ABSTRACT

Title:
COLON-AI: Evaluating the Reliability and Readability of AI Chatbots in Colonoscopy Counseling
Authors:
Luigiandrea Antuofermo[1,2], Pasquale Apolito[1], Stefano Landi[1], Ramona Schiumerini[1], Rachele Librizzi[1,2], Andreea Roxana Curta[1,2], Enrico Mussomeli[1,2], Vincenzo Cennamo[1]
Affiliations:
[1] Gastroenterology and Interventional Endoscopy Unit, AUSL Bologna, Surgical Department, Bologna, Italy.
[2] Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy.
Background and aims:
Colonoscopy is the routine endoscopic procedure for evaluating the lower gastrointestinal tract. Providing patients with adequate pre-procedural information through leaflets and individual counseling increases the likelihood of achieving proper bowel preparation and reduces the frequency of repeated examinations. Among the innovative approaches to gastroenterology patient counseling, the potential role of artificial intelligence (AI)-based chatbots has recently emerged in the literature. However, only limited data are currently available regarding the accuracy and readability of health information provided by such tools. The aim of this study was to assess the performance of the most widely available AI chatbots, that is to say ChatGPT-5, Google Gemini and Meta AI, in terms of accuracy and readability when delivering colonoscopy instructions.
Methods:
We conducted a prospective, single-center observational cohort study (Unit of Gastroenterology and Interventional Endoscopy, AUSL Bologna), following the METRICS checklist. Frequently asked questions (FAQ) on colonoscopy were collected and submitted to Meta AI, ChatGPT, and Google Gemini. Responses were independently evaluated by five gastroenterologists using the CLEAR tool (Likert scale 1–5), assessing accuracy, completeness, clarity, evidence-based content and absence of irrelevant information. Text readability was analyzed using Readability Formulas® (FRES score). Data collection was performed between August and September 2025.
Results:
Across the three chatbots, Google Gemini achieved the highest overall performance, with mean scores for accuracy (3.14), completeness (3.12), clarity (3.13), evidence-based content (2.91), and relevance (2.99) outperforming ChatGPT and Meta AI. The most satisfactory results were observed in responses to pre-procedural questions compared to intra- and post-procedural ones, particularly regarding accuracy, completeness, and relevance. ChatGPT ranked second, with slightly lower yet consistently distributed values for accuracy (2.96), completeness (2.93), and clarity (2.95). Conversely, Meta AI showed the weakest performance, with significantly lower scores for accuracy (2.62), completeness (2.56), and clarity (2.62). Regarding readability (FRES score, 0–100), ChatGPT (13.68) generated relatively more understandable texts compared with Gemini (6.35) and Meta AI (4.21). A recurrent trend was observed across all models, whereby post-procedural questions yielded

more accessible answers, while intra-procedural queries produced the least satisfactory results. Overall, Gemini provided the best quality of responses, ChatGPT offered slightly lower performance but greater readability and Meta AI showed globally less favorable results.

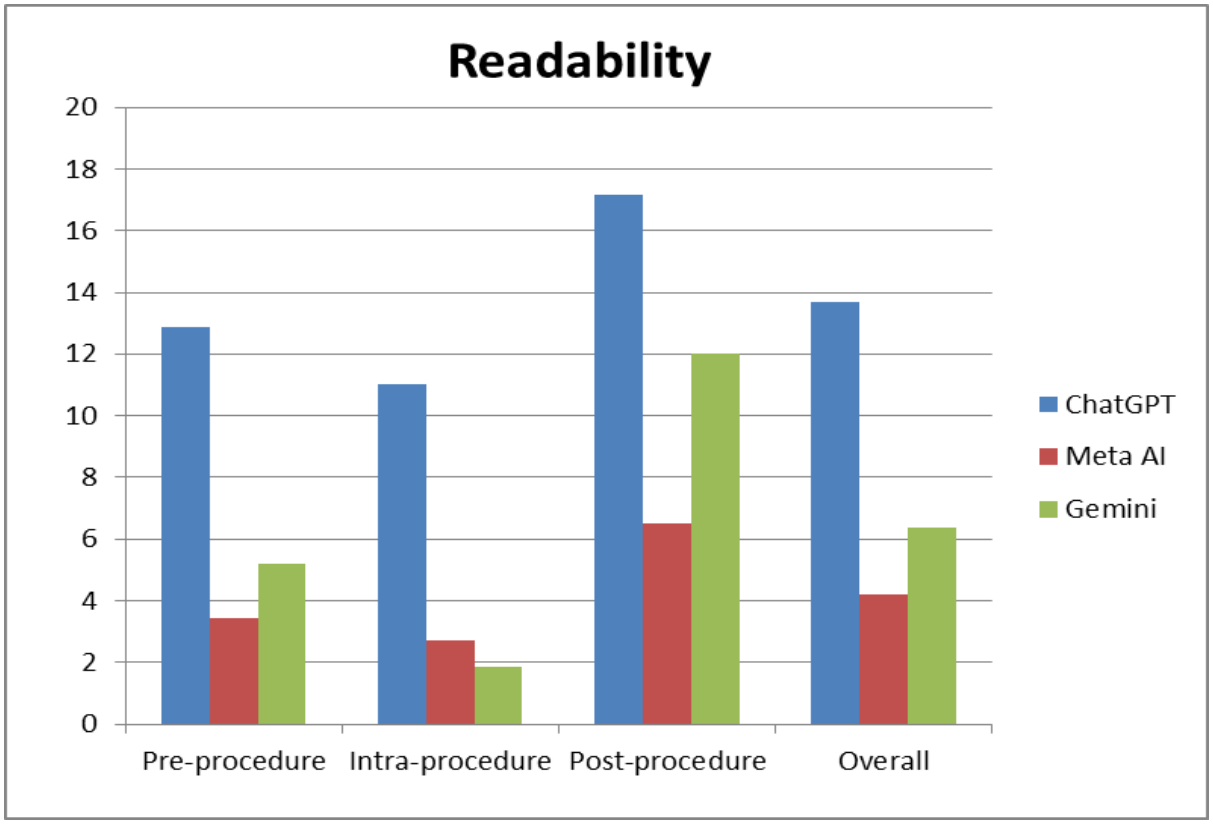
Conclusion:

This is the first single-center retrospective cohort study evaluating the performance of chatbots in pre-procedural medical counseling, providing patients with information that is accurate, complete, clear, relevant and accessible. Further multicenter studies involving patients are warranted to confirm the clinical applicability of these models.

Bibliography:

1) *Guo X, Yang Z, Zhao L, Leung F, Luo H, Kang X, Li X, Jia H, Yang S, Tao Q, Pan Y, Guo X. Enhanced instructions improve the quality of bowel preparation for colonoscopy: a meta-analysis of randomized controlled trials. Gastrointest Endosc. 2017 Jan;85(1):90-97.e6. doi: 10.1016/j. Gie .2016.05.012. Epub 2016 May 14. PMID: 27189659.*
2) *Donovan K, Manem N, Miller D, Yodice M, Kabbach G, Feustel P, Tadros M. The Impact of Patient Education Level on Split-Dose Colonoscopy Bowel Preparation for CRC Prevention. J Cancer Educ. 2022 Aug;37(4):1083-1088. doi: 10.1007/s13187-020-01923-x. Epub 2021 Jan 6. PMID: 33405208; PMCID: PMC7785930.*
3) *Giulio Calabrese, Roberta Maselli, Marcello Maida, Federico Barbaro, Rui Morais, Olga Maria Nardone, Emanuele Sinagra, Roberto Di Mitri, Sandro Sferrazza, Unveiling the effectiveness of Chat-GPT 4.0, an artificial intelligence conversational tool, for addressing common patient queries in gastrointestinal endoscopy, iGIE, 2025, ISSN 2949-7086, https://doi.org/10.1016/j.igie.2025.01.012. (https://www.sciencedirect.com/science/article/pii/S2949708625000123)*
4) *Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. Gastroenterology. 2023 Aug;165(2):509-511.e7. doi: 10.1053/j.gastro.2023.04.033. Epub 2023 May 5. PMID: 37150470.*
5) *Tariq R, Malik S, Khanna S. Evolving Landscape of Large Language Models: An Evaluation of ChatGPT and Bard in Answering Patient Queries on Colonoscopy. Gastroenterology. 2024 Jan;166(1):220-221. doi: 10.1053/j.gastro.2023.08.033. Epub 2023 Aug 26. PMID: 37634736.*
6) *Sallam M, Barakat M, Sallam M. A Preliminary Checklist (METRICS) to Standardize the Design and Reporting of Studies on Generative Artificial Intelligence-Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. Interact J Med Res. 2024 Feb 15;13:e54704. doi: 10.2196/54704. PMID: 38276872; PMCID: PMC10905357.*
7) *Sallam M, Barakat M, Sallam M. Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. Cureus. 2023 Nov 24;15(11):e49373. doi: 10.7759/cureus.49373. PMID: 38024074; PMCID: PMC10674084.*

		Pre-procedure	Intra-procedure	Post-procedure	Overall
ChatGPT	Accuracy	2,96	3,00	2,92	2,96
	Completeness	2,88	3,05	2,86	2,93
	Clarity	2,99	3,00	2,86	2,95
	Evidence-based content	2,79	2,79	2,83	2,80
	Relevance	2,97	2,95	2,86	2,93
Meta AI	Accuracy	2,49	2,67	2,69	2,62
	Completeness	2,37	2,62	2,69	2,56
	Clarity	2,49	2,67	2,69	2,62
	Evidence-based content	2,41	2,64	2,75	2,60
	Relevance	2,59	2,64	2,75	2,66
Gemini	Accuracy	3,22	3,12	3,08	3,14
	Completeness	3,18	3,17	3,03	3,12
	Clarity	3,18	3,19	3,03	3,13
	Evidence-based content	2,94	2,98	2,81	2,91
	Relevance	3,10	2,93	2,94	2,99


Readability
20
18
16
14
12
10
8
6
4
2
0
Pre-procedure
Intra-procedure
Post-procedure
Overall
ChatGPT
Meta AI
Gemini