Evaluating commercial large language models in delivering HIV-related health education

Divyan Moodley

Mildmay Hospital

MACH

Background:

Patient health education is a critical component in supporting sustained engagement in HIV care and treatment. Increasingly, people living with HIV are turning to Generative Artificial Intelligence **(Gen-AI)** specifically large language models **(LLMs)** and other online tools to access health information.

Consequently, it is essential for clinicians, healthcare managers, and policymakers to understand the nature and quality of the information patients receive from these sources and to assess whether such information is **safe, accurate, and appropriate** for patient recommendation.

Aims:

The aims of this study are to **evaluate HIV-related information** provided by commercial LLMs and to determine whether any single model outperforms others.

Methods:

1. **Four** popular commercial LLMs were assessed:

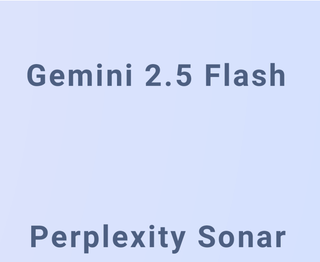
GPT-5 | Gemini 2.5 Flash | Claude Sonnet 4.5 | Perplexity Sonar

2. Each LLM was provided with the same 40-item questionnaire based on **commonly asked HIV-related questions** published by the UNAIDS. [1]

3. Model responses were **evaluated** using a combination of **human and LLM-based judges**.

4. **Human evaluation** was guided by clinically validated and standardised **DISCERN** [2], **EQIP** [3], and **PEMAT** [4] tools.

5. **LLM-as-a-judge evaluation** was based across **seven** dimensions:
   a. Understandability
   b. Actionability
   c. Reliability
   d. Readability
   e. Bias/Fairness
   f. Accuracy
   g. Overall Quality



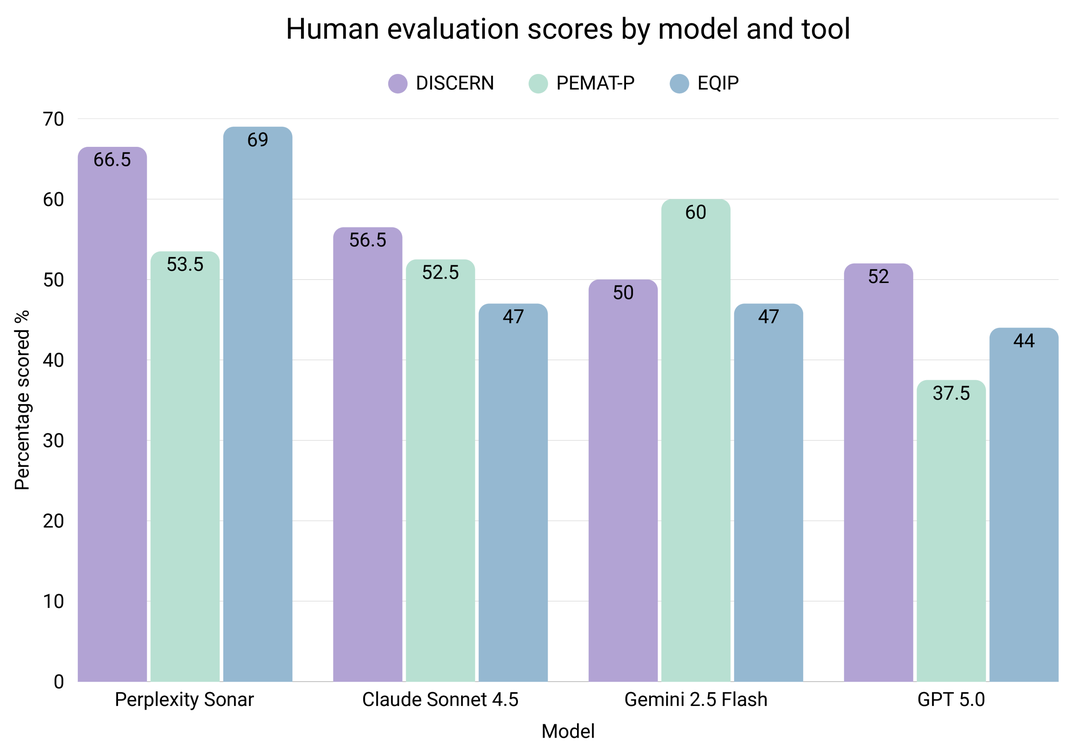
Results:

1. Human evaluation:
illustrated broadly comparable performance across models, with Perplexity and Gemini demonstrating marginally stronger results than Sonnet and GPT-5.

All four models produced highly accurate and unbiased HIV information but consistently lacked in providing visual aids, sufficient referencing, and explicit acknowledgement of uncertainty.

2. LLM-as-a-judge evaluation:
Five LLM judges (DeepSeek R1, Llama 3.3, Moonshot Kimi K2, Qwen 3, and Zai GLM 4.6) assessed the same set of responses. Metric-level ANOVA results showed consistently high model scores with p-values more than 0.05 across all seven dimensions. The internal consistency among LLM judges was acceptable, with a Cronbach's alpha of 0.7.

Overall, no singular model demonstrated statistically superior performance.



Conclusion:

Generative AI models (specifically LLMs) show tremendous potential as tools for HIV-related health education, producing accurate and unbiased information across a range of commonly asked questions. However, **no single commercial model outperformed the others**, and all four consistently fell short on referencing, visual support, and transparency around uncertainty.

Recommendations to LLM developers include **improved referencing, integration of visual content, and clearer communication of uncertainty** before these tools can be confidently advocated in a clinical setting.

References:

1. UNAIDS, "HIV and AIDS - Basic facts." [Online]. Available: https://www.unaids.org/en/frequently-asked-questions-about-hiv-and-aids. [Accessed: Nov. 01, 2025].
2. D. Charnock, S. Shepperd, G. Needham, and R. Gann, "DISCERN: an instrument for judging the quality of written consumer health information on treatment choices," J Epidemiol Community Health (1978), vol. 53, no. 2, p. 105, 1999.
3. B. Moult, L. S. Franck, and H. Brady, "Ensuring quality information for patients: Development and preliminary validation of a new instrument to improve the quality of written health care information," Health Expectations, vol. 7, no. 2, pp. 165–175, Jun. 2004.
4. E. Furukawa, T. Okuhara, M. Liu, H. Okada, and T. Kiuchi, "Evaluating Online and Offline Health Information With the Patient Education Materials Assessment Tool: Protocol for a Systematic Review," JMIR Res Protoc, vol. 14, p. e63489, 2025

Next Steps:

The gaps identified here (particularly around referencing, visual content, and acknowledging uncertainty) are informing the development of a targeted AI tool for clinical HIV education. A follow-up study will evaluate whether a domain-specific approach outperforms the general-purpose models tested here.

Scan the QR code to follow progress.