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.

Comparative Evaluation of Four Large Language Models in Answering Questions About Transperineal Prostate Biopsy and Biopsy Findings

Hao Chen
Shandong First Medical University
Bingyao Li
Shandong First Medical University
Rongyang Jin
Shandong First Medical University
Yong Zhao
zhaoyong6807@163.com
Shandong First Medical University
Xiaoliang Sun
Shandong First Medical University

Research Article

Keywords: artificial intelligence, large language models, patient communication, prostate biopsy, urology

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11484342/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

Patients undergoing transperineal prostate biopsy need clear explanations of the procedure and possible findings. This study compared four large language models (LLMs) for transperineal prostate biopsy counseling and result interpretation.

Methods

Thirty questions were submitted three times to four LLMs (360 responses). A three-member urologist panel assigned consensus ratings for Accuracy, Completeness, Consistency, Empathy, and Readability. Question-level means were compared using Friedman and Holm-corrected Wilcoxon signed-rank tests. ROUGE and TF-IDF assessed lexical similarity; BERTScore F1 assessed semantic similarity.

Results

Accuracy and Completeness did not differ significantly. Differences occurred for Consistency ($p = 0.015$), Empathy ($p < 0.001$), and Readability ($p < 0.001$), although no pairwise Consistency comparison remained significant after Holm correction. Gemini 3.1 had the highest mean Empathy and Readability scores. DeepSeek-V3 had the highest mean ROUGE scores, whereas ChatGPT-5 had the highest mean TF-IDF cosine similarity and BERTScore F1.

Conclusions

The models received high Accuracy and Completeness ratings, with clearer differences in Empathy and Readability. Gemini 3.1 had the highest mean scores in both communication dimensions, while rankings across reference-based similarity measures varied.

Trial registration:

Not applicable.

Background

Prostate cancer is a major cause of cancer morbidity and mortality [1]. Biopsy provides histopathological information for diagnosis and risk stratification [2]. The transperineal approach has a lower infectious risk than the transrectal approach [3], and infection after transperineal biopsy has been examined with and without antibiotic prophylaxis [4]. Patients considering biopsy need explanations of its purpose,

preparation, procedural risks, and possible outcomes. After biopsy, they must interpret terms such as Gleason score and Grade Group [5, 6], which can be difficult to understand [7]. Limited health literacy and the complexity of screening and treatment choices add to these communication demands [8–12].

Large language models (LLMs) have demonstrated medical knowledge [13, 14] and have been evaluated for responses to patient questions [15, 16]. Previous prostate cancer studies examined information quality and patient education, including radiotherapy counseling [17–19]. Nephrology research likewise evaluated models on defined patient-interaction and biopsy-interpretation tasks [20]. Transperineal prostate biopsy counseling spans both procedural explanations and interpretation of subsequent findings. A response can address the medical question yet differ in emotional acknowledgement, clarity, or the information it emphasizes. We therefore compared four LLMs using 30 standardized questions about transperineal prostate biopsy and hypothetical biopsy findings, combining clinician ratings with lexical and semantic similarity to expert reference answers.

Methods

The four models were accessed through their official web interfaces in March 2026, without model customization or fine-tuning.

Urologists developed 30 English-language questions across six domains: basic knowledge, indications and contraindications, biopsy procedure, risks and complications, result interpretation and subsequent management, and biopsy and imaging result interpretation (Additional file 1, Table S1). The final domain used hypothetical findings and imaging terminology rather than complete individual patient reports.

Four generative AI models were evaluated: ChatGPT-5 (OpenAI), DeepSeek-V3 (DeepSeek), Gemini 3.1 (Google), and Claude-Sonnet4.6 (Anthropic). Each question was submitted to each model in a new conversation without prior conversation history. Each question–model combination was queried on three separate occasions (Run 1, Run 2, and Run 3), yielding 360 responses (4 models × 30 questions × 3 runs).

Two senior physicians and one attending physician in urology evaluated all responses. Model identities were withheld, and responses were presented in randomized order. Following discussion, the panel assigned one consensus score per response for Accuracy, Completeness, Consistency, Empathy, and Readability, each on a 1–5 scale (Additional file 1, Table S2). Consistency was defined as internal coherence within an individual response.

Text similarity to one expert reference answer per question was assessed using TF-IDF cosine similarity [21], ROUGE-1, ROUGE-2, ROUGE-L [22], and BERTScore F1 [23]. Reference answers were drafted by a research team member and reviewed jointly by the three urologists. Clinical accuracy, completeness, and alignment with guidelines and standard practice were checked before agreement on the final answers. Calculation and preprocessing details are provided in Additional file 1. The question set, reference

answers, response texts, consensus scores and row-level similarity results are supplied in Additional file 2.

Results are presented as mean ± standard deviation (SD) across 90 responses per model. For inferential analysis, the three scores within each question–model combination were averaged, yielding 30 matched observations per model. Models were compared using the Friedman test [24], with Kendall's W as the effect size. Significant omnibus tests were followed by two-sided Wilcoxon signed-rank tests [25], implemented in Python 3.12 with SciPy [26]. Holm correction was applied across the six model pairs within each dimension [27]. Adjusted p values < 0.05 were considered significant. Implementation details are provided in Additional file 1. The composite score was the unweighted mean of the five dimension scores. Composite scores and text similarity metrics were summarized descriptively.

Results

Accuracy and Completeness did not differ significantly between models (Table 1). Consistency showed an omnibus difference, but none of its pairwise comparisons remained significant after Holm correction. Empathy and Readability showed larger effect sizes (Table 1; Fig. 1). Gemini 3.1 had the highest descriptive composite score (4.38), followed by ChatGPT-5 (4.29), DeepSeek-V3 (4.18), and Claude-Sonnet4.6 (4.04).

Table 1 Clinician ratings and matched between-model comparisons

Dimension	ChatGPT-5	DeepSeek-V3	Gemini 3.1	Claude-Sonnet4.6	χ^2	p	W
Accuracy	4.71 ± 0.67	4.79 ± 0.46	4.74 ± 0.46	4.80 ± 0.62	2.862	0.413	0.032
Completeness	4.39 ± 0.70	4.38 ± 0.53	4.28 ± 0.60	4.30 ± 0.84	2.869	0.412	0.032
Consistency	4.89 ± 0.32	5.00 ± 0.00	4.83 ± 0.43	4.94 ± 0.23	10.484	0.015	0.116
Empathy	3.12 ± 0.47	2.67 ± 0.52	3.38 ± 1.03	2.23 ± 0.54	46.495	<0.001	0.517
Readability	4.34 ± 0.56	4.07 ± 0.29	4.67 ± 0.50	3.92 ± 0.50	37.692	<0.001	0.419

Values are mean ± SD across 90 responses per model. Friedman tests used 30 question-level means per model (df = 3). χ^2, Friedman statistic; W, Kendall's coefficient of concordance. $W = \chi^2/[30 \times (4 - 1)]$.

Fig. 1 Clinician ratings across five evaluation dimensions. Bars represent means and error bars represent SD across 90 responses per model. Ratings range from 1 to 5; mean ± SD intervals may extend beyond these bounds. Statistical tests used 30 matched question-level means (Table 1). Model identity is indicated by both color and hatch pattern

For Empathy, Gemini 3.1 and ChatGPT-5 did not differ significantly (adjusted p = 0.109); both scored higher than DeepSeek-V3 and Claude-Sonnet4.6. Claude-Sonnet4.6 scored lower than each other model. For Readability, Gemini 3.1 scored higher than ChatGPT-5 (adjusted p = 0.036), DeepSeek-V3 and Claude-Sonnet4.6 (both adjusted p < 0.001). ChatGPT-5 also scored higher than DeepSeek-V3 (adjusted p = 0.023) and Claude-Sonnet4.6 (adjusted p = 0.004). DeepSeek-V3 and Claude-Sonnet4.6 did not differ significantly (adjusted p = 0.120). All pairwise statistics are provided in Additional file 1, Table S3.

Reference-based metrics were compared descriptively (Fig. 2; Additional file 1, Table S4). DeepSeek-V3 had the highest mean ROUGE-1 (0.735), ROUGE-2 (0.285), and ROUGE-L (0.549). ChatGPT-5 had the highest mean TF-IDF cosine similarity (0.334) and BERTScore F1 (0.842). Mean BERTScore F1 values ranged from 0.834 to 0.842. Additional profiles and distributions are shown in Additional file 1, Figs. S1–S3.

Fig. 2 Reference-based text similarity across models. Bars represent means and error bars represent SD across 90 responses per model. ROUGE-1, ROUGE-2 and ROUGE-L are recall-based measures; TF-IDF denotes cosine similarity; BERTScore denotes F1 without baseline rescaling. Model identity is indicated by both color and hatch pattern

Discussion

The clearest differences between models concerned how information was communicated. Accuracy means ranged from 4.71 to 4.80 and Completeness means from 4.28 to 4.39, with small between-model effects for both dimensions (Kendall's W = 0.032). Empathy and Readability showed wider separation, with W values of 0.517 and 0.419, respectively. Within this question set and rating framework, communication dimensions therefore provided greater discrimination between models. Alongside previous work on medical knowledge and broader LLM performance [13,14,28,29], these findings support evaluating both the clinical content of an answer and how that content is presented to patients.

The individual model profiles clarify the implications for model selection. Claude-Sonnet4.6 had the highest numerical Accuracy mean (4.80), although the overall Accuracy comparison was not significant, and the lowest Empathy mean (2.23). Gemini 3.1 received higher Readability ratings than every other model. Its Empathy mean was also the highest, but its difference from ChatGPT-5 was not significant after correction (adjusted p = 0.109); both scored higher than DeepSeek-V3 and Claude-Sonnet4.6. These profiles provide information that an unweighted composite ranking can obscure. For patient education, selection should consider the communication demands of the intended task alongside clinical accuracy. This complements studies evaluating readability and usefulness in patient education and report explanation [30,31]. It also identifies a testable development direction: whether prompts that encourage plain-language explanations and acknowledgement of patient concerns improve communication ratings while preserving accuracy.

Previous prostate cancer studies have evaluated general information and patient education [17,18,32], radiotherapy communication [19,33], doctor-patient communication [34], the PROSCA chatbot [35], and

preoperative prostatectomy counseling [36]. Other specialty evaluations have used knowledge tests or specific clinical tasks [20,37]. Our question set brings procedural information, complications, and hypothetical findings into the same biopsy setting. The high content ratings provide a rationale for testing clinician-reviewed LLM explanations as supplementary educational materials before and after transperineal biopsy. Potential uses include explaining preparation instructions, describing the procedure, and clarifying report terminology. The communication differences observed here suggest that reviewing such materials should address both factual content and how clearly and sensitively it is conveyed.

Reference-based similarity measures produced a different ordering from communication ratings. DeepSeek-V3 had the highest mean ROUGE scores, ChatGPT-5 had the highest mean TF-IDF cosine similarity and BERTScore F1, and Gemini 3.1 had the highest Readability mean. At the model level, greater similarity to an expert reference did not consistently identify the highest-rated communication performance. These measures capture lexical or semantic similarity, whereas clinician ratings address additional features of the explanation. BERTScore means were also closely grouped (0.834–0.842), making their ranking a narrow numerical distinction. A broader implication for medical LLM evaluation is that assessment should match the intended use: reference-based metrics can complement expert content review, communication assessment, and tests of patient understanding. Whether the model profiles observed here extend to other specialties requires evaluation in those settings.

The next step is to determine whether the communication differences identified by clinicians matter to patients. Building on patient-based approaches such as PROSCA [35] and MedEduChat [38], prospective studies could compare routine biopsy education with clinician-reviewed LLM-assisted materials. Priority outcomes include patients' understanding and recall of preparation instructions, aftercare advice, and report terminology, together with perceived clarity and whether their concerns were addressed. This would connect answer-level evaluation to the practical goals of biopsy counseling and establish which aspects of model performance are most useful in patient-facing care.

Consistency ratings were concentrated near the scale maximum, with DeepSeek-V3 scoring 5.00 on every response. The overall difference was significant ($p = 0.015$), but its effect size was smaller than those for Empathy and Readability ($W = 0.116$), and no pairwise comparison remained significant after Holm correction. These findings concern internal coherence within individual answers. The limited score variation also suggests that more demanding questions or more granular criteria may be useful for examining this dimension in future evaluations.

Several limitations affect interpretation. Consensus scoring did not permit estimation of independent inter-rater agreement, and Accuracy ratings were also concentrated near the scale maximum. The 30 English-language questions and March 2026 model snapshot limit generalizability to other prompts, languages, and model versions. A single reference answer per question may not represent all clinically acceptable responses, and reference wording and response length can influence similarity metrics.

Finally, clinician ratings of isolated answers do not establish patient understanding or subsequent decision quality.

Conclusions

In this question set, the four models received high Accuracy and Completeness ratings, with clearer differences in Empathy and Readability. Gemini 3.1 received higher Readability ratings than each other model, while both Gemini 3.1 and ChatGPT-5 scored higher for Empathy than DeepSeek-V3 and Claude-Sonnet4.6. These findings support further evaluation of clinician-reviewed LLM explanations for transperineal prostate biopsy education. Model selection and assessment should consider clinical content and communication quality alongside reference-based similarity. Prospective patient studies can establish whether these differences translate into better understanding and communication experiences.

Abbreviations

AI
Artificial intelligence
BERTScore
BERT-based text similarity score
LLM
Large language model
ROUGE
Recall-Oriented Understudy for Gisting Evaluation
SD
Standard deviation
TF-IDF
Term frequency–inverse document frequency.

Declarations

Ethics approval and consent to participate

Not applicable. This study evaluated AI-generated responses to hypothetical clinical questions and did not involve patient recruitment, identifiable patient data, or clinical interventions. The clinicians contributed as members of the research team.

Consent for publication

Not applicable. No identifiable individual information is reported.

Availability of data and materials

The data supporting the conclusions of this study are included in the article and its additional files. The question set, expert reference answers, model responses, consensus ratings, text similarity results, and statistical comparisons are provided in Additional file 2. Scoring criteria and calculation methods are provided in Additional file 1.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no financial support for this study.

Authors' contributions

HC conceived and designed the study, collected and analyzed the data, developed the software, prepared the figures, and drafted the manuscript. BL contributed to methodology, data collection, curation, and validation. RJ contributed to study conception, data collection, validation, and supervision. YZ and XS helped develop the expert reference answers and scored model responses. YZ also contributed to methodology, supervision, and project administration. All authors critically revised the manuscript, approved the final version, and agreed to be accountable for the work.

Acknowledgements

We thank Haiyang Zhang for contributing to the expert scoring and providing constructive comments on the manuscript.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63. https://doi.org/10.3322/caac.21834.
2. Cornford P, van den Bergh RCN, Briers E, EAU-EANM-ESTRO-ESUR-ISUP-SIOG, et al. Guidelines on Prostate Cancer-2024 Update. Part I: Screening, Diagnosis, and Local Treatment with Curative Intent. Eur Urol. 2024;86:148–63. https://doi.org/10.1016/j.eururo.2024.03.027.
3. Hu JC, Assel M, Allaf ME, et al. Transperineal versus transrectal magnetic resonance imaging-targeted and systematic prostate biopsy to prevent infectious complications: the PREVENT

randomized trial. Eur Urol. 2024;86:61–8. https://doi.org/10.1016/j.eururo.2023.12.015.

4. Castellani D, Pirola GM, Law YXT, et al. Infection rate after transperineal prostate biopsy with and without prophylactic antibiotics: results from a systematic review and meta-analysis of comparative studies. J Urol. 2022;207:25–34. https://doi.org/10.1097/JU.0000000000002251.
5. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol. 2016;40:244–252. https://doi.org/10.1097/PAS.0000000000000530
6. Epstein JI, Amin MB, Reuter VE, Humphrey PA. Contemporary Gleason Grading of Prostatic Carcinoma: An Update With Discussion on Practical Issues to Implement the 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol. 2017;41:e1-e7. https://doi.org/10.1097/PAS.0000000000000820
7. Tagai EK, Miller SM, Kutikov A, et al. Prostate cancer patients' understanding of the Gleason scoring system: implications for shared decision-making. J Cancer Educ. 2019;34:441–5. https://doi.org/10.1007/s13187-018-1320-1.
8. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med. 2011;155:97–107. https://doi.org/10.7326/0003-4819-155-2-201107190-00005.
9. DeWalt DA, Berkman ND, Sheridan S, Lohr KN, Pignone MP. Literacy and health outcomes: a systematic review of the literature. J Gen Intern Med. 2004;19:1228–39. https://doi.org/10.1111/j.1525-1497.2004.40153.x.
10. US Preventive Services Task Force, Grossman DC, Curry SJ, Owens DK, et al. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2018;319:1901–13. https://doi.org/10.1001/jama.2018.3710.
11. van Stam MA, Pieterse AH, van der Poel HG, Bosch JLHR, Tillier C, Horenblas S, et al. Shared Decision Making in Prostate Cancer Care-Encouraging Every Patient to Be Actively Involved in Decision Making or Ensuring the Patient Preferred Level of Involvement? J Urol. 2018;200:582–9. https://doi.org/10.1016/j.juro.2018.02.3091.
12. Martínez-González NA, Plate A, Senn O, et al. Shared decision-making for prostate cancer screening and treatment: a systematic review of randomised controlled trials. Swiss Med Wkly. 2018;148:w14584. https://doi.org/10.4414/smw.2018.14584.
13. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. https://doi.org/10.1038/s41586-023-06291-2.
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29:1930–40. https://doi.org/10.1038/s41591-023-02448-8.
15. 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;183:589–96. https://doi.org/10.1001/jamainternmed.2023.1838.

16. Chen D, Parsa R, Hope A, et al. Physician and artificial intelligence chatbot responses to cancer questions from social media. JAMA Oncol. 2024;10:956–60. https://doi.org/10.1001/jamaoncol.2024.0836.
17. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med. 2023;21:269. https://doi.org/10.1186/s12967-023-04123-5.
18. Alasker A, Alsalamah S, Alshathri N, Almansour N, Alsalamah F, Alghafees M, et al. Performance of large language models (LLMs) in providing prostate cancer information. BMC Urol. 2024;24:177. https://doi.org/10.1186/s12894-024-01570-0.
19. Trapp C, Schmidt-Hegemann N, Keilholz M, Brose SF, Marschner SN, Schönecker S, et al. Patient- and clinician-based evaluation of large language models for patient education in prostate cancer radiotherapy. Strahlenther Onkol. 2025;201:333–42. https://doi.org/10.1007/s00066-024-02342-3.
20. Lin SY, Jiang CC, Law KM, et al. Evaluation of generative AI assistance in clinical nephrology: assessing GPT-4, GPT-4o, Gemini 1.0 Ultra, and PaLM 2 in patient interaction and renal biopsy interpretation. Digit Health. 2025;11:20552076251342067. https://doi.org/10.1177/20552076251342067.
21. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. 1988;24:513–23. https://doi.org/10.1016/0306-4573(88)90021-0.
22. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the ACL Workshop on Text Summarization Branches Out. 2004. pp. 74–81. https://aclanthology.org/W04-1013/
23. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations. 2020. https://openreview.net/forum?id=SkeHuCVFDr
24. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32:675–701. https://doi.org/10.1080/01621459.1937.10503522.
25. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1:80–3. https://doi.org/10.2307/3001968.
26. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
27. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70. https://www.jstor.org/stable/4615733.
28. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med. 2024;177:210–20. https://doi.org/10.7326/M23-2772.
29. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3:141. https://doi.org/10.1038/s43856-023-00370-1.

30. Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: can you prepare my patients for [18F]FDG PET/CT and explain my reports? J Nucl Med. 2023;64:1876–9. https://doi.org/10.2967/jnumed.123.266114.
31. Hershenhouse JS, Mokhtar D, Eppler MB, Rodler S, Storino Ramacciotti L, Ganjavi C, et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. Prostate Cancer Prostatic Dis. 2025;28:394–9. https://doi.org/10.1038/s41391-024-00826-y.
32. Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N, et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res. 2024;26:e55939. https://doi.org/10.2196/55939.
33. Ozturk GA, Tumer EG, Çoban FSK, Ozturk H, Gurlek U, Ozturk M, et al. AI chatbots for patient education in localized prostate cancer radiotherapy: a comparative quality and readability analysis. Patient Educ Couns. 2026;149:109594. https://doi.org/10.1016/j.pec.2026.109594.
34. Geantă M, Bădescu D, Chirca N, Nechita OC, Radu CG, Rascu S, et al. The potential impact of large language models on doctor-patient communication: a case study in prostate cancer. Healthc (Basel). 2024;12:1548. https://doi.org/10.3390/healthcare12151548.
35. Baumgärtner K, Byczkowski M, Schmid T, Muschko M, Woessner P, Gerlach A, et al. Effectiveness of the medical chatbot PROSCA to inform patients about prostate cancer: results of a randomized controlled trial. Eur Urol Open Sci. 2024;69:80–8. https://doi.org/10.1016/j.euros.2024.08.022.
36. Wang L, Liu H, Yang J, Jiao P, Chen S, Si X, et al. Guideline-based evaluation of large language models for patient counselling before robot-assisted radical prostatectomy: accuracy, safety, and readability. J Robot Surg. 2026;20:888. https://doi.org/10.1007/s11701-026-03859-8.
37. Miao J, Thongprayoon C, Garcia Valencia OA, Krisanapan P, Sheikh MS, Davis PW, et al. Performance of ChatGPT on nephrology test questions. Clin J Am Soc Nephrol. 2024;19:35–43. https://doi.org/10.2215/CJN.0000000000000330.
38. Hao Y, Holmes J, Waddle MR, Davis BJ, Yu NY, Vickers KS, et al. Personalizing prostate cancer education for patients using an EHR-Integrated LLM agent. npj Digit Med. 2025;8:770. https://doi.org/10.1038/s41746-025-02166-0.

Figures

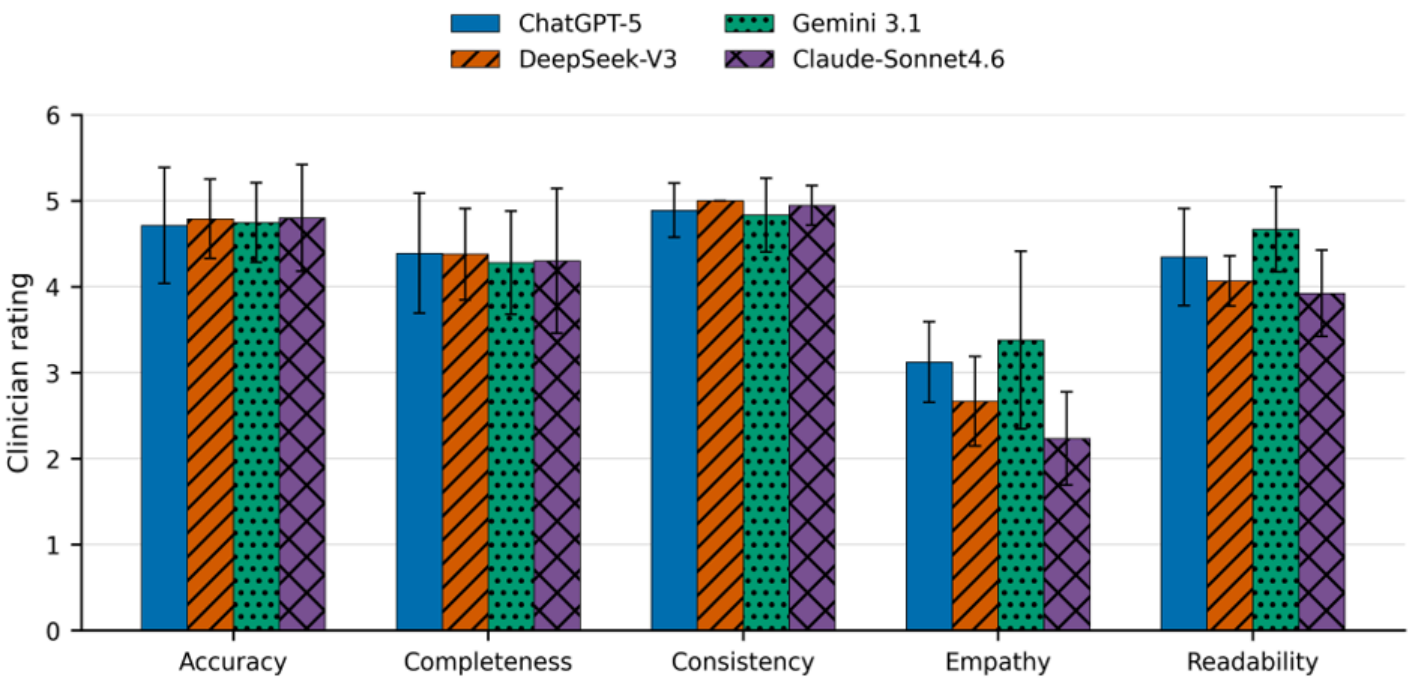

Figure 1

Clinician ratings across five evaluation dimensions. Bars represent means and error bars represent SD across 90 responses per model. Ratings range from 1 to 5; mean ± SD intervals may extend beyond these bounds. Statistical tests used 30 matched question-level means (Table 1). Model identity is indicated by both color and hatch pattern

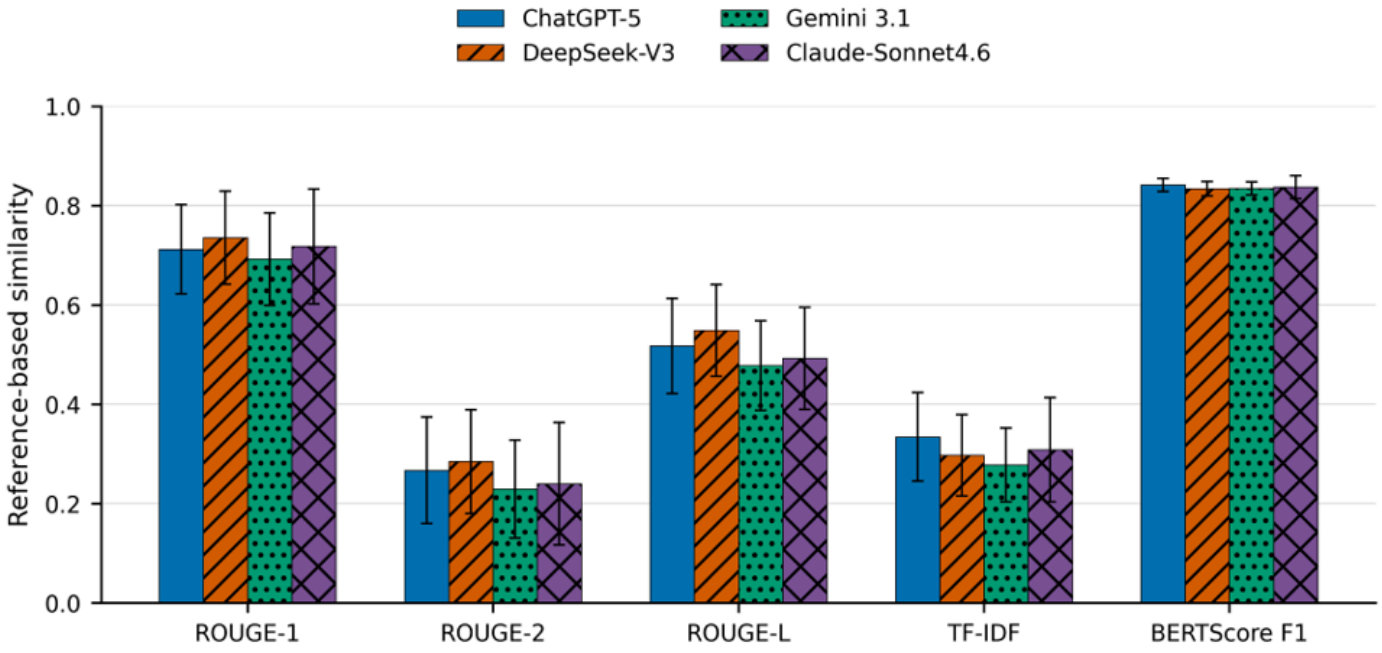

Figure 2

Reference-based text similarity across models. Bars represent means and error bars represent SD across 90 responses per model. ROUGE-1, ROUGE-2 and ROUGE-L are recall-based measures; TF-IDF denotes cosine similarity; BERTScore denotes F1 without baseline rescaling. Model identity is indicated by both color and hatch pattern

Supplementary Files

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

- Additionalfiles.docx