Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Effect of Emotional Prompt on the Quality of ChatGPT Based Discharge Instructions After Laparoscopic Cholecystectomy Using ERAS Derived Benchmarks

Mariusz Panczyk[1], Tomasz Dawid Piątek[2], Piotr Małkowski[2], Marta Katarzyna Hreńczuk[2], Mariusz Jaworski[1], Ilona Cieślak[1], Joanna Gotlib-Małkowska[1]

[1]Department of Education and Research in Health Sciences, Faculty of Health Sciences, Medical University of Warsaw, Warsaw, Poland; [2]Department of Surgical and Transplantation Nursing and Extracorporeal Treatment, Faculty of Health Sciences, Medical University of Warsaw, Warsaw, Poland

Correspondence: Mariusz Panczyk, Department of Education and Research in Health Sciences, Faculty of Health Sciences, Medical University of Warsaw, Litewska 14/16 St, Warsaw, 00-581, Poland, Tel +48 22 116 92 50, Email mariusz.panczyk@wum.edu.pl

Background: Nurse-led education is a crucial form of discharge communication for ensuring safe postoperative in-home patient recovery. However, discharge instructions are often incomplete or inadequately tailored to patients' needs. Generative large language models (LLMs), such as ChatGPT, may support postoperative patient education.

Objective: To evaluate the completeness, accuracy, and quality of post-cholecystectomy discharge instructions, generated in Polish by ChatGPT-4 Omni, comparing a primary prompt (PP) with an emotional prompt (EP), in which motivational cues were added to the PP structure.

Methods: In this evaluation study, ChatGPT 4.0 Omni generated 60 discharge instruction 30 outputs per prompt type. Two blinded experts assessed content completeness against a 43-item benchmark derived from Enhanced Recovery After Surgery (ERAS) protocols and professional guidelines, and accuracy defined as absence of clinically incorrect, unsafe, or guideline inconsistent statements. Inter rater agreement was quantified with Cohen kappa and between prompt type differences in domain level completeness were examined with the Brunner Munzel test. Communication quality was evaluated using directed qualitative content analysis and a dictionary-based text mining approach that counted affirmatives, modal verbs, and empathetic expressions and combined them into a Composite Relationality Index (CRI) for each output.

Results: Inter expert agreement for benchmark ratings was very high mean kappa 0.859 (95% confidence interval 0.841 to 0.878) with an overall agreement rate of 93.14%. Core domains such as analgesia, wound care, low fat diet, vital sign monitoring, and emergency procedures were present in all outputs 100%. Several elements were underrepresented, including full antibiotic prescription details 25.0%, gradual fibre reintroduction 41.7%, specification of an adaptation period 30.0%, monitoring of gastrointestinal or other red flag symptoms 33.3%, and individual dietary product groups each below 25%. No statistically significant differences in completeness were detected between primary and emotional prompts for any of the 43 features, and expert review identified no factual inaccuracies, guideline inconsistencies, or potentially harmful statements. Emotional prompts did, however, markedly increase relational linguistic density almost all primary prompt outputs had a CRI of zero, whereas emotional prompt outputs showed frequent use of modal verbs, affirmatives, and empathetic expressions with index values extending up to 5.28.

Conclusion: ChatGPT 4.0 Omni can generate Polish language discharge instructions after laparoscopic cholecystectomy that are accurate and largely complete for core postoperative domains, but with important gaps in antibiotic related information and structured dietary guidance. Emotional prompting substantially enhances the relational and motivational tone without improving content completeness. Nurse led discharge counselling and clinician review therefore remain essential when using large language model generated materials, ideally supported by standardised prompts, ERAS aligned benchmarks, and human in the loop oversight.

Keywords: ChatGPT, discharge instructions, postoperative care, prompt engineering, patient education, large language models

https://doi.org/10.2147/JMDH.S566966
Received: 12 September 2025
Accepted: 16 December 2025
Published: 13 January 2026

Introduction

Cholecystectomy is one of the most commonly performed abdominal procedures worldwide, with laparoscopic techniques considered the standard of care due to shorter hospital stays and compatibility with enhanced recovery protocols.[1,2] In 2020 the average cholecystectomy rate across thirty European health systems was 134.6 per 100,000 inhabitants, with more than 86% performed laparoscopically. In Poland the rate was 136.0 per 100,000 inhabitants 51,552 procedures of which 85.5% were laparoscopic.[3,4] In the United States approximately 750,000 cholecystectomies are carried out annually, equivalent to about 230 procedures per 100,000 inhabitants, and nearly 90% are laparoscopic.[5] Given this volume of surgery, effective and reliable discharge education is essential to support safe recovery at home.

The effectiveness of hospital discharge depends largely on the quality of education provided to patients. Inadequate or poorly structured discharge instructions can contribute to preventable complications, higher rates of emergency department visits, reduced patient confidence, and delayed recovery, especially among individuals with limited health literacy. Patients require clear and actionable guidance on wound care, analgesia, activity progression, dietary changes, bowel function, warning signs, and follow up. Educational content should align with evidence based recommendations, including Enhanced Recovery After Surgery (ERAS) protocols and guidance from professional organisations such as the European Association for Endoscopic Surgery (EAES) and the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES).[6–8] Yet conventional discharge practices frequently fall short of these standards and are often insufficiently tailored to the needs and literacy levels of patients.

Many written discharge materials are lengthy and difficult to comprehend, particularly for patients with reading skills below a high school level. To address these limitations, several studies have evaluated digital and multimodal interventions. Randomised trials have shown that educational videos, virtual reality animations and nurse led telephone consultations can improve postoperative knowledge, reduce pain and nausea, and enhance adherence to recovery instructions.[9–12] Multimodal programmes that combine visual, auditory and textual elements across perioperative stages further improve satisfaction and selected clinical outcomes,[13] but still require careful adaptation to health literacy and cultural context to be effective.

Large language models (LLMs) offer new opportunities to streamline the development of discharge instructions that are both comprehensive and accessible. Health authority guidelines recommend plain language, clear structure, and a reading level not exceeding approximately grade six, which can be supported by model generated drafts that clinicians subsequently refine.[14,15] Early evaluations suggest that such models can produce content with acceptable clinician rated clarity, accessibility and completeness, with potential applications in automated drafting, patient centred question answering and multilingual translation.[16] Beyond text based applications, recent work in ophthalmology has shown that multimodal in context learning with ChatGPT 4o can achieve high diagnostic accuracy for selected retinal, external ocular and facial image tasks, and can translate schematic diagnostic algorithms into simple browser based decision support tools using example based prompts without additional model training.[17] At the same time, important limitations have been identified, including variable accuracy across languages and particular vulnerability in medication related content, such as dosing details and safety warnings.[16,18–20] As a result, safe deployment of LLMs in clinical communication requires a human in the loop approach, checklist based verification of content domains and explicit readability assessment.

Beyond these technical considerations, several broader research and methodological gaps remain. First, although completeness, accuracy and communication quality are all critical to effective patient education, relatively few studies evaluate model generated post discharge materials using a multidimensional framework that explicitly combines these dimensions.[21] Second, prompt design is often under reported and insufficiently standardised, which hinders reproducibility and limits the interpretability of results.[22] Third, heterogeneity in study designs and reporting standards makes it difficult to compare interventions and to build cumulative evidence.[23] These limitations have led to calls for standardised prompt templates, transparent reporting, and outcome oriented quality assurance that can be applied across clinical contexts.

Prompt engineering, understood as the systematic design of model inputs to optimise performance, has emerged as a key factor in the behaviour of LLMs.[24] Within this field, emotional prompting introduces brief motivational or affective

cues into prompts in order to influence the tone and structure of outputs. The EmotionPrompt framework developed by Li et al augments standard prompts with predefined emotional stimuli and has demonstrated substantial improvements in benchmark performance and in human ratings of generated text quality.[25]

From a behavioural and communication perspective, emotional prompts can be interpreted through the lens of Social Cognitive Theory and patient centred counselling approaches such as motivational interviewing.[26] These frameworks emphasise self-efficacy, outcome expectations, empathy and support for patient autonomy as determinants of health behaviour and treatment adherence.[27,28] Short affirming or normalising statements in the prompt may function as social persuasion that strengthens patients perceived capability to manage wound care, physical activity or dietary changes, while emphasising personal agency and shared responsibility for recovery.[29] In this sense, emotional prompting provides a theoretically grounded mechanism by which the wording of model inputs can shape not only the relational tone of the generated message but also patients' confidence and intention to follow discharge instructions.

However, existing evidence on emotional prompting comes mainly from general purpose tasks and benchmark datasets rather than clinically oriented communication scenarios. A specific research gap therefore concerns whether and how emotional cues embedded in prompts can quantitatively influence the quality and completeness of model generated discharge instructions in real clinical contexts. To our knowledge, no previous study has systematically compared neutral and emotional prompts for LLMs in the setting of postoperative discharge education, using an explicit ERAS derived benchmark and a combined quantitative and qualitative evaluation, particularly in a non-English language such as Polish.

Building on this gap, the present study evaluates the effect of an emotional prompt on the quality of discharge instructions generated in Polish by ChatGPT 4.0 Omni for patients after laparoscopic cholecystectomy. We apply a multidimensional framework that assesses completeness against a 43-item ERAS based benchmark, accuracy with respect to evidence-based guidelines and communication quality through directed qualitative content analysis. Grounded in Social Cognitive Theory, patient centred communication models and evidence that EmotionPrompt can enhance LLM performance,[25] we hypothesised that the emotional prompt (EM) would lead to measurable improvements in content completeness and in relational and motivational aspects of communication compared with a primary prompt (PP) that does not contain explicit emotional cues.

Materials and Methods

Study Design

We conducted an evaluation study aligned with the METRICS checklist developed by Sallam et al[22,30] which provides a structured framework for transparent and rigorous reporting of research involving large language models. The checklist comprises nine core elements: model specification, evaluation approach, transparency, reproducibility, interpretation, clinical relevance, safety, ethics, and stakeholder engagement. The present study builds on our earlier conference work that compared post cholecystectomy home care instructions generated by ChatGPT using advanced prompt engineering reference.[31]

Model Specification

This study used ChatGPT, model label GPT 4o (May 2024) as displayed in the ChatGPT Plus user interface, accessed through the web application at https://chatgpt.com. The model was operated in the Polish language user interface in its default configuration, with no personalisation, custom instructions, or additional plug-in features enabled at any stage of data generation.

Evaluation Approach

The study adopted a comparative approach to assess the outputs generated with two distinct prompt types: Primary Prompt (PP) and Emotional Prompt (EP). Thirty independent outputs were generated in Polish for each prompt type. Each output was evaluated, by two blinded surgical nursing experts, against a 43-item benchmark derived from ERAS guidelines and international standards.

Prompt Development

Prompt design followed well-established prompt engineering principles provided by Microsoft Azure OpenAI documentation, ensuring transparency, reproducibility, and consistency in model interaction (see Supplementary Materials 1).[32] The PP specified a clear persona (surgical nurse), target audience (post-cholecystectomy patient), and professional yet supportive tone. It also adhered to a five-component structure (Table 1). The design provided a controlled framework to evaluate the incremental effect of emotional augmentation.

The EP was constructed as a direct extension of the PP by embedding short emotional stimuli within the instructional text ("To jest naprawdę ważne dla mnie i dla mojego pacjenta" *[This really matters to me and to my patient]*). The design followed principles of the EmotionPrompt framework proposed by Li et al,[25] which demonstrated that affective cues embedded in prompts can motivate LLMs to generate outputs more efficiently and accurately. In our study, the emotional component was grounded in well-established psychological models, including Social Cognitive Theory (outcome expectations).

Both prompts were created in Polish. The EP incorporated an additional motivational phrase, while maintaining an identical informational structure to the PP, thereby isolating the specific effect of emotional augmentation.

In addition, the full English translations wording of both prompts, and the complete informational scaffold used to generate the outputs are reproduced verbatim in Supplementary Materials 2 to facilitate exact replication of the interaction.

Benchmark Development

To enable a structured and reproducible evaluation, a detailed benchmark instrument was developed to assess ChatGPT generated outputs. The checklist encompassed six patient care domains subdivided into 43 specific criteria, each with explicit eligibility rules. The criteria evaluated the accuracy and specificity of medication names, dosages and administration schedules; guidance on vital sign monitoring and the recognition of red flag symptoms such as fever, jaundice, dyspnoea, severe or persistent abdominal pain, wound infection or urinary abnormalities; dietary progression with allowed and restricted foods; wound inspection and dressing change protocols; staged reintroduction of physical activity; and escalation protocols stratified by severity of postoperative complications see Supplementary Materials 1.

The benchmark was developed by an expert panel comprising surgical nurses, a gastrointestinal surgeon and a clinical dietitian. Draft criteria were generated from ERAS protocols and professional guidance and then refined through a series of moderated discussions in which panel members reviewed, merged or reworded items until consensus was reached that

Table 1 Structural Components of the Primary Prompt (PP-Type) and Emotional Prompt (EP-Type) Used to Generate AI-Based Discharge Instructions for Post-Cholecystectomy Patients

Component	PP-Type	EP-Type
1. Persona	Healthcare professional (nurse) specialising in postoperative education.	Same as PP-type.
2. Audience	Average patient after gallbladder removal surgery (cholecystectomy).	Same as PP-type.
3. Tone	Friendly, supportive, and accessible, while maintaining professionalism.	Same as PP-type.
4. Fine-tuning criteria	Accurate, detailed, and aligned with current clinical guidelines for post operative care.	Same as PP-type.
5. Content areas	(1) Medication management; (2) General health monitoring; (3) Dietary recommendations; (4) Wound care; (5) Daily life organisation; (6) Emergency contact information.	Same as PP-type.
6. Additional emotional layer	Not applicable.	Short emotional stimulus embedded in the prompt text (eg, "To jest naprawdę ważne dla mnie i dla mojego pacjenta" / "This really matters to me and to my patient").

Abbreviations: PP-type, primary prompt; EP-type, emotional prompt.

each criterion was clinically accurate and educationally relevant. The structure and content of the checklist were explicitly aligned with ERAS protocols and informed by guidance from professional organisations such as the EAES and SAGES, thereby ensuring that the criteria reflected up to date, evidence based standards for evaluating instructions for post cholecystectomy care.[6–8]

At this developmental stage, the validation of the benchmark focused on expert-based content validity and did not include the calculation of formal psychometric indices, such as item or scale level content validity indices or quantitative measures of agreement among panel members. The high inter rater agreement reported in the Results section therefore reflects the reliability with which two independent raters applied the benchmark in this study, rather than a full psychometric validation of the instrument itself.

Sample Size

The target sample size was 60 outputs, with 30 discharge instruction outputs generated for each prompt type PP and EP. From a quantitative perspective, the study was designed to be sensitive to moderate to large differences in completeness between the two prompt types. For the Brunner-Munzel test applied to domain level completeness scores, the effect size can be expressed as the relative effect, defined as the probability that an emotional prompt output has a higher score than a primary prompt output.[33] With 30 observations per group and a two-sided alpha level of 0.05, normal approximation for rank based tests indicates that the design provides moderate power for detecting a relative effect of about equal 0.70 that is, a 70% probability that an emotional prompt output exceeds a primary prompt output, while substantially smaller differences would require much larger samples to be detected with high power.[34]

The sample size of 30 outputs per prompt type total n equal 60 was also selected reflecting generative AI evaluation research. Prior benchmarking studies have demonstrated that thematic saturation in the content analysis of LLM outputs typically occurs within 20 to 30 instances per usage configuration, particularly when the prompts are deterministic and domain specific.[35] Moreover, a review of early phase LLM validation trials suggests that $N \geq 25$ per prompt group is usually sufficient to detect qualitative and distributional differences without overfitting interpretative frameworks.[21] In its original technical meaning, overfitting refers to a model that explains the training data too precisely, capturing noise or idiosyncratic details rather than generalisable patterns. In our context, the chosen sample size therefore balances statistical sensitivity to relevant differences in completeness, empirical guidance on thematic saturation, and the feasibility of detailed mixed methods analysis.

Data Generation Procedure

Three researchers, all co-authors of this study (MP, IC, and MJ), independently participated in the data generation process. Using ChatGPT-4.0 Omni (May 2024 release) accessed via the ChatGPT Plus interface, each researcher generated 20 outputs (10 per prompt type), yielding a total of 60 outputs. To avoid carryover effects and ensure the independence of the responses, every output was generated in a new chat session. The model operated under the default provider settings with no personalisation, parameter tuning, or temperature manipulation applied. All 60 outputs were exported as plain text, assigned numeric identifiers, pooled across prompt types, and then randomised using a computer generated sequence so that the two raters received a single anonymised list with no indication of whether a given text originated from the PP or EP condition.

Outcome Measures

The study focused on three primary outcomes: content completeness, accuracy, and communication quality.

Content Completeness

Content completeness was defined as the proportion of benchmark criteria satisfied by each ChatGPT-generated output, calculated both as an overall score and within each of the six predefined domains of postoperative care. Two independent surgical nursing experts (co-authors of this study (TP and PM)), blinded to prompt type and output sequence, applied the benchmark checklist to each output, recording a binary judgement for each criterion (Fulfilled/Not Fulfilled). Any disagreements were resolved through consensus discussion.

Accuracy

Accuracy was defined as the absence of statements that were clinically incorrect, inconsistent with evidence-based practice or established clinical guidelines, or that posed a potential safety risk. Any such inaccuracies were documented verbatim by the expert panel, consisting of two co-authors of this study (TP and PM).

Communication Quality

The quality of communication was assessed using a directed qualitative content analysis approach.[36] Two trained coders (MP and MJ), both co-authors of this study, independently reviewed all ChatGPT-generated outputs based on EP-type prompts for predefined features related to relational and motivational communication. The analysis focused on six aspects: motivational language, emphasis on patient agency, normalisation strategies, tone of guidance, relational connection, and clarity of instruction. These categories captured both linguistic devices (eg, pronoun use, modality, exclamatives, conditional phrasing) and broader communicative strategies relevant to patient-centred written education (see Supplementary Materials 3).

To complement the qualitative assessment, a quantitative text-mining analysis was conducted to measure the frequency of affirmatives, modal verbs, and empathetic expressions across all ChatGPT-generated outputs. A dictionary-based approach was applied to ensure transparency and interpretability. Frequencies were extracted at the document level and normalised per 1000 words.

Based on these counts, a Composite Relationality Index (CRI) was constructed for each output as the sum of affirmatives, empathetic expressions, and modal verbs divided by total word count and multiplied by 1000. The CRI served as a continuous indicator of overall linguistic relational load. Detailed preprocessing procedures, dictionary construction, and the formal definition of the CRI are provided in the Supplementary Materials 4. No inter-coder agreement coefficients or inferential statistical tests were calculated; quantitative findings were used descriptively and interpreted in parallel with the qualitative coding.

Statistical Analysis

Quantitative analysis was performed on 60 outputs (30 per type) with Jamovi statistical software (version 2.5; The Jamovi Project).[37] No between-session temperature manipulation or parameter tuning was undertaken beyond the default provider settings. Stochastic variation was accounted for through replication across independent sessions and non-parametric inference methods.

The inter-expert reliability for categorical ratings of 43 detailed features of the AI-generated outputs was assessed using Cohen's kappa coefficient, with corresponding 95% confidence intervals (CIs) calculated to quantify the precision of the agreement estimates. Feature completeness between outputs generated with EP-type and PP-type prompts was compared using Fisher's exact test, selected for its suitability in contexts involving small sample sizes and low expected cell counts. Analyses of completeness for the 43 individual benchmark features were regarded as exploratory and p values for these item level comparisons were not adjusted for multiple testing; interpretation therefore focuses on the pattern and magnitude of effect estimates rather than on isolated p values. Differences in domain-level summary scores across outputs were examined using the Brunner–Munzel test, a nonparametric alternative to the *t*-test that does not require the assumption of equal variances or identical distribution shapes between groups. In addition, for each Brunner Munzel comparison we calculated the relative effect of EP versus PP together with its 95% CIs. All statistical tests were two-tailed, and the level of statistical significance was set at $\alpha = 0.05$.

Ethical Considerations

This study did not involve human participants, patient records, or any identifiable personal data. All analysed materials were generated by ChatGPT and evaluated by expert reviewers serving in a professional capacity, with no personal or health-related information collected. As such, these activities did not constitute research involving human participants and therefore did not require review or approval by a research ethics committee.

Results

Inter-Expert Reliability

Inter-expert reliability was found to be exceptionally high across the dataset. The mean Cohen's kappa coefficient (κ) calculated across all 43 detailed features was 0.859 (95% CI [0.841, 0.878]), indicating a very high level of agreement between experts. Of the 2580 comparisons made in total, only 177 (6.86%) were identified as disagreements, yielding an overall agreement rate of 93.14%.

When disaggregated by prompt type, the mean kappa for outputs generated with the PP-type was 0.857 (95% CI [0.834, 0.880]), whereas for outputs generated with the EP-type it was 0.856 (95% CI [0.823, 0.889]). The near-identical values and overlapping confidence intervals indicate that the type of prompt used did not significantly affect inter-expert reliability.

Inter-expert agreement was analysed within six educational domains. High levels of reliability were observed across all the domains, with the highest kappa reported for Medication Management (κ = 0.897 (95% CI [0.746, 1.047])), and the lowest for Dietary Recommendations (κ = 0.762 (95% CI [0.588, 0.936])). Furthermore, the inter-expert rating variability (SD) was greatest for Dietary Compliance.

Content Completeness

Table 2 shows content completeness of 43 detailed postoperative self-care features in the ChatGPT-generated outputs, confirming high overall coverage with minimal variation between outputs generated using different prompt types. Core

Table 2 Completeness of Content Across 43 Detailed Self-Care Features in ChatGPT-4 Omni Outputs Generated with Primary Prompt (PP) and Emotional Prompt (EP) Types

No.	Detailed Feature	Overall [%]	EP-Type [%]	PP-Type [%]	Δ	*p*-value*
	Medication Management					
1	Paracetamol or NSAIDs (name, dose, frequency)	100.0	100.0	100.0	0.0	1.000
2	Gradual tapering of analgesics	95.0	96.7	93.3	3.4	1.000
3	Antibiotics (amoxicillin / metronidazole)	25.0	26.7	23.3	3.4	1.000
4	Advice not to discontinue antibiotic therapy	96.7	96.7	96.7	0.0	1.000
5	As-needed antiemetics	85.0	90.0	80.0	10.0	0.472
6	Instructions for missed doses	91.7	90.0	93.3	−3.3	1.000
7	No self-adjustment of antibiotic/antiemetic doses	91.7	90.0	93.3	−3.3	1.000
	General Health Monitoring					
8	Body-temperature monitoring twice daily	91.7	93.3	90.0	3.3	1.000
9	Morning blood-pressure monitoring	96.7	96.7	96.7	0.0	1.000
10	Morning pulse monitoring	96.7	96.7	96.7	0.0	1.000
11	Abdominal-pain self-assessment (NRS)	86.7	86.7	86.7	0.0	1.000
12	Monitoring of respiratory status	86.7	90.0	83.3	6.7	0.706
13	Monitoring of urinary output	100.0	100.0	100.0	0.0	1.000
14	Monitoring of scleral colour (jaundice)	91.7	93.3	90.0	3.3	1.000
15	Monitoring of skin condition (jaundice / pruritus)	86.7	90.0	83.3	6.7	0.706

(*Continued*)

Table 2 (Continued).

No.	Detailed Feature	Overall [%]	EP-Type [%]	PP-Type [%]	Δ	*p*-value*
	Dietary Recommendations					
16	Easily light, low-fat diet	100.0	100.0	100.0	0.0	1.000
17	Small, frequent meals	100.0	100.0	100.0	0.0	1.000
18	Bread & cereal— recommended/restricted	3.3	3.3	3.3	0.0	1.000
19	Fats — recommended/restricted	23.3	23.3	23.3	0.0	1.000
20	Meat & cold cuts — recommended/restricted	18.3	23.3	13.3	10.0	0.506
21	Fish — recommended/restricted	13.3	13.3	13.3	0.0	1.000
22	Dairy products — recommended/restricted	16.7	20.0	13.3	6.7	0.731
23	Vegetables — recommended/restricted	3.3	3.3	3.3	0.0	1.000
24	Fruit — recommended/restricted	5.0	6.7	3.3	3.4	1.000
25	Beverages — recommended/restricted	31.7	43.3	20.0	23.3	0.095
26	Desserts & sweets — recommended/restricted	0.0	0.0	0.0	0.0	1.000
27	Condiments & spices — recommended/restricted	3.3	3.3	3.3	0.0	1.000
28	Food-preparation techniques	85.0	76.7	93.3	−16.6	0.146
29	Gradual reintroduction of fibre	41.7	40.0	43.3	−3.3	1.000
30	Fat-soluble vitamin supplementation	0.0	0.0	0.0	0.0	1.000
31	Adaptation period	30.0	30.0	30.0	0.0	1.000
32	Monitoring of gastrontestinal symptoms	33.3	33.3	33.3	0.0	1.000
33	Hydration	90.0	90.0	90.0	0.0	1.000
	Wound Care					
34	Monitoring of wound pain / surrounding skin	96.7	96.7	96.7	0.0	1.000
35	Monitoring of surgical site (wound)	96.7	96.7	96.7	0.0	1.000
36	When / where to change dressing or remove sutures	95.0	96.7	93.3	3.4	1.000
	Daily Life Organisation					
37	Limit physical exertion for ≤ 2 weeks post-op	100.0	100.0	100.0	0.0	1.000
38	Gradual increase in physical activity (weeks 2–6)	100.0	100.0	100.0	0.0	1.000
39	Full activity resumption after 6 weeks	100.0	100.0	100.0	0.0	1.000
40	Return-to-work recommendations (manual / office-based)	85.0	86.7	83.3	3.4	1.000
	Emergency Contact Information					
41	Moderate symptoms → primary care	85.0	83.3	86.7	−3.4	1.000
42	Severe complications → surgical ward / clinic	98.3	96.7	100.0	−3.3	1.000
43	Life-threatening symptoms → ED / 112	100.0	100.0	100.0	0.0	1.000

Notes: Δ represents the difference in completeness between outputs generated with EP and PP types, expressed in percentage points EP minus PP. *p values from Fisher exact test.

components, such as guidance on analgesic use, low-fat diet, vital sign monitoring, and emergency procedures, were present in all outputs (100%). However, several elements were included inconsistently. Only 25.0% of outputs addressed full antibiotic prescription details, while specific dietary components (eg, vegetables, fruits, cereals) were each covered in fewer than 25.0% of cases. Similarly, information on fibre introduction (41.7%), adaptation periods (30.0%), and symptom monitoring (33.3%) was only occasionally provided. Wound care and physical activity guidance were nearly comprehensive across both prompt-generated output types, whereas return-to-work instructions and recommendations for managing moderate symptoms, such as low-grade fever, mild gastrointestinal disturbances, or non-severe pain requiring primary care consultation, appeared in 85.0% of cases. No statistically significant differences in completeness were found between outputs generated with EP-type prompts and those generated with PP-type prompts for any feature (all $p \geq .05$, Fisher's exact test).

Brunner-Munzel inferential testing revealed no statistically significant differences in outputs generated with EP- or PP-type prompts across any of the assessed domains. As shown in Table 3, all p-values exceeded the conventional alpha level of 0.05, and the relative effect estimates were consistently close to 0.50, indicating a lack of substantial differences attributable to the type of prompt used.

Accuracy Analysis

An expert review, conducted using a structured benchmark form to record any concerns, revealed no factual inaccuracies, recommendations inconsistent with evidence-based guidelines, or potentially harmful instructions in any of the ChatGPT-generated outputs, regardless of the prompt type. Notably, no evidence of hallucination was observed in ChatGPT-4.0 Omni (May 2024 release). Across all outputs reviewed, no content was flagged as problematic beyond the previously described omissions in completeness. This suggests that, although certain informational was lacking, the output was consistently accurate, safe, and conforming to current clinical standards.

Table 3 Comparison of ChatGPT-4 Omni Outputs Generated with Emotional Prompt (EP) and Primary Prompt (PP) Types Across Self-Care Domains

Domain	Prompt Type	Mean	SD	Statistic	p-value*	Relative Effect	95% CI	
							Lower	Upper
Medication adherence	EP	5.90	0.662	−0.644	0.501	0.46	0.33	0.59
	PP	5.80	0.664					
General condition monitoring	EP	7.47	0.730	−1.050	0.340	0.43	0.30	0.56
	PP	7.27	0.785					
Dietary compliance	EP	6.10	1.749	−0.641	0.517	0.45	0.31	0.60
	PP	5.87	1.717					
Wound care	EP	2.90	0.305	−0.396	0.432	0.48	0.40	0.57
	PP	2.87	0.346					
Daily life organization	EP	3.87	0.346	−0.356	0.472	0.48	0.40	0.57
	PP	3.83	0.379					
Contact emergency information	EP	2.80	0.407	0.684	0.737	0.53	0.43	0.64
	PP	2.87	0.346					
Overall	EP	29.03	2.312	−0.819	0.414	0.44	0.29	0.59
	PP	28.50	2.446					

Note: *Brunner-Munzel test.
Abbreviations: SD, standard deviation; 95% CI, 95 confidence interval.

Communication Quality

As summarised in Table 4, the directed qualitative content analysis showed that EP-type prompts enhanced the relational and motivational tone of the ChatGPT-generated educational materials by introducing motivational language, reinforcing patient agency, and employing normalising statements that framed recovery in a supportive and patient-centred manner. Although the outputs remained procedurally accurate, EP-type prompting systematically strengthened the relational framing of the guidance.

The complementary quantitative text-mining results are presented in Table 5 (EP-type prompt). In these outputs, modal verbs constituted the dominant structural element of the discourse, while affirmatives and empathetic expressions occurred less frequently but contributed to substantial variability in relational density, as reflected by the CRI (range 0.00–5.28). In contrast, 26 of 30 PP-type outputs showed CRI = 0.00, indicating an almost complete absence of lexically detectable relational features. Only four outputs generated with PP-type prompts contained any relational markers, all with low CRI values (maximum 3.14). Together, the qualitative and quantitative findings consistently demonstrate that EP-type prompting increases both the presence and configurational complexity of relational linguistic features in the generated materials. A detailed description of the text-mining procedures and dictionary construction is provided in the Supplementary Materials 4.

Table 4 Linguistic and Communicative Markers of Agency, and Relational Tone in ChatGPT-4 Omni Outputs Generated with Primary Prompt (PP) and Emotional Prompt (EP) Types

Aspect	PP-Type	EP-Type	Linguistic Markers
Motivational language	Absent or mainly neutral	Present (eg, "You've got this!", "I believe in your strength".)	Exclamative, present tense verbs (*zdrowiejesz* [*you are getting better*]), affirmative modality (*Masz wszystkie narzędzia* [*you have all the tools*])
	"This leaflet provides detailed information about your recovery after gallbladder removal surgery".	"I believe that following these recommendations will help you feel better with each passing day".	
Emphasis on agency	Instructional, externally directed	Empowering (eg, "You can manage this recovery successfully".)	Use of 2nd person pronouns (*Ci* [you]), affirmatives (*jesteś w stanie* [*you can*]), repetition of *pamiętaj* [*remember*] as a supportive reminder
	"Activity progression: from week 2 to week 6 you may gradually increase the intensity of your walks and introduce gentle stretching and breathing exercises".	"Remember, you are capable of taking care of your health and returning to full strength".	
Normalisation strategies	Minimal and embedded in predominantly informational content	Frequent reassurance (eg, "It's normal to feel tired".)	Conditional structures (*Jeśli…* [*if…*]), epistemic modality (*pomoże Ci* [*will help you*]), normalising statements contextualising recovery
	"The following recommendations describe what to expect during the recovery period after your surgery".	"Remember that recovery is a process that takes time, but I believe that with your determination you will start to feel better soon".	
Tone of guidance	Procedural, sometimes technical	Supportive and patient-centred (eg, "Taking your medication… is part of caring for yourself")	Softened deontic modality (*warto* [*it is worth*]), normative phrasing (*ważne jest* [*it is important to*]), relational sequencing (affirmative opening/closing).
	"For the first two weeks avoid strenuous activities, such as lifting heavy objects, doing major cleaning or performing other intense household chores".	"Reduce physical effort during the first two weeks after the procedure. Avoid lifting heavy objects, doing strenuous housework, or making sudden vigorous movements".	
Relational connection	Functional, detached	Engaging and partnership-oriented	Direct address with personal pronouns, expressions of appreciation (*Gratuluję* [*congratulations*]), declarations of trust (*wierzę w Twoją determinację* [*I believe in your determination*])
	"Dear Patient, you have undergone surgery to remove your gallbladder".	"Dear Patient, congratulations on having your gallbladder removal surgery. Now it is crucial that you take care of yourself so that you can make a full recovery".	
Instruction clarity	Clear but impersonal	Clearer with relational and motivational framing	Use of simple, colloquial but medically accurate phrasing; alignment of instructions with self-care (*część dbania o siebie* [*part of taking care of yourself*])
	"Paracetamol: take no more than eight 500 mg tablets in 24 hours, with doses spaced every four to six hours".	"You can take paracetamol every 4 to 6 hours, but do not exceed a total dose of 4 g per day, that is, a maximum of eight 500 mg tablets in 24 hours".	

Table 5 Frequency of Relational Linguistic Features and Composite Relationality Index (CRI) in EP-Type ChatGPT-Generated Outputs

Output	Words	Affirmatives	Empathetic Expressions	Modal Verbs	CRI (per 1000 Words)
31	661	1	0	1	3.03
32	702	0	0	2	2.85
33	729	0	1	0	1.37
34	728	0	0	0	0.00
35	731	0	1	1	2.74
36	740	0	0	1	1.35
37	760	0	0	1	1.32
38	745	0	0	2	2.68
39	734	0	0	1	1.36
40	746	0	0	2	2.68
41	750	0	0	1	1.33
42	756	0	0	2	2.65
43	742	0	0	2	2.70
44	737	0	0	1	1.36
45	748	0	0	2	2.67
46	751	0	0	2	2.66
47	755	0	0	1	1.32
48	747	0	0	2	2.68
49	743	0	0	1	1.35
50	754	0	0	2	2.65
51	742	0	0	1	1.35
52	739	0	0	2	2.71
53	751	0	0	2	2.66
54	748	0	0	1	1.34
55	746	0	0	2	2.68
56	753	0	1	2	3.99
57	749	0	0	1	1.34
58	744	0	0	2	2.69
59	752	0	0	1	1.33
60	758	0	2	2	5.28

Note: Higher values od CRI indicate greater overall density of relational linguistic features.
Abbreviations: CRI, Composite Relationality Index = (affirmatives + empathetic expressions + modal verbs) / word count × 1000.

https://doi.org/10.2147/JMDH.S566966

Discussion

This study examined whether the quality of discharge instructions generated by ChatGPT-4.0 Omni for patients after laparoscopic cholecystectomy could be improved by emotional prompt engineering. The outputs demonstrated high content completeness consistently covering essential domains such as analgesia, wound care, and recognition of concerning postoperative symptoms (eg, fever, jaundice, wound infection, severe or persistent abdominal pain, or urinary abnormalities). However, detailed dietary advice and antibiotic guidance were underrepresented. Importantly, there was no difference in content between outputs generated with PP- and EP-type prompts. Similarly, expert review confirmed that all materials were factually correct, aligned with evidence-based guidelines, and free from potentially harmful statements, with no systematic differences observed between the two prompt types in terms of accuracy. By contrast, the analysis of communication quality revealed clear distinctions: outputs generated with EP-type prompts incorporated a relational and motivational tone, with reinforcement of patient agency, normalising strategies, and clearer, more supportive phrasing, than outputs generated with PP-type prompts.

At the same time, the quantitative analysis of individual criteria showed marked variability in completeness across domains. While core elements such as analgesia, basic wound care and general advice on mobilisation were present in all outputs, only one quarter of the instructions contained full antibiotic prescription details, fewer than half explicitly mentioned gradual fibre reintroduction, and less than one third specified an adaptation period or systematic monitoring of gastrointestinal or other warning symptoms. Coverage of individual dietary product groups was below one quarter of outputs. These findings indicate that, despite overall high completeness at the level of broad domains, the model frequently omitted clinically relevant details that are important for safe self-management, particularly in the areas of pharmacotherapy and structured dietary guidance. From a clinical perspective, incomplete information on antibiotic regimens may contribute to suboptimal adherence or inappropriate use of leftover medication, and insufficiently specified dietary advice may increase the risk of gastrointestinal discomfort or anxiety in patients who are unsure which foods are safe in the early recovery period.

The absence of measurable improvements in the completeness of outputs generated with EP-type prompts is consistent with previous findings that informational accuracy is primarily model-driven.[21,38] While the addition of these relational features did not increase informational completeness, a function largely determined by model architecture and training data, it contributed to the clarity, accessibility, and motivational impact of the discharge instructions. In the context of patient education, these findings suggest that, while they do not alter factual accuracy, outputs generated with EP-type prompts may foster stronger patient engagement and encourage adherence to postoperative guidance.

Importantly, the Brunner Munzel analysis of domain level completeness did not reveal statistically significant differences between PP and EP outputs, with relative effect estimates close to 0.50 across domains. This pattern supports the interpretation that emotional prompting primarily reshapes the relational and motivational framing of information rather than the probability that specific ERAS based criteria are included. In other words, EP improved how the information was said but did not change what was said, leaving the observed gaps in antibiotic and dietary guidance quantitatively unchanged between prompt types.

Building on earlier research into postoperative education, our findings show that nurse-led, multimodal strategies embedded in ERAS protocols, such as video animations, telephone coaching, and teach-back, improve comprehension, reduce complications, and shorten hospital stays.[9–13] In this study, outputs generated by ChatGPT-4.0 Omni were comparable in terms of completeness across key domains, including analgesia, wound care, mobilisation, diet, and recognition of red-flag symptoms. Although readability was not formally assessed, the relational and motivational features more frequently observed in outputs generated with EP-type prompts, such as motivational language, reinforcement of patient agency, and normalising statements, suggest a potential role in overcoming comprehension difficulties related to high literacy demands.

Beyond these qualitative impressions, the text mining analysis provided quantitative evidence that emotional prompting systematically increased the relational density of the outputs. The CRI, combining the normalised frequencies of affirmatives, modal verbs, and empathetic expressions, showed a wide dispersion in EP outputs with CRI values extending up to 5 points, indicating frequent and varied use of relational language. In contrast, almost all PP outputs had

a CRI of zero, and only a small minority contained any detectable relational markers, all at low levels. This pattern confirms that emotional prompting not only changed the perceived tone of the text but also measurably increased the presence and complexity of linguistically identifiable relational features, in line with theoretical expectations derived from Social Cognitive Theory and patient centred communication models.

However, when these results are interpreted alongside the domain specific completeness data, it becomes clear that LLM generated instructions cannot currently substitute for nurse led education in areas where detailed, individualised guidance is required. In particular, the low frequency of explicit statements on antibiotic regimens and stepwise diet advancement contrasts with ERAS recommendations that emphasise clear instructions on medication use, early but structured oral intake and explicit warning signs that should prompt patients to seek help. This gap reinforces the need for human review to supplement and correct model outputs in precisely those domains where omissions are most likely to have clinical consequences.

These results are consistent findings from a recent pilot study, in which ChatGPT-generated brochures were generally well-received, but required clinical editing to ensure accuracy and readability.[14] This highlights both the potential and the limitations of LLMs: while they can generate substantial foundational content, they cannot replace human oversight.[20,39] Importantly, unlike the demonstrated benefits of nurse-led relational communication in improving recall and trust, outputs generated with emotional prompts did not demonstrate enhanced content completeness. Therefore, future research should examine whether qualitative differences in relational and motivational features observed in AI-generated outputs translate into improved patient-reported outcomes, such as engagement, understanding, and adherence. It should also involve direct patient feedback, particularly from individuals with low health literacy, to validate the usability and clarity of Polish-language LLM-generated outputs. In addition, the protocol should be extended with readability metrics and testing (eg, Flesch–Kincaid Test, Gunning Fog Index) in other inflectional languages to assess the generalisability of these findings.

This study has important implications for both clinical practice and the evaluation of LLMs in patient education. Clinically, the ability of ChatGPT-4.0 Omni to generate outputs that consistently address core postoperative domains suggests its potential as a tool in settings with limited resources, brief discharge encounters, or after-hours communication.[20,39,40]

In practical terms, these findings argue for positioning LLMs as a drafting aid within nurse led discharge workflows rather than as an autonomous source of patient instructions. A plausible pathway would involve generating a preliminary, ERAS aligned discharge leaflet from a standardised prompt template, followed by structured review and tailoring by the responsible nurse, who would verify antibiotic regimens, diet progression and red flag symptoms against institutional protocols before co-signing the final version with the patient. Such use embedded in electronic health record templates or patient portals and constrained by locally approved prompt libraries could support consistency and efficiency, while clear documentation that the nurse has checked and adapted the AI generated text would maintain professional accountability and medico legal traceability.[19,38,41,42] This supervised integration mirrors existing models of nurse led education, with the LLM providing a starting point for content generation and the nurse retaining responsibility for clinical accuracy, individualization and confirmation of patient understanding.

Yet, as emotional prompts influenced the relational and motivational features of the outputs rather than their informational completeness, content-driven prompt design remains critical. Relational-motivational phrasing should be explicitly integrated with clinically essential domains to minimise the risk of omission of key details, such as antibiotic regimens or dietary restrictions. Standardised prompt templates, aligned with institutional guidelines and embedded in electronic health record systems or patient-facing applications, may provide a scalable way of producing accurate and adaptable educational materials.[41,42]

Methodologically, the study demonstrates the value of benchmark-based evaluation as a framework for structured assessment of LLM-generated outputs. Such protocols provide a reproducible framework for early-phase testing across surgical contexts and complement, but do not replace, direct patient feedback.[41,42] Effective implementation will depend on interdisciplinary collaboration among clinicians, communication experts, and AI specialists to ensure both clinical adequacy and high standards of relational and motivational communication. Ultimately, the integration of generative AI into health care should follow a human-in-the-loop model, whereby AI-generated outputs are reviewed and

contextualised by healthcare professionals. This hybrid approach may strike an optimal balance between automation, personalisation, and patient safety.[21]

Limitations

This study has several limitations. It relied exclusively on expert evaluations of AI generated outputs assessed against a benchmark derived from ERAS protocols, without direct input from patients; thus, perceived clarity, relational motivational impact, and practical utility remained unassessed. A key limitation is that the 43-item ERAS based benchmark was developed using expert consensus focused mainly on content validity and has not yet undergone full psychometric validation, including formal indices of content validity or dimensional structure. Future research should extend the evaluation of this instrument by assessing additional reliability metrics, examining construct validity in diverse clinical settings, and confirming its stability across different surgical populations. The study examined only two prompt formats and used a modest sample size n equal 30 per type, limiting sensitivity to subtle effects and reducing generalisability to other clinical contexts. Furthermore, all outputs were generated using a single version of the model ChatGPT 4.0 Omni, May 2024 release, at one point in time. As commercial LLMs are continuously updated, temporal reproducibility is uncertain and future outputs may vary due to model drift, even when identical prompts are used. A further limitation is that we did not conduct formal readability analyses or direct comprehension testing with patients, so the ease with which patients can understand and act on the generated discharge instructions remains uncertain. Finally, because additional dimensions such as cultural appropriateness and real-world usability were not assessed, future studies should adopt a broader, multidimensional approach.

Conclusion

In conclusion, this study demonstrated that ChatGPT-4.0 Omni has the potential to generate discharge instructions for patients after laparoscopic cholecystectomy, with high completeness across key domains. Nonetheless, important gaps remain, particularly regarding detailed dietary recommendations, antibiotic use, and structured guidance on monitoring concerning symptoms. These findings highlight that LLMs should be considered supportive tools rather than substitutes for professional oversight. To ensure safe and effective patient care, LLM-generated content must be integrated into structured, nurse-led education, which constitutes a fundamental component of postoperative discharge communication.

Generative AI Statement

Generative artificial intelligence tools (ChatGPT, OpenAI, San Francisco, CA, USA) were used to assist in language editing of this manuscript. All content was subsequently reviewed and verified by the authors, who take full responsibility for the final version of the text.

Data Sharing Statement

Consensus rating data for completeness assessments are openly available in the Zenodo repository at https://doi.org/10.5281/zenodo.16886947, under the Creative Commons Attribution (CC BY 4.0) licence. All other data generated or analysed during this study are available from the corresponding author on reasonable request.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Li ZZ, Guan LJ, Ouyang R, Chen ZX, Ouyang GQ, Jiang HX. Global, regional, and national burden of gallbladder and biliary diseases from 1990 to 2019. *World J Gastrointestinal Surg*. 2023;15(11):2564–2578. doi:10.4240/wjgs.v15.i11.2564
2. Shaffer EA. Epidemiology and risk factors for gallstone disease: has the paradigm changed in the 21st century? *Curr Gastroenterol Rep*. 2005;7 (2):132–140. doi:10.1007/s11894-005-0051-8
3. Eurostat surgical operations and procedures performed in hospitals by ICD-9-CM. Available from: https://ec.europa.eu/eurostat/databrowser/view/HLTH_CO_PROC2__custom_3199068/default/table?lang=en. Accessed August 14, 2025.
4. Nawacki Ł, Kozłowska-Geller M, Wawszczak-Kasza M, Klusek J, Znamirowski P, Głuszek S. Iatrogenic injury of biliary tree-single-centre experience. *Int J Environ Res Public Health*. 2022;20(1):781. doi:10.3390/ijerph20010781
5. Fletcher R, Cortina CS, Kornfield H, et al. Bile duct injuries: a contemporary survey of surgeon attitudes and experiences. *Surg Endosc*. 2020;34 (7):3079–3084. doi:10.1007/s00464-019-07056-7
6. Nair A, Al-Aamri HHM, Borkar N, Rangaiah M, Haque PW. Application of enhanced recovery after surgery pathways in patients undergoing laparoscopic cholecystectomy with and without common bile duct exploration: a systematic review and meta-analysis. *Sultan Qaboos Univers Med J*. 2023;23(2):148–157. doi:10.18295/squmj.1.2023.005
7. Morales-Conde S, Peeters A, Meyer YM, et al. European association for endoscopic surgery (EAES) consensus statement on single-incision endoscopic surgery. *Surg Endosc*. 2019;33(4):996–1019. doi:10.1007/s00464-019-06693-2
8. Society of American Gastrointestinal and Endoscopic Surgeons. Gallbladder removal surgery (Cholecystectomy) patient information from SAGES. Available from: https://www.sages.org/publications/patient-information/patient-information-for-laparoscopic-gallbladder-removal-cholecystectomy-from-sages/. Accessed August 14, 2025.
9. Maheta B, Shehabat M, Khalil R, et al. The effectiveness of patient education on laparoscopic surgery postoperative outcomes to determine whether direct coaching is the best approach: systematic review of randomized controlled trials. *JMIR Perioper Med*. 2024;7:e51573. doi:10.2196/51573
10. Tian YQ, Lv XF, Zhang M, Gao M, Zhao L. Implementation of pathway-based care for patients undergoing daytime cholecystectomy. *Patient Prefer Adherence*. 2025;19:383–392. doi:10.2147/ppa.S492763
11. Bollschweiler E, Apitzsch J, Obliers R, et al. Improving informed consent of surgical patients using a multimedia-based program? Results of a prospective randomized multicenter study of patients before cholecystectomy. *Ann Surg*. 2008;248(2):205–211. doi:10.1097/SLA.0b013e318180a3a7
12. Da Silva Schulz R, Santana RF, Dos Santos CTB, et al. Telephonic nursing intervention for laparoscopic cholecystectomy and hernia repair: a randomized controlled study. *BMC Nurs*. 2020;19(1):38. doi:10.1186/s12912-020-00432-y
13. Sadeghi N, Salari N, Jalali R. Effect of multimedia education on anxiety and pain in patients undergoing laparoscopic cholecystectomy: a Solomon four-group randomized controlled trial. *Sci Rep*. 2025;15(1):9357. doi:10.1038/s41598-024-77207-x
14. De Rouck R, Wille E, Gilbert A, Vermeersch N. Assessing artificial intelligence-generated patient discharge information for the emergency department: a pilot study. *Int J Emerg Med*. 2025;18(1):85. doi:10.1186/s12245-025-00885-5
15. Parente DJ. Generative artificial intelligence and large language models in primary care medical education. *Fam Med*. 2024;56(9):534–540. doi:10.22454/FamMed.2024.775525
16. Bui D, Nakamura C, Bray BE, Zeng-Treitler Q. Automated illustration of patients instructions. *AMIA Annu Symp Proc*. 2012;2012:1158–1167.
17. Choi JY, Yoo TK. Evaluating ChatGPT-4o for ophthalmic image interpretation: from in-context learning to code-free clinical tool generation. *Informatics Health*. 2025;2(2):158–169. doi:10.1016/j.infoh.2025.07.002
18. Halaseh FF, Yang JS, Danza CN, Halaseh R, Spiegelman L. ChatGPT's role in improving education among patients seeking emergency medical treatment. *West J Emerg Med*. 2024;25(5):845–855. doi:10.5811/westjem.18650
19. Ye C, Malin BA, Fabbri D. Leveraging medical context to recommend semantically similar terms for chart reviews. *BMC Med Inform Decis Mak*. 2021;21(1):353. doi:10.1186/s12911-021-01724-2
20. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Network Open*. 2024;7(3):e240357. doi:10.1001/jamanetworkopen.2024.0357
21. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7(1):258. doi:10.1038/s41746-024-01258-7
22. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. *Interact J Med Res*. 2024;13: e54704. doi:10.2196/54704
23. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. *Front Med Lausanne*. 2024;11:1477898. doi:10.3389/fmed.2024.1477898
24. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *Npj Digital Med*. 2024;7(1):41. doi:10.1038/s41746-024-01029-4
25. Li C, Wang J, Zhang Y, et al. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:230711760*. 2023. doi:10.48550/arXiv.2307.11760.
26. Miller WR, Rollnick S. *Motivational Interviewing: Helping People Change*. Guilford press; 2012.
27. Bandura A. Health promotion by social cognitive means. *Health Educ Behav*. 2004;31(2):143–164. doi:10.1177/1090198104263660
28. Resnicow K, McMaster F. Motivational Interviewing: moving from why to how with autonomy support. *Int J Behav Nutr Phys Act*. 2012;9(1):19. doi:10.1186/1479-5868-9-19
29. Cole SA, Sannidhi D, Jadotte YT, Rozanski A. Using motivational interviewing and brief action planning for adopting and maintaining positive health behaviors. *Prog Cardiovasc Dis*. 2023;77:86–94. doi:10.1016/j.pcad.2023.02.003

30. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6). doi:10.3390/healthcare11060887
31. Panczyk M, Piątek T, Małkowski P, Jaworski M, Cieślak I, Gotlib-Małkowska J. Post-cholecystectomy home care information: optimizing with advanced prompt engineering using ChatGPT – a comparative study of empathy, accuracy, and content completeness. Presented at: *International Council of Nurses Congress*. Helsinki, Finland; 2025. Session Potential of AI in nursing.
32. Microsoft Learn. Prompt engineering techniques. Available from: https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/prompt-engineering?tabs=chat. Accessed August 16, 2025.
33. Karch JD. Psychologists should use brunner-munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Adv Methods Pract Psychol Sci*. 2021;4(2):2515245921999602. doi:10.1177/2515245921999602
34. Happ M, Bathke AC, Brunner E. Optimal sample size planning for the Wilcoxon-Mann-Whitney test. *Stat Med*. 2019;38(3):363–375. doi:10.1002/sim.7983
35. Yan LK, Niu Q, Li M, et al. Large language model benchmarks in medical tasks. *arXiv preprint arXiv:241021348*. 2024. doi:10.48550/arXiv.2410.21348.
36. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277–1288. doi:10.1177/1049732305276687
37. Şahin M, Aybek E. Jamovi: an easy to use statistical software for the social scientists. *Int J Assessment Tools Educ*. 2019;6(4):670–692. doi:10.21449/ijate.661803
38. Marey A, Saad AM, Tanas Y, et al. Evaluating the accuracy and reliability of AI chatbots in patient education on cardiovascular imaging: a comparative study of ChatGPT, gemini, and copilot. *Egypt J Radiol Nucl Med*. 2025;56(1):37. doi:10.1186/s43055-025-01452-x
39. Rust P, Frings J, Meister S, Fehring L. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Commun Med*. 2025;5(1):208. doi:10.1038/s43856-025-00927-2
40. Huang T, Safranek C, Socrates V, et al. Patient-representing population's perceptions of GPT-generated versus standard emergency department discharge instructions: randomized blind survey assessment. *J Med Internet Res*. 2024;26:e60336. doi:10.2196/60336
41. Tung JYM, Gill SR, Sng GGR, et al. Comparison of the quality of discharge letters written by large language models and junior clinicians: single-blinded study. *J Med Internet Res*. 2024;26:e57721. doi:10.2196/57721
42. Ganzinger M, Kunz N, Fuchs P, et al. Automated generation of discharge summaries: leveraging large language models with clinical data. *Sci Rep*. 2025;15(1):16466. doi:10.1038/s41598-025-01618-7