

ARTÍCULO DESTACADO DEL MES



The performance of ChatGPT-4 and Bing Chat in frequently asked questions about glaucoma

Doğan L, Yılmaz İE.



COMENTARIOS

El objetivo de este estudio fue evaluar la **idoneidad** y la **legibilidad** de las respuestas generadas por dos “chatbots” ante diversas preguntas frecuentemente planteadas sobre glaucoma.

Se diseñaron 34 preguntas, en consenso con profesionales clínicos, obtenidas a partir de las 10 páginas web no patrocinadas más populares de Google para la búsqueda "preguntas frecuentes sobre glaucoma". Estas preguntas abarcaban diversos aspectos de la patología: definición, tipos, factores de riesgo, prevalencia, impacto en la visión, prevención, recuperación visual, opciones de tratamiento médico y quirúrgico, posibles efectos secundarios de los fármacos, etc.

Las preguntas fueron planteadas en las plataformas disponibles en línea de dos conocidos “chatbots”, que integran modelos de lenguaje de gran tamaño (“large language models”, LLM): **ChatGPT-4** (OpenAI, California, EE. UU.) y **Bing Chat** (Microsoft Corp., Washington, EE. UU.). Se tuvieron en consideración las posibles variaciones en las respuestas debidas a la propia naturaleza de los LLM, por lo que cada pregunta se repitió 3 veces.

Cada conjunto de respuestas fue revisado exhaustivamente por dos especialistas en glaucoma independientes, quienes evaluaron y calificaron cada contestación según su experiencia clínica, clasificándolas como:

- "*Apropiada*": respuesta correcta que se ajustaba estrechamente a las recomendaciones que el revisor solía proporcionar a los pacientes.
- "*Inapropiada*": respuesta que se consideró inexacta o que se desviaba de las recomendaciones clínicas del revisor.
- "*Incompleta*": se definió como una respuesta que, siendo relevante y precisa, carecía de información suficiente para ser considerada completa.

Cuando las categorizaciones, según la decisión de los dos especialistas, eran las mismas para las 3 respuestas a la misma pregunta, se empleó esa evaluación como la categoría final de **idoneidad**. Sin embargo, cuando existía una discrepancia entre al menos 2 respuestas a la misma pregunta repetida, el conjunto de respuestas se consideraba *incoherente*.

Además, la **precisión** de las respuestas se evaluó mediante la taxonomía de la Estructura del Resultado de Aprendizaje Observado (SOLO) por un tercer experto. Esta estrategia comprende 5 niveles estructurales distintos, que categorizan los resultados de aprendizaje de acuerdo a su complejidad

creciente: “preestructural”, “uniestructural”, “multiestructural”, “relacional” y “abstracto extendido”.

El análisis de **legibilidad**, que cuantifica la facilidad con la que se puede leer un texto determinado, se realizó en la aplicación *Readable*. Se aplicaron cinco fórmulas que incluyen escalas que otorgan puntuaciones numéricas ordenadas en función de varios criterios: la facilidad de lectura para un adulto, un estudiante escolar, o un graduado universitario; el nivel educativo requerido para comprender un texto; el número de letras y oraciones; la densidad silábica; etc.

Resultados:

- **Idoneidad**: los resultados de la categorización de los dos revisores independientes mostraron una concordancia del 98%. ChatGPT-4 proporcionó respuestas *apropiadas* a las 3 preguntas repetidas en un 88% de casos, y Bing Chat en un 82% ($p > 0,05$). Ambos chatbots tuvieron solamente una respuesta *inapropiada*. La tasa de respuestas *incompletas* fue 6% y 12% para ChatGPT-4 y Bing Chat, respectivamente. Se identificaron respuestas *incoherentes* en dos de las preguntas en ambos chatbots.

- **Precisión**: según los resultados de la prueba SOLO, ambos LLM proporcionaron predominantemente respuestas que, en promedio, se acercaban a la categoría “relacional”, sin diferencias significativas entre los dos chatbots.

- **Legibilidad**: los niveles de legibilidad de los dos LLM no resultaron sencillos. Sin embargo, las respuestas de ChatGPT-4 fueron significativamente más largas y más difíciles/complejas en comparación con las generadas por Bing Chat.

Otros datos interesantes que aporta el trabajo:

- En sus respuestas, los chatbots suelen recomendar la confirmación por parte de un oftalmólogo.
- Las webs institucionales suelen tener información más completa que las páginas web estándar.
- Se aconseja a los usuarios que siempre evalúen críticamente y verifiquen de forma independiente la información obtenida de los chatbots.

Este trabajo presenta algunas limitaciones y posibles debilidades:

- No se explica en base a qué se eligieron las preguntas.
- Podría ser interesante leer alguna respuesta real aportada por los chatbots evaluados, para ejemplarizar los resultados, analizar diferencias e ilustrar el desempeño de ambos LLM.
- El sistema de calificación empleado fue totalmente subjetivo.
- Parece que existe alguna errata o confusión entre distintas versiones de ChatGPT, tanto en el resumen como en el cuerpo del manuscrito.

En conclusión, ChatGPT-4 y Bing Chat fueron capaces de proporcionar respuestas bastante **adecuadas** a las preguntas frecuentes sobre glaucoma seleccionadas, y lo hicieron de forma concordante con clínicos expertos. No obstante, a pesar de la alta **pertinencia** y **precisión** de dichas respuestas, la **legibilidad** (comprensión) podría ser reducida para una alta proporción de usuarios comunes y/o pacientes.

Comentario realizado por el **Dr. Ignacio Rodríguez Uña**.
Instituto Oftalmológico Fernández-Vega, Fundación de Investigación
Oftalmológica, Instituto de Investigación Sanitaria del Principado de Asturias
(ISPA). Oviedo.

ABSTRACT

Purpose

To evaluate the appropriateness and readability of the responses generated by ChatGPT-4 and Bing Chat to frequently asked questions about glaucoma.

Method

Thirty-four questions were generated for this study. Each question was directed three times to a fresh ChatGPT-4 and Bing Chat interface. The obtained responses were categorised by two glaucoma specialists in terms of their appropriateness. Accuracy of the responses was evaluated using the Structure of the Observed Learning Outcome (SOLO) taxonomy. Readability of the responses was assessed using Flesch Reading Ease (FRE), Flesch Kincaid Grade Level (FKGL), Coleman-Liau Index (CLI), Simple Measure of Gobbledygook (SMOG), and Gunning-Fog Index (GFI).

Results

The percentage of appropriate responses was 88.2% (30/34) and 79.2% (27/34) in ChatGPT-4 and Bing Chat, respectively. Both the ChatGPT-4 and Bing Chat interfaces provided at least one inappropriate response to 1 of the 34 questions. The SOLO test results for ChatGPT-3.5 and Bing Chat were 3.86 ± 0.41 and 3.70 ± 0.52 , respectively. No statistically significant difference in performance was observed between both LLMs ($p=0.101$). The mean count of words used when generating responses was $316.5 (\pm 85.1)$ and $61.6 (\pm 25.8)$ in ChatGPT-4 and Bing Chat, respectively ($p < 0.05$). According to FRE scores, the generated responses were suitable for only 4.5% and 33% of U.S. adults in ChatGPT-4 and Bing Chat, respectively ($p < 0.05$).

Conclusions

ChatGPT-4 and Bing Chat consistently provided appropriate responses to the questions. Both LLMs had low readability scores, but ChatGPT-4 provided more difficult responses in terms of readability.