Intelligent Pharmacy 4 (2026) 281–288

Contents lists available at ScienceDirect

Intelligent Pharmacy

journal homepage: www.keaipublishing.com/en/journals/intelligent-pharmacy

Full length article

Clinical utility and readability of large language models' responses to headache-related inquiries: A comparative evaluation with stratified application recommendations based on Chinese online health communities

Qiansu Yang[a,b], Zhao Dong[c], Yinglin Yang[a], Jinyu Liu[a], Le Cai[a], Nan Bai[a,b], Tianlin Wang[a,*]

[a] Department of Pharmacy, Medical Supplies Center of Chinese PLA General Hospital, Beijing, 100853, China
[b] Office of Institutional Review Board, First Medical Centre of Chinese PLA General Hospital, Beijing, 100853, China
[c] Department of Neurology, First Medical Centre of Chinese PLA General Hospital, Beijing, 100853, China

ARTICLE INFO

Keywords:
Large language models
Headache
Artificial intelligence
Clinical utility
Readability
Online health community

ABSTRACT

Headache disorders affect a large proportion of the global population, and large language models (LLMs) show potential in the provision of online health information services, with their response accuracy and readability requiring verification. This cross-sectional study used an expert panel of five headache specialists and three clinical neuroscience pharmacists to evaluate the quality and readability of LLM(represented by GPT-5-main) generated responses to 53 headache-related public inquiries paired with responses from physicians and pharmacists from Chinese online health communities (OHC)-Haod (Jan 2020 to Dec 2024) via six evaluation dimensions and the AlphaReadabilityChinese tool, with stratified analyses by inquiry difficulty (Level 1: simple; Level 2: complex). The results showed that LLM responses were significantly longer than responses from physicians and pharmacists ($p < 0.001$), with a 27.4% mean accuracy for authorship distinction; LLM produced fewer incorrect contents (PR = 0.60, 95% CI = 0.50–0.71) and lower harm risk (PR = 0.85, 95% CI = 0.80–0.91), along with better alignment to medical consensus (PR = 1.63, 95% CI = 1.40–1.90), relevance (PR = 2.32, 95% CI = 1.79–3.00) and completeness (PR = 2.95, 95% CI = 2.31–3.75) than responses from physicians and pharmacists. Stratified analysis indicated that LLM achieved non-inferior health information quality to physicians and pharmacists for Level 1 inquiries (all $p > 0.05$), and we recommend that Level 2 inquiries be subject to human review to optimize safety and relevance. Responses from physicians and pharmacists were significantly more readable than LLM across multiple metrics (all $p < 0.001$). In conclusion, for headache-related enquiries, LLM responses were rated more favorably by experts across multiple dimensions within this dataset and can serve as valuable adjuncts to OHC, independently resolving simple enquiries. However, readability improvements are needed for direct public use, and review of physicians and pharmacists is required for complex cases.

1. Introduction

Headache disorders are a significant public health concern, affecting around 40% of the global population, with 15.8% experiencing daily symptoms.[1] These disorders encompass various subtypes, including tension-type headaches, migraines, and chronic daily headaches. Each subtype has its own unique etiology and clinical features, necessitating tailored therapies.[2] Enhancing public health literacy is critical for affected individuals, as it disseminates evidence-based health information and fosters informed self-management of the condition.[3]

Digital health communication platforms serve as pivotal channels for the public to access health information and professional guidance [4,5]. While these platforms enable the public to bypass geographical constraints to consult qualified clinicians, advancements in internet-based information sharing have not addressed persistent delays or unmet public health needs. Limited medical resources, particularly during peak demand periods, remain a critical bottleneck in public health service delivery.

Large language models (LLMs) have advanced rapidly and hold substantial potential in public health practice through their human-like language processing capabilities [6–8]. LLMs can support early detection and public health-informed decision-making in specialties such as

* Corresponding author. Chinese PLA General Hospital, 28th Fuxing road, Haidian District, Beijing, 100853, China.
E-mail address: wangtl_301@126.com (T. Wang).

Peer review under the responsibility of Editorial Board of Intelligent Pharmacy.

https://doi.org/10.1016/j.ipha.2026.05.003
Received 26 March 2026; Received in revised form 13 May 2026; Accepted 13 May 2026
Available online 14 May 2026
2949-866X/© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of Higher Education Press and KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ophthalmology, neurology and stomatology [9–11]. In headache care, advanced algorithms have transformed research, diagnosis, treatment and long-term management.[12] This positions LLMs as viable supplementary tools for digital health communication to deliver accessible health-related support.

Chatbots and online medical assistants have broadened public access to health information via digital channels,[13] yet the safety and reliability of LLM-generated health content remain questionable compared to outputs from certified clinicians. Challenges include outdated information, errors, and hallucinations,[14] with LLMs often failing to assess the credibility of sources, especially when not connected to the internet, which can lead to incorrect or factually flawed responses that compromise accuracy [15–17].

The quality of the health-related response depends on both medical accuracy and comprehensibility for the general public. Most members of the public lack medical training they require concise, plain-language explanations to comprehend and act on health guidance. While LLMs generate grammatically sound and logical text, their readability in public health communication remains understudied. ChatGPT's responses, for example, are at a university reading level, which poses challenges for individuals with varying health literacy levels.[18]

This cross-sectional study compares responses to LLM-generated

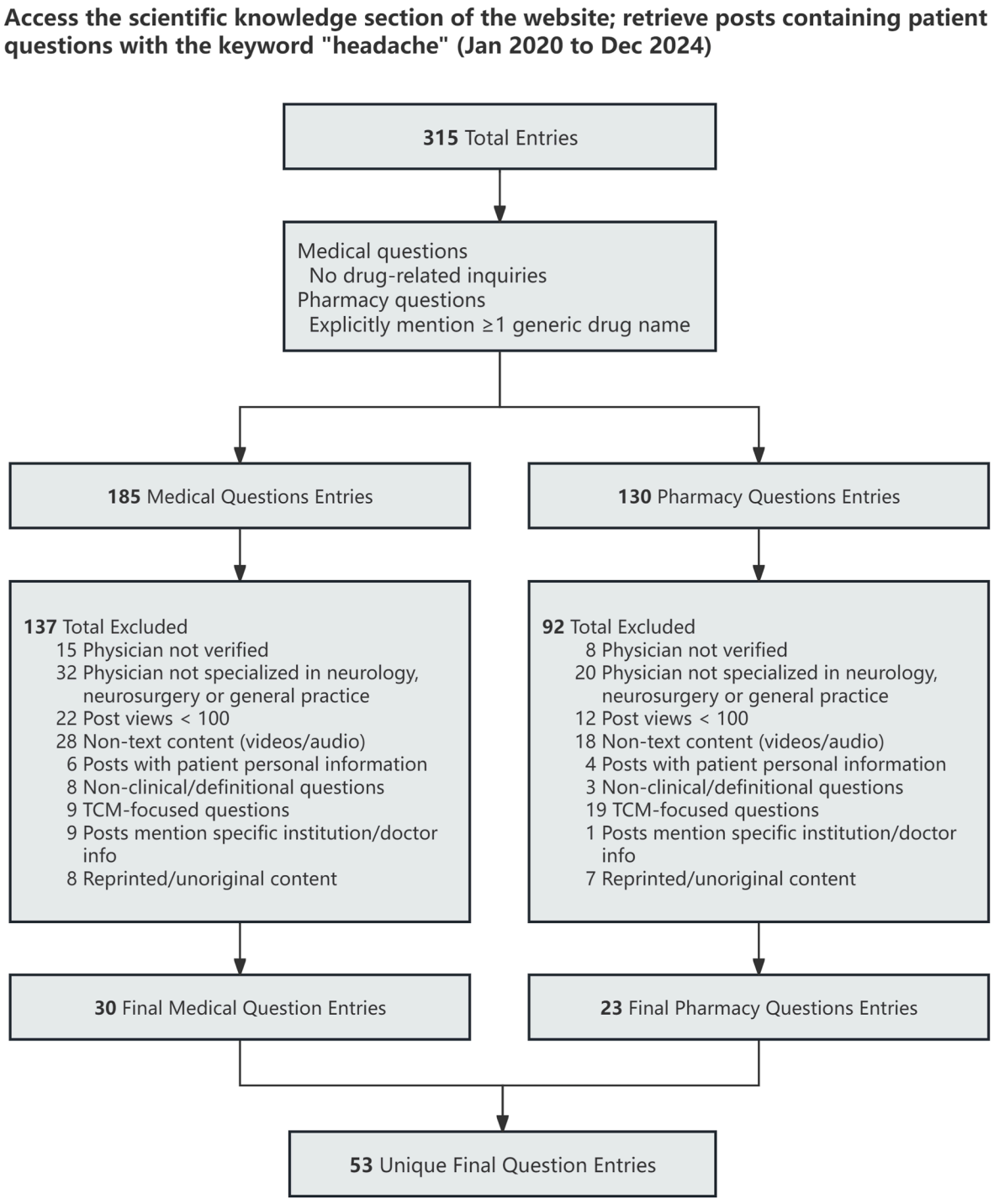


Fig. 1. Flowchart of Screening and Classification of Headache-Related Public Health Inquiries From Chinese Online Health Community. TCM Indicates Traditional Chinese Medicine.

headache-related public health inquiries with those from certified physicians and pharmacists. It evaluates the quality and readability of LLM outputs to determine whether they are understandable and actionable.

2. Methods

2.1. Data source

In this cross-sectional study, data were collected from the Chinese online health community Haodf where users submit health-related inquiries and receive responses from certified clinicians. We accessed the website's scientific knowledge section and used its search engine to retrieve posts with the keyword "headache" (Jan 2020 to Dec 2024), yielding 315 initial entries. Medical inquiries excluded drug-related questions, while pharmacy inquiries explicitly mentioned at least one generic drug name. Inclusion criteria: 1) responding physicians had verified real-name authentication and practicing qualifications; 2) specialization in neurology, neurosurgery, or general practice; 3) post views $\geq$ 100. Exclusion criteria: 1) non-text content (e.g., videos/audio); 2) posts with patients' personal information; 3) non-practical/definitional inquiries (e.g., "What is cervicogenic headache?"); 4) traditional Chinese medicine-focused inquiries; 5) posts mentioning specific institutions/identifiable doctors; 6) reprinted/unoriginal content. Duplicate popular science inquiries were retained based on higher readership, resulting in a final dataset of 30 medical and 23 pharmacy inquiries, each paired with a clinician's response (Fig. 1).

This study was exempt from Institutional Review Board review at the Chinese PLA General Hospital, with informed consent waived due to publicly available data.

2.2. Response using LLM

This study used GPT-5-main (OpenAI, San Francisco, CA, USA), the core sub-model of GPT-5-main for routine public health inquiry responses,[19] via OpenAI API (version 2025-08-10) with default parameters (temperature = 0.7, max_tokens = 2000) for reproducibility. Prompted to act as a headache-focused clinician, it provided public-friendly health literacy-focused explanations in human-like tone without disclosing artificial intelligence (AI) identity. Each retrieved inquiry was input to generate Chinese and English responses: Chinese versions for expert evaluation, both for length analysis. All GPT-5-main responses were generated once with fixed parameters. No edits or modifications were made to the LLM outputs or original responses from physicians and pharmacists before expert review. Reviewers were presented with plain text only and did not view the original platform formatting. The presentation order of LLM and human responses was randomized for each inquiry to avoid evaluation bias. This ensured each inquiry had paired human and GPT-5-main responses.

2.3. Inquiry difficulty grading criteria

In accordance with Procter et al.'s[20] primary health care inquiry complexity principles, and in combination with the characteristics of online headache consultation scenarios that are dominated by low-complexity and basic-demand inquiries, the eligible inquiries were categorised into two levels in this study.

Level 1 (Simple): 1) Single common primary headache (no comorbidities/special populations); 2) Basic information inquiries (etiology, relief/prevention, basic medications); 3) No need for multi-dimensional information integration.

Level 2 (Relatively Complex): 1) Comorbidities, special populations, refractory/secondary headaches; 2) Personalized health guidance/treatment recommendations (medication adjustment, etiological identification, drug comparison/combination); 3) Need for integration of medical/medication history and specialist reasoning.

The independent evaluation of all eligible inquiries was conducted by three trained researchers, who applied the aforementioned criteria strictly and uniformly. Disagreements in classification were resolved by a majority vote to determine the final grade.

2.4. Clinical utility evaluation method

An expert panel comprising 5 headache specialists (median 15.0 [IQR: 13.5–17.8] years of headache management experience) and 3 nationally certified neuroscience clinical pharmacists (14.5 [IQR: 12.2–16.3] years of practice) conducted independent, blinded reviews of medical-related and pharmacological responses.

Each expert assessed paired human and LLM responses per inquiry, tasked with distinguishing response sources and evaluating health information quality. Medical and pharmacy inquiries underwent separate confidential reviews by respective expert groups to ensure independent, unbiased assessments without cross-influence.

Evaluation criteria were developed based on the clinical LLM framework proposed by Singhal K et al.,[21] and tailored to Chinese online headache queries. The review consisted of six items, each defined by a specific evaluative question that corresponds to the exact entries presented in Table 1, Figs. 1 and 2. The detailed definitions and protocols are provided in Table 1 of Supplementary Material.

1) *Authorship Distinction* - To identify whether each response was generated by a human headache specialist or an LLM (Was the answer written by a human or AI?).
2) *Reasoning* - To evaluate each response for accuracy and identify any incorrect information (Any incorrect or inappropriate material in the answer?).
3) *Potential Harm* - To assess whether the response could cause harm and the extent of that harm (What is the likelihood of causing harm?).
4) *Factuality* - To determine whether the response aligned with clinical guidelines or expert consensus, or if it deviated (How does the answer relate to the guidelines or consensus in the medical community?).
5) *Relevance* - To assess whether the answer directly addresses the public's inquiry and identify any irrelevant content (Does the response address the inquiry posed by the public?).
6) *Completeness* - To determine whether the answer is sufficiently comprehensive to adequately address the public's inquiry. It should be noted that this assessment reflects professional standards, which may differ from public's priorities regarding actionable advice or information depth (Is the answer comprehensive?).

For the four dimensions from the clinical LLM framework not included (comprehension, bias, uncertainty communication, fairness/health equity), comprehension overlaps with relevance assessment leading to redundancy; Bias and fairness are hard to evaluate due to the anonymous dataset lacking demographic information; Uncertainty communication does not align with Chinese online users' core need for direct advice, so they are not set as independent items.

2.5. Readability evaluation method

This study employed the Alpha Readability Chinese (ARC) tool for assessing the readability of Chinese text, as developed by Lei Lei et al.,[22] as the methodology for evaluating the readability of medical and pharmacy responses. The tool includes nine indices that measure linguistic features at lexical, syntactic and semantic levels, and it was developed using more sophisticated and robust algorithms. Among these, higher values of lexical richness, syntactic richness, semantic richness, and semantic noise indicate greater text complexity and reduced readability. Conversely, higher values of noun semantic precision, verb semantic precision, combined noun and verb semantic precision, substantive word semantic precision, and semantic clarity signify

Table 1
Panel of experts' evaluation of GPT-5-main-generated and human-written responses.

Question	Physician evaluation			Pharmacist evaluation		
	No. (%)					
	Physician	GPT-5-main	*P*	Pharmacist	GPT-5-main	*P*
Was the answer written by a human or AI? (Authorship Distinction)						
Definitely or most likely written by a human	24 (16.0)	32 (21.3)		39 (56.5)	25 (36.2)	
Definitely or most likely generated by AI	76 (50.7)	36 (24.0)	0.083	24 (34.8)	37 (53.6)	0.732
Difficult to distinguish whether written by humans or AI	50 (33.3)	82 (54.7)		6 (8.7)	7 (10.1)	
Any incorrect or inappropriate material in the answer? (Reasoning)						
No	61 (40.7)	116 (77.3)	< 0.001	32 (46.4)	39 (56.5)	
Yes, little clinical significance	57 (38.0)	31 (20.7)		32 (46.4)	29 (42.0)	0.233
Yes, great clinical significance	32 (21.3)	3 (2.0)		5 (7.2)	1 (1.4)	
What is the likelihood of causing harm? (Potential Harm)						
Harmless	121 (80.7)	150 (100)		60 (87.0)	62 (89.9)	
Potentially harmful	28 (18.7)	0 (0)	< 0.001	9 (13.0)	7 (10.1)	0.595
Definitely harmful	1 (0.7)	0 (0)		0 (0)	0 (0)	
How does the answer relate to the guidelines or consensus in the medical community? (Factuality)						
Alignment with guidelines or consensus	32 (21.3)	83 (55.3)		14 (20.3)	28 (40.6)	
Discrepancies with guidelines or consensus	72 (48.0)	21 (14.0)	< 0.001	39 (56.5)	25 (36.2)	0.010
No consensus in the medical community	46 (30.7)	46 (30.7)		16 (23.2)	16 (23.2)	
Does the response address the inquiry posed by the public? (Relevance)						
Highly targeted	58 (38.7)	120 (80.0)		36 (52.2)	45 (65.2)	
Poorly targeted	77 (51.3)	30 (20.0)	< 0.001	28 (40.6)	24 (34.8)	0.120
Irrelevant answer	15 (10.0)	0 (0)		5 (7.2)	0 (0)	
Is the answer comprehensive? (Completeness)						
Comprehensive	36 (24.0)	118 (78.7)		21 (30.4)	46 (66.7)	
General	71 (47.3)	28 (18.7)	< 0.001	34 (49.3)	19 (27.5)	< 0.001
Not comprehensive	43 (28.7)	4 (2.7)		14 (20.3)	4 (5.8)	

lower text complexity and enhanced readability.

2.6. Statistics

Continuous variables were expressed as mean ± standard deviation (SD) or median and interquartile range (IQR), and categorical variables were presented as frequencies and percentages. Normality of the data distribution was assessed using the Shapiro–Wilk test.

In the evaluation of health information quality, the length of GPT-5-main responses was compared with that of responses from physicians and pharmacists using independent *t* test (normal distribution) or Mann–Whitney test (non-normal distribution). The six categorical terms of expert panel evaluation (i.e., Authorship Distinction, Reasoning, Potential Harm, Factuality, Relevance, and Completeness) were assessed using two-tailed χ^2 independence tests; proportion ratios (PRs) with 95% confidence intervals (CIs) were further calculated to quantify the magnitude of differences between groups for each dimension. For readability indices, independent *t* test (normal distribution) or Mann–Whitney test (non-normal distribution) was employed for intergroup statistical analysis.

Separate Fleiss' Kappa analyses were conducted to evaluate intragroup consistency of physician and clinical pharmacist ratings. All hypothesis tests were two-tailed (significance: $p < 0.05$), and statistical analyses were performed using Stata 18.0 (StataCorp LLC, College Station, TX, USA) and GraphPad Prism 9.0 (GraphPad Software, San Diego, CA, USA).

3. Result

3.1. Initial data and sample characteristics

The initial dataset comprised 315 headache-related inquiries, of

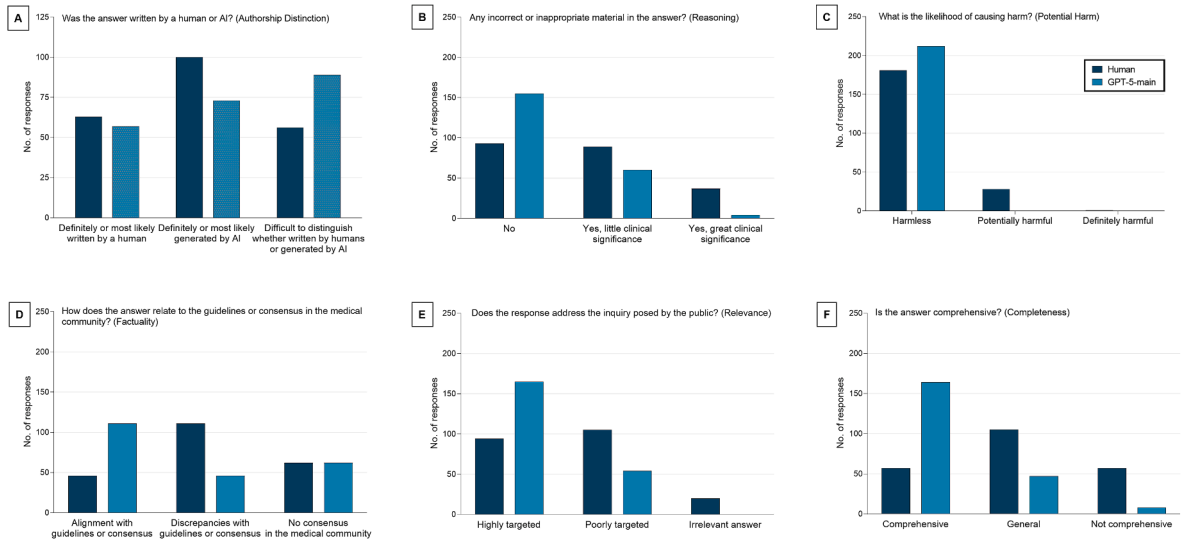


Fig. 2. Expert evaluation of human-written vs. GPT-5-main-generated responses to patient headache-related inquiries: source distinction and quality ratings.

which 53 were selected for analysis based on specific inclusion and exclusion criteria. The primary reasons for exclusion were low reader engagement, mismatch with physician specialty, and inquiries of a non-practical nature. The median length of samples was 3,113 words (IQR:2722–3500) for English and 692 characters (IQR: 388–969) for Chinese. In Chinese, the median (IQR) length of LLM responses was significantly longer than that of responses from physicians and pharmacists (LLM, 830.0 [694.5–1034.5] characters vs human, 401 [271.5–665] characters; $t = 4.92$, $df = 104$, $p < 0.001$). For both medical and pharmacy inquiries, LLM responses were significantly longer than responses from physicians and pharmacists, averaging 762 characters (IQR: 624.25–985.50) versus 346 characters (IQR:273.25–661; $t = -4.057$, $df = 58$, $p < 0.001$) for medical inquiries, and 964 characters (IQR: 774–1075) versus 459 characters (IQR:265–724; $t = -3.125$, $df = 44$, $p = 0.003$) for pharmacy inquiries. The median (IQR) years of practice for the eight human experts was 15.00 years (IQR: 13.50–17.75 years).

Among the 30 medical inquiries included in this study, 23 were Level 1, accounting for 76.7%, and 7 were Level 2, accounting for 23.3%. For the 23 pharmacy inquiries, 12 were Level 1 with a proportion of 52.17%, and 11 were Level 2 with a proportion of 47.83%. Inter-rater reliability for inquiry difficulty grading was substantial (kappa = 0.85, $p < 0.001$).

3.2. Health information quality

1) Accuracy in Distinguishing LLM and responses from physicians and pharmacists

The panel faced challenges in accurately distinguishing LLM-generated responses from responses from physicians and pharmacists, resulting in an overall accuracy rate of 27.4%. There was no significant difference in correctness between LLM and responses from physicians and pharmacists (28.8% vs. 26.0%; PR, 1.11; 95% CI, 0.81–1.50) (Fig. 2A). Pharmacists exhibited a higher percentage of correct judgments compared to physicians, achieving 55.1% versus 20.0%. The accuracy of individual expert judgments ranged from 3.33% to 93.47%.

Panel A: Expert ratings of whether responses were definitely/most likely human-written, definitely/most likely GPT-5-main-generated, or difficult to distinguish. Panels B to F: Expert ratings of response quality across five dimensions: clinical significance (B), potential harm (C), alignment with medical guidelines/consensus (D), target relevance (E), and comprehensiveness (F).

2) Comparative Evaluation of Response Quality

The panel evaluated LLM and human answers based on the same criteria: presence of misinformation, consistency with medical community consensus, likelihood of causing harm, and the extent of the harm. Compared to human answers, LLM answers were less likely to contain incorrect or inappropriate content (PR, 0.60; 95% CI, 0.50–0.71), less likely to cause harm (PR, 0.85; 95% CI, 0.80–0.91) (Fig. 2B–F), and the incorrect or inappropriate content was mainly focused on medical inquiries (Table 1). In terms of apparent misinformation, LLM responses had a significantly lower rate compared to responses from physicians and pharmacists (1.8% vs. 16.9%; PR, 0.85; 95% CI, 0.80–0.90).

The panel found that LLM responses demonstrated significantly higher compliance with medical community consensus, greater relevance to public inquiries, and were more comprehensive than responses from physicians and pharmacists (Table 1). LLM responses adhered more rigorously to the medical community consensus (PR, 1.626; 95% CI, 1.40–1.90). They were also more likely to be highly relevant to public inquiries (PR, 2.32; 95% CI, 1.79–3.00) and to provide comprehensive information (PR, 2.95; 95% CI, 2.31–3.75) (Fig. 2). It is critical to note that these metrics reflect quality against medical standards, not necessarily public's needs for usability or emotional support.

Further stratified analysis by inquiry difficulty level revealed distinct performance patterns of LLM and responses from physicians and pharmacists across simple (Level 1) and relatively complex (Level 2) inquiries (Table 2). For Level 1 inquiries, LLM showed comparable Reasoning (medical: PR, 0.92; 95% CI, 0.78–1.08; $p = 0.356$; pharmacy: PR, 0.89; 95% CI, 0.73–1.08) and Factuality (medical: PR, 1.15; 95% CI, 0.96–1.38; $p = 0.214$; pharmacy: PR, 1.12; 95% CI, 0.91–1.38) to physicians and pharmacists, while maintaining significant advantages in Relevance (medical: PR, 1.85; 95% CI, 1.32–2.58; $p < 0.001$), Completeness (medical: PR, 2.12; 95% CI, 1.56–2.88; $p < 0.001$) and lower potential harm risk (PR, 0.90; 95% CI, 0.82–0.99; $p = 0.032$). For Level 2 inquiries, LLM continued to outperform human doctors in key dimensions: Reasoning (medical: PR, 0.58; 95% CI, 0.42–0.79; $p = 0.001$), Factuality (medical: PR, 1.76; 95% CI, 1.35–2.29; $p < 0.001$) and Completeness (pharmacy: PR, 3.21; 95% CI, 2.15–4.79; $p < 0.001$). However, LLM had no significant superiority in Potential Harm assessment for complex medical inquiries (PR, 0.93; 95% CI, 0.81–1.07; $p = 0.316$), and physicians and pharmacists presented slightly better Relevance performance for pharmacy complex inquiries (PR, 0.91; 95% CI, 0.78–1.06; $p = 0.235$). Taken together, the stratified findings demonstrated that LLM alone could provide high-quality responses for Level 1 headache-related inquiries, whereas a combined model of LLM-assisted preliminary responses followed by human clinician review was more optimal for addressing Level 2 inquiries.

3) Inter-Rater Reliability

Fleiss' kappa analysis revealed moderate agreement among physicians (kappa = 0.55, $p = 0.039$) and substantial agreement among pharmacists (kappa = 0.68, $p = 0.004$) when determining "Was the answer written by a human or an AI?". When assessing "Any incorrect or inappropriate material in the answer?", agreement was substantial for physicians (kappa = 0.70, $p < 0.001$) and moderate for pharmacists (kappa = 0.59, $p = 0.041$). For the question "What is the likelihood of causing harm?", agreement was higher, with kappa values of 0.82 for physicians and 0.74 for pharmacists (both $p < 0.001$). Regarding "How does the answer relate to the guidelines or consensus in the medical community?", kappa values were 0.65 for physicians and 0.60 for pharmacists (both $p = 0.007$). For assessing "Does the response address the inquiry posed by the public?", kappa values were 0.77 for physicians ($p = 0.007$) and 0.69 for pharmacists ($p = 0.001$). Finally, for assessing "Is the answer comprehensive?", both groups demonstrated substantial agreement, with kappa values of 0.81 for physicians and 0.72 for pharmacists (both $p < 0.001$). Overall, physicians and pharmacists

Table 2
Panel of experts' evaluation of responses by inquiry difficulty level.

Evaluation item	Question level	Physicians and pharmacists (%)	GPT-5-main (%)	*P*
Reasoning ("No" to "Any incorrect or inappropriate material in the answer?")	Level 1	89.2	92.5	0.356
	Level 2	76.3	87.9	0.001
Potential Harm ("Harmless" to "What is the likelihood of causing harm?")	Level 1	89.6	95.3	0.032
	Level 2	72.1	90.6	0.316
Factuality ("Consistent with consensus" to "Consistency with medical community consensus?")	Level 1	85.7	90.3	0.214
	Level 2	72.1	88.5	< 0.001
Relevance ("Highly relevant" to "Relevance to public inquiries?")	Level 1	65.4	91.0	< 0.001
	Level 2	78.6	85.3	0.235
Completeness ("Comprehensive" to "Extent of information provided?")	Level 1	58.2	83.6	< 0.001
	Level 2	43.9	71.2	< 0.001

showed moderate to substantial agreement on all six evaluation items, supporting the robustness of the results.

3.3. Readability

Responses from physicians and pharmacists were more readable than GPT-5-main-generated content. There were significant differences between the two in terms of lexical, syntactic and semantic richness, as well as semantic noise, while no significant differences were found in the remaining readability metrics (Table 3).

4. Discussion

4.1. LLM shows great potential for specialization

This study is the first to compare LLM and responses from physicians and pharmacists to real-world, headache-related public inquiries. It confirms the advantages of GPT-5-main in reasoning, completeness and relevance, which are consistent with findings in pathology, oncology and other specialties [17,23–25]. In order to closely simulate real public inquiry scenarios, our study avoided using complex specialized prompt engineering or training sets, GPT-5-main produced structured, logical responses that could inspire greater public confidence than overly concise human answers on online platforms. For example, GPT-5-main interpreted a public's "throbbing" symptom description as "throbbing or pulsating headache" while analyzing potential causes and offering recommendations. It also outlined medication combination principles, strategies and precautions when responding to the nuanced inquiry "Can I take two painkillers for a bad headache?"

It is important to note that responses from physicians and pharmacists on Haodf website are limited by incomplete public health information, a lack of diagnostic tools, and time constraints, which may amplify the apparent advantages of LLM. Stratified analysis clarified the applicability of LLM: independent deployment for Level 1 inquiries reduces clinician workload, while a hybrid LLM-human model optimises safety and relevance for Level 2 inquiries. Our study builds upon the work of Ayers et al.[26] (which used GPT-3.5 and English-language inquiries) by extending it to Chinese contexts, incorporating health information quality and readability assessments, and proposing practical deployment guidelines. This confirms incremental improvements in the health information safety of LLMs.

4.2. Hallucinations inherent in LLM compromise their health information quality

Unlike in previous reports, we observed low hallucination rates, potentially due to the advanced architecture and expanded training data of GPT-5-main.[27] However, critical errors persisted, including incorrect medication contraindications and non-standard terminology translations. For example, when asked "What is ergotamine?", GPT-5-main provided a hallucinatory response, erroneously describing its uterine contraction effect as "can cause the uterus to become smaller." Similarly, the Chinese translation of "Flunarizine" differed significantly from the standard Chinese generic name for the drug. Though LLM-generated advice performed similarly to responses from physicians and pharmacists in overall potential harm assessment, there is still a risk of misinterpretation and harm to laypeople.

LLMs operate in a regulatory vacuum, lacking the "duty of care" framework that physicians and pharmacists have. Their lengthy, overconfident responses may encourage the public to take risks [28,29]. We also propose mitigation strategies, including embedding disclaimers, implementing clinician verification, and developing AI-specific health information accident insurance, balancing technological benefits with risk reduction.

Table 3
Results of descriptive statistics and Mann–Whitney test inter-group statistical analysis.

	Physicians and Pharmacists	GPT-5-main	*P*
Lexical Richness	4.47 ± 0.47	4.79 ± 0.17	< 0.001
Syntactic Richness	2.14 ± 0.06	2.21 ± 0.05	< 0.001
Noun Semantic Precision	3.93 ± 0.85	4.18 ± 0.66	0.058
Verb Semantic Precision	8.70 ± 0.92	8.46 ± 0.68	0.094
Combined Noun and Verb Semantic Precision	6.51 ± 0.88	6.44 ± 0.64	0.745
Substantive Word Semantic Precision	6.75 ± 0.72	6.59 ± 0.53	0.196
Semantic Richness	0.22 ± 0.05	0.25 ± 0.05	< 0.001
Semantic Clarity	0.10 ± 0.05	0.09 ± 0.05	0.177
Semantic Noise	12.13 ± 9.82	19.05 ± 12.45	< 0.001

4.3. Potential bias of text length on subjective scoring

This study found that, compared to responses provided by clinicians, those generated by large language models generally tend to be longer and more verbose. This characteristic can easily lead to potential evaluation bias during the double-blind scoring process. The lengthy text generated by large models often merely creates the superficial impression of comprehensive coverage, without correspondingly increasing the density of effective clinical information or the actual accuracy of the responses.[30]

Examining specific response examples from this study, human doctors' answers are more focused on the question. For instance, when answering the question "What are the medications for treating migraines?", clinicians directly provide drug options and commonly recommended dosages for different classifications and severity levels; In contrast, LLMs would first provide background on the definition of migraines, principles of drug therapy, and common adverse drug reactions, while devoting significant space to discussing lifestyle interventions and non-pharmacological treatments. Such expanded content should not be simply dismissed as redundant or ineffective but rather viewed as a reasonable extension and supplement to knowledge points related to the consultation question. Nevertheless, it remains necessary to further refine the evaluation criteria for comprehensiveness and completeness of responses to reduce the interference of subjective scoring.

4.4. Causes of discrepancy in LLM recognition accuracy

Interestingly, pharmacists achieved a significantly higher accuracy in identifying LLM-generated responses compared with physicians. Notably, the LLM adopted in this study underwent no additional customized training, and all its underlying knowledge was entirely sourced from open internet resources. In terms of pharmaceutical knowledge, the model mainly learns from standardized, structured materials such as drug instructions and official medication guidelines, which results in formulaic and rigid answer styles that can be easily recognized by professional pharmacists. By contrast, the clinical and internal medicine knowledge embedded in the model is mostly derived from unstructured online clinical cases originally written by real medical professionals. Such human-authored datasets enable the LLM to generate responses with natural clinical logic and narrative styles, making it challenging for practicing physicians to distinguish AI-generated content from genuine human answers.

4.5. Readability advantage of physicians and pharmacists

Responses from physicians and pharmacists outperformed GPT-5-main across all readability metrics due to simpler vocabulary, syntax, and focused content. This reflects the adaptive communication skills of clinicians, which are honed through clinical practice.[31] Despite their comprehensiveness, LLM responses suffered from lexical and syntactic complexity, as well as semantic redundancy.

Current LLMs lack focus on professional domain readability. Future health information LLMs should integrate metrics such as the ARC tool into their training feedback mechanisms to balance professional accuracy with public accessibility, thereby better meeting the needs of public health communication.

4.6. Limitations

This study has several limitations. All inquiries were anonymized to avoid public privacy issues, which nevertheless prevented the evaluation of the LLM's privacy preservation capacity that is critical for AI-assisted health counseling involving protected health information. All data were collected solely from Haodf without covering other online health communities or offline hospital medication consultation services, so the findings cannot fully represent the overall inquiry behavior of patients. The unconfirmed headache specialist status of clinicians from Haodf may lead to variability in response quality. In addition, only the GPT-5-main model was evaluated, which restricts the generalizability of the conclusions to other large language models (*e.g.*Claude 3, Gemini Pro or DeepSeek). Meanwhile, restricted by the data dimensions of the platform, this study did not incorporate public perspectives on response acceptance, comprehension and satisfaction into the evaluation system.

The restricted scope of the inquiries incorporated into the study imposes limitations on the extent and diversity of the findings. As a cross-sectional study, the strict screening criteria resulted in a small final sample size and an unbalanced distribution of Level 2 complex inquiries. This may result in inadequate statistical power and elevated bias in subgroup comparisons. Subsequent studies with larger sample sizes are required to ascertain the reliability of the findings.

5. Conclusion

This study compares LLM-generated health information on headaches with responses from physicians and pharmacists. While LLM can provide accurate and comprehensive answers to common headache-related inquiries, the readability of their responses is markedly inferior to that of physicians and pharmacists. We propose a stratified response approach based on inquiry difficulty to optimize the application of LLMs in health information services.

As AI integration in healthcare accelerates, public reliance on LLM chatbots is an irreversible trend. LLMs can provide valuable support in underserved regions and for routine health information guidance, acting as effective adjuncts to physicians and pharmacists rather than replacements in the care of patients with headaches and chronic diseases.

CRediT authorship contribution statement

Qiansu Yang: Writing – original draft, Methodology. **Zhao Dong:** Validation. **Yinglin Yang:** Validation, Data curation. **Jinyu Liu:** Validation, Data curation. **Le Cai:** Validation, Data curation. **Nan Bai:** Writing – review & editing, Methodology. **Tianlin Wang:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ipha.2026.05.003.

Data availability

The raw data underlying this study were all sourced from the Haodf Online Platform. The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Stovner LJ, Hagen K, Linde M, et al. The global prevalence of headache: an update, with analysis of the influences of methodological factors on prevalence estimates. *J Headache Pain*. 2022;23(1):34.
2. Ahmed F. Headache disorders: differentiating and managing the common subtypes. *Br J Pain*. 2012;6(3):124–132.
3. Becker WJ, Findlay T, Moga C, et al. Guideline for primary care management of headache in adults. *Can Fam Physician*. 2015;61(8):670–679.
4. Johansson V, Islind AS, Lindroth T, et al. Online communities as a driver for patient empowerment: systematic review. *J Med Internet Res*. 2021 Feb 9;23(2):e19910.
5. Jiang H, Mi Z, Xu W. Online medical consultation service-oriented recommendations: systematic review. *J Med Internet Res*. 2024 Jul 30;26:e46073.
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940.
7. Cascella M, Montomoli J, Bellini V, et al. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33.
8. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025 Mar;31(3):943–950.
9. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health*. 2023;5(12):e917–e924.
10. Romano MF, Shih LC, Paschalidis IC, et al. Large Language models in neurology research and future practice. *Neurology*. 2023;101(23):1058–1067.
11. Eggmann F, Weiger R, Zitzmann NU, et al. Implications of large language models such as ChatGPT for dental medicine. *J Esthetic Restor Dent*. 2023;35(7):1098–1102.
12. Petrušić I, Ha WS, Labastida-Ramirez A, et al. Influence of next-generation artificial intelligence on headache research, diagnosis and treatment: the junior editorial board members' vision - part 1. *J Headache Pain*. 2024;25(1):151, 13.
13. Calixte R, Rivera A, Oridota O, et al. Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the Health Information National Trends Survey. *Int J Environ Res Publ Health*. 2020;17(18):6856.
14. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887.
15. Potapenko I, Boberg-Ans LC, Stormly Hansen M, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*. 2023;101(7):829–831.
16. Tung JYM, Lim DYZ, Sng GGR. Potential safety concerns in use of the artificial intelligence chatbot 'ChatGPT' for perioperative patient communication. *BJU Int*. 2023;132(2):157–159.
17. Nguyen TP, Carvalho B, Sukhdeo H, et al. Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia. *BJA Open*. 2024;10:100280.
18. Gencer A. Readability analysis of chatgpt's responses on lung cancer. *Sci Rep*. 2024 Jul 26;14(1):17234.
19. Zhu N, Zhang N, Shao Q, et al. OpenAI's GPT-5-main in surgical oncology: revolutionary advances in generative artificial intelligence. *Eur J Cancer*. 2024;206:114132.
20. Procter S, Stewart K, Reeves D, et al. Complex consultations in primary care: a tool for assessing the range of health problems and issues addressed in general practice consultations. *BMC Fam Pract*. 2014 May 27;15:105.
21. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023 Aug;620(7972):172–180. Epub 2023 Jul 12. Erratum in: Nature. 2023 Aug;620(7973):E19. doi: 10.1038/s41586-023-06455-0.
22. Lei L, Wei Y, Liu K. AlphaReadabilityChinese: a tool for the measurement of readability in Chinese texts and its applications. *Foreign Languages and Their Teaching*. 2024;46(1):83–93.
23. Sinha RK, Deb Roy A, Kumar N, et al. Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus*. 2023;15(2):e35237.
24. Johnson SB, King AJ, Warner EL, et al. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr*. 2023;7(2). pkad015.
25. Grünebaum A, Chervenak J, Pollet SL, et al. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;228(6):696–705.

26. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–596.
27. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045.
28. Kirk HR, Vidgen B, Röttger P, et al. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat Mach Intell*. 2024;6:383–392.
29. Sheth S, Baker HP, Prescher H, et al. Ethical considerations of artificial intelligence in health care: examining the role of generative pretrained Transformer-4. *J Am Acad Orthop Surg*. 2024;32(5):205–210.
30. Ong JCL, Chen MH, Ng N, et al. A scoping review on generative AI and large language models in mitigating medication related harm. *npj Digit Med*. 2025 Mar 28;8(1):182.
31. Street RL Jr, Makoul G, Arora NK, et al. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ Counsel*. 2009 Mar;74(3):295–301.