ASSESSING THE EFFICACY OF LARGE LANGUAGE MODELS IN RESPONDING TO QUERIES IN GASTROINTESTINAL PROCEDURES: A SYSTEMATIC REVIEW

Sebastian Zeki

UEG Week Berlin 2025 Poster

October 7, 2025

A systematic review of 14 studies evaluating seven large language models for gastrointestinal procedure queries found that while LLMs produce readable responses, none achieve the 95% accuracy threshold expected for clinical use and performance varies substantially across procedures and models.

KEY POINTS

- On binary accuracy scales Google Bard scored 82% and on three-point scales Claude scored 2.44 out of 3, while ChatGPT ranked first on five-point scales at 3.6 out of 5, though sample sizes and scoring methods varied widely across studies.
- Colonoscopy and endoscopy questions achieved accuracy above 80% in some assessments, whereas complex surgical interventions such as PEG placement and GERD management sometimes fell below 60%.
- Readability did not correlate with accuracy, and some models produced easy-to-read but factually inaccurate text, raising concern that patients may be misled by confident but incorrect information.
- All included studies displayed high risk of bias with heterogeneous scoring methods and variable question counts, and standardised benchmarking is needed before LLMs can be safely used in clinical gastroenterology.
- This review is relevant for gastroenterologists and informaticians evaluating the reliability of LLM-generated patient education and clinical decision support content.

VIEW THIS CONTENT ON GUTFLIX

https://gutflix.eu/search/d/d4a6f89c-a813-11f0-b633-0242ac140006

AI-generated summary

This summary was generated by an AI large language model based on the content transcript. It is for informational purposes only and should not be considered a substitute for clinical judgment. Always rely on your professional expertise and the full clinical context when making clinical decisions.