Supplementary Material

Supplementary material 1. Physician Evaluation Rubric

Each large language model (LLM)–generated response was independently evaluated by three board-certified gastroenterologists blinded to model identity. Assessment followed a standardized, pre-defined rubric encompassing four domains derived from established patient-communication and medical-education frameworks. Scores were assigned on ordinal scales, with higher values indicating superior quality.

1. Accuracy (1–4)

Assessed the factual correctness and evidence base of each response. Reviewers determined whether information was medically valid, current, and aligned with international GERD guidelines (American College of Gastroenterology [ACG], American Gastroenterological Association [AGA], Indian Society of Gastroenterology [ISG], and World Gastroenterology Organisation [WGO]).

- **1 = Incorrect or misleading content:** major factual errors or outdated advice
- **2 = Partially correct:** minor factual errors or missing context
- **3 = Mostly accurate:** correct core content but lacking full guideline alignment
- **4 = Entirely accurate and evidence-based:** factually correct, current, and guideline-concordant

2. Comprehensiveness (1–4)

Evaluated how completely each response addressed the question's key clinical dimensions (etiology, pathophysiology, diagnosis, management, and follow-up).

- **1 = Very incomplete:** omits essential aspects
- **2 = Partially complete:** covers some but not all relevant points
- **3 = Substantially complete:** addresses major elements with minor omissions
- **4 = Fully comprehensive:** holistic, balanced, and adequately detailed

3. Empathy / Tone (1–4)

Captured the emotional quality and patient-centredness of communication, reflecting language warmth, reassurance, and avoidance of jargon.

- **1 = Cold or impersonal tone**
- **2 = Neutral or technical tone:** minimal acknowledgement of patient concerns
- **3 = Moderately empathetic:** acknowledges emotion with reassurance
- **4 = Highly empathetic and reassuring:** conveys compassion and understanding throughout

4. Actionability (1–5)

Measured the clarity and practicality of guidance provided, focusing on next steps a patient could reasonably follow (lifestyle measures, warning signs, and when to seek medical attention).

- **1 = No practical advice:** theoretical or purely descriptive
- **2 = Minimal guidance:** vague or generic suggestions
- **3 = Some actionable points:** partially usable advice
- **4 = Clear and practical:** specific, implementable recommendations
- **5 = Very actionable:** explicit, step-by-step instructions with escalation cues

Each rater applied this rubric independently; inter-rater reliability was assessed with a two-way random-effects intraclass correlation coefficient (ICC [2,1]). Composite mean domain scores were computed per model for descriptive and inferential analysis.

Supplementary material 2. Patient Evaluation Rubric and Rating Protocol

To incorporate the end-user perspective, we recruited 20 English-speaking adults (≥18 years) with diverse educational backgrounds ranging from high school to postgraduate. Participants were either patients with prior heartburn or individuals familiar with online health information. All participants provided informed consent.

Evaluation Process

Patients were blinded to the AI model identities. The 120 model-generated responses were presented in randomized order (distinct from physicians' order) using an electronic survey platform. Participants were told that each answer represented an AI-generated reply to a common acidity-related question written in a conversational tone. They were asked to rate each answer on three patient-centred aspects:

1. Comprehensibility (1–4)

Assessed how easy the answer was to understand.

- **1 = Very hard to understand:** overly technical or confusing
- **2 = Moderately difficult:** contains medical jargon or complex phrasing
- **3 = Understandable:** mostly clear but may include occasional jargon
- **4 = Very easy to understand:** written in simple, clear lay terms

2. Empathy (1–4)

Captured whether the response conveyed care, reassurance, and emotional awareness.

- **1 = Not at all empathetic:** detached or purely factual tone
- **2 = Mildly empathetic:** limited recognition of patient emotion
- **3 = Empathetic:** supportive and polite tone
- **4 = Very empathetic and supportive:** warm, compassionate, and patient-affirming

3. Actionability (1–5)

Evaluated whether the answer offered useful and feasible next steps.

- **1 = Not actionable:** no practical advice
- **2 = Minimally actionable:** vague or generic suggestions
- **3 = Moderately actionable:** some clear next steps
- **4 = Actionable:** provides clear, feasible advice
- **5 = Extremely actionable:** concrete, specific guidance on what to do next

Patients were instructed not to assess medical accuracy but rather to focus on clarity, trustworthiness, tone, and usability. Optional free-text fields allowed qualitative comments about strengths or weaknesses of the answers. All participants completed the full set of evaluations, providing a comprehensive dataset reflecting layperson impressions of AI-generated health information.

Supplementary material 3. Readability Assessment Framework

Because readability influences health literacy, we analyzed each AI-generated response using five established readability indices. These complementary metrics assess linguistic complexity, average sentence length, and syllable density, providing objective measures of text difficulty.

1. Flesch Reading Ease (FRE)

Scores range from 0 to 100 (higher = easier). FRE incorporates sentence length and syllables per word.

- 90–100 = very easy (≈ 5th grade)
- 60–70 = plain English (≈ 8th–9th grade)
- 30–50 = difficult (college level)
- < 30 = very difficult (graduate level)
  Health-communication guidelines recommend scores ≥ 60 for patient-facing content.

2. Simple Measure of Gobbledygook (SMOG)

Estimates grade level based on polysyllabic word frequency.

- **Formula:** Grade = 1.043 × √(polysyllabic words × (30 / sentences)) + 3.1291
  A SMOG grade of 6–8 is ideal for general audiences.

3. Automated Readability Index (ARI)

Calculates grade level from characters per word and words per sentence.

- **Formula:** ARI = 4.71 × (characters / words) + 0.5 × (words / sentences) – 21.43
  Scores align with U.S. school grades (e.g., 8 ≈ 8th grade).

4. Gunning Fog Index (GFI)

Assesses readability using sentence length and proportion of complex (≥3-syllable) words.

- **Formula:** GFI = 0.4 × [(words / sentences) + 100 × (complex words / words)]
  Values > 12 = college-level difficulty; 8–10 = general-public readability.

5. Automated Readability Count (ARC)

A composite measure integrating sentence length, word length, and syllabic density, producing a grade-equivalent index. Lower values denote easier text; scores > 12 indicate university-level complexity.

Implementation

All indices were computed programmatically using validated natural-language processing libraries in Python and cross-verified manually. Scores were summarized as mean ± standard deviation per model, and inter-model comparisons employed ANOVA with Holm-adjusted pairwise tests. These metrics quantified linguistic accessibility and complemented the physician and patient subjective ratings.

Overall Purpose:
Together, these appendices define the methodological backbone for quantitative and qualitative evaluation of AI-generated GERD patient-education responses. They ensure reproducibility, transparency, and adherence to STROBE and EQUATOR reporting standards for observational studies evaluating language-model outputs.

Supplementary material

Appendix 4: Full List of Frequently Asked Questions (FAQs) Evaluated in the Study

The 40 patient-facing questions used to evaluate the large language models (LLMs) were categorized into six domains. These questions were selected based on their frequency in clinical practice, prominence on gastroenterology education websites, and relevance to patient concerns.

Domain 1: General Understanding of Dyspepsia (n = 6)

1. What is dyspepsia?
2. Is dyspepsia the same as indigestion?
3. Can dyspepsia be cured permanently?
4. How is dyspepsia different from acidity or gas?
5. Is dyspepsia a serious condition?
6. Why does dyspepsia keep coming back even with treatment?

Domain 2: Symptoms and Diagnosis (n = 9)

7. What are the common symptoms of dyspepsia?
8. How do I know if I have dyspepsia or something more serious?
9. What tests are used to diagnose dyspepsia?
10. Do I need an endoscopy for dyspepsia?
11. What is the role of ultrasound or CT scan in dyspepsia?
12. Can dyspepsia cause chest pain or palpitations?
13. What are the red flag symptoms in dyspepsia?
14. Can dyspepsia be diagnosed without any tests?
15. Is H. pylori infection always present in dyspepsia?

Domain 3: Pathophysiology and Mechanisms (n = 6)

16. What causes dyspepsia if all my tests are normal?
17. Is dyspepsia caused by stress or anxiety?
18. What is functional dyspepsia?
19. What is visceral hypersensitivity and how does it relate to dyspepsia?

20. How does the gut-brain axis affect my digestion?
21. Can hormones or menstrual cycle affect dyspepsia?

Domain 4: Diet and Lifestyle Modifications (n = 7)

22. What foods should I avoid if I have dyspepsia?
23. Is coffee or tea bad for dyspepsia?
24. Can alcohol or smoking worsen dyspepsia symptoms?
25. Does eating late at night cause dyspepsia?
26. Are there specific diets like FODMAP for dyspepsia?
27. How should I change my eating habits to reduce symptoms?
28. Can regular exercise help with dyspepsia?

Domain 5: Pharmacologic Management (n = 6)

29. What are the best medicines for dyspepsia?
30. Are proton pump inhibitors (PPIs) safe for long-term use?
31. When should I stop taking antacids?
32. Are antibiotics needed to treat dyspepsia?
33. What are prokinetics and how do they work?
34. Can antidepressants help with dyspepsia symptoms?

Domain 6: Psychological and Complementary Approaches (n = 6)

35. Does stress management help reduce dyspepsia?
36. Is psychological therapy recommended for dyspepsia?
37. Can yoga or meditation improve my symptoms?
38. Are there herbal or natural remedies for dyspepsia?
39. What is the role of hypnotherapy in dyspepsia?
40. How do I deal with the anxiety and fear related to my symptoms?

Supplementary material 5: Prompt Format Used for LLM Querying

Each question was submitted in the following standardized prompt format:

"As a patient with indigestion or dyspepsia, I want to understand: [insert question]. Please explain in simple terms."

Prompts were entered using the "new chat" feature for both ChatGPT-4 (May 2024) and Gemini-1.5 (June 2024). No contextual or follow-up prompts were allowed to preserve consistency.

Supplementary material 6: Scoring Rubric for Response Evaluation

Comprehensiveness and Accuracy (4-point scale):

1. Comprehensive and accurate
2. Correct but inadequate or incomplete
3. Mixed (some correct, some vague/inaccurate)
4. Incorrect or misleading

Empathy (4-point scale):

1. Highly empathetic
2. Moderately empathetic
3. Minimally empathetic
4. Not empathetic

Readability Tools Used:

- Flesch Reading Ease (FRE)
- Gunning Fog Index (GFI)
- SMOG Index

These metrics were recorded for each response using ReadabilityFormulas.com.

Supplementary material 7: Reviewer Roles

- Two blinded gastroenterologists independently rated all responses.
- A third senior clinician resolved any rating discrepancies.
- Three independent reviewers evaluated the empathy scores.

This supplementary appendix is intended to enhance the transparency and reproducibility of this study's methodology.