Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability and readability of large language model chatbot responses to public-facing questions about hypertension, antihypertensive therapy, and erectile dysfunction: a cross-sectional comparative study

Junke Luo
Fengcheng People's Hospital
Shaoyong Chen
Csy19890301@163.com

Fengcheng People's Hospital

Research Article

Keywords: large language models, hypertension, antihypertensive therapy, erectile dysfunction, patient education, readability, reliability

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11332190/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Large language model (LLM) chatbots are increasingly used as public-facing sources of health information. This study compared the reliability, information quality, source transparency, and readability of four publicly accessible LLM chatbots responding to questions about hypertension, antihypertensive therapy, and erectile dysfunction (ED).

Methods

In this cross-sectional web-based study, 37 questions were retained from a 99-item pool informed by MeSH/PubMed, Google Trends, and recurring themes on publicly accessible Chinese online platforms. On May 22, 2026, each question was submitted once in a new conversation to ChatGPT 5.5, Microsoft Copilot, Gemini 3.5 Flash, and Perplexity, yielding 148 responses. Two blinded clinicians independently rated responses using DISCERN, EQIP, JAMA, and GQS; readability was assessed with six indices.

Results

Inter-rater agreement was excellent for DISCERN and EQIP (ICC(2,1) = 0.982 and 0.991) and almost perfect for JAMA and GQS (quadratic-weighted kappa = 0.851 and 0.92). Perplexity had the highest median DISCERN (47.00), EQIP (75.00), and JAMA (1.00) scores, with significant between-model differences (all $P < 0.001$). GQS did not differ significantly ($P = 0.055$; all medians 4.00). No model met the prespecified sixth-grade readability target; Copilot was descriptively easiest to read, but its median FKGL was 11.65 and FRES 38.60.

Conclusion

In this single-date benchmark, Perplexity achieved higher structured quality and transparency scores, whereas readability remained poor across all models. The findings support cautious use of general-purpose LLMs as supplementary rather than standalone patient-education sources and highlight the need for clearer source attribution and plain-language optimization.

1. Introduction

Hypertension is a major modifiable cardiovascular risk factor that requires sustained treatment and long-term self-management.[1,2] Erectile dysfunction (ED) is common in men with hypertension and is relevant to cardiovascular care because the two conditions share vascular and metabolic risk pathways. ED may also precede clinically apparent cardiovascular disease in some men, and current sexual-medicine

guidance places ED within cardiovascular risk assessment rather than treating it as an isolated symptom.[3–10]

Antihypertensive therapy adds a clinically important layer of uncertainty. Patients may attribute sexual symptoms either to hypertension itself or to blood pressure-lowering drugs, although antihypertensive classes differ in their associations with erectile function.[6–8] Questions about sildenafil or tadalafil, nitrate interactions, sexual activity, exercise, weight reduction, smoking, alcohol, psychological factors, and sleep apnea further link cardiovascular care with sexual health.[3,4,11–19] Because these topics are sensitive, patients may prefer to raise them through private online searches rather than during routine clinical encounters.

Large language model (LLM) chatbots now provide immediate, conversational health information, but fluent presentation does not ensure reliable sourcing, balanced content, or accessible language.[20–23] This distinction is important in cardiovascular patient education, where materials frequently exceed recommended reading levels and may be difficult for patients to understand.[24–28] The Chatbot Assessment Reporting Tool (CHART) emphasizes transparent reporting of model versions, prompting, access conditions, and evaluation procedures in studies of chatbot health advice.[29,30]

Existing studies have examined LLM-generated information for hypertension, home blood pressure monitoring, gestational hypertension, and ED, but few have evaluated the combined information needs created by hypertension, antihypertensive therapy, and ED within a single public-facing question set.[31–36] This intersection includes medication adverse-effect concerns, phosphodiesterase type 5 inhibitor use, nitrate interactions, sexual activity, and cardiovascular risk communication. We therefore compared four publicly accessible LLM chatbots on reliability, information quality, source transparency, and readability when answering these questions. Factual accuracy, guideline concordance, and clinical safety were outside the prespecified scope.

2. Methods

2.1. Study Design and Reporting Framework

This cross-sectional, web-based comparative study evaluated four publicly accessible LLM chatbots. Data were collected on May 22, 2026, and reporting was guided by CHART.[29,30] The analytic sample was determined by the question-development process: all 37 retained questions were submitted to all four models, yielding 148 response units (37 per model). Each chatbot response to one question constituted one response unit.

2.2. Development of the Public-Facing Question Set

MeSH terms centered on Hypertension, Antihypertensive Agents, Erectile Dysfunction, and Sexual Dysfunction, Physiological were used to define the topic. Public information needs were mapped with Google Trends (Worldwide, preceding five years) and recurring lay themes from Xiaohongshu, Baidu

Tieba, Weibo, and Zhihu; PubMed and current clinical guidance were used to check topic coverage. Public posts were used only to identify recurring themes and wording, without retaining usernames, direct links, identifiable information, or verbatim content.

The resulting 99-item candidate pool was independently screened by clinicians with complementary cardiology and andrology backgrounds. Questions were excluded if they were duplicate or near-duplicate, outside the study scope, overly technical for a public-facing set, commercial, requested unsupervised prescription changes, or required highly individualized clinical decisions. The 37 retained questions were harmonized into concise public-facing English and grouped into four domains: high blood pressure and erection problems (n = 9), blood pressure medicines and sexual function (n = 10), treating erection problems when high blood pressure is present (n = 9), and sex, lifestyle, and long-term health (n = 9). The complete screening record, including inclusion/exclusion decisions and reasons, is provided in Supplemental Digital Content 1 (Table S1).

2.3. Large Language Model Selection and Standardized Data Collection

ChatGPT 5.5, Microsoft Copilot, Gemini 3.5 Flash, and Perplexity were accessed through their public web interfaces with default settings on the study date. Each of the 37 questions was submitted once to each chatbot in a new conversation, producing 148 responses. The prompt consisted only of the question; no role assignment, length instruction, citation request, formatting instruction, or follow-up prompt was added. Model-specific system prompts, proprietary training data, and generation parameters were unavailable to the investigators and were not modified.

2.4. Reliability and Information-Quality Assessment

Reliability and information quality were assessed with four established instruments. DISCERN evaluates treatment information on a 16–75 scale and was classified as excellent (63–75), good (51–62), fair (39–50), poor (27–38), or very poor (16–26).[37] EQIP evaluates content and presentation across 20 binary items and was converted to a percentage score.[38] JAMA benchmark criteria assess authorship, attribution, disclosure, and currency on a 0–4 scale.[39] GQS provides a 1–5 global rating of quality, flow, and usefulness.[40] These instruments were not interpreted as direct measures of factual accuracy, guideline concordance, or clinical safety.

2.5. Evaluation Procedure

Two physicians from tertiary hospitals independently rated all 148 responses: a cardiologist with approximately 10 years of experience in hypertension management and an andrologist with approximately 10 years of experience in ED management. Model identity was masked during scoring. Approximately 10% of the question set was used for pilot calibration before full scoring. Disagreements were resolved by discussion, with senior adjudication when needed. Inter-rater reliability was calculated from the two independent ratings using ICC(2,1) for DISCERN and EQIP and quadratic-weighted Cohen's kappa for JAMA and GQS; 95% confidence intervals were estimated from 5,000 bootstrap resamples.

2.6. Readability Assessment and Text Preprocessing

For readability analysis, outputs were converted to plain text; formatting symbols, list markers, URLs, citation markers, and standalone reference lists were removed without changing substantive wording. Readability was measured with ARI, GFI, FKGL, Coleman-Liau Index (CL), SMOG, and FRES using Python 3.10.4 and textstat 0.7.13. The prespecified target was approximately sixth-grade readability (< 6 for the grade-level/complexity indices and > = 80 for FRES).[22,24–28,35,41,42] Readability differences between models were interpreted descriptively rather than as inferential rankings.

2.7. Statistical Analysis

Scores were summarized as medians with first and third quartiles. Between-model differences were tested with the Kruskal-Wallis test, with epsilon-squared as the omnibus effect size. Significant omnibus tests were followed by Dunn pairwise comparisons with Bonferroni adjustment. Readability was compared descriptively with the prespecified sixth-grade thresholds. Tests were two-sided with $P < 0.05$ considered statistically significant.

2.8. Use of generative artificial intelligence during manuscript preparation

The authors declare that generative artificial intelligence tools were used only for limited support in improving the language and translation of the manuscript. They were not used for data collection, response scoring, statistical analysis, interpretation of results, or formulation of scientific conclusions. All content was independently reviewed and verified by the authors, who assume full responsibility for the accuracy and integrity of the work.

3. Results

3.1. Question Set and Analytic Dataset

Screening retained 37 of 99 candidate questions and excluded 62. The four domains contained 9, 10, 9, and 9 questions, respectively. Submitting all retained questions to the four LLMs yielded 148 responses for analysis (Table 1).

Table 1
Public-facing questions included in the evaluation

Issue list
High blood pressure and erection problems
1. Can high blood pressure cause problems getting or keeping an erection?
2. Why can high blood pressure make it harder to get or keep an erection?
3. Are erection problems more common in men with high blood pressure?
4. Can erection problems be an early warning sign of heart or blood vessel problems?
5. Does having higher blood pressure, or having it for a long time, make erection problems more likely?
6. Can getting my blood pressure under better control improve my erections?
7. Can I have erection problems from high blood pressure even if I feel well otherwise?
8. How can I know whether my erection problems are related to high blood pressure or something else?
9. If I have both high blood pressure and erection problems, should I also be checked for things like diabetes, high cholesterol, or smoking-related risk?
Blood pressure medicines and sexual function
10. Can blood pressure medicine cause erection problems?
11. How can I tell whether my erection problem is from high blood pressure itself or from the medicine I take for it?
12. Are some blood pressure medicines more likely than others to affect erections?
13. Can medicines such as metoprolol affect erections or sexual performance?
14. Can water pills used for high blood pressure, such as hydrochlorothiazide, affect erections?
15. Can blood pressure medicines such as lisinopril or losartan affect erections?
16. Can amlodipine affect erections or sexual function?
17. If my erection problems started after I began a new blood pressure medicine, what should I ask my doctor?
18. Can changing my blood pressure medicine or its dose help with erection problems?
19. What should I do if I think my blood pressure medicine is affecting my erections?
Treating erection problems when you have high blood pressure
20. Can a man with high blood pressure take sildenafil (Viagra) for erection problems?
21. Can a man with high blood pressure take tadalafil (Cialis) for erection problems?

Issue list
22. Can sildenafil or tadalafil make my blood pressure drop too low when I am also taking blood pressure medicine?
23. Are there any heart or blood pressure medicines that should not be taken with sildenafil or tadalafil?
24. Why is it dangerous to take sildenafil or tadalafil with chest pain medicines such as nitroglycerin?
25. Is it okay to use medicine for erection problems if my blood pressure is still high or not well controlled?
26. Should I have my heart health checked before I start medicine for erection problems?
27. What other treatments are available if sildenafil or tadalafil do not work or I cannot take them?
28. Are over-the-counter or online 'male enhancement' products safe if I have high blood pressure?
Sex, lifestyle, and long-term health
29. Is sex safe for most men with high blood pressure?
30. Does sex temporarily raise blood pressure or heart rate?
31. Are there times when a man with high blood pressure should wait before having sex?
32. Can regular exercise help both high blood pressure and erection problems?
33. Can losing weight improve erection problems in men with high blood pressure?
34. Can smoking make both high blood pressure and erection problems worse, and can quitting help?
35. How does alcohol affect blood pressure and erections?
36. Can stress, anxiety, or poor sleep contribute to both high blood pressure and erection problems?
37. Can sleep apnea be related to both high blood pressure and erection problems?

3.2. Reliability, Information Quality, and Transparency

Inter-rater agreement was excellent for DISCERN (ICC(2,1) = 0.982, 95% CI 0.970–0.991) and EQIP (ICC(2,1) = 0.991, 95% CI 0.983–0.997), and almost perfect for JAMA (quadratic-weighted kappa = 0.851, 95% CI 0.770–0.931) and GQS (kappa = 0.92, 95% CI 0.661-1.000) (Supplemental Digital Content 2 (Table S2)).

Perplexity had the highest median DISCERN score, 47.00 (45.00–49.00), followed by ChatGPT 44.00 (38.00–47.00), Copilot 41.00 (37.00–45.00), and Gemini 39.00 (36.00–44.00); all medians were in the fair range (epsilon-squared = 0.195, $P < 0.001$). Perplexity scored higher than ChatGPT ($P = 0.034$), Copilot ($P < 0.001$), and Gemini ($P < 0.001$) (Tables 2 ; Tables 3; Fig. 1).

Table 2
Reliability, information-quality, transparency, and global-quality scores across large language model chatbots

Measure	ChatGPT	Copilot	Gemini	Perplexity	epsilon-squared	P
DISCERN	44.00 (38.00, 47.00)	41.00 (37.00, 45.00)	39.00 (36.00, 44.00)	47.00 (45.00, 49.00)	0.195	< 0.001
EQIP	60.00 (55.00, 75.00)	60.00 (60.00, 75.00)	55.00 (55.00, 60.00)	75.00 (75.00, 80.00)	0.307	< 0.001
JAMA	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	1.00 (1.00, 1.00)	0.517	< 0.001
GQS	4.00 (4.00, 4.00)	4.00 (4.00, 4.00)	4.00 (4.00, 4.00)	4.00 (4.00, 4.00)	0.032	0.055

Values are median (Q1, Q3). P values were obtained with the Kruskal-Wallis test; epsilon-squared denotes the Kruskal-Wallis effect-size estimate. DISCERN = DISCERN instrument; EQIP = Ensuring Quality Information for Patients; GQS = Global Quality Score; JAMA = Journal of the American Medical Association benchmark criteria; Q1 = first quartile; Q3 = third quartile.

EQIP also favored Perplexity, with a median of 75.00 (75.00–80.00) versus 60.00 (55.00–75.00) for ChatGPT, 60.00 (60.00–75.00) for Copilot, and 55.00 (55.00–60.00) for Gemini (epsilon-squared = 0.307, P < 0.001). Perplexity scored higher than all three comparators (all P < = 0.001), and Copilot scored higher than Gemini (P = 0.015). JAMA showed the largest between-model separation (epsilon-squared = 0.517, P < 0.001): Perplexity had a median of 1.00 (1.00–1.00), ChatGPT and Copilot 0.00 (0.00–1.00), and Gemini 0.00 (0.00–0.00). By contrast, GQS did not differ across models; all medians were 4.00 (4.00–4.00), corresponding to good overall quality (epsilon-squared = 0.032, P = 0.055).

Table 3
Post hoc pairwise comparisons across large language model chatbots

Pairwise comparison	DISCERN	EQIP	JAMA	GQS
ChatGPT-Copilot	1.000	1.000	0.818	1.000
ChatGPT-Gemini	0.065	0.125	< 0.001	0.113
ChatGPT-Perplexity	0.034	< 0.001	< 0.001	1.000
Copilot-Gemini	1.000	0.015	0.066	1.000
Copilot-Perplexity	< 0.001	0.001	< 0.001	1.000
Gemini-Perplexity	< 0.001	< 0.001	< 0.001	0.113

Cells show Bonferroni-adjusted P values from Dunn post hoc tests. GQS pairwise results are shown for completeness but are not interpreted inferentially because the omnibus Kruskal-Wallis test was not significant (P = 0.055). DISCERN = DISCERN instrument; EQIP = Ensuring Quality Information for Patients; GQS = Global Quality Score; JAMA = Journal of the American Medical Association benchmark criteria.

3.3. Readability

Readability was poor across all models (Table 4; Fig. 2). Every median grade-level index exceeded the target of < 6, and every FRES median was below 80. Copilot was descriptively the easiest to read, with FKGL 11.65 (11.27–12.39) and FRES 38.60 (35.74–43.59). ChatGPT was the most difficult on most metrics, with FKGL 14.99 (13.55–15.53) and FRES 24.66 (18.50–32.78). Gemini and Perplexity were intermediate. No model met the prespecified sixth-grade target.

Table 4
Readability scores across large language model chatbots

Metric	ChatGPT	Copilot	Gemini	Perplexity	6th-grade threshold
ARI	17.67 (17.02, 18.52)	14.39 (13.29, 15.21)	16.42 (15.18, 18.00)	15.66 (14.64, 17.06)	< 6
GFI	17.40 (16.19, 18.74)	14.31 (13.39, 15.08)	16.57 (15.09, 18.17)	15.83 (14.79, 16.79)	< 6
FKGL	14.99 (13.55, 15.53)	11.65 (11.27, 12.39)	13.57 (12.81, 15.49)	12.74 (11.58, 13.43)	< 6
CL	17.03 (16.01, 17.70)	14.88 (13.98, 16.06)	15.75 (14.77, 16.49)	15.68 (14.68, 16.68)	< 6
SMOG	15.71 (15.02, 16.53)	13.47 (13.10, 13.82)	15.02 (14.15, 16.36)	14.55 (13.90, 15.20)	< 6
FRES	24.66 (18.50, 32.78)	38.60 (35.74, 43.59)	31.97 (26.82, 37.97)	34.95 (30.45, 41.44)	>=80

Values are median (Q1, Q3). Lower ARI, GFI, FKGL, CL, and SMOG scores indicate easier reading; higher FRES scores indicate easier reading. Prespecified sixth-grade thresholds were < 6 for grade-level/complexity indices and > = 80 for FRES. ARI = Automated Readability Index; CL = Coleman-Liau Index; FKGL = Flesch-Kincaid Grade Level; FRES = Flesch Reading Ease Score; GFI = Gunning Fog Index; Q1 = first quartile; Q3 = third quartile; SMOG = Simple Measure of Gobbledygook.

3.4. Question-Level Variability

Query-level heatmaps showed within-model variation in DISCERN and EQIP scores and a platform-specific JAMA pattern consistent with differences in visible attribution features (Fig. 3). GQS values clustered around the good-quality range. Readability remained consistently difficult across questions; Copilot was relatively easier, but the heatmaps did not show a question group that consistently approached the sixth-grade target.

4. Discussion

4.1. Principal Findings

In this single-date cross-platform benchmark, Perplexity achieved the highest DISCERN, EQIP, and JAMA scores, whereas all four models had similarly high GQS medians. Readability was poor across every model; even Copilot, which was descriptively easiest to read, had a median FKGL of 11.65 and FRES of 38.60. The divergence between global quality ratings and structured quality or transparency measures indicates that fluent, well-organized responses should not be equated with well-sourced or easily accessible patient information.

4.2. Reliability, Information Quality, and Source Transparency

The instruments capture different properties of health information. DISCERN assesses reliability and balance of treatment information, EQIP addresses content and presentation, JAMA focuses on provenance and currency, and GQS reflects overall flow and usefulness.[37–40] Perplexity's largest separation was on JAMA, which is consistent with more visible source-attribution features in its responses but should not be interpreted as proof of greater factual accuracy. Similar cross-instrument differences and limited source transparency have been reported in LLM studies of sexually transmitted disease, maintenance hemodialysis, autoimmune hepatitis, chronic heart failure, breast-cancer fertility counseling, and perinatal depression.[41–46] The maintenance hemodialysis study most similar to our design likewise reported relatively strong reliability for Perplexity alongside difficult text.[42]

Our question set targeted a clinical intersection rather than generic hypertension or ED education. Prior hypertension studies have examined broad education, home monitoring, gestational hypertension, and retrieval-augmented systems, while ED studies have evaluated general search queries.[31–36] Here, questions linked antihypertensive adverse effects, phosphodiesterase type 5 (PDE5) inhibitor use, nitrate interactions, cardiovascular assessment, sexual activity, and modifiable risk factors. These topics are connected in current EAU and Princeton IV guidance.[3,4] The present design therefore extends prior benchmarking by examining a set of questions that routinely cross cardiovascular and sexual-medicine boundaries, where source context is especially important.

4.3. Readability as a Persistent Barrier

Readability was consistently poor. Similar findings have been reported for ED, sexually transmitted disease, hemodialysis, perinatal depression, and other AI-generated patient materials.[24,35,36,41,42,46] The problem also predates generative AI: cardiovascular educational resources frequently exceed recommended reading levels.[25–28] Explanations in this field often require technical terms, interaction warnings, and caveats, which can increase sentence length and lexical complexity. Information depth and linguistic accessibility should therefore be evaluated separately.[22,42–46]

4.4. Clinical and Practical Implications

The findings support a cautious, supplementary role for general-purpose LLMs in patient education rather than use as substitutes for individualized assessment, particularly for medication changes, nitrate-PDE5 inhibitor interactions, or ED as a cardiovascular risk marker.[3–8,16,17] Clinicians may need to anticipate that patients use AI for sensitive sexual-health questions and provide reviewed, plain-language resources when concerns about antihypertensive adverse effects arise. Developers could improve patient-facing utility by making guideline-linked sources more visible and offering explicit controls for reading level. Retrieval-augmented hypertension systems and readability-focused work suggest that provenance and linguistic accessibility can be improved without changing the underlying clinical question.[22,24,32]

4.5. Strengths and Limitations

This study used a focused public-facing question set, standardized single-turn prompting, blinded specialty review, four established information-quality instruments, six readability metrics, adjusted post-hoc comparisons, and query-level heatmaps. Inter-rater agreement was excellent or almost perfect. Several limitations warrant emphasis. First, this was a single-date snapshot of four rapidly changing proprietary web services, and each question was asked once; stochastic response variability and later model updates were therefore not captured. Second, final prompts were in English although Chinese platforms informed question themes, which limits linguistic and cultural generalizability. Third, 37 questions cannot represent every individualized scenario, and source-level provenance was not retained for each candidate question because public posts were used only to identify recurring themes and wording. Fourth, DISCERN, EQIP, JAMA, and GQS were developed for written health information and do not directly measure factual accuracy or clinical safety; we did not independently assess guideline concordance, harmfulness, or bias. Fifth, formula-based readability scores estimate linguistic complexity rather than actual comprehension, and no patient usability testing was performed. Future studies should repeat sampling across dates and model versions, include multilingual and patient-based comprehension testing, and separately evaluate factual accuracy, guideline concordance, and safety.

5. Conclusions

In this single-date benchmark, Perplexity achieved higher DISCERN, EQIP, and JAMA scores, while GQS was similar across models. None approached the prespecified sixth-grade readability target. These results identify source transparency and reading complexity as distinct limitations of current general-purpose LLM responses in this cardiovascular-sexual health context. The findings support a supplementary rather than standalone role for these tools and indicate that future evaluations should pair readability and information-quality measures with direct assessments of factual accuracy, safety, and patient comprehension.

Declarations

Authors' contributions

S.Y.C: Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. J.K.L: Data curation, Software, Writing – original draft, Writing – review & editing. Both authors contributed to the article and approved the submitted version.

Funding

None.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study involved no human participants, human data, human tissue, animals, clinical interventions, or identifiable private information. Public online material was used only to identify topic-level themes and lay wording; no usernames, direct links, identifiable information, or verbatim posts were retained. The analytic dataset comprised researcher-developed questions and chatbot-generated text. Ethical approval and informed consent were therefore not required.

Clinical trial number

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The final question set, aggregate results, and inter-rater reliability data are included in this article and its Supplemental Digital Content. The chatbot response dataset and derived scoring data generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

None.

Supplementary information

None.

References

1. McEvoy JW, McCarthy CP, Bruno RM, et al. 2024 ESC Guidelines for the management of elevated blood pressure and hypertension. Eur Heart J. 2024;45(38):3912–4018. 10.1093/eurheartj/ehae178.
2. Mancia G, Kreutz R, Brunström M, et al. 2023 ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Hypertension: Endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA). J Hypertens. 2023;41(12):1874–2071. 10.1097/HJH.0000000000003480.
3. Salonia A, Capogrosso P, Boeri L, et al. European Association of Urology Guidelines on Male Sexual and Reproductive Health: 2025 Update on Male Hypogonadism, Erectile Dysfunction, Premature Ejaculation, and Peyronie's Disease. Eur Urol. 2025;88(1):76–102. 10.1016/j.eururo.2025.04.010.
4. Kloner RA, Burnett AL, Miner M, et al. Princeton IV consensus guidelines: PDE5 inhibitors and cardiac health. J Sex Med. 2024;21(2):90–116. 10.1093/jsxmed/qdad163.
5. Mostafaei H, Mori K, Hajebrahimi S, et al. Association of erectile dysfunction and cardiovascular disease: an umbrella review of systematic reviews and meta-analyses. BJU Int. 2021;128(1):3–11. 10.1111/bju.15313.
6. Lou IX, Chen J, Ali K, et al. Relationship Between Hypertension, Antihypertensive Drugs and Sexual Dysfunction in Men and Women: A Literature Review. Vasc Health Risk Manag. 2023;19:691–705. 10.2147/VHRM.S439334.
7. Corona G, Vena W, Pizzocaro A, et al. Anti-hypertensive medications and erectile dysfunction: focus on β-blockers. Endocrine. 2025;87(1):11–26. 10.1007/s12020-024-04020-x.
8. Farmakis IT, Pyrgidis N, Doundoulakis I, et al. Effects of Major Antihypertensive Drug Classes on Erectile Function: a Network Meta-analysis. Cardiovasc Drugs Ther. 2022;36(5):903–14. 10.1007/s10557-021-07197-9.
9. Rashidi M, Kiscak N, Kaya Meral D, et al. Assessment of erectile dysfunction prevalence and associated factors in hypertensive men. PeerJ. 2024;12:e18596. 10.7717/peerj.18596.
10. Gebrie E, Chane E, Mengistie BA, et al. Prevalence of erectile dysfunction and its associated factors among hypertensive men in Africa: a systematic review and meta-analysis. Sex Med Rev. 2026;14(3):qeag046. 10.1093/sxmrev/qeag046.
11. Khera M, Bhattacharyya S, Miller LE. Effect of aerobic exercise on erectile function: systematic review and meta-analysis of randomized controlled trials. J Sex Med. 2023;20(12):1369–75.

10.1093/jsxmed/qdad130.

12. Li H, Xu W, Wang T, et al. Effect of weight loss on erectile function in men with overweight or obesity: a meta-analysis of randomised controlled trials. Andrologia. 2022;54(1):e14250. 10.1111/and.14250.
13. Sivaratnam L, Selimin DS, Abd Ghani SR, et al. Behavior-Related Erectile Dysfunction: A Systematic Review and Meta-Analysis. J Sex Med. 2021;18(1):121–43. 10.1016/j.jsxm.2020.09.009.
14. Li S, Song JM, Zhang K, et al. A Meta-Analysis of Erectile Dysfunction and Alcohol Consumption. Urol Int. 2021;105(11–12):969–85. 10.1159/000508171.
15. Siddiqui ZA, McNicholas DP, Spacey A, et al. The Relationship between Erectile Dysfunction and Obstructive Sleep Apnoea and the Impact on Severity of ED following Treatment of OSA: A Systematic Review. Urol Int. 2026;110(2):176–85. 10.1159/000546690.
16. Soulaidopoulos S, Terentes-Printzios D, Ioakeimidis N, et al. Long-term effects of phosphodiesterase-5 inhibitors on cardiovascular outcomes and death: a systematic review and meta-analysis. Eur Heart J Cardiovasc Pharmacother. 2024;10(5):403–12. 10.1093/ehjcvp/pvae029.
17. Seidu S, Cebrian A, Kunutsor SK, et al. Erectile dysfunction, phosphodiesterase-5 inhibitor use and risk of cardiovascular disease and mortality in people with diabetes: A systematic review and meta-analysis. Prim Care Diabetes. 2022;16(5):601–13. 10.1016/j.pcd.2022.09.004.
18. Atallah S, Haydar A, Jabbour T, et al. The effectiveness of psychological interventions alone, or in combination with phosphodiesterase-5 inhibitors, for the treatment of erectile dysfunction: A systematic review. Arab J Urol. 2021;19(3):310–22. 10.1080/2090598X.2021.1926763.
19. Longoni M, Bertini A, Schifano N, et al. A review on pharmacological options for the treatment of erectile dysfunction: state of the art and new strategies. Expert Opin Pharmacother. 2023;24(12):1375–86. 10.1080/14656566.2023.2221785.
20. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930–40. 10.1038/s41591-023-02448-8.
21. Aydin S, Karabacak M, Vlachos V, et al. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898. 10.3389/fmed.2024.1477898.
22. Nasra M, Jaffri R, Pavlin-Premrl D, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Intern Med J. 2025;55(1):20–34. 10.1111/imj.16607.
23. Gendler M, Nadkarni GN, Sudri K, et al. Large Language Models in Cardiology: Systematic Review. JMIR Cardio. 2026;10:e76734. 10.2196/76734.
24. Rouhi AD, Ghanem YK, Yolchieva L, et al. Can Artificial Intelligence Improve the Readability of Patient Education Materials on Aortic Stenosis? A Pilot Study. Cardiol Ther. 2024;13(1):137–47. 10.1007/s40119-023-00347-0.
25. Wasir AS, Volgman AS, Jolly M. Assessing readability and comprehension of web-based patient education materials by American Heart Association (AHA) and CardioSmart online platform by

American College of Cardiology (ACC): How useful are these websites for patient understanding? Am Heart J Plus. 2023;32:100308. 10.1016/j.ahjo.2023.100308.

26. Bhatt C, Lin E, Ferreira-Legere LE, et al. Evaluating Readability, Understandability, and Actionability of Online Printable Patient Education Materials for Cholesterol Management: A Systematic Review. J Am Heart Assoc. 2024;13(8):e030140. 10.1161/JAHA.123.030140.
27. Sharma S, Latif Z, Makuvire TT, et al. Readability and Accessibility of Patient-Education Materials for Heart Failure in the United States. J Card Fail. 2025;31(1):154–7. 10.1016/j.cardfail.2024.06.015.
28. Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. J Patient Exp. 2021;8:2374373521998847. 10.1177/2374373521998847.
29. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. Br J Surg. 2025;112(8):znaf142. 10.1093/bjs/znaf142.
30. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ. 2025;390:e083305. 10.1136/bmj-2024-083305.
31. Hu M, Wang Z, Zhang Z, et al. Challenges of patient-facing generative artificial intelligence in hypertension care: A cross-platform evaluation of the quality, readability, and actionability of LLM-generated patient education materials. Digit Health. 2026;12:20552076261464724. 10.1177/20552076261464724.
32. Wang Y, Luan Y, Cheng S, et al. A multi-layer retrieval-augmented large language model framework for enhancing hypertension education. Hypertens Res. 2026;49(4):1428–40. 10.1038/s41440-025-02481-9.
33. Motaghi Niko M, Karbasi Z, Kazemi M, et al. Comparing ChatGPT and Bing, in response to the Home Blood Pressure Monitoring (HBPM) knowledge checklist. Hypertens Res. 2024;47(5):1401–9. 10.1038/s41440-024-01624-8.
34. Liu Z, Tong P, Jiang Y, et al. Comparative evaluation of the quality, reliability, and readability of five large language model responses to frequently asked questions on gestational hypertension. Front Public Health. 2026;14:1833611. 10.3389/fpubh.2026.1833611.
35. Sahin MF, Ates H, Keles A, et al. Responses of Five Different Artificial Intelligence Chatbots to the Top Searched Queries About Erectile Dysfunction: A Comparative Analysis. J Med Syst. 2024;48(1):38. 10.1007/s10916-024-02056-0.
36. Razdan S, Siegal AR, Brewer Y, et al. Assessing ChatGPT's ability to answer questions pertaining to erectile dysfunction: can our patients trust it? Int J Impot Res. 2024;36(7):734–40. 10.1038/s41443-023-00797-z.
37. Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105–11. 10.1136/jech.53.2.105.

38. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7(2):165–75. 10.1111/j.1369-7625.2004.00273.x.
39. Sood A, Hui YJ, Turnbull S, et al. A critical analysis of online patient-directed resources on catheter ablation for ventricular arrhythmias. J Arrhythm. 2025;41(2):e70026. 10.1002/joa3.70026.
40. Bernard A, Langille M, Hughes S, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102(9):2070–7. 10.1111/j.1572-0241.2007.01325.x.
41. Yildiz HA, Sogutdelen E. AI Chatbots as Sources of STD Information: A Study on Reliability and Readability. J Med Syst. 2025;49(1):43. 10.1007/s10916-025-02178-z.
42. Cao J, Wu Z, Wu Z, et al. The reliability and readability of large language models in answering patient questions on maintenance hemodialysis: A comparative study. Digit Health. 2026;12:20552076261435836. 10.1177/20552076261435836.
43. Li H, Lin Y, Lv L. Performance evaluation of mainstream large language models in autoimmune hepatitis patient education: a comparative study of readability, quality, and reliability. Front Public Health. 2026;14:1805848. 10.3389/fpubh.2026.1805848.
44. Wang Z, Li X, Ma C, et al. Challenges of using generative AI for patient education in chronic heart failure: an evaluation of content quality, readability, and actionability in cross-platform LLM-generated texts. Front Public Health. 2026;14:1801829. 10.3389/fpubh.2026.1801829.
45. Wang J, Zhan YY, Shen Y, et al. A centralized decision-making support consultation response for fertility preservation in breast cancer patients: benchmark performance of generative large language models in terms of reliability and readability. Front Public Health. 2026;14:1845389. 10.3389/fpubh.2026.1845389.
46. Huang J, Yu H, Chen J, et al. Can large language models be trusted? Reliability and readability of responses to perinatal depression FAQs. Front Public Health. 2026;14:1760872. 10.3389/fpubh.2026.1760872.

Figures

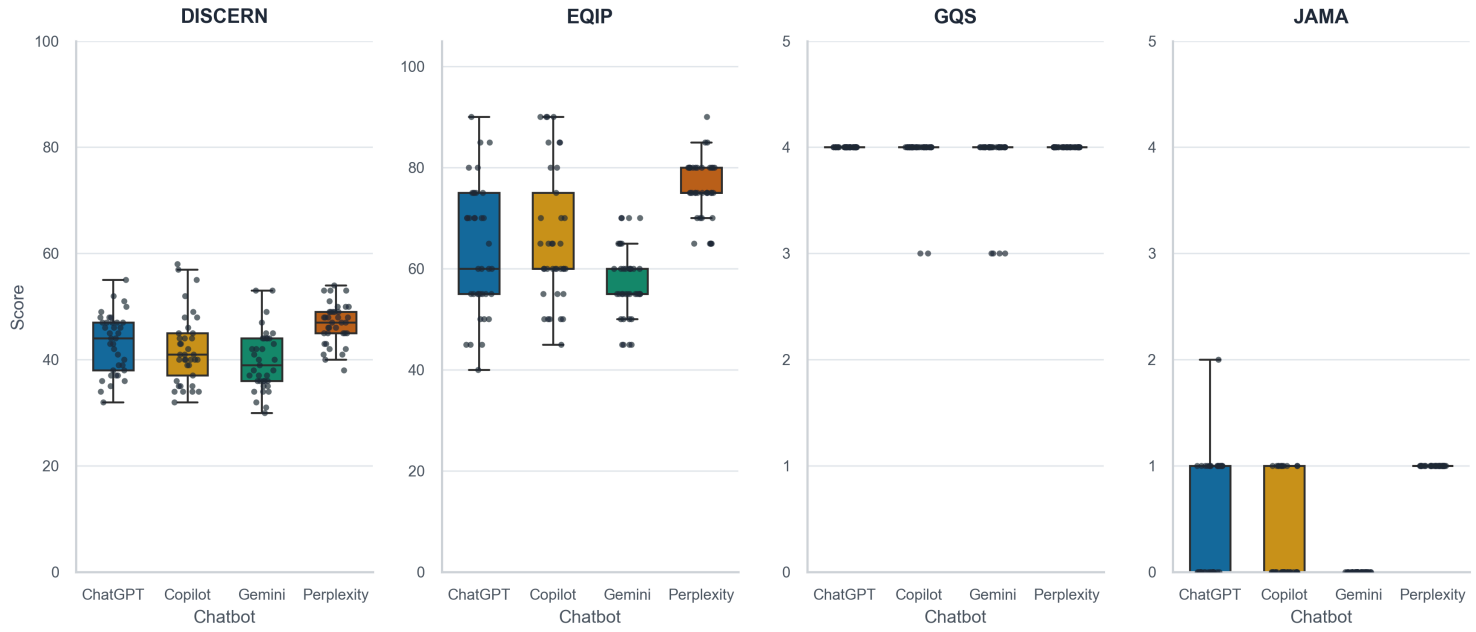


Figure 1

Distributions of DISCERN, Ensuring Quality Information for Patients (EQIP), Global Quality Score (GQS), and Journal of the American Medical Association (JAMA) benchmark scores across the four large language model chatbots. Boxes represent the interquartile range, center lines the median, and points question-level observations.

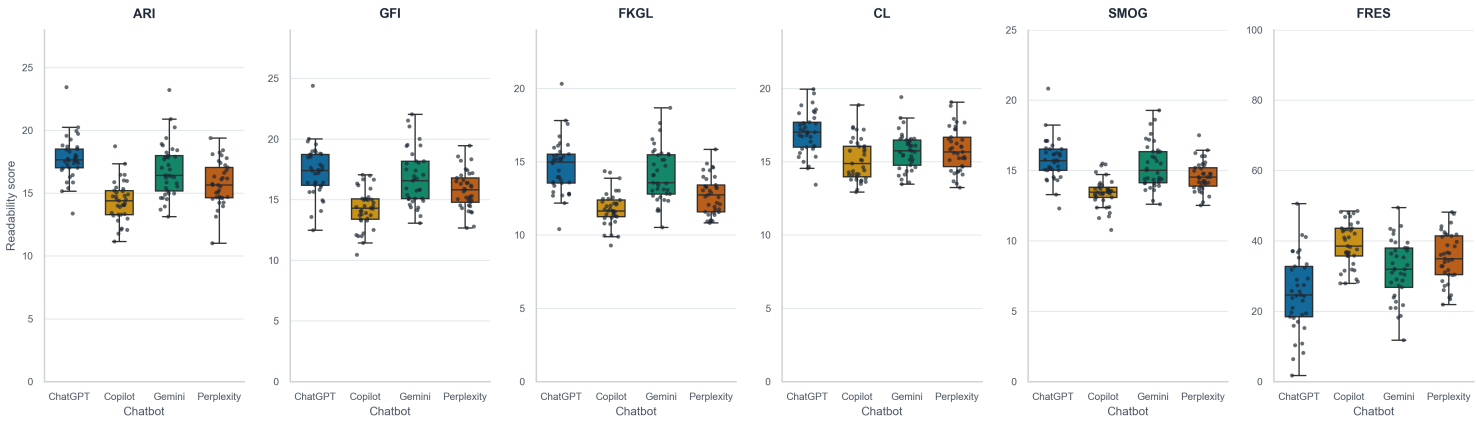


Figure 2

Distributions of six readability indices across the four large language model chatbots. Lower values indicate easier reading for the Automated Readability Index (ARI), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CL), and Simple Measure of Gobbledygook (SMOG); higher values indicate easier reading for the Flesch Reading Ease Score (FRES).

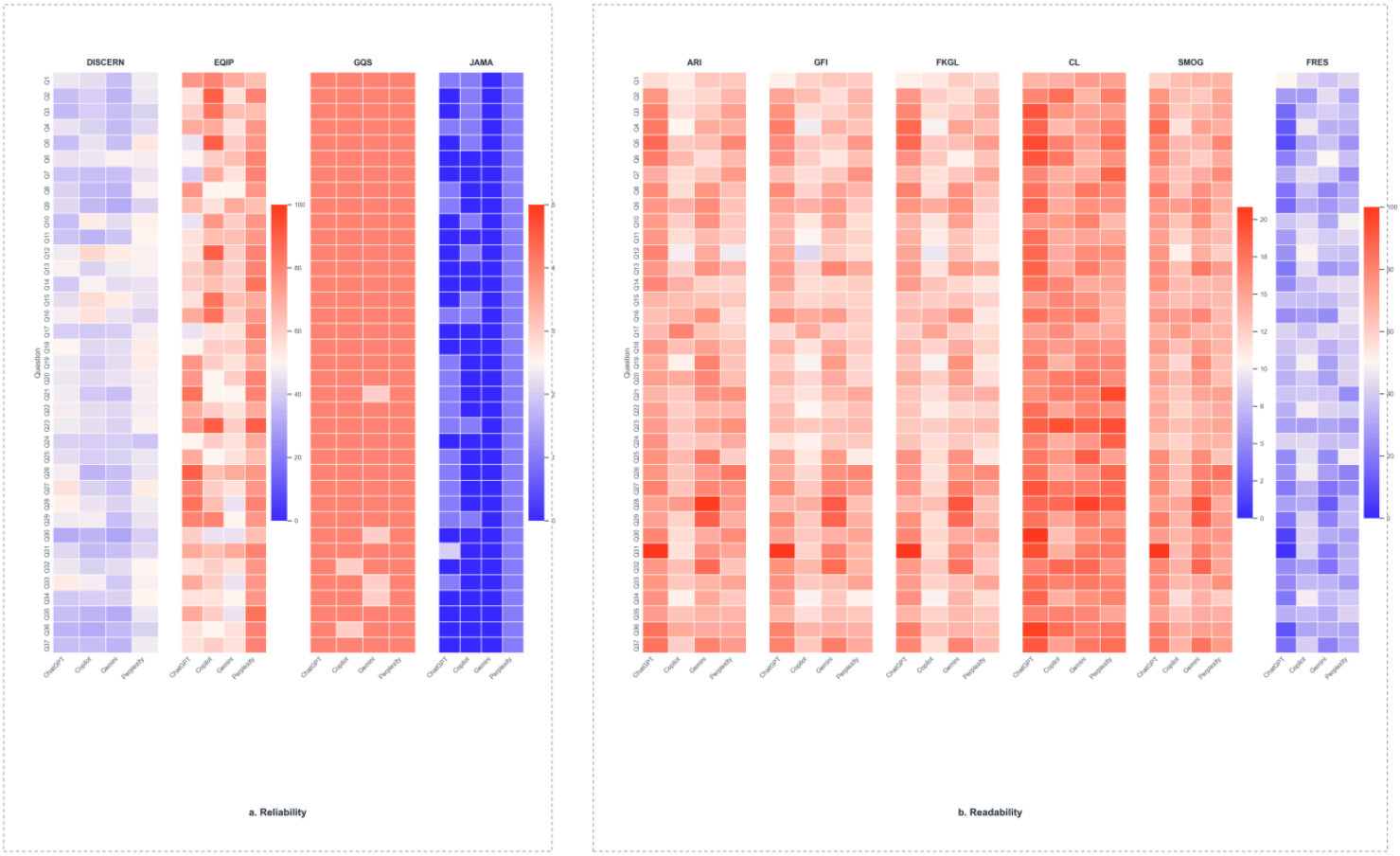


Figure 3

Query-level heatmaps of reliability and readability across the 37 retained questions. (A) DISCERN, Ensuring Quality Information for Patients (EQIP), Global Quality Score (GQS), and Journal of the American Medical Association (JAMA) benchmark scores. (B) Automated Readability Index (ARI), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CL), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES) values. Color scales are metric-specific; color intensity should not be compared directly across different instruments.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalDigitalContent1TableS1.docx
- SupplementalDigitalContent2TableS2.docx