Comparative Analysis of Prompted and Unprompted ChatGPT Responses to Frequently Asked Questions About Pediatric Glaucoma

Deniz Goodman, Alana L. Grajewski, Elena Bitrian, Eleonore J. Savatovsky, Ta Chen Peter Chang, on behalf of the Study Initiative on Generative AI Health Technologies: Pediatric Glaucoma Study Group

Bascom Palmer Eye Institute, University of Miami Health System | Samuel & Ethel Balkan International Pediatric Glaucoma Center

Background

- Large language models (LLM) such as ChatGPT:
  - Human-like responses to questions
  - Search engine functionality[1]
  - Problem: Unable to assess reliability of sources[1]

ChatGPT

Purpose

- Evaluate the accuracy and readability of ChatGPT in generating responses to parents' frequently asked questions (FAQ) about pediatric glaucoma as graded by clinical experts

Methods

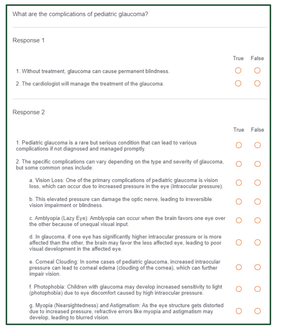
- 10 FAQs
- Question categories
  - General Pediatric Glaucoma Information
  - Diagnosis, Signs, and Symptoms
  - Pathophysiology
  - Prognosis and Treatment
- Website Sources
  - Glaucoma Research Foundation
  - Children's Hospital of Philadelphia
  - St. Louis Children's Hospital
  - American Association for Pediatric Ophthalmology and Strabismus



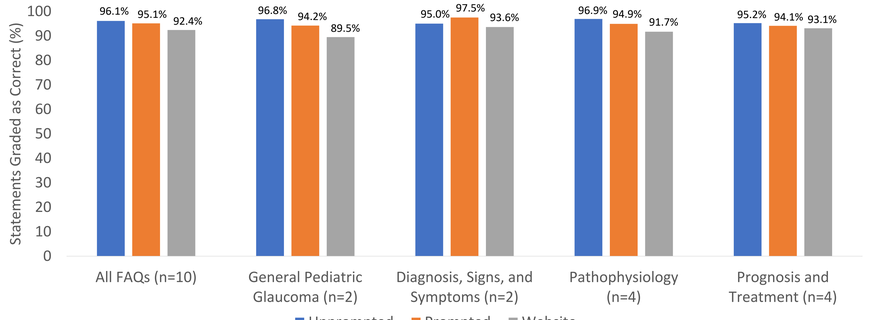
- Prompt: "Be specific and incorporate any applicable medical guidelines."[2]
- Response types: unprompted ChatGPT, prompted ChatGPT, website
- Questions and responses entered in a Qualtrics survey
- Reviewers: ophthalmologists working at tertiary hospitals with expertise in pediatric glaucoma as identified with online clinical profiles
- Readability: paired t-tests used to compare the Flesch-Kincaid Grade Level (FKGL) and Simple Measure of Gobbledygook (SMOG) index
- Chi-square tests to compare the proportion of statements graded as correct between sources

Pediatric Glaucoma FAQs and Survey

FAQ Number	FAQ
1	What is childhood glaucoma?
2	What causes childhood glaucoma?
3	How is childhood glaucoma treated?
4	What are the complications of pediatric glaucoma?
5	How does pediatric glaucoma affect the eye?
6	Is glaucoma in children hereditary?
7	What are the types and causes of glaucoma in children?
8	How is childhood glaucoma evaluated and diagnosed?
9	How does glaucoma affect your child's vision?
10	What are the signs and symptoms of pediatric glaucoma?



Accuracy Results



Readability Results

	FKGL*		SMOG Index**	
	Mean Grade Level	P-value	Mean Years of Education	P-value
Unprompted ChatGPT vs. Prompted ChatGPT	13.6 vs. 14.6	.02	15.4 vs. 16.2	.01
Unprompted ChatGPT vs. Website	13.6 vs. 11.2	.006	15.4 vs. 13.2	.009
Prompted ChatGPT vs. Website	14.6 vs. 11.2	<.001	16.2 vs. 13.2	.001

*FKGL: Reading grade level of a text[3]
**SMOG Index: Years of education required to understand a text[4]
Higher values indicate greater reading difficulty for both measures

Results

- Overall ChatGPT accuracy > 90%
- Proportion of correct statements among all FAQs:
  - Unprompted ChatGPT ≈ prompted ChatGPT > website responses
- Among all FAQs, the response type was significantly associated with the proportion of correct statements (p<.001)
- Readability
  - Website > ChatGPT
  - Unprompted ChatGPT > prompted ChatGPT

Discussion

- ChatGPT is generally more accurate but less readable than websites
- Websites may contain outdated information
  - Unknown author and publication/date last updated
- ChatGPT continually updated over time
  - Results may vary with future updates
- Process of editing website information per readability recommendations could have compromised accuracy
  - American Medical Association recommends ≤ 6th grade[5]

Conclusion

- Chatbots are fairly accurate overall
  - Possibly advantaged by continual updates
- Exercise caution when using online resources for medical information

ChatGPT ≠



Future Directions

- Evaluate the relevance and comprehensiveness of information provided by LLMs

You can reach me via email at dgm@bu.edu

References

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023 May 4;6:1169595. doi: 10.3389/frai.2023.1169595.
2. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. Published online February 28, 2023. doi:10.21203/rs.3.rs-2566942/v1
3. Flesch Reading Ease and the Flesch Kincaid Grade Level. Readable. Published July 9, 2021. https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/
4. The SMOG Index. Readable. Published July 9, 2021. Accessed April 28, 2024. https://readable.com/readability/smog-index/
5. Weiss BD. Health Literacy: A Manual for Clinicians. AMA Foundation http://lib.ncfh.org/pdfs/6617.pdf
6. https://blog.logomyway.com/chatgpt-logo-history/
7. Image by jcomp on Freepik. https://www.freepik.com/free-vector/thoughtful-woman-with-laptop-looking-big-question-mark_13330330.htm#query=question%20mark&position=1&from_view=search&track=ais&uuid=b17cdbb3-d602-4a32-9363-933008bdd4fe"
8. https://www.istockphoto.com/vector/a-senior-male-doctor-wearing-a-white-robe-in-their-cabinet-a-medical-checkup-gm1468861068-500263781?clickid=R5a2gKW3xyPT56xaIV8UxUnUkH2I8RVR2FmUg0&irgwc=1&cid=IS&utm_medium=affiliate_SP&utm_source=Freelmages&clickid=R5a2gKW3xyPT56xaIV8UxUnUkH2I8RVR2FmUg0&utm_term=doctor&utm_campaign=srp_freeillustrations_null-inline-popup&utm_content=270498&irpid=246195