Revista Española de Enfermedades Digestivas
The Spanish Journal of Gastroenterology

Title:

Evaluation of the clinical reasoning of GPT-4o, a multimodal generative artificial intelligence model, in 18 public gastroenterology case studies

Authors:
Alejandro García-Rudolph, Elena Hernández-Pena, Nuria del Cacho, Claudia Teixido-Font, Marc Navarro-Berenguel, Eloy Opisso

DOI: 10.17235/reed.2025.11369/2025

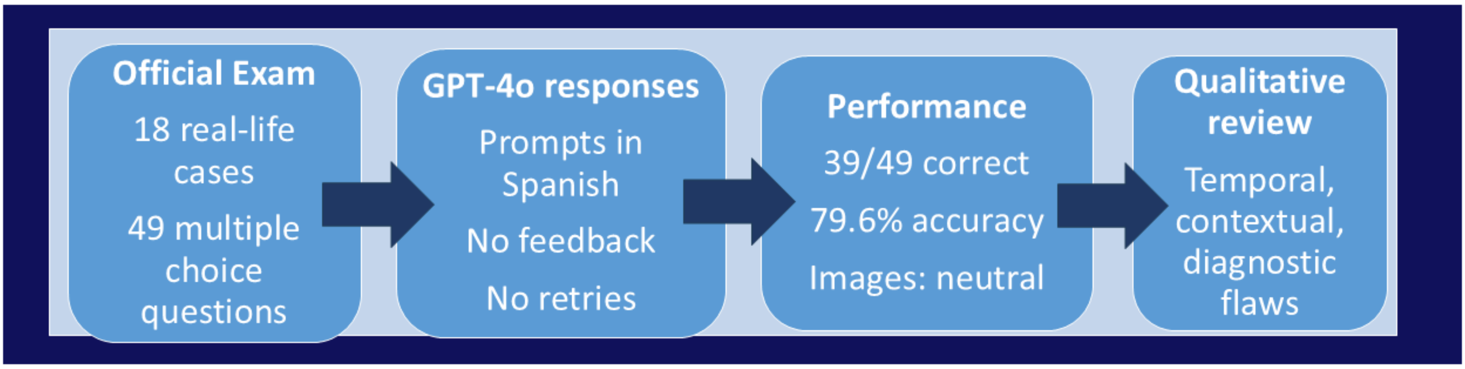
Link: PubMed (Epub ahead of print)

Please cite this article as:
García-Rudolph Alejandro, Hernández-Pena Elena, del Cacho Nuria, Teixido-Font Claudia, Navarro-Berenguel Marc, Opisso Eloy. Evaluation of the clinical reasoning of GPT-4o, a multimodal generative artificial intelligence model, in 18 public gastroenterology case studies. Rev Esp Enferm Dig 2025. doi: 10.17235/reed.2025.11369/2025.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Revista Española de Enfermedades Digestivas
The Spanish Journal of Gastroenterology

Evaluation of the clinical reasoning of GPT-4o, a multimodal generative artificial intelligence model, in 18 public gastroenterology case studies



Official Exam

18 real-life cases

49 multiple choice questions

GPT-4o responses

Prompts in Spanish

No feedback

No retries

Performance

39/49 correct

79.6% accuracy

Images: neutral

Qualitative review

Temporal, contextual, diagnostic flaws

80% **Correct answers by GPT-4o in the Spanish Digestive Medicine Board Exam (2023)**

20% **Errors in clinical reasoning: Therapeutic overgeneralizations, diagnostic or procedural confusion, contextual gaps, missed contraindications, failure to apply timing criteria.**

Garcia-Rudolph, et al.

Revista Española de Enfermedades Digestivas (REED)

The Spanish Journal of Gastroenterology

Revista Española de Enfermedades Digestivas
The Spanish Journal of Gastroenterology

Revista Española de Enfermedades Digestivas
The Spanish Journal of Gastroenterology

Evaluation of the clinical reasoning of GPT-4o, a multimodal generative artificial intelligence model, in 18 public gastroenterology case studies

Alejandro García-Rudolph, PhD[1,2,3*], Elena Hernandez-Pena, RN[1,2,3], Nuria del Cacho, RN[1,2,3], Claudia Teixidó-Font, MD[1,2,3], Marc Navarro-Berenguel, BSc[1,2,3], Eloy Opisso, PhD[1,2,3]

1-Departmento de Investigación e Innovación, Institut Guttmann, Institut Universitari de Neurorehabilitació adscrit a la UAB, Badalona, Barcelona, Spain;

2-Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain;

3-Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Barcelona, Spain

[*]Corresponding Author

Alejandro García-Rudolph,

Departmento de Investigación e Innovación, Institut Guttmann – Hospital de Neurorehabilitació,

Cami Can Ruti s/n 08916 – Badalona, Barcelona, Spain

Email: alejandropablogarcia@gmail.com

ORICD: 0000-0003-0853-8334

Conflict of interest declaration

None to declare.

Authors' contributions (CRediT)

AG-R: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Programming; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

EH-P: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing – review & editing.

NdC: Conceptualization; Data curation; Methodology; Validation; Writing – review & editing.

CT-F: Conceptualization; Data curation; Investigation; Methodology; Validation; Visualization; Writing – review & editing.

MN-B: Data curation; Programming; Formal analysis; Writing – review & editing.

EO: Conceptualization; Formal analysis; Project administration; Resources; Programming; Supervision; Writing – review & editing.

Abstract

Introduction and aim: Although generative language models have been extensively studied in the field of digestive diseases, further progress requires addressing underexplored aspects such as linguistic bias, the evaluation of clinical reasoning underlying model responses, and the use of realistic clinical material in non-English-speaking contexts. The aim of this study was to evaluate the accuracy of GPT-4o in answering clinical questions in Spanish and to qualitatively analyze its errors.

Methods: We used the most recent official board examination for Specialist in Gastroenterology (Spain, 2023), focusing on its practical section, which includes 18 real clinical cases described through text and images, totaling 50 multiple-choice questions (200 options in total). Forty-nine valid questions were analyzed, excluding one withdrawn by the organizing committee. GPT-4o answered 39 questions correctly (79.6%). No significant differences were observed between questions with clinical images (22/29 correct) and those without images (17/20 correct).

Results: Twenty percent of the answers were incorrect. In those cases, the model was prompted to provide its reasoning, which was then qualitatively analyzed by a team of experts. Errors were associated with inappropriate therapeutic generalizations, confusion regarding diagnostic or therapeutic sequencing, poor integration of contextual information, unawareness of contraindications, and omission of key temporal criteria in clinical decision-making.

Conclusions: Clinical images did not increase the error rate; however, the observed failures revealed that the model tends to omit information already provided (such as clinical context or temporal criteria), thereby compromising the quality of its reasoning.

Keywords: Artificial intelligence. Language models. ChatGPT. Clinical reasoning. Gastroenterology.

Lay abstract

Artificial intelligence systems that generate text, such as large language models, are being increasingly used in healthcare. They can help explain diseases or answer clinical questions, but their reliability—especially in languages other than English—is still uncertain.

This study evaluated how well one of the most advanced models, GPT-4o, performed on a real medical specialty exam in Digestive Diseases held in Spain in 2023. The exam included clinical cases based on real patients, described using both written information and diagnostic images, followed by multiple-choice questions similar to those faced by doctors in training. GPT-4o answered 80% of the questions correctly. There were no significant differences between questions with or without medical images. However, when analyzing the incorrect answers, the research team identified several types of errors that could be clinically relevant if the model were used without medical supervision. These included confusing diagnostic or treatment steps, failing to consider important clinical details, overlooking contraindications, or ignoring time-sensitive criteria.

The results suggest that, although these models may be helpful for medical learning or educational purposes, they are not yet reliable enough for use in clinical decision-making without professional oversight. The study also highlights the need to test these tools in different languages and using realistic clinical materials to better understand their limitations and safety risks in diverse healthcare settings.

Declaration on the use of artificial intelligence

ChatGPT-4o, a large language model developed by OpenAI, was used to assist in the drafting and editing of this manuscript. All AI-generated content was carefully reviewed and validated by the authors to ensure its accuracy and integrity.

Data availability statement

The data supporting the findings of this study are available from official public sources corresponding to the 2023 national board examination for Specialist in Gastroenterology. Direct links to the call and the official examination materials have been included in the References section. The complete set of GPT-4o responses to the 50 examination questions can be obtained from the corresponding author upon reasonable request.

1. Introduction

The growing interest in language models within gastroenterology has led to multiple systematic reviews. Early reviews highlighted their potential in health education and clinical communication, while also pointing out limitations such as inconsistent or inaccurate responses (1). More recent studies have extended their application to real clinical contexts, although challenges remain, including methodological heterogeneity and the lack of standardization (2). A more specific review, published in July 2024, assessed the clinical accuracy of these models in diagnostic and therapeutic tasks in gastroenterology (3). The authors emphasized the absence of methodological criteria and standardized evaluation metrics across the studies reviewed, which hinders comparability and limits the generalizability of findings.

Beyond gastroenterology-focused research, standardized medical examinations such as the USMLE, which include gastroenterology content, have been used to evaluate generative models. A recent review of 45 studies (4) reported that GPT-4 surpassed the passing threshold in most sections; however, performance varied widely in questions incorporating clinical images (ranging from 13.1% to 100%) and was consistently superior in English, highlighting a significant language bias (4). In parallel, models such as GPT-3.5 have been tested on frequently asked questions regarding endoscopic procedures (5), digestive diseases (6), and Helicobacter pylori infection (7). While showing some degree of accuracy and reproducibility, these studies primarily addressed educational content.

Despite the growing interest in generative models in gastroenterology, important methodological gaps remain. Many studies rely on patient-directed educational material, open-ended questions predominate—limiting objective performance evaluation—and incorrect responses are rarely analyzed in terms of reasoning. In addition, most of the questions used are not publicly available, as they are either subject to copyright restrictions or drawn from proprietary item banks. Furthermore, linguistic bias remains insufficiently explored, since the vast majority of evaluations are conducted in English, limiting the understanding of model performance in other linguistic and cultural contexts. Finally, the ability of these models to interpret clinical images—an essential skill in this specialty—has been scarcely investigated.

To address these underexplored aspects, the present study examines the performance of GPT-4o in a set of 50 single-answer multiple-choice questions derived from 18 real clinical cases included in the 2023 public board examination for Specialist in Gastroenterology organized by the Government of Spain (8). We employed official material, written in Spanish and freely accessible, thereby ensuring transparency and reproducibility. In addition to quantifying overall model accuracy, we conducted a qualitative analysis of clinical reasoning in incorrectly answered items to identify error patterns. Finally, given that a substantial proportion of the questions included clinical images, we also assessed the model's ability to operate in multimodal settings.

2. Methods.

2.1. Experimental design

This study was conducted in May 2025 at the Department of Research and Innovation, Institut Guttmann Hospital, Spain. The performance of the GPT-4o model was evaluated (the "o" refers to omni, reflecting its ability to process text and images in an integrated manner) (9). A ChatGPT Plus subscription account, providing full access to all model functionalities, was used. At the time of the study, GPT-4o was the most advanced publicly available version released by OpenAI. All responses were generated on site through the public chat interface, using predefined standardized instructions, and were recorded without any editing or modification of the original prompt.

2.2. Prompting

GPT-4o was accessed through OpenAI's official web interface in its standard end-user configuration. The API was not employed, as the study aimed to simulate real-world usage conditions similar to those typically encountered by patients or healthcare professionals when interacting with the model via the public platform. Default parameters were maintained, including a temperature setting of 1.0, with no further modifications.

The same prompt was applied consistently to all questions and was formulated as follows: *"You will be provided with a clinical case followed by one or more multiple-choice questions. Each question has four answer options (A, B, C, D), with only one correct answer. For each question, indicate only the letter corresponding to the correct option. Do not repeat the question stem or provide explanations. When a new clinical case is indicated, a new set of questions will begin."*

More than half of the 50 questions included clinical images, which were uploaded alongside the case description or question stem using GPT-4o's integrated image upload feature.

2.3. Multiple-choice questions

The questions used in this study were selected from the official examination for Specialist in Gastroenterology convened by the Government of Spain in 2023 (10). This examination, prepared and graded by an evaluation committee appointed by the Andalusian Health Service, forms part of the open competitive selection process regulated by the Resolution of December 22, 2022, within the framework of the Public Employment Offer for the stabilization of temporary positions in public healthcare centers. The composition of the qualifying committee, made public through an official resolution, included 10 clinical experts in Gastroenterology with recognized healthcare and academic experience. This institutional framework reinforces the validity of the instrument as a representative and up-to-date model for assessing specialized clinical competencies.

The content of the examination is based on an extensive official syllabus covering all fundamental areas of the specialty. Correct answers were published together with the official examination booklet, and the set is considered a standardized reference comparable to medical specialization examinations in other European Union countries.

In total, 50 questions distributed across 18 clinical cases were analyzed. All questions were independently reviewed by two authors of the study (AG-R and MN-B) and were presented to the model individually, one at a time, along with their corresponding answer options. The responses generated by GPT-4o were systematically recorded in a spreadsheet for subsequent analysis, following common methodological practices in studies evaluating automated clinical reasoning (11).

The complete set of cases, questions, correct answers, and the official booklet with the associated clinical images are publicly available on the corresponding municipal government website (8).

2.4. Analysis of reasoning in incorrect answers.

For questions answered incorrectly, GPT-4o was prompted to provide a structured explanation of its reasoning using a standardized instruction applied consistently across cases. Prior to this request, the model was re-exposed to all relevant material: the clinical description, the original question with its four options, the correct answer, and the option chosen by the model. The following prompt was then used:

"This was your answer: [B]. The correct answer is: [C]. Can you explain why you chose option [B]? Explain your reasoning step by step as if you were a physician reasoning through the clinical case."

During this phase, the model did not receive any feedback, corrections, or additional guidance from the researchers regarding the quality or validity of its explanations. It was allowed to reason freely based solely on the information previously provided.

2.5. Expert evaluation and qualitative analysis of reasoning.

The qualitative analysis of the explanations generated by GPT-4o for incorrect answers was conducted by a team of three professionals with clinical and teaching experience in digestive diseases. Based on the textual content provided by the model, each explanation was independently reviewed by two evaluators, and any discrepancies were resolved by consensus with the participation of the third reviewer.

For each incorrectly answered question, four elements were documented: the clinical topic involved, the type of error made by the model (e.g., incorrect or incomplete clinical reasoning), a synthesis of the reasoning expressed in the generated explanation, and a probable cause of the error, interpreted from the semantic and clinical content of the response.

This analysis enabled the construction of a summary table (see Table 1), in which each row corresponds to a question incorrectly answered by GPT-4o.

2.6. Data analysis

The quantitative analysis first involved calculating the overall proportion of correct answers, from which the incorrectly answered questions were identified for subsequent qualitative evaluation. In addition, the performance of GPT-4o was compared between questions that included clinical images and those that did not. For this comparison, Pearson’s chi-square test was applied. In cases where expected counts were low, Fisher’s exact test was used. The threshold for statistical significance was set at $p < 0.05$. All analyses were performed using R statistical software.

The qualitative analysis of reasoning in incorrect answers is described in detail in Section 2.5 and includes expert coding of the content generated by the model (see Table 1).

3. Results

Of the 50 multiple-choice questions from the practical section of the official examination, one was excluded as it had been withdrawn by the organizing committee. GPT-4o answered 39 questions correctly (79.6%).

Of the 49 questions analyzed, 29 (59.2%) included an associated clinical image, either in the initial case description or in the specific question. The remaining 20 (40.8%) did not contain any visual content.

GPT-4o correctly answered 75.9% of the questions with images (22/29) and 85.0% of those without images (17/20). The difference in accuracy between the two groups was not statistically significant according to Fisher's exact test (p = 0.496). These results do not support the existence of performance differences based on the presence of clinical images, although the limited sample size warrants cautious interpretation.

A total of 10 questions were answered incorrectly by GPT-4o. For each of these, the model was asked to provide its clinical reasoning. The qualitative analysis of these errors (Table 1) revealed several clinically relevant shortcomings. The most frequent errors included inappropriate therapeutic generalizations (e.g., Question 108), confusion in diagnostic or therapeutic sequencing (Questions 120, 125), and failures in integrating the clinical context (Questions 126, 138). In addition, omissions of important contraindications (Question 135) and of key temporal aspects in clinical decision-making (Question 137) were observed.

[Table 1]

4. Discussion.

The results of our study can be contextualized in light of recent systematic reviews on the performance of generative models in gastroenterology. A review published in July 2024 reported wide variability in the accuracy of ChatGPT (3): ranging from 6.4% to 45.5% with GPT-3.5, and from 40% to 91.4% with GPT-4. This dispersion reflects the lack of methodological criteria and standardized metrics, which hinders comparability across studies. In this context, the performance observed in our study (79.6%) is consistent with the highest reported values and underscores the importance of employing real clinical cases with objective, unadapted questions.

Nevertheless, some of the errors identified were clinically relevant and even potentially dangerous, raising concerns from a patient safety perspective. These findings reinforce the notion that generative models should be used solely as supportive tools under professional supervision, and not as autonomous systems for clinical decision-making.

The lack of linguistic diversity in the existing literature restricts the generalizability of findings and may conceal important limitations in non-English-speaking settings. Moreover, the considerable variability observed even among studies conducted in English (1–4) suggests that results depend not only on language, but also on task type, content, and methodological design. Nevertheless, this variability also underscores the need to evaluate models in other languages under controlled conditions.

The consistency of responses has been investigated in previous studies. Balta et al. (12) reported high variability when repeating questions multiple times in the field of critical care, with consistency rates of 40% in GPT-4. In contrast, Suárez et al. (13) evaluated endodontics questions with GPT-4 and found high consistency (85%) across runs. These findings suggest that model stability may depend on the type of clinical content and the study design. In our case, we simulated a typical clinical-use scenario, in which a single query is submitted to the model and a direct answer is expected. Although each question was not repeated multiple times, we performed a second complete administration of the examination two months after the first, in an independent session, and GPT-4o produced identical responses for all 49 items. This suggests greater stability in this version of the model and strengthens the validity of the results reported, in contrast with the high variability described in previous studies (12).

Despite the relevance of the findings, this study has several limitations. First, a single brief and uniform prompt was employed, without exploring variations in its formulation. Although this approach allowed conditions to be standardized, the analysis of more complex or ambiguous prompts would require a specific study design. Research focused on this issue should consider readability indicators (such as Flesch–Kincaid (14,15)) and other linguistic variables (length, ambiguity, level of detail), which were beyond the scope of the present work.

Second, the analysis was based on a single official examination, which limits the generalizability of the results. Nevertheless, it represents the most recent publicly accessible examination and is a representative benchmark of the knowledge required at both the national and European levels. Moreover, the study specifically targeted the practical section, which includes real clinical cases and multimodal resources, particularly useful for assessing clinical reasoning.

Finally, aspects related to the adaptation of language to different user profiles (16) were not addressed. Future studies could incorporate these communicative dimensions.

5. Conclusions

In this study, the performance of GPT-4o was evaluated on 18 real clinical cases, demonstrating a high overall accuracy but also clinically relevant errors in 20% of responses. The qualitative analysis revealed reasoning failures that, although plausible in form, were incorrect in substance. Studies conducted in languages other than English, such as this one in Spanish, contribute to advancing the understanding and mitigation of linguistic bias that remains present in generative models.

References

1. Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. Therap Adv Gastroenterol. 2023 Dec 25;16:17562848231218618. doi: 10.1177/17562848231218618.
2. Giuffrè M, Kresevic S, You K, Dupont J, Huebner J, Grimshaw AA, Shung DL. Systematic review: The use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther. 2024 Jul;60(2):144-166. doi: 10.1111/apt.18058.
3. Gong EJ, Bang CS, Lee JJ, Park J, Kim E, Kim S, Kimm M, Choi SH. Large Language Models in Gastroenterology: Systematic Review. J Med Internet Res. 2024 Dec 20;26:e66648. doi: 10.2196/66648.
4. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024 Jul 25;26:e60807. doi: 10.2196/60807.
5. Ali, Hassam, Pratik Patel, Itegbemie Obaitan, Babu P. Mohan, Amir Humza Sohail, Lucia Smith-Martinez, Karrisa Lambert, Manesh Kumar Gangwani, Jeffrey J. Easler, and Douglas G. Adler. "Evaluating the performance of ChatGPT in responding to questions about endoscopic procedures for patients." iGIE 2, no. 4 (2023): 553-559.
6. Kerbage A, Kassab J, El Dahdah J, Burke CA, Achkar JP, Rouphael C. Accuracy of ChatGPT in Common Gastrointestinal Diseases: Impact for Patients and Providers. Clin Gastroenterol Hepatol. 2024 Jun;22(6):1323-1325.e3. doi: 10.1016/j.cgh.2023.11.008.

7. Lai Y, Liao F, Zhao J, Zhu C, Hu Y, Li Z. Exploring the capacities of ChatGPT: A comprehensive evaluation of its accuracy and repeatability in addressing helicobacter pylori-related queries. Helicobacter. 2024 May-Jun;29(3):e13078. doi: 10.1111/hel.13078.

8. Andalusian Health Service. Examination booklet. Competitive examination 2023 for vacant staff positions: Specialist in Gastroenterology (FEA) https://www.sspa.juntadeandalucia.es/servicioandaluzdesalud/profesionales/ofertas-de-empleo/oferta-de-empleo-publico-puestos-base/oep-extraordinaria-decreto-ley-122022-centros-sas/cuadro-de-evolucion-concurso-oposicion-centros-sas/fea-aparato-digestivo (Accessed 21/07/2025)

9. OpenAI. GPT-4o Technical Report. OpenAI; 2025. Available at: https://openai.com/index/gpt-4o (Accessed 28/05/2025)

10. Andalusian Health Service. Examination booklet. Competitive examination 2023 for vacant staff positions: Specialist in Gastroenterology (FEA)-Exam https://www.sspa.juntadeandalucia.es/servicioandaluzdesalud/sites/default/files/sincfiles/wsas-media-ope_fichero/2023/revisado_56007_fea_aparato_digestivo_final.pdf (Accessed 21/07/2025)

11. Li DJ, Kao YC, Tsai SJ, Bai YM, Yeh TC, Chu CS, Hsu CW, Cheng SW, Hsu TW, Liang CS, Su KP. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. Psychiatry Clin Neurosci. 2024 Jun;78(6):347-352. doi: 10.1111/pcn.13656.

12. Balta KY, Javidan AP, Walser E, Arntfield R, Prager R. Evaluating the Appropriateness, Consistency, and Readability of ChatGPT in Critical Care Recommendations. J Intensive Care Med. 2024 Aug 8:8850666241267871. doi: 10.1177/08850666241267871.

13. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024 Jan;57(1):108-113. doi: 10.1111/iej.13985.

14. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, Mansour HA, Abishek RM, Xu D, Sridhar J, Yonekawa Y, Kuriyan AE. Appropriateness and Readability of ChatGPT-4-Generated Responses for Surgical Treatment of Retinal Diseases. Ophthalmol Retina. 2023 Oct;7(10):862-868. doi: 10.1016/j.oret.2023.05.022.

15. Gencer A. Readability analysis of ChatGPT's responses on lung cancer. Sci Rep. 2024 Jul 26;14(1):17234. doi: 10.1038/s41598-024-67293-2.

16. Fazilat, A.Z., Brenac, C., Kawamoto-Duran, D. et al. Evaluating the quality and readability of ChatGPT-generated patient-facing medical information in rhinology. Eur Arch Otorhinolaryngol 282, 1911–1920 (2025). https://doi.org/10.1007/s00405-024-09180-0

Table 1. Qualitative analysis of the 10 questions answered incorrectly

Question Id	Clinical topic	Type of error	Reasoning synthesis	Probable cause of error
108	Esophageal motility disorder	Incorrect clinical reasoning	Assumed similar efficacy to typical achalasia with generalized benefit	Failure to distinguish specific symptomatic response; therapeutic generalization
111	Hilar obstructive cholangitis	Incomplete clinical reasoning	Preferred plastic stent drainage in the setting of diagnostic uncertainty without considering the high suspicion of malignancy	Did not adjust the indication to the hilar location or the clear oncological suspicion
120	Primary biliary cholangitis	Incorrect therapeutic sequence	Considered obeticholic acid appropriate without first initiating ursodeoxycholic acid	Confused first-line and second-line therapy in PBC
122	Primary biliary cholangitis	Confusion between histological finding and diagnostic criterion	Gave priority to the histological finding without considering that the current diagnosis is clinical–serological	Did not distinguish between morphological findings and current diagnostic criteria
125	Suspected pancreatic tumor	Incorrect diagnostic sequencing	For a suspicious pancreatic lesion, added CT angiography before confirming malignancy	Prematurely combined histological diagnosis with resectability assessment
126	Advanced pancreatic cancer	Lack of contextual integration	Assumed the case was resectable without recognizing that palliative management had already been indicated	Ignored prior findings that established the disease as unresectable
135	Active ulcerative colitis	Lack of awareness of procedural risks	Recommended full colonoscopy and standard treatment without considering the risk of perforation	Omitted the fact that full colonoscopy is contraindicated in severe acute phase
137	Fulminant	Premature	In the setting of	Ignored that

	ulcerative colitis	intervention	severe clinical deterioration, recommended immediate colectomy without awaiting disease progression	response to infliximab should be assessed at 72 hours, not earlier

		Premature intervention		
138	Mild ulcerative colitis with infection	Incorrect assessment of disease activity and inappropriate immunosuppression	Classified the colitis as moderate and proposed intensifying infliximab despite active infection	Overestimated clinical activity and failed to adjust treatment in the presence of an intercurrent infection
150	Suspected severe complication in celiac disease	Diagnostic over-interpretation	Interpreted TCRγ rearrangement as a definitive diagnosis of lymphoma	Failed to consider that ulcerative jejunoileitis may present with similar findings

PBC: Primary biliary cholangitis