The Use of Artificial Intelligence Software in Cornea Clinics

Prem A. H. Nichani[1], MD MSc, Stephan Ong Tone[1], MDCM PhD FRCSC, Sara M. AlShaker[1,] MD FRCSC DABO, Joshua C. Teichman[1], MD MPH FRCSC, Clara C. Chan[1], MD FRCSC DABO

[1] Department of Ophthalmology & Vision Sciences,
University of Toronto

Introduction: Artificial intelligence (AI), deep learning, and large language models (LLM) have gained significant attention for their potential applications in various domains, including streamlining efficiency, providing education, and advancing research. This study focuses on the use of LLMs in ophthalmology, particularly in managing cornea-related scenarios. Following the success of OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT) and the popularity of other platforms such as Writesonic, Google Gemini, and Bing Chat, this study aims to assess LLMs' ability to respond to prompts related to counseling, management, and advocacy for patients with corneal disease.

Methodology: Overarching topics and prompts were generated in collaboration with cornea specialists to identify areas where LLMs, namely ChatGPT, Writesonic, Google Bard, and Bing Chat, could streamline clinic efficiency. Using a rigorous scoring system, three independent cornea specialists, blinded to the LLM used to generate each response, graded the responses on accuracy, comprehension, compassion, professionalism, humanness, comprehensiveness of treatment options, and overall quality. Scores were equally weighted and averaged to generate means which were in turn ranked from highest to lowest score. Subgroup analyses were performed to identify the LLM which responded best to each prompt and based on each rubric criterion.

Results: Five categories of prompts were curated (clinic administration, patient counselling, treatment algorithms, surgical management, and research) under which a total of 11 prompts were constructed to produce 66 unique responses across LLMs. ChatGPT consistently outperformed other LLMs across various criteria, achieving an overall response score of approximately 83.8% versus Writesonic scoring 75%, Google Bard 62%, and Bing Chat 55.8%. Subgroup analyses further confirmed ChatGPT's superiority in responding to individual prompts and criteria compared to other LLMs. While ChatGPT's responses were highly scored, Bing Chat's responses included references to scientific literature, potentially enhancing credibility. No LLM-generated response was identified to pose a risk of harm to patients.

Conclusion: ChatGPT demonstrated a robust ability to respond to cornea-related prompts, outperforming other LLMs in terms of accuracy and comprehensiveness. The study emphasizes the potential of LLMs, particularly ChatGPT, in streamlining cornea-related clinical, administrative, and research tasks. Future research should involve patient feedback and repeated data collection to assess LLM-generated response improvement longitudinally. While LLMs show promise, caution is advised in their deployment, emphasizing the need for ongoing scrutiny by medical professionals to ensure patient safety while maximizing benefits and minimizing risks.