Review

Educational Materials for *Helicobacter pylori* Infection: A Comparative Evaluation of Large Language Models Versus Human Experts

Giulia Ortu , Elettra Merola , Giovanni Mario Pes , Maria Pina Dore *

Posted Date: 16 October 2025

doi: 10.20944/preprints202510.1258.v1

Keywords: *Helicobacter pylori;* large language models (LLMs); patient education; gastroenterology; artificial intelligence in healthcare

 doi:10.20944/preprints202510.1258.v1

Review

Educational Materials for *Helicobacter pylori* Infection: A Comparative Evaluation of Large Language Models Versus Human Experts

Giulia Ortu [1], Elettra Merola [1], Giovanni Mario Pes [1] and Maria Pina Dore [1,2,*]

[1] Dipartimento di Medicina, Chirurgia e Farmacia, University of Sassari, Clinica Medica, Viale San Pietro 8, 07100 Sassari, Italy

[2] Department of Medicine, Baylor College of Medicine, One Baylor Plaza Blvd, Houston, TX 77030, USA

* Correspondence: mpdore@uniss.it; Tel.: +39-079-229886

Abstract

Helicobacter pylori infects about half of the global population and is a major cause of peptic ulcer disease and gastric cancer. Improving patient education can increase screening participation, enhance treatment adherence, and help reduce gastric cancer incidence. Recently, large language models (LLMs) such as ChatGPT, Gemini, and DeepSeek-R1 have been explored as tools for producing patient education materials, yet their performance compared to expert gastroenterologists remains under evaluation. This review analyzed seven peer-reviewed studies (2024-2025) assessing LLMs' ability to answer *H. pylori*-related questions or generate educational content, evaluated against physician- and patient-rated benchmarks across six domains: accuracy, completeness, readability, comprehension, safety, and user satisfaction. LLMs demonstrated high accuracy, up to 92% in some studies, comparable to or exceeding that of general gastroenterologists and approaching senior specialist levels. However, their responses were often judged incomplete, described as "correct but insufficient." Readability exceeded the recommended sixth-grade level, though comprehension remained acceptable. Occasional inaccuracies in treatment advice raised minor safety concerns. Experts and medical trainees rated LLM outputs positively, while patients found them less clear and helpful. Overall, LLMs show strong potential to provide accurate, scalable *H. pylori* education. Enhancing completeness, simplifying language, and ensuring clinical safety are key for their effective integration into gastroenterology patient education.

Keywords: *Helicobacter pylori*; large language models (LLMs); patient education; gastroenterology; artificial intelligence in healthcare

1. Introduction

Helicobacter pylori infection can lead to gastritis, peptic ulcer, mucosa-associated lymphoid tissue (MALT) lymphoma, and is a well-established risk factor for gastric adenocarcinoma [1]. Given that *H. pylori* contributed to approximately 4.8% of global cancer incidence in 2018 (excluding non-melanoma skin cancers, controlling this infection is a public health priority. Early detection and eradication of *H. pylori* have been shown to reduce the incidence of gastric cancer [2]. The success of eradication programs partly depends on public awareness and engagement. Unfortunately, knowledge of *H. pylori* in the general population is often poor, and individuals with low awareness are less likely to undergo testing or adhere to treatment. This gap highlights the importance of effective patient education in improving disease outcomes [3].

Patient education initiatives for *H. pylori* aim to convey information about transmission, risks, testing, and treatment clearly and understandably to encourage informed decision-making and adherence [4]. Traditional patient education materials (pamphlets, websites) require significant effort by experts to create content that is accurate, comprehensive, and pitched at the right literacy level [5].

 doi:10.20944/preprints202510.1258.v1

In recent years, advances in artificial intelligence (AI) have led to the development of new tools for generating health information. Large language models (LLMs) like OpenAI's ChatGPT can produce human-like texts and have been explored in various medical applications (e.g. drafting medical notes and assisting clinical decision support) [6]. In the field of gastroenterology, there is growing interest in utilizing LLMs to address patient questions and provide guidance on diseases such as inflammatory bowel disease and *H. pylori* infection. LLMs offer the potential for an on-demand, scalable approach to disseminate health information, but their reliability and quality must be rigorously evaluated before clinical integration [7].

Early studies evaluated LLM-generated information on *H. pylori*. They examined content quality on multiple fronts, including factual accuracy, completeness of information, readability for the average patient, and safety (absence of misleading or harmful advice). They also gauge how well patients understand the AI-provided information and how satisfied users are with the answers, compared to traditional expert-derived content. To date, the findings are mixed: while ChatGPT and similar models can produce mostly correct answers to *H. pylori* questions, concerns have been raised about occasional errors, omissions, or the use of complex language that exceeds the recommended reading level. Given the rapid adoption of LLMs by the public, clinicians and informaticians must understand the strengths and limitations of these tools in patient education.

This review provides a summary analysis of LLM-generated *H. pylori* educational content in comparison with content written by gastroenterologists. We focus on six key quality domains: (i) accuracy (correctness of information); (ii) completeness (breadth and depth of content); (iii) readability (reading level and ease of understanding); (iv) patient comprehension (how well target audiences understand the material); (v) safety (freedom from harmful or misleading information); and (vi) user satisfaction. By synthesizing results from recent studies, we aim to determine whether LLMs can serve as reliable patient education tools in *H. pylori* infection and identify any necessary improvements or safeguards to enhance their usefulness in clinical practice.

2. Methods

A literature review was conducted to identify studies evaluating LLM-generated educational content on *H. pylori* in comparison to human experts' (physicians) content. We searched for English-language, peer-reviewed articles published between 2023 and 2025 that assessed ChatGPT or other LLMs in answering patient-oriented *H. pylori* questions or creating patient education materials about *H. pylori*. Relevant studies were those that directly compared LLM outputs to content provided by clinicians (e.g., gastroenterologists) or established references, and that evaluated outcomes related to content quality or user experience. A total of seven studies met these criteria, comprising six original research articles and one letter (Table 1).

Table 1. Studies included in the review.

Study	Type of Study	AI Models Analyzed
"Assessing Accuracy of ChatGPT on Addressing Helicobacter pylori Infection-Related Questions: A National Survey and Comparative Study" (Hu et al., 2024)	National Survey and Comparative Study	ChatGPT3.5 and ChatGPT4
"Comparative analysis of large language models in medical counseling: A focus on Helicobacter pylori infection" (Kong et al., 2024)	Comparative Analysis	ChatGPT 4, ChatGPT 3.5, and ERNIE Bot 4.0
"Exploring the capacities of ChatGPT: A comprehensive evaluation of its accuracy and repeatability in addressing helicobacter pylori- related queries" (Lai et al., 2024)	Observational Study	ChatGPT-3.5

 doi:10.20944/preprints202510.1258.v1

"Artificial Intelligence- Generated Patient Education Materials for Helicobacter pylori Infection: A Comparative Analysis" (Zeng et al., 2024)	Comparative Analysis	Bing Copilot, Claude 3 Opus, Gemini Pro, ChatGPT-4, and ERNIE Bot 4.0
"Assessing the Capabilities of Novel Open-Source Artificial Intelligence—DeepSeek in Helicobacter pylori- Related Queries" (Du et al., 2025)	Letter to the Editor (Comparative Analysis)	DeepSeek (versions V3 and R1) and ChatGPT (versions 4o and o1)
"Potential application of ChatGPT in Helicobacter pylori disease relevant queries" (Gao et al., 2024)	Evaluation Study	ChatGPT-4
"Comparative evaluation of the accuracy and reliability of ChatGPT versions in providing information on Helicobacter pylori infection" (Ye et al., 2025)	Comparative Evaluation Study	ChatGPT-3.5, ChatGPT-4, and ChatGPT-4o

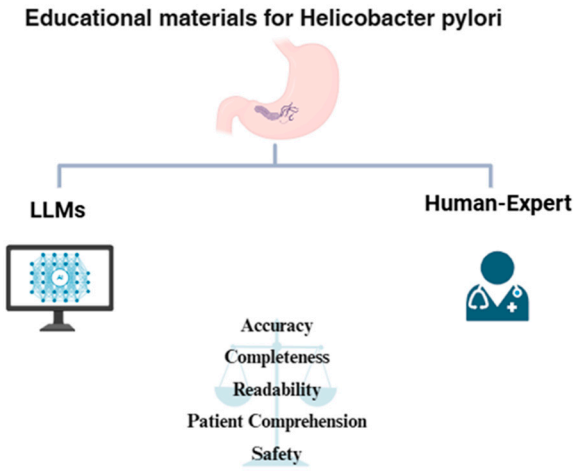

Figure 1. *Comparison framework for educational materials on Helicobacter pylori.* The figure illustrates the evaluation of patient educational content generated by large language models (LLMs) versus human gastroenterology experts.

These studies collectively examined multiple LLM platforms (including various versions of OpenAI's ChatGPT, such as GPT-3.5, GPT-4, GPT-4.5, GPT4o, and other models like ERNIE Bot and DeepSeek, Gemini, etc.). They often included multiple languages (typically English and Chinese). For each study, we extracted data pertaining to the six predefined domains of interest: accuracy, completeness, readability, patient comprehension, safety, and user satisfaction (for both patients and providers). Due to heterogeneity in study designs and measurement scales, a meta-analytic quantitative synthesis was not feasible; instead, results were integrated qualitatively. Key findings from each domain were compared and summarized, emphasizing consistent patterns or notable discrepancies between LLM- and expert-generated content. We ensured that the source data from these studies supported any interpretative claims. All data presented are reported as originally stated in the studies, with representative examples and statistical results cited to illustrate each point. No additional experimental data were generated for this review.

 doi:10.20944/preprints202510.1258.v1

3. Results

3.1. Accuracy

Across all studies, LLM-generated content on *H. pylori* demonstrated a high level of factual accuracy, often approaching or matching the performance of expert gastroenterologists (Table 2). In a national-based survey, ChatGPT (both GPT-3.5 and GPT-4 versions) answered *H. pylori*-related clinical questions correctly 92% of the time (median accuracy), outperforming the average accuracy of 1,279 gastroenterologists (≈ 80%) on the same questions. Notably, ChatGPT's accuracy was comparable to that of senior subspecialists for many topics. Several independent evaluations corroborated GPT-4 as the most accurate LLM [8]. For example, one study scoring answers on a 5-point scale found that ChatGPT-4o had the highest average accuracy score (~4.7/5), significantly above older versions [9]. Similarly, the Chinese LLM DeepSeek-R1, in a letter-based study by Du et al., achieved 95.2% accuracy, outperforming ChatGPT-4o and OpenAI-o1 [10]. LLM accuracy did vary by content area. Certain knowledge domains (e.g. basic facts, indications for testing) were handled quite accurately by ChatGPT, whereas nuanced clinical management questions were more challenging. In Lai et al, 61.9% of ChatGPT's answers were rated completely correct, and an additional 33.3% accurate but inadequate, leaving only ~5% of answers outright incorrect or containing errors. The few errors primarily involved outdated treatment recommendations or misinterpretation of complex scenarios [11]. In addition, variations were also observed according to the language chosen for the study. In one investigation analyzing three models (ChatGPT-4, ChatGPT-3.5, and ERNIE Bot 4.0) across English and Chinese tasks, both languages achieved satisfactory performance (~90%), yet a discrepancy was noted (91.1% in English vs. 88.9% in Chinese). Despite this difference, no statistically significant variation was found among the LLMs. Notably, ChatGPT-3.5 recommended serological testing for post-treatment follow-up, which is in discordance with current clinical guidelines [12]. Similarly, Zeng et al. assessed Patient Educational Materials (PEMs) generated in both English and Chinese by five LLMs (ChatGPT-4, ChatGPT-3.5, Claude 3 Opus, Gemini Pro, ERNIE Bot) and by a physician. All responses were considered acceptable across both LLMs and physicians, with the sole exception of the Chinese outputs produced by Claude 3 Opus [13]. Consistent with Kong et al., English-language outputs demonstrated overall superior performance [12]. Overall, these data indicate that current LLMs can deliver predominantly accurate *H. pylori* information, echoing findings that ChatGPT performs at or above the level of practicing clinicians on knowledge-based queries. However, "accuracy" here assumes static factual queries; dynamic clinical decision accuracy (e.g., choosing optimal therapy) remains an area for caution, as discussed under the safety section

Table 2. Evaluation of Accuracy. Comparison between educational materials on *Helicobacter pylori* generated by large language models (LLMs), such as ChatGPT, and those produced by human gastroenterology experts.

LLM Model	Study	Description of the Accuracy Metric (Scale/Threshold)	Mean score (SD)	Value/Score %
ChatGPT-3.5	Hu et al.	4-point Likert scale; (threshold ≥3)	3.44, average of responses from 3 attemps	92% Highest percentage of answers over 3 attempts
	Lai et al.	4-point Likert scale; (threshold ≥3)	Overall score 3.57 (0.13)	~95.23% (61.9% completely correct+ 33.33% correct but not complete)
	Kong et al.	6-point Likert scale; (threshold ≥4)	English 4.84 (1.07). Chinese 4.76 (0.86)	90% (overall) 91.1% English, 88.9% Chinese. Summation score of all LLMs
	Ye et al.	5-point Likert scale; (threshold ≥4)	3.94 (0.75)	Not reported

 doi:10.20944/preprints202510.1258.v1

ChatGPT-4	Hu et al.	4-point Likert scale; (threshold ≥3)	3.55 average of responses from three trials/questions	92% (Highest percentage of answers over 3 attempts
	Kong et al.	6-point Likert scale; (threshold ≥4)	English 4.87 (1.01) Chinese 4.84 (1.00)	90% (overall) 91.1% English, 88.9% Chinese. Summation score of all LLMs
	Zeng et al.	6-point Likert scale; (threshold ≥4)	English: 4.00 (1.00) Chinese: 4.40 (0.89)	
	Gao et al.	5-point Likert scale; (threshold ≥4)	Overall mean for experts: 4.58 (0.50)	
	Ye et al.	5-point Likert scale; (threshold ≥4)	4.14 (0.75)	
ChatGPT-4o	Du et al.	Average accuracy (objective) as percentage of correct answers over total		77.4%
	Ye et al.	5-point Likert scale; (threshold ≥4)	4.49 (0.74)	
DeepSeek-V3	Du et al.	Average accuracy (objective) as percentage of correct answers over total		90.4%
Claude 3 Opus	Zeng et al.	6-point Likert scale; (threshold ≥4)	English: 4.30 (1.30). Chinese: 3.40 (0.89)	
DeepSeek-R1	Du et al.	Average accuracy (objective) as percentage of correct answers over total		95.2%
OpenAI o1	Du et al.	Average accuracy (objective) as percentage of correct answers over total		87.0%
Bing Copilot	Zeng et al.	6-point Likert scale; (threshold ≥4)	English: 4.40 (0.89). Chinese: 4.20 (0.84)	
Gemini Pro	Zeng et al.	6-point Likert scale; (threshold ≥4)	English: 5.40 (0.55). Chinese: 5.00 (0.71)	
ERNIE Bot 4.0	Kong et al.	6-point Likert scale; (threshold ≥4)	English 5.07 (0.89). Chinese 4.42 (1.18)	90% (overall) 91.1% inglese, 88.9% cinese. Summation score of all LLMs
	Zeng et al.	6-point Likert scale; (threshold ≥4)	English: 5.20 (0.45). Chinese: 5.20 (0.45)	

3.2. *Completeness*

In contrast to accuracy, the completeness of LLM-generated educational content was consistently identified as a weakness (Table 3). Completeness refers to whether the content covers all relevant aspects of the topic with sufficient depth and context. Multiple studies have found that LLM responses often omit specific details or caveats that expert-written answers would typically include. Kong et al. observed that while 90% of LLM answers met the accuracy thresholds, only 45.6% were judged to be sufficiently complete (using a ≥ 2/3 Likert completeness criterion) [12]. Similarly, expert raters in Zeng et al. reported that LLM-generated patient education materials were frequently

 doi:10.20944/preprints202510.1258.v1

missing some content elements, rating most LLM outputs as "unsatisfactory" in completeness by gastroenterologist standards. For instance, in the Chinese-language materials, four of five AI-generated brochures had mean completeness scores <2 (on a 3-point scale) when evaluated by specialists, indicating that important information was lacking. Interestingly, the physician-written material was also not perfectly comprehensive in every case. In the English versions, completeness scores for the doctors' and AI's brochures were statistically comparable, suggesting that even experts might condense information. Patients tended to perceive the information as more complete than the experts did. In this study, lay patients received slightly higher completeness scores on average than gastroenterologists for the same AI outputs [13]. This discrepancy suggests that non-expert readers may not recognize the missing nuances. Nonetheless, from a clinical perspective, specific LLM answers were too superficial. Common gaps included a lack of detail on *H. pylori* transmission prevention, incomplete explanation of diagnostic steps, or insufficient emphasis on follow-up and antibiotic resistance issues [8,12,14]. In Gao et al., experts still gave ChatGPT-4 high completeness marks (~2.8/3), likely because the prompts in that study were designed around guideline topics [14]. Yet, even there, the consensus was that more exhaustive coverage would be beneficial. Overall, ensuring adequacy of content remains a challenge. LLMs may require more effective prompting or iterative querying to elicit all key information for patients. Improving completeness is critical because patients require not just correct facts, but a whole picture of their condition and care.

Table 3. Evaluation of Completeness. Comparison between educational materials on *Helicobacter pylori* generated by large language models (LLMs), such as ChatGPT, and those produced by human gastroenterology experts.

LLM Model	Study	Completeness Metric/Scale	Mean score (SD)	Value/Score %
ChatGPT-3.5	Kong et al.	3-point Likert scale (threshold ≥2);	Overall score: English 1.82 (0.78) Chinese 1.67 (0.77)	45.6%; English: 57.8% Chinese 33.3% (composite performance score of the three LLMs evaluated in the study)
	Lai et al.	Completeness assessed via 4-point accuracy scale (4= Comprehensive)	3.57 (0.13)	61.9%
	Ye et al.	Assessed via 5-point Likert accuracy. (threshold ≥4)	Overall score: 3.94	Not considered fully complete. Completeness mentioned but not quantitatively evaluated
	Hu et al.	Completeness mentioned but not quantitatively evaluated	Not reported	Not reported
ChatGPT-4	Kong et al.	3-point Likert scale (threshold ≥2);	Overall score: English 2.11 (0.68) Chinise 1.78 (0.74)	45.6%; English: 57.8% Chinse 33.3% (composite performance score of the three LLMs evaluated in the study)
	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists' rating: English 1.60 (0.55), Chinese 1.80 (0.45) Non-Expert rating: Chinese 2.44 ± 0.54	Not specified
	Gao et al.	3-point Likert scale; (threshold ≥2)	Overall score: 2.79 (0.41)	Not specified
	Ye et al.	Assessed via 5-point Likert accuracy. (threshold ≥4)	Overall score: 4.14	Considered complete Completeness mentioned but not quantitatively evaluated
	Hu et al.	Completeness mentioned but not quantitatively evaluated	Not reported	Not reported

 doi:10.20944/preprints202510.1258.v1

ChatGPT-4o	Ye et al.	Assessed via 5-point Likert accuracy. (threshold ≥4)	Overall score: 4.94	Considered complete Completeness mentioned but not quantitatively evaluated
Claude 3 Opus	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists' rating: English 1.80 (0.84), Chinese 1.20 (0.45) Non-Expert rating: Chinese 1.82 (0.72)	Not specified
Bing Copilot	Zeng et al.	3-point Likert scale; (threshold ≥2);	Gastroenterologists' rating: English 2.00 (0.00), Chinese 1.80 (0.45) Non-Expert rating: Chinese 2.20 (0.70)	Not specified
Gemini Pro	Zeng et al.	3-point Likert scale; (threshold ≥2);	Gastroenterologists' rating: English 2.40 (0.55), Chinese 2.20 (0.45) Non-Expert rating: Chinese 2.70 (0.58)	Not specified
ERNIE Bot 4.0	Kong et al.	3-point Likert scale (threshold ≥2)	Overall score: English 1.84 (0.80) Chinese 1.71 (0.87)	45.6%; English: 57.8% Chinse 33.3% (composite performance score of the three LLMs evaluated in the study)
	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists' rating: English 1.80 (0.45), Chinese 1.80 (0.45) Non-Expert rating: Chinese 2.12 (0.69)	Not specified

3.3. Readability

The readability of the educational content, specifically the reading grade level and textual clarity, was another key point of comparison (Table 4). An ideal patient handout should be written at roughly a 6th-grade reading level (as per American Medical Association and other health literacy guidelines) to be easily understood by the majority of adults. Both LLM- and expert-generated materials often failed to meet this benchmark [13]. Zeng et al. explicitly tested for reading level and found that none of the patient education documents (neither AI-generated nor physician-written) achieved the 6th-grade level; all were more complex than recommended. This held even though the prompt to the LLMs had specifically requested a "sixth-grade reading level" [13]. Quantitatively, in Ye et al. the readability of responses generated by ChatGPT 3.5, 4, and 4o was evaluated through expert review and quantitative analysis. Outputs were assessed based on word count and standard readability metrics, including the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). ChatGPT 4o produced the longest responses, followed by ChatGPT 4 and 3.5. While ChatGPT 4 showed numerically "superior" (FRE ~25), the differences among models were not statistically significant [9]. Similarly, an open-source model, DeepSeek, was found to give particularly verbose and complex answers that were harder to read. Evaluators noted DeepSeek's responses "required significant simplification for layperson comprehension" [10]. In general, while LLMs employ a conversational tone that avoids overly technical jargon, they still often produce sentences and vocabulary that exceed the level of a middle-school reader. This highlights the need to simplify language further or utilize health literacy tools. No matter how accurate an educational passage is, if many patients find it linguistically challenging, its utility is diminished. Both AI developers and medical communicators should prioritize readability optimization to ensure that *H. pylori* educational content is accessible to patients with varying literacy levels. "You do not really understand something unless you can explain it to your grandmother," as the apocryphal aphorism sound.

 doi:10.20944/preprints202510.1258.v1

Table 4. Evaluation of Readability. Comparison between educational materials on *Helicobacter pylori* generated by large language models (LLMs), such as ChatGPT, and those produced by human gastroenterology experts.

LLM Model	Study	Readability Value/Score	Scores for Each Scale	Word Count/Length Metric
ChatGPT-3.5	Ye et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL). Word count	FRE: 20.24 (9.44) FKGL: 15.19 (1.60).	Word count: 155.5 (67.09)
ChatGPT-4	Zeng et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL). Simple Measure of Gobbledygook (SMOG) Word count	FRE: 75.47 FKGL: 7th grade SMOG: 9.87	Word count: 578 (Chinese) 433 (English)
	Ye et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL) Word count	FRE: 24.88 (8.04) FKGL: 14.82 (1.48)	Word count: 199.2 (75.51)
	Gao et al.	Word count		Word count: 195.94 (52.96)
ChatGPT-4o	Ye et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL) Word count	FRE: 21.64 (10.54) FKGL: 15.01 (2.09)	Word count: 230.9 (104.5)
	Du et al.	Hemingway Editor, Grammarly	Hemingway Readability Score: poor. Hemingway Very Hard Sentences: 219/299. Grammarly Readability Score: 20. Grammarly Text Score: 85%	Word length: 5.5 Sentence length:24
OpenAI o1	Du et al.	Hemingway Editor, Grammarly	Hemingway Readability Score: poor. Hemingway Very Hard Sentences: 209/282. Grammarly Readability Score: 15. Grammarly Text Score: 85%.	Word length: 5.7 Sentence length: 22.8
DeepSeek-V3	Du et al.	Hemingway Editor, Grammarly	Hemingway Readability Score: poor. Hemingway Very Hard Sentences: 248/309. Grammarly Readability Score: 9. Grammarly Text Score: 88%	Word length: 5.8 Sentence length: 23.7
Claude 3 Opus	Zeng et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL). Simple Measure of Gobbledygook (SMOG) Word count	FRE: 60.87 FKGL: 8th and 9th grade SMOG: 11.73	Words (English version): 319 Words (Chinese version): 354
ERNIE Bot 4.0	Zeng et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL). Simple Measure of Gobbledygook (SMOG) Word count	FRE: 71.23 FKGL: 7th grade SMOG: 10.01	Words (English version): 309 Words (Chinese version): 509
Gemini Pro	Zeng et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL).	FRE: 72.44 FKGL: 7th grade SMOG: 9.46	Words (English version): 474 Words (Chinese version): 748

 doi:10.20944/preprints202510.1258.v1

		Simple Measure of Gobbledygook (SMOG) Word count		
Bing Copilot	Zeng et al.	Flesch Reading Ease (FRE) Flesch–Kincaid Grade Level (FKGL). Simple Measure of Gobbledygook (SMOG) Word count	FRE: 55.55 FKGL: 10th to 12th grade SMOG: 11.94	Words (English version): 323 Words (Chinese version): 519

3.4. Patient Comprehension

Readability metrics offer an objective measure of text complexity, but actual patient comprehension remains the ultimate test of practical education (Table 5). Several studies directly assessed how well people (patients or non-experts) understood the information provided by LLMs versus experts. Overall, initial evidence suggests that properly written content, whether generated by AI or physicians, can be understood by patients at least in a controlled setting; however, comprehension can vary depending on the audience's background knowledge. In Zeng et al., 50 patients were asked to rate the comprehensibility of *H. pylori* brochures (without knowing which source they came from). All materials, including those generated by LLMs, were rated as satisfactorily understandable (≥2 on a 3-point ease-of-understanding scale). There was virtually no difference in median comprehension scores between AIs and human-written brochures in that study, indicating that, from a patient's perspective, the clarity was adequate [13]. These findings are consistent with other reports. Kong et al. documented high performance in this domain, with ChatGPT 3.5, ChatGPT 4, and ERNIE Bot achieving 100% of responses rated as sufficiently comprehensible in both English and Chinese. However, it is essential to note that in this study, evaluators were not patients or laypersons, but rather physicians [12]. Similarly, Lai et al. confirmed the practical applicability of ChatGPT 3.5 by stating that it "can provide correct answers to the majority of *H. pylori*-related queries." Moreover, they highlighted that "It exhibited good reproducibility and delivered responses that were easily comprehensible to patients." Notably, this study did not apply a dedicated evaluation scale [11]. Gao et al., however, highlighted a critical nuance: when "ordinary people" (14 laypersons) were asked to evaluate ChatGPT-4's answers, their scores for comprehension were significantly lower than the scores given by medical experts reviewing the same answers. In that study, experts rated ChatGPT's responses nearly perfect for comprehension (mean ~2.95/3), whereas non-medical participants gave lower ratings (mean ~2.08/3) for how easy the answers were to understand. Medical students' ratings fell in between, closer to the experts' views. Consistently, the authors noted that "Some individuals thought that these answers were too obscure and lacked significance" and further stated that "those who have medical knowledge, who have a certain knowledge base, can more easily understand the answers." [14] This suggests that individuals with some medical knowledge found the AI explanations clear. Still, those without such a background had more difficulty. The likely reason is that although the text is grammatically clear, truly grasping concepts like "antibiotic resistance" or "urea breath test" requires some baseline knowledge. Another qualitative finding was that patients with lower levels of education or those with low health literacy might struggle with key terms or longer answers. Du et al., in a study comparing the Chinese LLM DeepSeek (versions R1 and V3) with ChatGPT-4o and OpenAI o1, noted that despite improvements in the completeness of information in Chinese AI, outputs remain "overly complex, limiting usability for non-expert audiences" [10]. Thus, while patient comprehension of LLM-generated content can be good in many cases (especially when patients are relatively educated or the content is simplified), there is a risk that a subset of patients will misinterpret or not fully absorb the information. Ensuring comprehension may involve supplementing AI-generated text with illustrations, utilizing interactive Q&A to clarify misunderstandings, or tailoring explanations based on patient feedback. It will be important in future

 doi:10.20944/preprints202510.1258.v1

trials to measure knowledge gain or recall after patients use AI-provided education to quantify comprehension outcomes directly.

Table 5. Evaluation of Comprehensibility. Comparison between educational materials on *Helicobacter pylori* generated by large language models (LLMs), such as ChatGPT, and those produced by human gastroenterology experts.

LLM Model	Study	Comprehensibility Metric/Scale	Mean score (SD)	Value/Score %
ChatGPT-3.5	Kong et al.	3-point Likert scale; (threshold ≥2)	English 2.96 (0.21) Chinese 2.93 (0.25)	100% of responses reached an acceptable level of comprehensibility
	Lai et al.	Qualitative assessment of concision	Responses were "coherent and easy to understand," "in language easily understandable for patients."	Not expressed
	Hu et al.	Qualitative assessment of concision	Not directly assess with specific scores	Not expressed
	Ye et al.	Comprehensibility was evaluated through measures of readability		
ChatGPT-4	Kong et al.	3-point Likert scale; (threshold ≥2)	English 2.93 (0.33) Chinese 2.80 (0.40)	100% of responses reached an acceptable level of comprehensibility
	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists: English 2.60 (0.55) Chinese 3.00 (0.00) Patients: Chinese 2.76 (0.43)	100% of responses reached an acceptable level of comprehensibility
	Gao et al.	3-point Likert scale; (threshold ≥2)	Overall mean for experts: 2.95 (0.21) Medical students 2.68 (0.54) scored higher than non-medical participants 2.16 (0.79)	
	Ye et al.	Comprehensibility was evaluated through measures of readability.		
	Hu et al.	Qualitative assessment of concision	Not directly assess with specific scores	
ChatGPT-4o	Du et al.	Comprehensibility was evaluated through measures of readability.		
	Ye et al.	Comprehensibility was evaluated through measures of readability.		
OpenAI o1	Du et al.	Comprehensibility was evaluated through measures of readability.		
DeepSeek-V3	Du et al.	Comprehensibility was evaluated through measures of readability.		

 doi:10.20944/preprints202510.1258.v1

DeepSeek-R 1	Du et al.	Comprehensibility was evaluated through measures of readability.		
Claude 3 Opus	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists: English 2.80 (0.45) Chinese 3.00 (0.00) Patients: Chinese 2.56 (0.67)	100% of responses reached an acceptable level of comprehensibility
Bing Copilot	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists: English 2.80 (0.45) Chinese 3.00 (0.00) Patients: Chinese 2.68 (0.55)	100% of responses reached an acceptable level of comprehensibility
Gemini Pro	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists: English 2.60 (0.89) Chinese 3.00 (0.00) Patients: Chinese 2.86 (0.40)	100% of responses reached an acceptable level of comprehensibility
ERNIE Bot 4.0	Kong et al.	3-point Likert scale; (threshold ≥2)	English 2.93 (0.33); Chinese 2.83 (0.53)	100% of responses reached an acceptable level of comprehensibility
	Zeng et al.	3-point Likert scale; (threshold ≥2)	Gastroenterologists: English 3.00 (0.00) Chinese 3.00 (0.00) Patients: Chinese 2.72 (0.50)	100% of responses reached an acceptable level of comprehensibility

3.5. Safety

Any tool providing medical information must be evaluated for safety, i.e., the absence of dangerous misinformation or advice that could harm patients if followed. The studies reviewed did not identify overtly dangerous instructions in response to *H. pylori* queries; nonetheless, they did report subtle inaccuracies and misleading information that could pose risks if left uncorrected. Zeng et al. had gastroenterologists conduct a structured safety review of each educational material, evaluating the likelihood and severity of potential harm arising from content errors. While the materials produced by both physicians and several LLMs (e.g., Bing, Gemini, ERNIE Bot) were generally judged as safe and unlikely to cause serious patient harm, some outputs raised more significant concerns. In particular, Ernie Bot was considered the safest among the LLMs, recording 100% of responses classified as "No harm", followed by Gemini Pro, with risk classifications of “Potentially mild to moderate harm” identified in only 20% of its Chinese responses. In contrast, Claude 3 Opus exhibited the most significant risk, with 60% of its English responses classified as “Potentially mild to moderate harm” and 20% of its Chinese responses classified as “Definitely mild to moderate harm.” These findings were attributed mainly to inaccuracies and insufficient precision in the generated content, underscoring the importance of systematic safety evaluation across different models [13]. Lai et al. reported that 16.6% of ChatGPT’s answers in the treatment domain contained a mix of correct and outdated information, such as recommending antibiotic regimens no longer considered first-line due to resistance patterns [11]. Similarly, Ye et al. highlighted notable errors, including ChatGPT-3.5 suggesting symptom relief as evidence of eradication and ChatGPT-4/4o recommending amoxicillin-containing regimens despite penicillin allergy [9]. Such inaccuracies, while not always overtly harmful, could nonetheless be dangerous and underscore the importance of up-to-date, guideline-consistent information in patient care. All authors stressed the importance of human oversight. Therefore, LLMs should be used with caution and ideally have their medical content reviewed or augmented by clinicians to catch subtle mistakes. As LLM deployment expands, implementing safety checks (for example, integrating medical knowledge bases or citing sources in answers) will be key to maintaining a high safety profile.

User Satisfaction: Ultimately, user satisfaction with the educational content is a crucial outcome, as it may influence whether patients trust and utilize the information. Because LLMs can provide information in a conversational format, one hypothesis is that patients might find this format engaging. Empirical data on satisfaction are still limited, but early indications are generally positive

 doi:10.20944/preprints202510.1258.v1

among healthcare professionals and mixed among patients. In Gao et al., expert gastroenterologists rated their satisfaction with ChatGPT-4's answers at 4.55 out of 5 on average, indicating that the specialists were very satisfied with the quality of the AI-generated responses. These experts also rated the content's usefulness highly (mean ~2.83/3), suggesting they felt the information provided would be helpful to patients. In the same study, medical students also gave positive evaluations of ChatGPT's answers, aligning with the notion that the content was educationally valuable. However, among ordinary laypeople, satisfaction-related metrics were more tempered. Founding that non-medical participants gave significantly lower scores for the "usefulness" of ChatGPT's answers compared to experts [14]. This could reflect differences in expectations or understanding; if parts of the answer were not fully grasped, a layperson might not feel satisfied. Kong et al. did not formally measure patient satisfaction; however, in their discussion, they emphasized that real-world patient satisfaction remains to be studied and may depend on factors such as a person's educational background and health context [12]. No study to date has reported on long-term satisfaction (for instance, whether patients would choose an AI tool again or recommend it to others), as most were one-time evaluations. It's also worth noting that none of the studies provided patients with a choice between an AI-generated brochure and a doctor-written brochure to determine which they preferred; such comparisons in the future could be enlightening. In summary, initial satisfaction levels with LLM-provided *H. pylori* information appear high when judged by content experts and reasonably good, but not uniformly excellent, when judged by lay users. Bridging this gap by improving content tailoring and addressing comprehension issues could enhance patient satisfaction. After all, an accurate brochure is only valuable if the patient feels it answered their concerns helpfully. As LLMs become more user-aware (through improved prompt engineering or interactive clarification), we expect user satisfaction to improve; however, it will be vital to continue capturing patient feedback in deployments.

4. Discussion

This comparative analysis suggests that LLMs have significant potential to complement gastroenterologists in providing patient education about *H. pylori* infection; however, critical challenges must be addressed before LLM-generated content can be adopted in practice. Accuracy emerged as a clear strength of modern LLMs. The reviewed studies uniformly show that these models can deliver factually correct answers to *H. pylori* questions at a level comparable to clinicians. This high accuracy is consistent with reports that most of the AIs under study have passed medical exams, suggesting that, on average, patients querying an advanced LLM are likely to receive generally reliable information. Such a capability could be invaluable in settings where physicians are not readily available to answer every question. For instance, patients often have numerous concerns about *H. pylori* (ranging from transmission to diet to treatment side effects) that they may not fully address during a brief clinic visit. An LLM-based tool could provide immediate, accurate answers as a supplement to the physician's advice, potentially improving patient understanding and reducing anxiety. However, accuracy alone is not sufficient. Completeness of information is where LLM responses currently fall short in comparison to a thorough consultation or a well-crafted pamphlet by an expert. In practice, an incomplete answer can be as problematic as an incorrect one when critical guidance is omitted. The observation that LLMs frequently provide only partially complete answers (e.g., lacking detail on follow-up testing or omitting certain risk factors) highlights the risk that patients using these tools might encounter knowledge gaps. A patient might recognize that *H. pylori* causes ulcers and can be treated with antibiotics (information that an LLM is likely to provide), but not realize that family members should also be tested, or that antibiotic resistance could affect therapy, nuances that an expert would emphasize [14]. One strategy to improve completeness is better prompting: clinicians or developers could design structured prompts that ensure all key topics are covered. Another approach is an interactive Q&A, where the AI can prompt the user to learn about related issues (for example, "Would you like to hear about how to prevent reinfection?"). Until such solutions are implemented, it may be advisable for any AI-derived content to undergo review

 doi:10.20944/preprints202510.1258.v1

by a healthcare provider who can identify and address any missing issues before the material is provided to patients [12].

Regarding readability and comprehension, our review underscores that current AI-generated content is not yet optimized for all patient populations. The fact that none of the evaluated materials met the target reading level for 6th graders is a call to action for both AI technology and health communication practices. Even the gastroenterologist-written materials were too advanced linguistically, which is a known challenge in patient education; explaining medical concepts in elementary language is difficult [13]. LLMs, with proper training or constraints, might be capable of simplifying language more consistently than busy clinicians. Future LLM development could focus on a "patient-friendly mode" that prioritizes shorter sentences, familiar words, and clear definitions of medical terms. Additionally, the multilingual capabilities of LLMs are a considerable asset: these models can instantly produce content in multiple languages, a task that would require significant human resources and time. This can help bridge language barriers and reach patient groups who speak different languages. Ensuring readability in each language (not just direct translation) is important [12,13]. Our findings on patient comprehension, especially the gap between experts and laypeople in perceiving clarity, suggest that involving actual patients in the development and testing of LLM-based tools is vital. By observing where non-experts get confused, developers can tweak the AI's explanations. The AI might also incorporate visual aids and analogies to enhance understanding [15].

In terms of safety, although most LLMs generally adhered to clinical guidelines, some studies did report a certain level of "harmful" responses. This was primarily due to outdated or incomplete information, such as treatment recommendations that are no longer considered effective (for example, using an antibiotic regimen that has become ineffective), which can lead to treatment failure. Nonetheless, in several studies, ChatGPT consistently included a recommendation to consult a physician in addition to the information provided, which may help mitigate potential risks. Thus, minimizing the dissemination of outdated or partial content remains crucial [8,11,14]. Thus, minimizing the risk of outdated information is crucial. One solution is to continually update LLM knowledge bases with current clinical guidelines, although models like GPT-4 are not easily updatable in real-time [14]. Future systems may incorporate live data or retrieve information from trusted databases. Another safety measure is transparency: if the AI provides citations or sources (as some LLM-based medical assistants are starting to do), patients and providers can verify the information against reputable references. Ultimately, we envision that LLM-generated patient education will not operate in isolation but rather under a framework of human-AI collaboration: clinicians could supervise the content, or the AI could triage questions and draft answers that a clinician then reviews for accuracy and safety. Such a model would harness the efficiency of AI while prioritizing patient well-being.

User satisfaction and engagement are essential for the practical success of any patient-facing tool. The early positive feedback from physicians and trainees suggests that, if the content is of high quality, healthcare professionals are willing to trust and even recommend these AI resources. This is important, as doctors could serve as facilitators; a doctor might guide a patient to use a vetted chatbot for follow-up questions at home. However, the lukewarm responses from some lay users indicate a need for improvement in the user experience. Patients will be satisfied not just by correct answers, but by feeling that their concerns are addressed [12]. LLMs are capable of a conversational style, which can be friendly and empathetic, but they currently lack the true personalization and emotional intelligence of a human provider. Future development could incorporate more adaptive responses, where the AI asks the user if the answer was helpful or if they have other concerns, thereby mimicking a dialogue with a doctor. Moreover, some patients might distrust information from an "algorithm." Building trust will require showing that the AI's information is endorsed or co-developed by medical experts and that it has been tested in real patient populations with good outcomes. Over time, as patients become more accustomed to digital health tools, their satisfaction is likely to increase, provided the information is reliable and comprehensible. An often-mentioned

 doi:10.20944/preprints202510.1258.v1

benefit is the 24/7 availability of LLM-based assistance; patients can get answers at any time, which could improve satisfaction in the context of anxiety (e.g., a patient worrying at night about their *H. pylori* test results might consult the AI for immediate information on what to expect).

There are some limitations to consider in this analysis. The studies reviewed mainly evaluated static text outputs or single-session Q&A with LLMs. Real-world use may involve more interactive conversations that can either enhance understanding (by allowing follow-up questions) or introduce new errors. Also, the patient populations in these studies were relatively small (dozens of patients at most, often with at least a high school education), so the findings may not be generalized to all demographic groups. Individuals with very low literacy or from a different cultural background may react differently to the content. Additionally, the rapid evolution of LLMs means that models available today could become significantly more advanced (or specialized for healthcare) in the near future, potentially altering performance on each domain. Nonetheless, the core areas identified — accuracy, completeness, readability, comprehension, safety, and user satisfaction —will remain relevant benchmarks for any such technology.

In conclusion, LLM-generated educational content for *H. pylori* demonstrates high accuracy and acceptable patient comprehensibility, suggesting that these AI tools can effectively address many patient questions about the infection. They can greatly expand access to standardized, evidence-based information on *H. pylori* across different languages and regions. However, to fully realize this potential, improvements are needed in the completeness of the information supplied and in tailoring the language to patient reading levels. Medical oversight remains essential to ensure safety and to update content in line with the latest clinical guidelines. With ongoing refinement and responsible implementation, LLMs could become a valuable adjunct in gastroenterology, empowering patients with knowledge, reinforcing physicians' advice, and ultimately contributing to a better management of *H. pylori* infection and its related diseases. Future research should focus on patient outcomes when using LLM-based education (such as increased knowledge retention, reduced decisional conflict, or improved treatment adherence) to establish the clinical benefits of this promising technology.

Author Contributions: Conceptualization, G.O. and M.P.D.; methodology, M.P.D. and G.M.P.; software, G.M.P.; validation, G.M.P., E.M. and G.M.P.; formal analysis, G.O. and G.M.P.; investigation, M.P.D.; resources, M.P.D.; data curation, M.P.D.; writing—original draft preparation, M.P.D.; writing—review and editing, M.P.D, E.M., and G.M.P..; visualization, G.M.P.; supervision, M.P.D.; project administration, M.P.D.; funding acquisition, M.P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **2024**, *74*, 229-263, doi:10.3322/caac.21834.
2. de Martel, C.; Georges, D.; Bray, F.; Ferlay, J.; Clifford, G.M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* **2020**, *8*, e180-e190, doi:10.1016/S2214-109X(19)30488-7.

3. Zha, J.; Li, Y.Y.; Qu, J.Y.; Yang, X.X.; Han, Z.X.; Zuo, X. Effects of enhanced education for patients with the Helicobacter pylori infection: A systematic review and meta-analysis. *Helicobacter* **2022**, *27*, e12880, doi:10.1111/hel.12880.
4. Hafiz, T.A.; D'Sa, J.L.; Zamzam, S.; Visbal Dionaldo, M.L.; Aldawood, E.; Madkhali, N.; Mubaraki, M.A. The Effectiveness of an Educational Intervention on Helicobacter pylori for University Students: A Quasi-Experimental Study. *J Multidiscip Healthc* **2023**, *16*, 1979-1988, doi:10.2147/JMDH.S419630.
5. Association, A.M. Health literacy and patient safety: help patients understand. Available online: https://www.ama-assn.org/sites/ama-assn.org/files/corp/media-browser/public/health-literacy/ama-health-literacy-patient-safety-2007.pdf (accessed on
6. Iqbal, U.; Tanweer, A.; Rahmanti, A.R.; Greenfield, D.; Lee, L.T.; Li, Y.J. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci* **2025**, *32*, 45, doi:10.1186/s12929-025-01131-z.
7. Berry, P.; Dhanakshirur, R.R.; Khanna, S. Utilizing large language models for gastroenterology research: a conceptual framework. *Therap Adv Gastroenterol* **2025**, *18*, 17562848251328577, doi:10.1177/17562848251328577.
8. Hu, Y.; Lai, Y.; Liao, F.; Shu, X.; Zhu, Y.; Du, Y.Q.; Lu, N.H.; National Clinical Research Center for Digestive, D. Assessing Accuracy of ChatGPT on Addressing Helicobacter pylori Infection-Related Questions: A National Survey and Comparative Study. *Helicobacter* **2024**, *29*, e13116, doi:10.1111/hel.13116.
9. Ye, Y.; Zheng, E.D.; Lan, Q.L.; Wu, L.C.; Sun, H.Y.; Xu, B.B.; Wang, Y.; Teng, M.M. Comparative evaluation of the accuracy and reliability of ChatGPT versions in providing information on Helicobacter pylori infection. *Front Public Health* **2025**, *13*, 1566982, doi:10.3389/fpubh.2025.1566982.
10. Du, R.C.; Zhu, Y.C.; Xiao, Y.T.; Yang, B.N.; Lai, Y.K.; Zhou, Z.X.; Deng, H.; Shu, X.; Lu, N.H.; Zhu, Y.; et al. Assessing the Capabilities of Novel Open-Source Artificial Intelligence-DeepSeek in Helicobacter pylori-Related Queries. *Helicobacter* **2025**, *30*, e70045, doi:10.1111/hel.70045.
11. 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**, *29*, e13078, doi:10.1111/hel.13078.
12. Kong, Q.Z.; Ju, K.P.; Wan, M.; Liu, J.; Wu, X.Q.; Li, Y.Y.; Zuo, X.L.; Li, Y.Q. Comparative analysis of large language models in medical counseling: A focus on Helicobacter pylori infection. *Helicobacter* **2024**, *29*, e13055, doi:10.1111/hel.13055.
13. Zeng, S.; Kong, Q.; Wu, X.; Ma, T.; Wang, L.; Xu, L.; Kou, G.; Zhang, M.; Yang, X.; Zuo, X.; et al. Artificial Intelligence-Generated Patient Education Materials for Helicobacter pylori Infection: A Comparative Analysis. *Helicobacter* **2024**, *29*, e13115, doi:10.1111/hel.13115.
14. Gao, Z.; Ge, J.; Xu, R.; Chen, X.; Cai, Z. Potential application of ChatGPT in Helicobacter pylori disease relevant queries. *Front Med (Lausanne)* **2024**, *11*, 1489117, doi:10.3389/fmed.2024.1489117.
15. Dore, M.P.; Merola, E.; Pes, G.M. Advances and future perspectives in the pharmacological treatment of Helicobacter pylori infection: Taking advantage from artificial intelligence. *Clin Res Hepatol Gastroenterol* **2025**, *49*, 102689, doi:10.1016/j.clinre.2025.102689.