical variables were expressed as numbers and percentages. Differences in the length and readability of responses across the LLM-Chatbots were compared using One-Way ANOVA and Tukey's honest significance post-hoc test since the samples met parametric assumptions. Relationships between the data were evaluated with a two-tailed Pearson's χ^2 test. A p-value ess than 0.05 was considered statistically significant.

Results

Frequently asked Questions after Google Searches for 'Keratoconus'

Table 1 shows the distribution of website categories, and their JAMA Benchmark scores in terms of LLM accuracy in providing keratoconus-related information. Sixty-five percent (13) of the 20 websites were "Private practice or independent user" and 35% (7) were "Official patient education materials."

Table 1. Distribution of website category and a JAMA Benchmark criterion score for website.

Website Category	Number n (%)	JAMA Benchmarks (Mean score±SD)
Private practice or independent user	13 (65%)	1.5±0.68

 doi:10.20944/preprints202409.0257.v1

Official patient education materials published by a national organization	7 (35%)	1.57±0.75
Total	20(100%)	1.5±0.68
JAMA Benchmarks of Website	**Score**	**Number n (%)**
Authorship	5	
Attribution	3	
Disclosure	15	
Currency	9	
4.0		0(0%)
3.0		2(10%)
2.0		6(30%)
1.0		12(60%)

JAMA Accountability Scores for Webpages to Keratoconus-Related FAQs

The mean JAMA Benchmark score for all websites was 1.5 ± 0.68. The mean "Private practice or independent user" website score was 1.5 ± 0.68, while that for "Official patient education materials" was 1.57 ± 0.75. In terms of of JAMA scores, five websites met authorship criteria, three met attribution criteria, 15 met disclosure criteria, and nine currency criteria. Most websites (60%) scored only 1, indicating poor adherence to JAMA guidelines. Only two websites (10%) achieved scores of 3, representing moderate accountability. Six websites (30%) scored 2, and none scored 4.

Average Score for Each Question

Table 2 evaluates the performance of LLMs in answering keratoconus-FAQ using mDISCERN, GQS, and ARLC scores. The most frequently addressed question was "What is keratoconus?" (on 70% of websites), which received the highest mDISCERN score (49.33 ± 4.96), a high GQS score (3.50 ± 0.55), and an ARLC score of 13.17 ± 2.13. Other questions exhibited lower coverage. For example, "How do patients with keratoconus see?" (on15% of websites) received a mDISCERN score of 44.83 ± 3.43, a GQS score of 3 ± 0.63, and an ARLC score of 14.17 ± 2.85. Scores for "Can keratoconus go away on its own?" (on 15% of websites) were mDISCERN 44.5 ± 3.61, GQS 2.83 ± 0.41, and ARLC 14.5 ± 1.64. "What should be considered after keratoconus surgery?" (on 15% of websites) received a mDISCERN score of 46.33 ± 2.87, a GQS score of 3.17 ± 0.41, and an ARLC score of 13.67 ± 2.06.

Moderate positive correlations were observed between the number of websites and both mDISCERN (r=0.265, p=0.25) and GQS scores (r=0.453, p=0.05), indicating higher quality and reliability for FAQs. However, a weak negative correlation was found between the number of websites and ARLC scores (r=−0.151, p=0.55), suggesting that readability is not strongly correlated with the number of websites addressing a particular question.

Table 2. Mean scores for each question.

No	Question	Number of websites n (%)	mDISCERN (Mean ± SD)	GQS score (Mean ± SD)	ARLC (Mean ± SD
1.	What Is Keratoconus?	14 (70%)	49.33±4.96	3.5±0.55	13.17±2.13
2.	What Are the Symptoms of Keratoconus?	10 (50%)	43.5±2.58	3±0.00	12.33±2.33
3.	How do patients with keratoconus See ?	3 (15%)	44.83±3.43	3±0.63	14.17±2.85
4.	How Can Keratoconus Affect My Life?	6 (30%)	45±3.34	3.17±0.41	14.5±1.37
5.	How Common Is Keratoconus?	7 (35%)	45.17±3.18	3.17±0.75	13.67±1.96

 doi:10.20944/preprints202409.0257.v1

6.	What Causes Keratoconus?	14 (70%)	43.5±6.41	3±0.63	15.5±1.37
7.	Does Keratoconus Cause Blindness?	6 (30%)	47.17±5.74	3.5±0.84	14.83±1.47
8.	Can LASIK or RK Surgery Cause Keratoconus?	5 (25%)	42.67±3.93	3.17±0.75	16.33±2.65
9.	Are There Multiple Forms of Keratoconus?	4 (20%)	38.67±4.63	2.33±0.52	13.67±2.16
10.	How Is Keratoconus Diagnosed?	8 (40%)	42.5±2.42	3.17±0.41	15.33±1.50
11.	How Do You Measure the Severity of Keratoconus?	4 (20%)	43.33±1.96	3.17±0.41	15.5±3.39
12.	How Can I Treat My Keratoconus?	16 (80%)	47.5±4.32	3.67±0.52	13.67±1.63
13.	What is the Best Keratoconus Treatment?	5 (25%)	46.5±3.56	3.33±0.52	14±2.09
14.	How Can I Stop My Keratoconus From Getting Worse?	6 (30%)	44.83±4.57	3.33±0.52	13.67±1.75
15.	Is Keratoconus Always Progressive?	9 (45%)	43±3.34	3±0.00	13.67±1.21
16.	Does Keratoconus Cause Eye Pain?	4 (20%)	43.83±4.44	2.67±0.52	13.5±1.51
17.	Can Keratoconus Go Away On Its Own?	3 (15%)	44.5±3.61	2.83±0.41	14.5±1.64
18.	Can Keratoconus Cause Dry Eye?	4 (20%)	46.67±3.55	3.17±0.98	14.17±1.72
19.	What Do I Do If I Think I Have Keratoconus?	4 (20%)	45.5±3.27	3±0.63	13.83±1.72
20.	What Should Be Considered After Keratoconus Surgery?	3 (15%)	46.33±2.87	3.17±0.41	13.67±2.06

The score for each question was calculated by averaging the scores of the large language models(ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity) ARLC (The Average Reading Level Consensus Calculator); processes your text through 8 popular readability formulas (Linsear Write Formula, SMOG Index, Coleman-Liau Index, Flesch-Kincaid Grade Level, Gunning Fog Index, Flesch Reading Ease Formula, Automated Readability Index, FORCAST Readability Formula) and averages out the results to yield an approximate reading difficulty score.

Reliability

(mDISCERN Score)

All LLMs performed reasonably well, with Gemini, Copilot, and Perplexity exhibiting higher reliability. The lowest mDISCERN score was 42.9 ± 3.16 (ChatGPT-3.5) and the highest was 46.95 ± 3.53 (Copilot). The differences were statistically significant ($p < 0.05$).

(Quality GQS Score)

The lowest GQS score was 3.01±0.51, observed in the ChatGPT-3.5 model, and the highest was 3.3 ± 0.65, in the Gemini model. The differences between these models were not significant, with a p-value of 0.34. Gemini and Copilot again achieved higher scores, indicating better overall quality.

Readability Indices

 doi:10.20944/preprints202409.0257.v1

In terms of the readability indexes of all texts, Table 4 shows how the various models performed in the context of text comprehensibility. The texts were clearly generally difficult to read and were suitable for readers educated to high school or college level. The p-values for all indexes were <0.05, and it may therefore be concluded that the differences between the texts were statistically significant. Gemini emerged as the easiest readable text, having received the lowest score on most readability indexes. More specifically, the low scores on the FRE and FKGL indexes suggest that the texts were simpler and more comprehensible. ChatGPT-3.5 and Perplexity emerged as the most difficult readable texts, exhibiting the highest scores on most readability indexes. The high scores on the GFI and ARI emphasize that the texts required a more advanced reading proficiency level.

Table 4. Comparison of DISCERN, GQS, and Readability Results of Large Language Models.

	chatgpt3.5	Chatgpt4	Gemini	Copilot	Chatsonic	Perplexity	p-value
Reliability							
mDISCERN score (mean ± SD)	42.96±3.16	43.2±2.87	46.05±5.12	46.95 ±3.53	43.95±2.16	45.95±3.53	<0.05
Quality							
GQS (Mean ± SD)	3.01±0.51	3.05±0.44	3.3±0.65	3.25±0.55	3.15±0.67	3.06±0.68	.34
Readability indexes							
FRE (Mean ± SD)	21.43±7.10	28.85±8.44	34.7±8.79	29.6±9.04	24.4±8.98	22.3±12.75	<0.05
FKGL (Mean ± SD)	15.41±1.39	14.64±1.70	12.46±1.73	12.04±1.44	13.38±1.31	15.5±3.02	<0.05
GFI (Mean ± SD)	18.52±2.09	18.05±2.18	15.33±2.03	15.21±1.93	17.43±1.80	18.98±3.79	<0.05
CLI (Mean ± SD)	15.85±0.75	14.75±1.27	14.57±1.51	15.29±1.35	16.07±1.42	15.93±1.75	<0.05
ARI (Mean ± SD)	16.07±1.31	15.62±1.94	13.21±1.91	12.18±1.29	13.38±1.27	16.58±3.16	<0.05
SMOG (Mean ± SD)	13.77±1.23	13.16±1.51	11.11±1.46	10.25±1.27	11.89±1.11	13.69±2.64	<0.05
LINSEAR (Mean ± SD)	16.33±1.97	16.07±3.11	11.59±2.78	8.2±2.37	11.07±2.12	16.36±5.33	<0.05
FORCAST (Mean ± SD)	12.45±0.52	12.08±0.65	12.08±0.54	12.75±0.67	12.61±0.62	12.53±0.61	<0.05
ARLC	15.65±1.13	14.85±1.53	12.9±1.61	12.35±1.18	13.6±1.14	15.75±2.55	<0.05
Response length							
Sentences (Mean ± SD)	13.05±4.21	13.3±3.31	13±5	16±4.69	13.35±3.78	6.9±4.29	<0.05
Words (Mean ± SD)	264.95±78.48	285.9±60.76	237.05±71.85	231.3±50.69	283.4±70.53	127.8±51.96	<0.05
Characters (Mean ± SD)	1802.8±538.68	1919.6±425.43	1595.15±489.22	1585±363.31	1985.2±491.80	893.6±379.47	<0.05
Syllable (Mean ± SD)	510.9±144.98	534.6±118.61	435.1±129.66	442.7±98.51	564.05±133.18	251.6±106.22	<0.05
Word/sentence (Mean ± SD)	19.66±2.69	20.43±3.39	14.22±1.96	11.3±2.52	13.65±2.15	16.57±5.34	<0.05
Syllable/word	1.96±0.06	1.89±0.10	1.86±0.09	1.96±0.12	2.01±0.10	1.99±0.12	<0.05

GQS: Global Quality Score **FRE:** Flesch Reading Ease, **GFI:** Gunning Fog Index, **FKGL:** Flesch-Kincaid Grade Level, **SMOG:** Simple Measure of Gobbledygook, **CLI:** Coleman-Liau Index, **ARI:** Automated Readability Index, **LINSEAR:** Linsear Write Formula, **FORCAST:** Readability Formula, and the **ARLC:** Automated Readability Level Calculator.

Response Length

 doi:10.20944/preprints202409.0257.v1

Copilot generated the longest responses (16 ± 4.69 sentences) and Perplexity the shortest (6.9 ± 4.29). ChatGPT-4 produced the most words (285.9 ± 60.76) and Perplexity the fewest (127.8 ± 51.96). Chatsonic exhibited the highest character count (1985.2 ± 491.80) and syllable count (564.05 ± 133.18), the lowest values for both being determined in Perplexity (893.6 ± 379.47 characters and 251.6 ± 106.22 syllables). ChatGPT-4 exhibited the highest words per sentence ratio (20.43 ± 3.39) and Copilot the lowest (11.3 ± 2.52). Chatsonic registered the highest syllables per word ratio (2.01 ± 0.10) and Gemini the lowest (1.86 ± 0.09). All differences were statistically significant ($p < 0.05$).

Score Distributions of mDISCERN Scale and Quality Classification

Table 5 presents the mDISCERN score distribution and quality classification of keratoconus responses from the various different LLMs. Most models (75-95%) scored in the "Reasonable" range (39-50 points). Gemini and Copilot achieved the highest "Good" range scores (51-62 points) at 30% and 20%, respectively. However, no model achieved the "Excellent" range (63-75 points). Perplexity and Chatsonic exhibited the highest "Poor" range scores (27-38 points) at 5% and 10%, respectively.

GQS showed moderate quality, with most models scoring in the range of 3-3.5. Gemini achieved the highest "Good" quality responses (40% scoring 4-5 points). Copilot and Chatsonic also registered significant "Good" quality responses (30%). ChatGPT-3.5 and ChatGPT-4.0 exhibited the lowest "Good" quality responses at 15% and 10%, respectively.

Table 5. Score distribution of large language models responses according to the mDISCERN scale and quality classification. Categorical variables are presented as n (%) in the table.

mDISCERN criteria	chatgpt3.5	Chatgpt4	Gemini	Copilot	Chatsonic	Perplexity
	n=20 (%)	n=20 (%)	n=20 (%)	n=20 (%)	n=20 (%)	n=20 (%)
Excellent (63–75 points)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Good (51–62 points)	1 (5%)	1 (5%)	6 (30%)	4 (20%)	0 (0%)	5 (25%)
Reasonable (39–50 points)	17 (85%)	16 (80%)	12 (70%)	15 (75%)	19 (95%)	15 (75%)
Poor (27–38 points)	2 (10%)	3 (15%)	2 (10%)	1 (5%)	1 (5%)	0 (0%)
Very poor (15–26 points)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Quality classification						
Low quality	2 (10%)	2 (10%)	2 (10%)	1 (5%)	3 (15%)	4 (20%)
Moderate quality	15 (75%)	16 (80%)	10 (50%)	13 (65%)	11 (55%)	11 (55%)
High quality	3 (15%)	2 (10%)	8 (40%)	6 (30%)	6 (30%)	15(25%)

Discussion

This study evaluated the efficacy of six LLMs in terms of accurately responding to medical queries by comparing their performance on common keratoconus-related questions sourced from Google searches. The findings indicate that LLM-Chatbots have the potential to provide comprehensive responses to keratoconus-related inquiries.

LLMs can provide keratoconus patients with up-to-date, evidence-based information, facilitating rapid access to the latest therapeutic options and research findings. Patients can use LLMs for a better understanding of their condition and to make informed healthcare decisions. Accurate and comprehensible information from LLMs can enhance patient adherence to treatment plans and alleviate concerns, thereby improving their emotional well-being. LLMs also have the potential to empower keratoconus patients by equipping them with the knowledge required for active participation in their healthcare journeys.

This study investigated practical scenarios in which concerned patients might seek assistance from emerging resources. To the best of our awareness, this is the first study to evaluate LLM responses to keratoconus-related queries. The research builds on previous studies examining the applicability of LLM chatbots, such as ChatGPT, across various medical subspecialties. Prior research has explored the use of LLMs for providing medical information, patient education, and diagnostic and treatment recommendations, albeit with mixed results [2,3]. One significant cause for concern is the potential for misinformation in medical chatbots, which can be manipulated by special interest

 doi:10.20944/preprints202409.0257.v1

groups [4]. AI models are designed to generate content based on probable word sequences rather than producing factual answers. While generative AI chatbots can debunk misinformation, they can also spread falsehoods if not regularly updated with the latest scientific evidence [5]. For example, Lim et al. reported that ChatGPT-3.5 incorrectly stated that "Atropine eye drops are a new treatment for myopia and their optimal dosage has not yet been determined"[6]. Giuffrè et al. evaluated LLMs in the context of digestive diseases and concluded that, despite their potential, their current accuracy and reliability are inadequate for clinical use [7]. Conversely, LLMs such as ChatGPT and Google Bard have exhibited impressive medical knowledge and capabilities, proving beneficial for patient communication [8–10].

In the field of ophthalmology, LLM chatbots have exhibited promise in addressing common patient queries concerning eye health [11]. Cohen et al. determined that human responses to ophthalmology-based questions contained a similar rate of incorrect or inappropriate material (27%), as also reported by Bernstein et al. [12,13]. However, AI responses in the current study were more accurate (94%) than those provided by ChatGPT in Bernstein et al.'s study (77%) [13].

In this study, responses from private practice or independent organization websites (65%, n=13) registered a mean JAMA accountability score of 1.5±0.68, indicating infrequent adherence to JAMA criteria. Official patient education materials from national organizations (35%, n=7) had a slightly higher mean score of 1.57±0.75, suggesting marginally better compliance. The overall mean JAMA Benchmark score was 1.5±0.68, indicating low accountability across the websites. This trend is consistent with numerous previous studies. A comprehensive analysis of five studies regarding the readability and accountability of online ophthalmology patient education materials reported a mean JAMA accountability score of 1.13 with a standard deviation of 1.15, reflecting substantial deficiencies in both quality and accountability [14–18]. These findings highlight the need for improved standards in creating and disseminating online patient education materials.

When considering keratoconus-related FAQs, metrics such as mDISCERN, GQS, and ARLC scores provide insights into the performance of LLMs in the medical sphere. The question "What is keratoconus?," addressed by 70% of websites, registered the highest mDISCERN and GQS scores. In contrast, less frequently addressed questions such as "Are There Multiple Forms of Keratoconus?" (15%) received lower mDISCERN scores. Similarly, "Does Keratoconus Cause Eye Pain?" (20%) and "Can Keratoconus Go Away On Its Own?" (15%) registered lower GQS scores. The ARLC score, indicating readability, exhibited less variability, with most questions scoring between 12 and 16. For instance, "What is Keratoconus?" achieved an ARLC score of 13.17 ± 2.13, while "Can LASIK or RK Surgery Cause Keratoconus?" (25%) scored 16.33 ± 2.65. These findings highlight the importance of question frequency in determining the response quality and the potential of LLMs to provide high-quality, reliable medical information, especially for frequently asked questions. However, readability does not exhibit strong correlation with the number of websites addressing a question.

mDISCERN indices evaluate the performance of LLMs in providing medical information, assessing the informativeness, accuracy, and safety of the content. Wilhelm et al. identified significant quality differences among LLMs, with notable variability in mDISCERN scores. The Claude-instant-v1.0 model received the highest score, and Bloomz the lowest [19]. The present study indicates that although all LLMs performed reasonably well, their ability to provide accurate and reliable medical information differs significantly. Models such Gemini and Copilot scored higher, suggesting better performance. The significant variability in mDISCERN scores underscores the need for continuous improvement and validation. Standardized evaluation metrics and rigorous testing protocols are essential for assessing AI model performance and identifying potential areas for improvement.

The mDISCERN score distribution in this study reveals that Gemini and Copilot performed better in the "Good" range (51-62 points) compared to other LLMs, probably due to superior training data, fine-tuning, or algorithms. Conversely, Perplexity and Chatsonic registered the highest percentage of responses in the "Poor" range (27-38 points), indicating potential weaknesses due to less comprehensive training and suboptimal fine-tuning. These findings suggest that while LLMs can generate reasonably reliable medical information, there is a significant gap in achieving high reliability across all models. No model reached the "Excellent" range in mDISCERN scores,

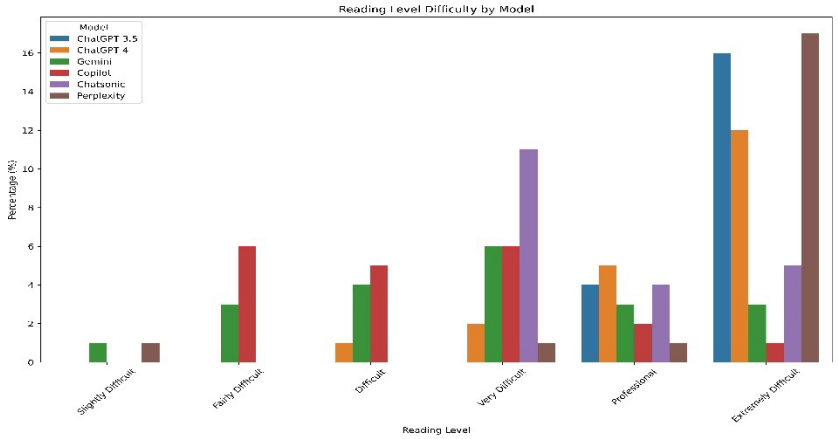
 doi:10.20944/preprints202409.0257.v1

indicating that current LLMs are not yet capable of providing highly reliable information for all questions. Onder et al. evaluated ChatGPT-4 responses concerning hypothyroidism during pregnancy using DISCERN tools, reporting that most responses were either Fair (78.9%) or Good (21.1%) [20]. This highlights the model's capability to generate dependable information in most instances. The performance differences among LLMs emphasize the need for ongoing research and development in order to enhance the reliability and quality of information generated by these models.

Evaluating LLMs using the GQS provided valuable insights into the quality of medical information generated by them. Although no significant differences were observed among LLMs, models such as Gemini and Copilot consistently scored higher, indicating better overall quality and more robust mechanisms for generating accurate content. Ostrowska et al. evaluated the reliability and safety GQS of LLMs in the context of laryngeal cancer, describing ChatGPT 3.5 as the most successful model [21]. This emphasizes the need for model-specific evaluations in order to identify the best-performing models for particular medical spheres.

GQS score analysis revealed varying levels of quality in medical information produced by LLMs. The majority of models scored in the 3-3.5 range on a five-point scale, indicating moderate quality. Gemini emerged as the top performer, with 40% of its outputs in the "Good" quality range (4-5 points). Copilot and Chatsonic also performed well, with 30% of their responses in the "Good" range. In contrast, ChatGPT models (3.5 and 4.0) achieved lower rates of "Good" quality responses (15% and 10%, respectively). In contrast to our findings, Onder et al. reported that 84.2% of ChatGPT-4's responses regarding hypothyroidism during pregnancy were of high quality, followed by 10.5% medium quality responses [20]. This discrepancy suggests that the specific medical sphere or the nature of the questions in the present study may have been particularly challenging for these models, a subject warranting further investigation at a later date.

Although our expert evaluators preferred chatbot responses, their readability frequently exceeded the American Medical Association's (AMA) recommendation of a sixth-grade reading level for patient education materials. Using eight popular readability formulae, the final ARLC scores indicated the following reading levels: ChatGPT-3.5, ChatGPT-4, and Perplexity were rated as extremely difficult, Gemini and Copilot as difficult, and Chatsonic as very difficult. The corresponding grade levels were ChatGPT-3.5, ChatGPT-4, Perplexity at the College Graduate level, Gemini and Copilot at the Twelfth Grade level, and Chatsonic at College Entry level (Figure 1). These findings align with previous research showing that chatbot-generated patient education information is frequently written at reading levels significantly exceeding the comprehension of the average patient [12,22].



 doi:10.20944/preprints202409.0257.v1

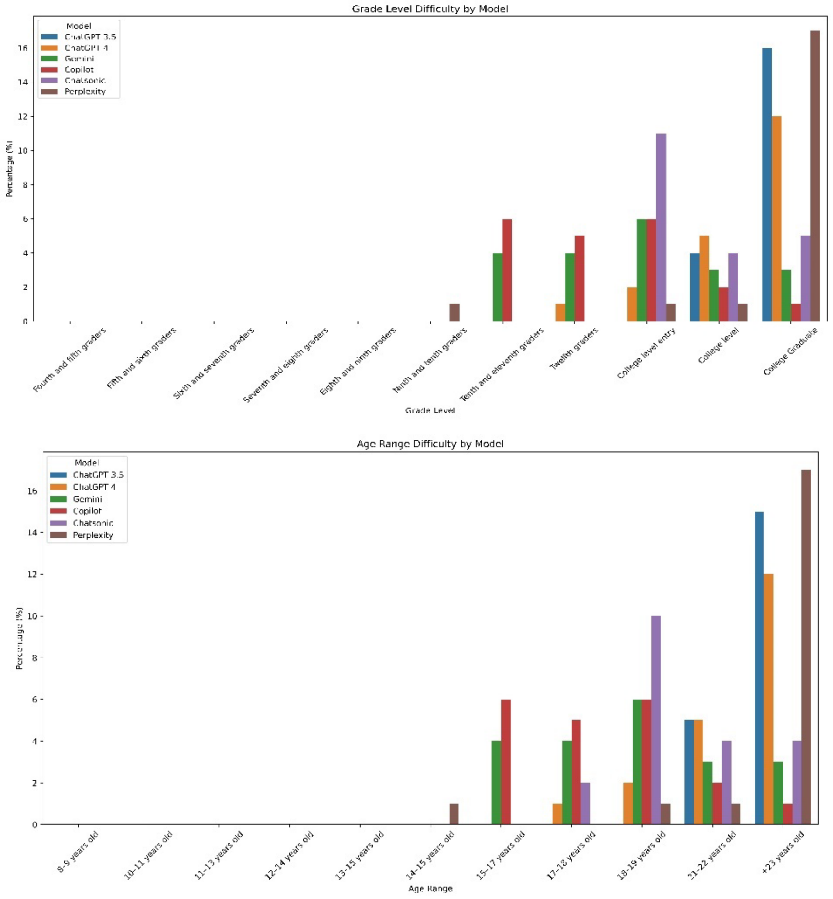


Figure 1. The graph of Reading Difficulty, Grade Level, and Age Range of large language models evaluated according to the Average Reading Level Consensus Calculator.

Research indicates that tailoring patient education materials to patients' health literacy levels can significantly enhance compliance and optimize health outcomes [23]. A scoping review of visual aids in health literacy reported that materials intended for individuals with low literacy levels significantly improved health literacy outcomes, including medication adherence and comprehension [24].

While some chatbots, such as ChatGPT-4 and Chatsonic, produce detailed and complex responses, others, including Perplexity, generate shorter and simpler answers. These differences in response length and complexity highlight the varying capabilities of LLM-Chatbots in addressing keratoconus-related FAQs. This information is crucially important for selecting an appropriate chatbot for specific informational needs, particularly in medical and educational contexts in which the depth and clarity of information are paramount.

The adaptability of chatbots to user requests is significant for their potential application in ophthalmology. Despite challenging reading levels, providing patient education materials remains highly beneficial. This study demonstrates the usefulness of chatbots in providing keratoconus-related information for patients. Ophthalmologists report a loss of efficiency due to excessive time spent on non-clinical tasks. Chatbots can help alleviate this burden. A semi-supervised model, in which the ophthalmologist reviews AI-generated responses, represents the future of AI and can be highly beneficial tool for ophthalmologists.

While this study provides insights into the differences in responses from six LLMs to common keratoconus-related questions, a number of limitations must also be considered. In particular, the questions were sourced from Google, and the manner in which patients interpret these responses was not investigated. When ophthalmologists provide information regarding keratoconus and advice on using AI tools such LLMs, it is essential that the patient's health literacy level be taken into account.

Conclusions

LLMs can provide comprehensive and accurate responses to keratoconus-related queries, enhancing patient adherence, decision-making, and emotional well-being. However, the performance of the different LLM chatbots varies in terms of quality, reliability, and readability. While all LLMs performed commendably, Gemini and Copilot emerged as superior in providing reliable and high-quality information, with Gemini demonstrating the best readability. In contrast, ChatGPT-3.5 and Perplexity produced the most difficult-to-read texts, potentially hindering patient comprehension. Tailoring information to patients' health literacy levels is crucial. Continuous improvement and validation of LLM chatbots is essential, together with standardized evaluation metrics and rigorous testing protocols. A semi-supervised model, in which an ophthalmologist reviews AI-generated responses, represents a promising approach to the integration of AI in ophthalmology, potentially reducing the burden on healthcare professionals.

Authorship Contribution Statement: A.H.R. *material preparation, data collection and analysis,* supervision, writing – review & editing. Ç.M. Conceptualization, Data curation, review & editing. F.Y. and İ.U. writing – review & editing. A.Ş. interpreted the results and revised critically the manuscript. All authors read and approved the final manuscript.

Funding: *The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.*

Disclosure Statement: No potential conflict of interest was reported by the author(s).

References

1. Santodomingo-Rubido J, Carracedo G, Suzaki A. et al. Keratoconus: An updated review. Cont Lens Anterior Eye 2022; 45:101559. doi: 10.1016/j.clae.2022.101559
2. Nazario-Johnson L, Zaki HA, Tung GA. Use of Large Language Models to Predict Neuroimaging. J Am Coll Radiol 2023; 20: 10004-09. doi:10.1016/j.jacr.2023.01.004
3. Kumari A, Kumari A, Singh A. et al. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus 2023; 15: e43861. doi:10.7759/cureus.43861
4. Meyrowitsch DW, Jensen AK, Sørensen JB.et al. AI chatbots and (mis)information in public health: impact on vulnerable communities. Front Public Health 2023; 11:1226776. doi:10.3389/fpubh.2023.1226776
5. Stephens LD, Jacobs JW, Adkins BD.et al. Battle of the (Chat)Bots: Comparing Large Language Models to Practice Guidelines for Transfusion-Associated Graft-Versus-Host Disease Prevention. Transfus Med Rev 2023; 37:150753. doi: 10.1016/j.tmrv.2023.150753
6. Lim ZW, Pushpanathan K, Yew SME.et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023; 95:104004. doi: 10.1016/j.ebiom.2023.104004
7. Giuffrè M, Kresevic S, You K.et al. Systematic review: The use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther 2024; 27. doi: 10.1111/apt.18058.
8. Neo JRE, Ser JS, Tay SS. Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers. Frontiers in Digital Health 2024; 6: 1395501. doi: 10.3389/fdgth.2024.1395501
9. Wu Y, Zhang Z, Dong X.et al. Evaluating the performance of the language model ChatGPT in responding to common questions of people with epilepsy. Epilepsy & Behavior 2024;151: 109645. doi: 10.1016/j.yebeh.2024.109645
10. Peng C, Yang X, Chen A. et al. A study of generative large language model for medical research and healthcare. NPJ digital medicine 2023;*6*: 210. doi: 10.1038/s41746-023-00760-6
11. Betzler BK, Chen H, Cheng CY. et al. Large language models and their impact in ophthalmology. Lancet Digit Health 2023;5: 917-24. doi: 10.1016/S2589-7500(23)00159-2
12. Cohen SA, Brant A, Fisher A.et al. Dr. Google vs. Dr. ChatGPT: Exploring the Use of Artificial Intelligence in Ophthalmology by Comparing the Accuracy, Safety, and Readability of Responses to Frequently Asked Patient Questions Regarding Cataracts and Cataract Surgery. Semin Ophthalmol 2024; 39:1-8. doi: 10.1080/08820538.2024.2193524
13. Bernstein IA, Zhang YV, Govil D.et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Common Patient Queries. JAMA *JAMA network open* 2023; *6*: e2330320. doi: 10.1001/jamaophthalmol.2023.1968.

14. Cohen SA, Fisher AC, Pershing S. Analysis of the readability and accountability of online patient education materials related to glaucoma diagnosis and treatment. Clin Ophthalmol 2023; 17:779-88. doi: 10.2147/OPTH.S395524
15. Martin CA, Khan S, Lee R.et al. Readability and suitability of online patient education materials for glaucoma. Ophthalmol Glaucoma 2022; 5:525-530. doi: 10.1016/j.ogla.2022.05.005
16. Patel P, Patel P, Ahmed H.et al. Content, Readability, and Accountability of Online Health Information for Patients Regarding Blue Light and Impact on Ocular Health. Cureus 2023;15: e43861. doi: 10.7759/cureus.43861
17. Redick DW, Hwang JC, Kloosterboer A.et al. Content, readability, and accountability of freely available online information for patients regarding epiretinal membranes. Semin Ophthalmol 2022; 37:67-70. doi: 10.1080/08820538.2021.1991688
18. Kloosterboer A, Yannuzzi N, Topilow N.et al. Assessing the quality, content, and readability of freely available online information for patients regarding age-related macular degeneration. Semin Ophthalmol 2021; 36:400-405. doi: 10.1080/08820538.2021.1912305
19. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 2023;25: e49324. doi: 10.2196/49324
20. Onder CE, Koc G, Gokbulut P.et al. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024; 14:243. doi: 10.1038/s41598-024-27524-4
21. Ostrowska M, Kacała P, Onolememen D.et al. To trust or not to trust: evaluating the reliability and safety of AI responses to laryngeal cancer queries. Eur Arch Otorhinolaryngol. 2024; 1-13. doi: 10.1007/s00405-023-08000-w.
22. Wu G, Zhao W, Wong A.et al. Patients with floaters: Answers from virtual assistants and large language models. Digit Health 2024; 10: 20552076241229933. doi: 10.1177/20552076241229933
23. Newman-Casey PA, Niziol LM, Lee PP.et al. The impact of the support, educate, empower personalized glaucoma coaching pilot study on glaucoma medication adherence. Ophthalmol Glaucoma. 2020; 3:228-37. doi: 10.1016/j.ogla.2020.01.009
24. Mbanda N, Dada S, Bastable K.et al. A scoping review of the use of visual aids in health education materials for persons with low-literacy levels. Patient Educ Couns 2021; 104:998-1017. doi: 10.1016/j.pec.2020.11.034