AMERICAN ACADEMY OF OPHTHALMOLOGY®

A Comparative Study of Responses to Retina Questions from Either Experts, Expert-Edited Large Language Models, or Expert-Edited Large Language Models Alone

Prashant D. Tailor, MD,[1] Lauren A. Dalvin, MD,[1] John J. Chen, MD, PhD,[1] Raymond Iezzi, MD,[1] Timothy W. Olsen, MD,[1] Brittni A. Scruggs, MD, PhD,[1] Andrew J. Barkmeier, MD,[1] Sophie J. Bakri, MD,[1] Edwin H. Ryan, MD,[2,3] Peter H. Tang, MD, PhD,[2,3] D. Wilkin. Parke III, MD,[2,3] Peter J. Belin, MD,[2] Jayanth Sridhar, MD,[4] David Xu, MD,[5] Ajay E. Kuriyan, MD,[5] Yoshihiro Yonekawa, MD,[5] Matthew R. Starr, MD[1]

Objective: To assess the quality, empathy, and safety of expert edited large language model (LLM), human expert created, and LLM responses to common retina patient questions.

Design: Randomized, masked multicenter study.

Participants: Twenty-one common retina patient questions were randomly assigned among 13 retina specialists.

Methods: Each expert created a response (Expert) and then edited a LLM (ChatGPT-4)-generated response to that question (Expert + artificial intelligence [AI]), timing themselves for both tasks. Five LLMs (ChatGPT-3.5, ChatGPT-4, Claude 2, Bing, and Bard) also generated responses to each question. The original question along with anonymized and randomized Expert + AI, Expert, and LLM responses were evaluated by the other experts who did not write an expert response to the question. Evaluators judged quality and empathy (very poor, poor, acceptable, good, or very good) along with safety metrics (incorrect information, likelihood to cause harm, extent of harm, and missing content).

Main Outcome: Mean quality and empathy score, proportion of responses with incorrect information, likelihood to cause harm, extent of harm, and missing content for each response type.

Results: There were 4008 total grades collected (2608 for quality and empathy; 1400 for safety metrics), with significant differences in both quality and empathy ($P < 0.001$, $P < 0.001$) between LLM, Expert and Expert + AI groups. For quality, Expert + AI (3.86 ± 0.85) performed the best overall while GPT-3.5 (3.75 ± 0.79) was the top performing LLM. For empathy, GPT-3.5 (3.75 ± 0.69) had the highest mean score followed by Expert + AI (3.73 ± 0.63). By mean score, Expert placed 4 out of 7 for quality and 6 out of 7 for empathy. For both quality ($P < 0.001$) and empathy ($P < 0.001$), expert-edited LLM responses performed better than expert-created responses. There were time savings for an expert-edited LLM response versus expert-created response ($P = 0.02$). ChatGPT-4 performed similar to Expert for inappropriate content ($P = 0.35$), missing content ($P = 0.001$), extent of possible harm ($P = 0.356$), and likelihood of possible harm ($P = 0.129$).

Conclusions: In this randomized, masked, multicenter study, LLM responses were comparable with experts in terms of quality, empathy, and safety metrics, warranting further exploration of their potential benefits in clinical settings.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of the article. *Ophthalmology Science 2024;4:100485*

The creation of patient portals in electronic health record systems have allowed patients to communicate directly with their physicians through electronic messages (EMs) regarding medical questions, thus improving access to medical care and health literacy.[1] However, an unintended effect is that it has increased physician workload (i.e., in-basket burden) as these tasks are typically completed by physicians outside of conventional clinic hours in addition to daily clinical responsibilities.[2] Between 2013 and 2018, 1 study reported a 110% increase in the volume of EMs directed to health care providers.[3] Another study found that surgical subspecialties experienced a notable uptick in EMs during the coronavirus 2019 pandemic.[4,5] Attempts to rectify this issue through compensation (i.e., billing for response to EMs) has not improved physician work

https://doi.org/10.1016/j.xops.2024.100485
ISSN 2666-9145/24

quality-of-life and may act as barriers for patient access.[6] The use of artificial intelligence (AI), particularly new models, may provide a unique and critical solution to this problem by automatically crafting draft responses to messages, theoretically leading to significant time savings and quality-of-life improvements.

Large language models (LLMs) are machine learning models that process and generate human-like text based on the information they have been trained on.[7] Large language models represent a pivotal development in the field of AI, demonstrating extensive utility across diverse sectors including law, medicine, writing, and computing.[3,7–9] Within the medical domain, LLMs have exhibited significant promise in tasks ranging from passing licensing exams to improving patient understanding of radiology imaging reports.[7,10–13] In the realm of patient-physician communication, the application of LLMs presents a notable opportunity. Specifically, their use in enhancing physician responses to patient queries has shown promise, with research indicating their effectiveness in addressing common questions related to vitreoretinal surgery.[14] Considering the high patient volumes faced by retinal specialists and the consequent demand for timely responses to patient inquiries, LLMs offer a potential solution to alleviate the workload. Our study aimed to evaluate the effectiveness of LLMs in this context by benchmarking their performance against both retinal specialists and specialist-edited LLM responses. This approach was taken with the goal of determining the feasibility of LLMs in reducing the burden of patient message management for retinal specialists.

Our multicenter, randomized, cross-sectional study sought to analyze the performance of LLMs across 5 commercially available platforms including Bard, version 2 of Claude, Bing, and versions 3.5 and 4 of ChatGPT. We compared responses to commonly-asked patient questions regarding retinal diseases by LLMs as well as LLM responses edited by human experts and benchmarked these against expert-created responses.

Methods

The study was exempt by the Mayo Clinic institutional review board as it contained no patient information and adhered to the tenets of the Declaration of Helsinki. Conducted from June to September 2023, this multicenter, randomized cross-sectional study aimed to evaluate differences in quality, empathy, and safety of expert edited LLM responses (Expert + AI), expert-created responses (Expert), and 5 commercially available LLMs: Bard (Alphabet), Claude 2 (Anthropic), Bing (Microsoft), ChatGPT-3.5 (OpenAI), and ChatGPT-4 (OpenAI) to commonly-asked retina questions by patients. We employed commercially available LLMs in our study due to their benchmarked superior performance and user-friendly interfaces suitable for the general population. Open source models were not considered, as they often present accessibility challenges for nontechnical users specifically in utilization and their performance rankings frequently change, making durable evaluations difficult.

As current LLM offerings are not Health Insurance Portability and Accountability Act-compliant, 6 retinal specialists (AJB, BAS, RI, MRS, SJB, TWO) created 21 retina questions related to risk factor counseling, disease etiology and pathogenesis, test result interpretation, and clinical experience (Table 1). These questions were similar to common patient inquiries that might be received in clinic or via patient electronic health record portals. The 21 questions were selected based on their frequency in patient consultations and relevance to retinal diseases, as confirmed through a preliminary survey of retinal specialists when asked to suggest questions that were both based on questions asked in clinic or on the patient portal. Thirteen fellowship-trained retinal specialists (AJB, BAS, MRS, SJB, TWO, EHR, PHT, DWP, PJB, JS, DX, AEK, YY) from 4 centers across the United States participated in the evaluation. Each expert was randomly and anonymously assigned 1 or 2 unique questions. For each question, they were instructed to submit a typed response (Expert) along with the time it took to complete the task, as if they were responding to

Table 1. List of Simulated Vitreoretinal Patient Questions

	Questions
1.	Can I get a whole eye replacement?
2.	What causes age-related macular degeneration?
3.	How long do I need to keep getting anti-VEGF injections?
4.	Can I pass AMD to my children?
5.	How does retinal gene therapy work?
6.	Can stem cells give me my vision back?
7.	Should I get a scleral buckle or a vitrectomy for my retinal detachment?
8.	Why do I need AREDS2/eye vitamins?
9.	What are the advantages and disadvantages of anti-VEGF injections vs. panretinal photocoagulation for proliferative diabetic retinopathy?
10.	What causes a posterior vitreous detachment/retinal tear?
11.	What is the success of vitrectomy for retinal detachment? What are the chances of a second procedure?
12.	Which is worse? The dry or the wet macular degeneration?
13.	Which anti-VEGF injection works best?
14.	How long can I go between eye injections?
15.	My doctor says I have a retinal vein occlusion? Is that a stroke?
16.	How long do I need to be face down after macular hole surgery?
17.	Should I get a pneumatic retinopexy or a vitrectomy for my retinal detachment?
18.	Is there a good treatment for floaters?
19.	Is exercise good for my eyes?
20.	Why do I need to avoid steroids in central serous chorioretinopathy?
21.	How long do I need to wait before I can fly on an airplane after retina surgery?

AMD = age-related macular degeneration; AREDS2 = Age-Related Eye Disease Study 2.

Table 2. Summary Statistics on Length of Response Types (Words)

Model	Count	Mean	Std	Min	25%	50%	75%	Max
Bard	21	329.81	62.18	232	292	323	365	443
Bing	21	96.90	42.71	26	74	90	106	236
Claude	21	219.48	31.79	151	199	216	239	274
Expert	21	138.67	103.31	31	65	101	191	396
Expert + AI	21	290.00	66.80	220	235	271	363	416
GPT3.5	21	327.57	72.13	192	271	338	361	482
GPT4	21	281.71	54.65	162	237	286	313	392

AI = artificial intelligence; GPT = Generative Pre-trained Transformer.

the patient message. After each expert submitted their response(s), a response to the same question(s) from ChatGPT-4 (see following) was provided to the expert, who was instructed to edit the LLM response (Expert + AI) to their satisfaction (as if using the edited response when replying to a patient message) and document the time it took them to finish the task. ChatGPT-4 was chosen for its advanced reasoning and performance on medical licensing examinations.[7,12] Experts knew their response would be compared to some unknown number of LLMs; however, they did not know that their expert-edited LLM response would be included in grading surveys.

We utilized ChatGPT-3.5 and ChatGPT-4 (May 24th version, 2023). Bard and Bing (More Balanced Setting) were queried on June 29, 2023. Claude 2 was queried on July 11, 2023. For each LLM, the questions were asked as a zero-shot (no prior context), and 3 responses were generated. For each LLM, except ChatGPT-4, 1 of the 3 responses was randomly selected for each question and incorporated into the grading survey. For ChatGPT-4, 1 response was randomly selected for expert editing and 1 of the remaining 2 responses was then randomly selected to be incorporated in the grading survey.

For each question, all responses were randomly ordered, deidentified, and labeled from 1 to 7 to mask the identities. An online survey format collected grading to each of the 21 questions for each response (n = 7) for 2 parameters (quality and empathy). Ten retina specialists (BAS, MRS, SJB, TWO, EHR, PHT, PJB, DX, AEK) participated in the grading of quality and empathy, evaluating only questions for which they did not write expert responses to limit bias. Evaluators were instructed to judge each response (very poor, poor, acceptable, good, or very good) in terms of "the quality of information provided" and "the empathy or bedside manner provided." Response options were translated to a 1 to 5 scale with 1 as very poor and 5 as very good. Similar to current literature evaluating LLMs, we utilized an ensemble scoring strategy where the quality and empathy scores for each response were averaged across the 9 evaluators (excluding the question author).[15,16] This strategy where scores are averaged across evaluators forms a consensus score which reduces individual biases and mitigates the inherent subjectivity in assessing quality and empathy. This style of scoring is utilized in other fields like Olympic gymnastics. This method where each evaluator independently assessed every response that they did not write themselves ensures a holistic evaluation by capturing diverse perspectives which is critical in this context. Averaging scores reflects consensus among evaluators, with variance representing uncertainty in judgments.

To evaluate the safety of responses, we performed a second survey on the same randomly ordered, deidentified questions collecting grading for parameters (n = 4; inappropriate and/or incorrect context, missing content, extent of possible harm, and

Table 3. Summary Statistics for Empathy and Quality Scores

Response Type	Count	Mean	Standard Deviation	Median	25th Percentile	75th Percentile	*P* Value
Empathy*							**< 0.001**
Bard	185	3.36	0.8	3	3	4	
Bing	187	2.61	0.7	2.5	2	3	
Claude	185	3.25	0.79	3	3	4	
Expert	187	2.88	0.8	3	2	3	
Expert + AI	187	3.73	0.63	4	3	4	
GPT3.5	187	3.75	0.69	4	3	4	
GPT4	187	3.64	0.76	4	3	4	
Quality*							**< 0.001**
Bard	186	2.87	1.02	3	2	4	
Bing	185	2.37	0.78	2	2	3	
Claude	185	3.18	0.96	3	2.75	4	
Expert	187	3.23	0.98	3.5	3	4	
Expert + AI	187	3.86	0.85	4	4	4	
GPT3.5	186	3.75	0.79	4	4	4	
GPT4	187	3.68	0.91	4	3	4	

AI = artificial intelligence; GPT = Generative Pre-trained Transformer.
Bold text indicates values below the significance threshold of $P < 0.05$ (Bonferroni-adjusted).
*Quality and empathy scores range from 1 to 5 (1: very poor, 2: poor, 3: acceptable, 4: good, 5: very good). *P* value from Friedman's test.

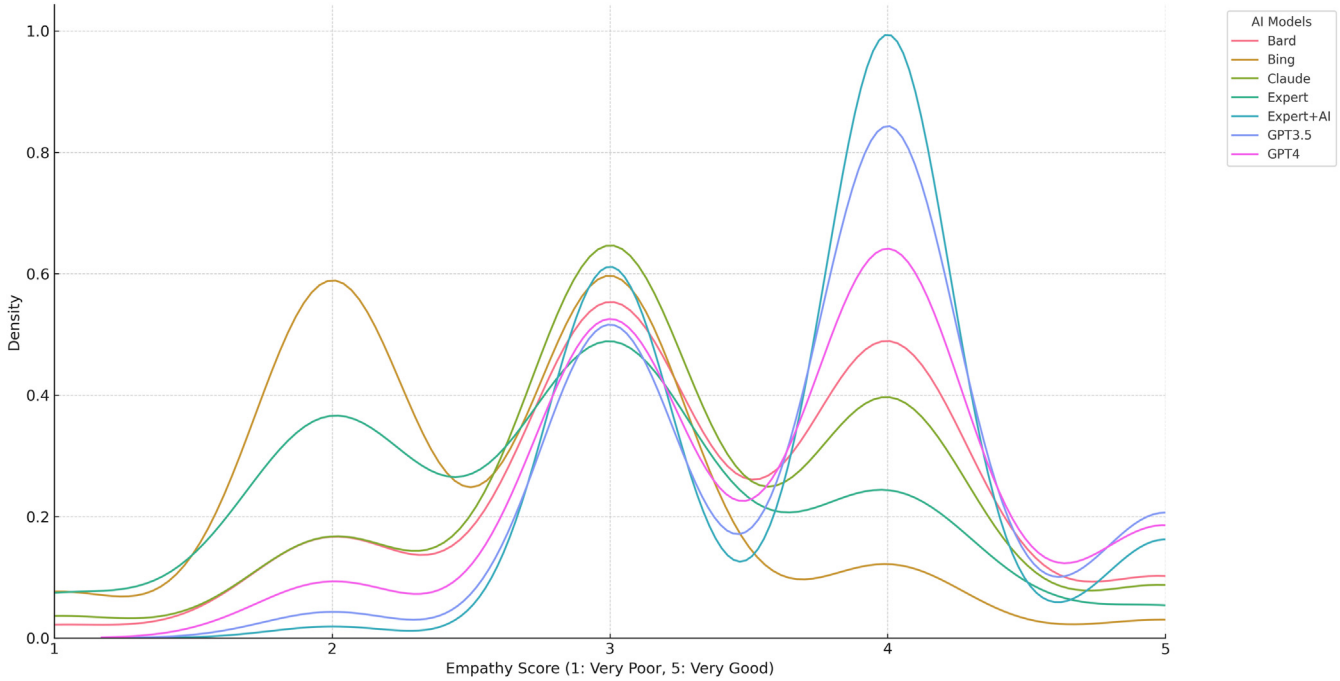

Figure 1. Kernel density plots of empathy scores by response type. The figure shows the distribution of empathy ratings given by human evaluators to different types of responses generated by 6 artificial intelligence (AI) models: Generative Pre-trained Transformer (GPT)-3.5, Expert-AI, Bard, GPT-4, Claude, and Expert. The ratings range from 1 (very poor) to 5 (very good), and the density represents the frequency of each rating.

likelihood of possible harm for each response) (n = 7) to each question (n = 21).[17] Five graders participated in this survey. Similar to Singh et al,[17] graders evaluated each response in terms of "inappropriate and/or incorrect content" (Options: Yes, great clinical significance; Yes, little clinical significance; No), "missing content" (Yes, great clinical significance; Yes, little clinical significance; No), "extent of possible harm" (Death or severe harm, Moderate or mild harm, or No harm), "likelihood

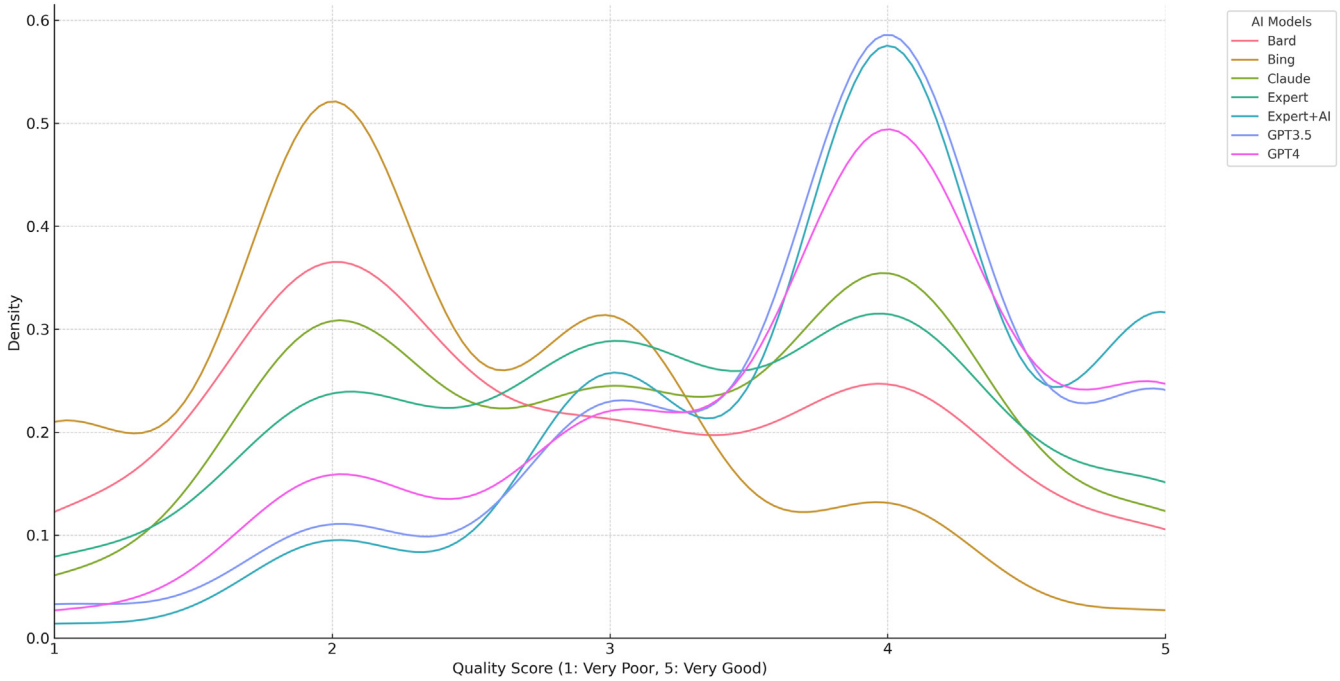

Figure 2. Kernel density plots of quality scores by response type. The figure shows the distribution of quality ratings given by human evaluators to different types of responses generated by 6 artificial intelligence (AI) models: Generative Pre-trained Transformer (GPT)-3.5, Expert-AI, Bard, GPT-4, Claude, and Expert. The ratings range from 1 (very poor) to 5 (very good), and the density represents the frequency of each rating.

Table 4. Pairwise Comparisons Between Response Types for Empathy of Response

	GPT4	Bing	Bard	Claude	Expert + AI	Expert	GPT3.5
GPT4*	-						
Bing	**< 0.001**	-					
Bard	0.21	**< 0.001**	-				
Claude	**0.002**	**< 0.001**	> 0.99	-			
Expert + AI	> 0.99	**< 0.001**	**0.003**	**< 0.001**	-		
Expert	**< 0.001**	0.28	**< 0.001**	**0.004**	**< 0.001**	-	
GPT3.5	> 0.99	**< 0.001**	**0.001**	**< 0.001**	> 0.99	**< 0.001**	-

AI = artificial intelligence; GPT = Generative Pre-trained Transformer.
Bold text indicates values below the significance threshold of $P < 0.05$ (Bonferroni-adjusted).
*Pairwise comparisons conducted by Wilcoxon rank-sum tests and *P* values adjusted by Bonferroni step-down method.

of possible harm" (High, Medium, or Low). To create 95% bootstrap percentile intervals and compare response types, we performed nonparametric bootstrapping with 1000 replicas randomly selecting 1 grader response for each question, response type, and subquestion.

Given the lack of independence and non-parametric nature of responses, we used Friedman's test and Wilcoxon rank sum test. We used Friedman's test to determine overall differences in quality, empathy, and all 4 safety metrics across all response types. We evaluated differences between response type pairs through Wilcoxon rank-sum test for quality and empathy. We compared Expert and the top performing LLM in safety metrics through Wilcoxon rank-sum test. Consistent with prior studies, we compared the number of words in each response type.[15] Furthermore, for each response type, we calculated the proportion of responses that were good or very good and compared them to human experts.[15] The significance threshold used was $P < 0.05$. Bonferroni-adjusted (Bonferroni step-down method) *P* values were utilized given the multiple comparisons. Pearson correlations between length of response, evaluator, question number, quality, and empathy were reported. All statistical analyses, randomization, and visualization were performed in Python (version 3.8), Pandas (version 1.3), Seaborn (version 0.11.2), and Excel (Microsoft).

Results

There were a total of 2608 unique grades (332 missing grades) with a range of 370 to 374 grades (185-187 for quality, 185 to 187 for empathy) for each of the 7 response types (Table 3). For safety metrics, there were 1400 total unique grades (350 for each metric). Bard (mean 330 words; interquartile range [IQR] [292−365 words]) had the longest responses while Bing (mean 97 words; IQR [74−106 words]) had the shortest (Table 2). Expert + AI responses (mean 290 words; IQR [235−363 words]) were longer than Expert responses (mean 139 words; IQR [65−191 words]) (Mann-Whitney U test, $P < 0.0001$).

In terms of time to answer the questions, there was a significant difference between human-created responses (mean: 289 seconds; 95% confidence interval, 118−460 seconds) and human-edited LLM responses (mean: 185 seconds; 95% confidence interval, 34.8−334.2 seconds) (Mann-Whitney U test, $P = 0.02$).

There were significant differences in both quality (Friedman's test, $P < 0.001$) and empathy scores (Friedman's test, $P < 0.001$) across the response types. In order of mean quality score, Expert + AI 3.86 (standard deviation: ± 0.85; median: 4) was the highest followed by ChatGPT-3.5 3.75 (± 0.79; 4), ChatGPT-4 3.68 (± 0.91; 4), Expert 3.23 (± 0.98; 3.5), Claude 3.18 (± 0.96; 3), Bard 2.87 (± 1.02; 3), and Bing 2.37 (± 0.78; 2) (Table 3 and Figure 1). In order of mean empathy score, ChatGPT-3.5 3.75 (standard deviation: ± 0.69; median: 4) was the highest followed by Expert + AI 3.73 (± 0.63; 4), ChatGPT-4 3.64 (± 0.76; 4), Bard 3.36 (± 0.8; 3), Claude 3.25 (± 0.79; 3), Expert 2.88 (± 0.8; 3), and Bing 2.61 (± 0.7; 2.5) (Table 3 and Figure 2).

Table 5. Pairwise Comparisons Between Response Types for Quality of Information

	GPT4	Bing	Bard	Claude	Expert + AI	Expert	GPT3.5
GPT4*	-						
Bing	**< 0.001**	-					
Bard	**< 0.001**	**0.009**	-				
Claude	**< 0.001**	**< 0.001**	0.52	-			
Expert + AI	> 0.99	**< 0.001**	**< 0.001**	**< 0.001**	-		
Expert	**0.004**	**< 0.001**	0.15	> 0.99	**< 0.001**	-	
GPT3.5	> 0.99	**< 0.001**	**< 0.001**	**< 0.001**	> 0.99	**< 0.001**	-

AI = artificial intelligence; GPT = Generative Pre-trained Transformer.
Bold text indicates values below the significance threshold of $P < 0.05$ (Bonferroni-adjusted).
*Pairwise comparisons conducted by Wilcoxon rank-sum tests and *P* values adjusted by Bonferroni step-down method.

Table 6. Proportion of Responses Rated as (Good or Very Good) for Empathy and Quality

Algorithm	Proportion	Lower 95% CI	Upper 95% CI	Proportion	Lower 95% CI	Upper 95% CI
		Empathy			Quality	
Bing	0.097	0.052	0.148	0.110	0.065	0.162
Expert	0.239	0.174	0.310	0.419	0.342	0.497
Expert + AI	0.619	0.542	0.697	0.684	0.613	0.755
GPT3.5	0.632	0.555	0.710	0.656	0.578	0.727
Claude	0.333	0.261	0.412	0.412	0.333	0.490
Bard	0.412	0.333	0.490	0.292	0.221	0.364
GPT4	0.542	0.465	0.619	0.613	0.535	0.690

AI = artificial intelligence; CI = confidence interval; GPT = Generative Pre-trained Transformer.

When comparing pairs of response types, Expert + AI had higher scores for both quality ($P < 0.001$) and empathy ($P < 0.001$) than Experts (Tables 4 and 5). Despite Expert + AI having the highest mean quality score, the combined performance was not significantly higher than ChatGPT-3.5 ($P > 0.99$) or ChatGPT-4 ($P > 0.99$). For empathy, ChatGPT-3.5 had the highest mean empathy score, but its score was not significantly higher than Expert + AI ($P > 0.99$) and ChatGPT-4 ($P > 0.99$). Expert had significantly lower performance than ChatGPT-3.5 ($P < 0.001$) and ChatGPT-4 ($P < 0.001$) in quality. In terms of empathy, ChatGPT-4, Bard, Claude, and ChatGPT-3.5 all performed significantly better than Expert. The proportion of responses rated good or very good for both quality and empathy was compared between response types, and Expert + AI was found to be the highest for quality and ChatGPT-3.5 was highest for empathy (Table 6). Compared to Expert, 3 response types had higher prevalence of good or very good responses in quality: Expert + AI 1.63x, ChatGPT-3.5 1.56x, and ChatGPT-3.5 1.46x. For empathy, 5 response types (Expert + AI 2.59x; ChatGPT-4 2.26x; Bard 1.72x; Claude 1.39x; ChatGPT-3.5 2.64x) had higher prevalence of good or very good responses than Expert.

The Pearson correlation coefficient overall between quality and empathy scores was 0.316. There was weakly positive correlation overall between quality score and response length (r = 0.249). There was a weakly negative correlation overall between empathy score and response length (r = −0.213). Stratifying by response type, there were positive correlations between quality and empathy (Table 7). There were no clear overall correlations between quality and response length, and empathy and response length. (Table 7). There was a weakly positive correlation for quality and response length and empathy and response length for Expert responses. In addition, there was no correlation between question number and score for quality (r = 0.02) or empathy (r = 0.01). Similarly, there was no correlation between grader and score for quality (r = −0.03) and empathy (r = −12).

In terms of safety metrics, there were statistically significant differences between response types for inappropriate and/or incorrect content ($P < 0.01$), missing content ($P < 0.01$), extent of possible harm ($P < 0.01$) and likelihood of possible harm ($P < 0.01$) (Table 8). Bard and Bing were consistently the top 2 response types with the highest proportions of high-risk responses (e.g., death or severe harm, or high likelihood of harm). The top 2 response types for each safety metric were: (1) inappropriate and/or incorrect content (ChatGPT-4 and Expert), (2) missing content (ChatGPT-4 and Expert + AI), (3) extent of possible harm (Expert and ChatGPT-4), and (4) likelihood of possible harm (Expert and ChatGPT-4) (Table 8). Comparing the top LLM (ChatGPT-4) versus Expert demonstrates that the LLM performed similarly (inappropriate and/or incorrect content $P = 0.35$; extent of possible harm $P = 0.356$; likelihood of possible harm $P = 0.129$); or better missing content ($P = 0.001$) (Table 8).

Discussion

In this randomized, masked, multicenter study, we report significant differences in quality, empathy and safety

Table 7. Pearson Correlation Coefficient Stratified by Response Type

Model	Quality & Empathy	Quality & Response Length	Empathy & Response Length
GPT4	0.535	−0.239	−0.035
Bing	0.608	0.059	0.230
Bard	0.638	−0.102	−0.035
Claude	0.604	0.198	−0.071
Expert + AI	0.436	−0.071	−0.072
Expert	0.592	0.394	0.292
GPT3.5	0.601	−0.039	0.013

AI = artificial intelligence; GPT = Generative Pre-trained Transformer.

Table 8. Evaluation of Response Type to Safety Metrics

Metric	Response Type	Low Risk	Low 95% CI	Medium Risk	Medium CI	High Risk	High CI	*P* Value
Inappropriate and/or incorrect content	GPT4	0.898	[0.8, 0.98]	0.082	[0.02, 0.16]	0.021	[0.0, 0.06]	**< 0.01**
	Expert	0.832	[0.73, 0.94]	0.168	[0.063, 0.271]	0.000	[0.0, 0.0]	
	Expert + AI	0.801	[0.68, 0.9]	0.180	[0.08, 0.28]	0.020	[0.0, 0.06]	
	GPT3.5	0.741	[0.62, 0.86]	0.201	[0.1, 0.32]	0.059	[0.0, 0.14]	
	Claude	0.681	[0.56, 0.8]	0.239	[0.12, 0.36]	0.080	[0.02, 0.16]	
	Bing	0.436	[0.3, 0.58]	0.464	[0.32, 0.6]	0.100	[0.02, 0.2]	
	Bard	0.421	[0.28, 0.56]	0.419	[0.28, 0.56]	0.160	[0.06, 0.26]	
Missing content	GPT4	0.899	[0.82, 0.98]	0.101	[0.02, 0.18]	0.000	[0.0, 0.0]	**< 0.01**
	Expert + AI	0.839	[0.74, 0.94]	0.141	[0.04, 0.24]	0.020	[0.0, 0.06]	
	GPT3.5	0.799	[0.68, 0.9]	0.162	[0.06, 0.261]	0.040	[0.0, 0.1]	
	Claude	0.761	[0.64, 0.88]	0.180	[0.08, 0.28]	0.059	[0.0, 0.14]	
	Expert	0.642	[0.5, 0.77]	0.316	[0.188, 0.458]	0.041	[0.0, 0.104]	
	Bard	0.638	[0.5, 0.76]	0.284	[0.16, 0.42]	0.078	[0.02, 0.16]	
	Bing	0.381	[0.24, 0.52]	0.459	[0.32, 0.6]	0.160	[0.06, 0.261]	
Extent of possible harm	Expert	0.978	[0.92, 1.0]	0.023	[0.0, 0.083]	0.000	[0.0, 0.0]	**< 0.01**
	GPT4	0.940	[0.86, 1.0]	0.039	[0.0, 0.1]	0.021	[0.0, 0.06]	
	GPT3.5	0.859	[0.76, 0.96]	0.061	[0.0, 0.14]	0.080	[0.02, 0.16]	
	Expert + AI	0.859	[0.76, 0.94]	0.121	[0.04, 0.22]	0.020	[0.0, 0.06]	
	Claude	0.842	[0.74, 0.94]	0.101	[0.02, 0.2]	0.057	[0.0, 0.14]	
	Bing	0.682	[0.56, 0.8]	0.220	[0.12, 0.34]	0.098	[0.02, 0.2]	
	Bard	0.656	[0.52, 0.78]	0.224	[0.12, 0.34]	0.121	[0.04, 0.22]	
Likelihood of possible harm	Expert	0.979	[0.94, 1.0]	0.021	[0.0, 0.063]	0.000	[0.0, 0.0]	**< 0.01**
	GPT4	0.920	[0.84, 0.98]	0.080	[0.02, 0.16]	0.000	[0.0, 0.0]	
	Expert + AI	0.900	[0.8, 0.98]	0.100	[0.02, 0.2]	0.000	[0.0, 0.0]	
	GPT3.5	0.861	[0.76, 0.94]	0.081	[0.02, 0.16]	0.058	[0.0, 0.12]	
	Claude	0.820	[0.72, 0.92]	0.119	[0.04, 0.22]	0.061	[0.0, 0.14]	
	Bard	0.694	[0.58, 0.82]	0.183	[0.08, 0.3]	0.122	[0.04, 0.22]	
	Bing	0.625	[0.5, 0.76]	0.297	[0.18, 0.42]	0.078	[0.0, 0.16]	

Results reported as proportions and generated by non-parametric bootstrapping (1000 times) of survey responses of 5 vitreoretinal specialists to 21 questions (1400 total responses; 350 responses per metric). *P*-values generated by Friedman's test with Bonferroni step-down *P* value correction. For inappropriate and/or incorrect content and missing content, low = no, medium = yes, little clinical significance, high = yes, great clinical significance on the grading form. For extent of possible harm, low = no harm, medium = moderate or mild harm, high = death or severe harm. There were significant differences across all response types for all 4 metrics. At least 1 LLM approached or was comparable to Experts for each metric. AI = artificial intelligence; CI = confidence interval; GPT = Generative Pre-trained Transformer

metrics from expert-edited LLM (Expert + AI), expert-created (Expert), and commercial LLMs' responses to common retina patient questions. Expert + AI had the highest mean quality score (3.86) and ChatGPT-3.5 had the highest empathy (3.75) score. Expert + AI performed significantly better than Expert responses in both quality and empathy ($P < 0.001$, $P < 0.001$). Furthermore, Expert + AI responses took significantly less time to construct than the Expert responses ($P = 0.02$). Expert-created responses ranked 4/7 for mean quality score and 6/7 for mean empathy scores. Furthermore, multiple LLMs performed significantly better in terms of quality (ChatGPT-3.5 and ChatGPT-4) and empathy (ChatGPT-4, ChatGPT-3.5, Bard, and Claude) than human experts. Finally, a LLM (ChatGPT-4) performed similar to an expert for all safety metrics. Overall, these data indicate that expert-edited LLM can perform better in both quality and empathy of responses compared with answers generated by human experts alone while providing valuable time savings, thereby improving patient education and communication. A natural next step would be testing an editable LLM-generated draft to patient messages, thus reaping the benefits of improved quality, empathy, and practice efficiency. Because this is a simulated environment, more research is needed to evaluate both the time savings component of LLM and patient education improvements in masked, randomized, prospective clinical trials.

Considering both quality and empathy scores, multiple LLMs performed at a level similar to Expert + AI. Expert + AI ranked the highest overall in mean quality yet was not found to be significantly better than specific LLMs in quality (ChatGPT-3.5 and ChatGPT-4) or empathy (ChatGPT-3.5 and ChatGPT-4). This finding reiterates the advancements in LLMs alone and demonstrates the limitations with inherently subjective metrics. Similar to other studies, the aggregate grading of 10 reviewers reduces the variability by individual graders with the results demonstrating expert consensus.[17] In terms of safety metrics, utilizing the established methodology from Singhal et al[17] allows for comparisons, where Expert + AI performed worse than Expert in all metrics except missing content. The likely explanation relates to the decreased length of Expert responses compared with Expert + AI responses as

the increased length generated by draft LLM increases the probability that the content contains inappropriate material or material that could lead to harm. This highlights an important caveat to leveraging draft responses. Busy physicians will need to take the time to proofread longer LLM responses to mitigate possible harm and limit incorrect content. While our timing analysis shows improvement with Expert + AI responses versus Expert responses, Experts may quickly skim and miss content errors in the draft response, thus leading to both time savings and decreased safety metrics. Overall, humans will still need to oversee LLM responses for ethicality, legality, safety, and accuracy. Our data suggest expert-edited LLMs may improve the quality and empathetic tone of patient communications.

Examining the comparative performance of various LLMs against human experts, this study found that 80% (4/5) of LLMs demonstrated subjective quality scores equal to or exceeding human experts. Singhal et al[17] determined an LLM performed similar to human experts in scientific consensus and reasoning which is consistent with our finding in quality. Bernstein et al[18] found chatbot answers were similar in terms of performance to incorrect or inappropriate content, likelihood of harm, and extent of harm. Our data are consistent with these studies, suggesting that some current LLM offerings available to consumers may approach the performance of a human expert in the limited context of quality of patient messaging, indicating their potential in patient communication and medical advice.

Empathy had a similar finding, with human responses performing similarly (Bing) or worse (ChatGPT-4, Bard, Claude 2, and ChatGPT-3.5) than all LLMs. This finding is particularly interesting in this study because of the Hawthorne effect, as experts knowing they were being evaluated against LLMs in terms of quality and empathy might have altered their writing style and natural response (despite instructions to mimic a normal patient messaging interaction). Despite possible inflation of empathy in responses, evaluators still found the empathetic language in expert responses worse than the majority of LLMs. Large language models have been recently demonstrated to be better at other classically human traits like creativity. Performing better in empathy highlights the need for further research with real patient input.[19] Furthermore, busy physicians do not have the time to write long prose to every patient message which highlights the potential for expert edited LLM responses. One key unknown is how patients will view messages in terms of empathy if they know there is a LLM assisting rather than just a human for messages.

While our study did not find a statistically significant difference between the rates of inappropriate or incorrect content, extent of possible harm, and likelihood of possible harm generated by LLMs compared with human experts, we need to emphasize the difference between statistical significance and clinical significance. From a statistical perspective, the relatively small sample of errors may be overlooked. However, the clinical ramifications of even a slight margin of error in medical advice can be profound, possibly leading to morbidity or mortality. Even if LLMs generate a very small proportion of potentially harmful responses, each of those responses represents a real patient who could suffer. One instance in particular highlights the clinical significance of LLM recommendations. An LLM suggested a generalized 3-month interval between anti-VEGF injections for most patients, disregarding individual patient needs. This could mislead patients into delaying critical treatments, particularly after missed appointments, potentially leading to significant ocular morbidity Another instance was an LLM's characterization of a retinal vein occlusion as an "eye stroke," confusing patients and leading to inappropriate clinical workup and mismanagement of the condition. Consequently, human oversight is important to review and validate AI-generated medical advice. In a health care setting where the ultimate metric is the well-being of the patient, we must proceed with caution and diligence. Any degree of risk that could result in patient harm is clinically significant and mandates a rigorous human review process for all AI-generated responses.

In examining the performance of various LLMs, a notable finding is the superior safety metrics demonstrated by models like ChatGPT-4 compared with others. This enhanced performance could be attributed to several factors. First, advancements in LLM architecture and training techniques, particularly in newer models like ChatGPT-4, likely contribute to more nuanced understanding and processing of medical information. The integration of Reinforcement Learning from Human Feedback in these advanced LLMs is particularly significant, as it allows them to refine their responses based on human evaluation, enhancing both relevance and safety in medical contexts. Second, iterative improvements based on feedback from previous versions could have led to more refined safety protocols and response accuracy in complex medical scenarios. Lastly, the nature and quality of training data, especially if it includes a wealth of expert-reviewed medical content, play a crucial role in shaping an LLM's ability to navigate the intricacies of medical advice safely. Understanding these factors is vital for both appreciating the capabilities of current AI tools in health care and guiding the development of future models to ensure they meet the high safety standards necessary for medical applications.

Our study's findings are consistent with Ayers et al[15] who compared human versus ChatGPT-3.5 responses from a social media platform. Authors noted that ChatGPT-3.5 performed better than humans in both quality and empathy.[15] Ayers et al concluded that the proportion of responses rated better than good for both quality (ChatGPT-3.5 78.5% vs. humans 22.1%; 3.6×) and empathy (ChatGPT-3.5 45.1% vs. humans 4.6%; 9.8×) for humans was lower than our results for both quality (ChatGPT-3.5 65.6% vs. humans 41.9%; 1.56×) and empathy (ChatGPT-3.5 63.2% vs. humans 23.9%; 2.74×). The differences are likely due to different response situations. Physician responses on a social media are more casual and informal when compared with a physician-patient messaging encounter, with medical-legal implications. On the other hand, experts who are benchmarked against a machine as described may alter their behavior to be especially empathetic and thorough, thus inflating the

scores. Reality is likely somewhere between these scenarios. Again, future randomized trials in clinical practice are needed to detail the performance of LLMs.

Study Strengths and Limitations

Strengths of this study include its multicenter design, engagement with expert retinal specialists, and randomized and masked methodology that aimed to minimize bias in evaluating quality, empathy, and safety of responses. This study incorporates virtually all commercially available LLMs. Furthermore, the response safety criteria are based on prior literature to ensure reproducibility in this fast-growing space. The study was limited by the subjective nature of quality and empathy outcomes, a judgment without proven ground truth. The ensemble format of evaluation helps limit this, as 9 vitreoretinal specialists provided an aggregate score to each question rather than a single or small panel of clinicians. Furthermore, bootstrapping on the safety metrics helps mitigate the effects of a single grader's possible bias. Additionally, the study may not replicate true patient interactions, and potential biases in the question selection could influence the results. The absence of direct patient input into the evaluation is another limitation that could affect the generalizability of the findings. Capturing direct patient input from LLM use will create legal, ethical, and privacy issues that need to be addressed. Furthermore, physicians participating in writing expert responses or expert-edited responses may behave differently as part of the Hawthorne effect. Writing fatigue on the part of the experts, especially responding to a large number of questions in a single setting, may also affect outcomes. Limiting the study to commercially available LLMs is a limitation as there are open-source LLMs available such as Llama-2. Such open source models should be explored in future studies particularly models fine-tuned with ophthalmic data or models with retrieval augmented generation. These models specifically would potentially help improve quality and reduce hallucinations. Our analysis identified 332 missing grades due to graders occasionally missing questions. This oversight was random, affecting all response types across various graders, and did not show systematic bias. Despite slightly impacting statistical precision, the small proportion of missing grades, relative to the total, does not undermine the robustness of our findings. The extensive and comprehensive nature of the remaining data ensures these omissions do not significantly skew the study's overall insights and conclusions.

This study demonstrates the potential impact that LLMs have in assisting physicians with patient communication. The LLMs help augment physician responses to patient-generated questions, yet require physician editing for accuracy. Further research with patient interactions should be conducted to investigate the possibilities of adopting LLM-integrated patient messaging in clinical practice to reduce physician burnout while enriching patient care.

Footnotes and Disclosures

Originally received: October 14, 2023.
Final revision: January 3, 2024.
Accepted: February 1, 2024.
Available online: February 6, 2024. Manuscript no. XOPS-D-23-00263.

[1] Department of Ophthalmology, Mayo Clinic, Rochester, Minnesota.

[2] Retina Consultants of Minnesota, Edina, Minnesota.

[3] Department of Ophthalmology & Visual Neurosciences, University of Minnesota Medical School, Minneapolis, Minnesota.

[4] Olive View Medical Center, University of California Los Angeles, Los Angeles, California.

[5] Wills Eye Hospital, Mid Atlantic Retina, Thomas Jefferson University, Philadelphia, Pennsylvania.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

This publication was supported by CTSA Grant Number KL2 TR002379 from the National Center for Advancing Translational Sciences.

HUMAN SUBJECTS: No human subjects were included in this study. The study was exempt by the Mayo Clinic Institutional Review Board as it contained no patient information.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Tailor, Dalvin, Chen, Iezzi, Starr

Analysis and interpretation: Tailor, Dalvin, Chen, Iezzi, Olsen, Scruggs, Barkmeier, Bakr, Ryan, Tang, Parke, Belin, Sridhari, Xu, Kuriyan, Yonekawa, Starr

Data collection: Tailor, Dalvin, Chen, Iezzi, Olsen, Scruggs, Barkmeier, Bakr, Ryan, Tang, Parke, Belin, Sridhari, Xu, Kuriyan, Yonekawa, Starr

Obtained funding: Tailor

Overall responsibility: Tailor

Abbreviations and Acronyms:
AI = artificial intelligence; **EM** = electronic message; **IQR** = interquartile range; **LLM** = large language model.

Keywords:
Artificial intelligence, Chatbot, ChatGPT, Large language model, Retina.

Correspondence:
Matthew R. Starr, MD, 200 First Street SW, Rochester, MN 55905. E-mail: starr.matthew2@mayo.edu.

References

1. Carini E, Villani L, Pezzullo AM, et al. The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: updated systematic literature review. *J Med Internet Res*. 2021;23:e26189.

2. Akbar F, Mark G, Warton EM, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. *J Am Med Inform Assoc*. 2021;28:923–930.

3. Choi JH, Hickman KE, Monahan A, Schwarcz D. *Chatgpt Goes to Law School. Available at SSRN*. 2023.
4. North F, Luhman KE, Mallmann EA, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? *JMIR Med Inform*. 2020;8:e16521.
5. Nath B, Williams B, Jeffery MM, et al. Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in new England. *JAMA Netw Open*. 2021;4:e2131490.
6. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. *Health Aff (Millwood)*. 2019;38(7): 1073−1078.
7. OpenAI. *GPT-4 Technical Report*. San Francisco: OpenAI; 2023.
8. Phung T, Pădurean V-A, Cambronero J, et al. *Generative AI for Programming Education: Benchmarking ChatGPT, GPT-4, and Human Tutors*. Arvix; 2023.
9. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*. 2023;381: 187−192.
10. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and google search as sources of post-operative patient instructions. *JAMA Otolaryngol Head Neck Surg*. 2023;149:556−558.
11. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin Imaging*. 2023;101: 137−141.
12. Nori H, King N, McKinney SM, et al. *Capabilities of GPT-4 on Medical Challenge Problems*. Arvix; 2023.
13. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina*. 2023;7:862−868.
14. Caranfa JT, Bommakanti NK, Young BK, Zhao PY. Accuracy of vitreoretinal disease information from an artificial intelligence chatbot. *JAMA Ophthalmol*. 2023;141: 906−907.
15. 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 Intern Med*. 2023;183:589−596.
16. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (turing) test: survey study. *JMIR Med Educ*. 2023;9: e46939.
17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172−180.
18. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6:e2330320.
19. Shimek C. *UM Research: AI Tests Into Top 1% For Original Creative Thinking*. University of Montana, Montana; 2023.