

Comparative Readability of Large Language Models Responses to Male Infertility Questions: Impact of Contextual Prompting

Ramy Abou Ghayda

Ramy.AbouGhayda@uhhospitals.org

Case Western Reserve University / University Hospitals Cleveland Medical Center <https://orcid.org/0000-0002-5170-3983>

Hachem Ziadeh

Case Western Reserve University / University Hospitals Cleveland Medical Center <https://orcid.org/0009-0009-8069-0964>

Thriaksh Rajan

Case Western Reserve University / University Hospitals Cleveland Medical Center

Article

Keywords: Comparative Readability, Large Language Models Responses, Male Infertility Questions, Contextual Prompting

Posted Date: February 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8577223/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: There is **NO** conflict of interest to disclose.

Abstract

Large language models (LLMs) are increasingly used by patients seeking medical information, yet the readability of LLM-generated content in male infertility remains insufficiently characterized. We evaluated the readability of responses generated by five widely used LLM platforms to frequently asked questions (FAQs) on male infertility collected from urology association and hospital websites. Fifty-four FAQs were submitted to OpenAI (ChatGPT-5/5-mini), Claude (Sonnet 4.5), Google Gemini (2.5 Flash), DeepSeek (V3), and Grok (V3/V4) in two conditions: (1) no additional context and (2) contextual prompting directing the model to explain to a lay patient/couple worried about male infertility. Readability was assessed using Flesch-Kincaid Reading Ease (FRE) and SMOG indices. In the non-prompted condition, DeepSeek generated the most readable responses (mean FRE 46.28 ± 6.86 ; mean SMOG 11.71 ± 0.76), whereas Claude produced the least readable outputs (mean FRE 23.83 ± 12.29 ; mean SMOG 16.37 ± 2.47). After prompting, Grok generated the most readable responses (mean FRE 69.71 ± 6.09 ; mean SMOG 9.96 ± 1.04), and readability improved across all models. These findings suggest that simple contextual prompting can substantially enhance readability of LLM-generated male infertility education; however, readability gains must be paired with ongoing verification of clinical accuracy to mitigate misinformation risk.

Introduction

Male infertility is defined as the inability to achieve conception after 12 months of regular, unprotected intercourse due to one or more abnormalities in sperm production, function, or delivery(1). It contributes to approximately half of all cases of couple infertility worldwide and affects an estimated 8–12% of reproductive-aged men globally(2, 3). In the United States, male factor infertility is similarly prevalent, with recent estimates suggesting that it contributes to nearly 40–50% of infertility evaluations among couples(4). Despite its frequency, male infertility remains a complex and multifactorial condition influenced by genetic, hormonal, environmental, and lifestyle factors(5). The diagnostic and therapeutic pathways can involve extensive evaluation, invasive testing, and uncertain prognoses, which collectively contribute to significant emotional, psychological, and financial burden for affected individuals(6). Patients must navigate numerous concerns, including fears related to masculinity, stigma, treatment expectations, and long-term reproductive planning(7). As a result, accessible and trustworthy patient education becomes a crucial component of care, enabling individuals to understand their condition, participate in decision-making, and engage with appropriate therapeutic options(8, 9).

Historically, patient education relied heavily on printed brochures, clinic-based counseling, and static online materials, which often varied in quality, readability, and cultural relevance(10). These traditional approaches provided limited interactivity and were not easily adaptable to individual learning needs. However, the rapid evolution of digital health resources has transformed how medical information is accessed. In the current digital era, the internet has become one of the most widely used platforms for acquiring health-related knowledge, particularly within medicine(11). A large proportion of patients now turn to online resources as their first step when attempting to understand symptoms, diagnoses, or treatment options(12). This shift has been further accelerated by the emergence of artificial intelligence (AI) driven technologies, such as large language models (LLMs)(13). LLMs represent a major advancement in this landscape. Unlike traditional search engines, which retrieve websites and require users to interpret and synthesize information themselves, LLMs such as ChatGPT, Gemini, and Claude can process natural language queries and generate synthesized, coherent, and contextually relevant responses(14, 15). Through their ability to integrate information from diverse sources, adapt output to user intent, and simplify medically complex topics, LLMs offer an appealing method for patient education(16). Their conversational nature allows users to ask follow-up questions, clarify misunderstandings, and engage in a more personalized form of information-seeking. Patients have already begun using conversational LLMs to obtain medical information and guidance, particularly regarding quality-of-life issues following treatment for various conditions(17). This growing adoption suggests that LLMs may serve as valuable tools for supplementing the role of clinicians and caregivers in disseminating reliable and understandable medical information.

An important advantage of LLMs is their ability to tailor responses through specific prompting strategies, often referred to as prompt engineering(18). By guiding the model to provide information at an appropriate literacy level or in a structured manner, clinicians and researchers can potentially improve the readability and clarity of the generated content(18). This consideration is particularly important in specialties where medical information is complex and often emotionally charged, including male infertility. Hence, ensuring readability not only facilitates comprehension but also improves treatment adherence, reduces anxiety, and supports shared decision-making(19).

Although early studies have examined LLM readability in the context of infertility, they have several limitations. Most evaluated older generations of LLMs(20), did not incorporate advanced prompting strategies(21), or did not focus exclusively on male factor infertility(22, 23). Investigations dedicated to male infertility are scarce, despite the significant differences in diagnostic evaluation, disease mechanisms, and patient concerns between male and female reproductive disorders. Additionally, comparative analyses across multiple

state-of-the-art LLMs remain limited, leaving a gap in understanding how these models perform relative to one another when delivering patient-facing reproductive health information.

To address these gaps, the primary aim of this study is to evaluate the readability and comparative performance of responses generated by the most widely used and up-to-date LLMs for patient queries related to male infertility. A secondary objective is to assess the influence of prompting versus non-prompting on the readability of LLM-generated content, thereby determining whether structured prompting improves the accessibility of information delivered to patients seeking to understand this complex condition.

Methods

This study evaluated the readability of responses produced by a selection of the most widely used large language models. The methodological design centered on identifying standardized male infertility FAQs, generating corresponding LLM responses, and analyzing these responses using validated readability metrics.

FAQs were collected from publicly available resources, including urological associations and academic hospital websites, and patient education portals. Questions related explicitly to male infertility evaluation, etiologies, treatment options, lifestyle considerations, diagnostic procedures, and prognostic factors were included. Questions that were duplicates, overly vague or insufficiently specific to male infertility, or outside the scope of the topic were excluded from the analysis. Questions were kept in their original format as presented on their respective websites to maintain consistency with the style and language typically posed by patients.

Each question was submitted to five different LLM models. The selected models were OpenAI (ChatGPT-5 & 5-mini), Claude (Sonnet 4.5), Google Gemini (2.5 Flash), DeepSeek (V3), and Grok (V3&V4). Submissions were performed in November 2025 using each platform's standard chat interface, using default settings, and without specialized plugins, to avoid altering response style. Each question was entered as a separate, independent prompt using the "New Chat" function and an anonymous window to minimize bias.

Two rounds of data collection were performed. In the first round, each FAQ was submitted to each LLM exactly as originally written, without additional context. In the second round, the identical FAQ was preceded by a brief contextual prompt: "Please explain like you would to a patient or a couple worried about male infertility, who does not have a medical background". The purpose of this prompt was to determine whether LLMs adjust their communication style toward greater readability when explicitly instructed to provide patient-oriented explanations that reflect real-world clinical interactions.

To evaluate the readability of every response generated by the LLMs in both conditions, two validated readability metrics were used; Flesch-Kincaid Reading Ease (FRE) score, which ranges from 0 to 100, with higher values indicating easier comprehension, and the Simple Measure of Gobbledygook (SMOG) index, which estimates the years of education needed to understand the text. Lower SMOG values represent more accessible writing. Descriptive statistics were used to summarize the readability findings. Mean Readability scores were calculated for each response using the *KoRpus* package in R software (R Foundation for Statistical Computing, Vienna, Austria; version 4.4.2), and scatter plots were generated to display the individual FRE and SMOG scores for prompted and non-prompted responses across all LLMs (Fig. 2–3). The primary outcome examined differences in readability among the five LLMs, with and without prompting. The secondary outcome focused on the variation in readability scores of each LLM before and after prompting. As the study did not involve any patient data, it did not require an ethical committee approval.

Results

A total of 81 patient-oriented questions related to male infertility were initially collected. After screening, 27 questions were excluded, resulting in a final set of 54 questions for analysis. The flowchart detailing the screening and exclusion process is shown in Fig. 1. All included questions were successfully submitted to the five LLMs under both non-prompted and prompted conditions, generating a complete dataset for readability evaluation.

In the non-prompted condition, not all LLMs produced content that was easily readable for a general audience, with several responses exhibiting relatively high complexity. Clear differences in readability were observed across the models, as indicated by variations in both Flesch-Kincaid Reading Ease (FRE) scores and SMOG indices, highlighting the variability in their ability to generate patient-friendly text. All mean readability scores and corresponding standard deviations for each LLM, under both prompted and non-prompted conditions, are provided in Table.1. Among the non-prompted outputs, DeepSeek produced the most readable responses, achieving the highest mean FRE score of 46.28 (SD $\pm$ 6.86), indicative of moderately accessible text, and the lowest mean SMOG index of 11.71 (SD $\pm$ 0.76) (Fig. 2–3). In contrast, Claude generated the least readable responses, with the lowest mean FRE score of 23.83 (SD $\pm$ 12.29) and the highest mean

SMOG index of 16.37 (SD $\pm$ 2.47) (Fig. 2–3). Grok ranked second lowest in readability among the non-prompted outputs, with a mean FRE of 29.97 (SD $\pm$ 10.01) and a mean SMOG index of 14.55 (SD $\pm$ 1.86) (Fig. 2–3). The OpenAI models and Gemini demonstrated intermediate readability, performing better than Claude and Grok but still producing responses that were more complex than those generated by DeepSeek (Fig. 2–3).

When contextual prompting was applied, readability improved substantially across all LLMs. The mean FRE scores increased, and mean SMOG indices decreased for every model, demonstrating that prompting had a consistent positive impact on making the text more readable. Following prompting, the order of models from most to least readable, based on both FRE and SMOG scores, was Grok, DeepSeek, Gemini, OpenAI, and Claude (Fig. 2–3). Grok showed the most pronounced improvement, achieving the highest mean FRE score of 69.71 (SD $\pm$ 6.09) and the lowest mean SMOG index of 9.96 (SD $\pm$ 1.04), indicating text that was accessible to readers at approximately a middle school to early high school reading level. DeepSeek continued to perform strongly, with mean FRE increasing to 67 (SD $\pm$ 5.57) and mean SMOG decreasing to 10.3 (SD $\pm$ 0.67). The OpenAI models and Gemini also showed marked improvements, reaching mean FRE scores of 57.25 (SD $\pm$ 7.77) and 59.3 (SD $\pm$ 7.8), respectively, alongside reductions in SMOG to 11.63 (SD $\pm$ 1.23) and 11.4 (SD $\pm$ 1.09). Although Claude improved with prompting, it remained the least readable model, with a post-prompt FRE of 51.59 (SD $\pm$ 10.24) and a SMOG index of 13.03 (SD $\pm$ 1.85). Overall, prompting improved readability across all models, though the degree of improvement varied, highlighting the beneficial effect of providing context-specific instructions (Fig. 2–3).

Discussion

The aim of this study was to evaluate the readability of responses generated by multiple large language models (LLMs) to frequently asked questions related to male infertility and to assess the effect of prompting on improving the clarity of patient-facing information. To our knowledge, this is the first study to examine readability specifically within the domain of male infertility while also comparing multiple contemporary LLMs and evaluating the influence of prompting on text accessibility.

Overall, our findings demonstrate meaningful variation in how effectively different LLMs communicate medical information to patients. The readability of LLM-generated responses varied widely across models, indicating that it's not uniform and depends heavily on the default output of each LLM. However, prompting consistently improved readability across all models, highlighting its value as a practical and easily implementable strategy to enhance the clarity of patient education materials produced by LLMs.

When examining general readability patterns, we found that non-prompted responses from most models corresponded to reading levels ranging from late high school (12th grade) to college or even graduate reading levels, which fall well above the recommended thresholds for patient-guided medical information. This is consistent with prior urologic LLM evaluations, where ChatGPT outputs for prostate cancer information and prostate cancer screening typically yielded Flesch–Kincaid grade levels around 11–13, corresponding to late high school or college reading levels, despite being intended for patients(24). Several studies have likewise shown that LLMs produce content at a reading level that exceeds recommended thresholds for patient education. Morrison et al. evaluated ChatGPT-3.5 responses to 30 patient-style urologic inquiries across multiple domains and reported an average Flesch-Kincaid Readability Grade Level of 18, corresponding to a college-graduate reading level, far above the 6th -8th grade level generally advised for patient educational materials(25). Similarly, Thia and Saluja assessed ChatGPT's responses to questions about prostate, bladder, and renal cancers using the Flesch-Kincaid reading ease score and found that readability scores were substantially lower (more difficult) than those of many institutional patient information sheets, again indicating content pitched above the average lay reading ability(26).

In contrast, our findings showed that readability of prompted responses frequently ranged from middle school to early high school reading levels. These ranges indicate that, while unprompted LLM responses may be too complex for many patients, prompting can markedly improve the accessibility of generated medical explanations.

Among all models evaluated, Grok produced the most readable content in the prompted condition, with text typically falling within the 7th -10th grade range. This aligns closely with the commonly recommended 6th -8th grade level for patient education materials(27). These observations suggest that, when appropriately guided, some contemporary LLMs, like Grok, may now achieve readability that is at least as good as, and in some cases better than, existing urology patient education resources.

Although Grok performed best post-prompting, DeepSeek closely followed, with readability levels only slightly higher. Its outputs typically aligned with late middle school to early high school levels when prompted, demonstrating that it can also generate patient-appropriate content with minimal prompting effort. The relatively small difference between Grok and DeepSeek suggests that both models may serve as effective resources for clinicians or patients seeking clear explanations of male infertility concepts. Notably, DeepSeek was the top performing model in the non-prompted condition, producing the most accessible spontaneous responses among LLMs, with outputs

generally falling within late high school reading levels. Although higher than the ideal recommended levels for patient-guided information, its output was substantially more readable than those generated by other models without prompting. This pattern of “too complex but relatively better than alternatives” is similar to findings in prostate cancer screening and general urology case scenarios, where newer LLM versions (for example GPT-4 vs GPT-3.5) showed modestly improved Flesch Reading Ease and lower grade levels, yet still remained above recommended patient targets(28, 29).

In contrast, Claude consistently exhibited the lowest readability in both prompted and non-prompted conditions, with responses frequently aligning with college or graduate-level reading categories, reflecting dense sentence structures and high linguistic complexity. These findings underscore the limitations of assuming that advanced LLMs naturally produce patient-friendly content; without targeted prompting or simplification, some models may generate text far beyond the comprehension level of the average patient.

OpenAI’s model and Gemini demonstrated intermediate performance. With prompting, both models achieved readability levels generally equivalent to early high school, though still not as accessible as those produced by Grok or DeepSeek. This parallels a renal cancer education study comparing ChatGPT-4 content with patient education materials from major urological associations, where AI-generated texts showed higher numeric readability scores, indicating greater complexity than the most readable guideline-derived materials, again pointing to residual barriers despite improvement(30). In clinical contexts, the present results suggest these models may still be useful for generating informative and reasonably clear explanations but may require more explicit prompting to reach patient-friendly readability levels.

Beyond the question-answer formats, prior studies has also evaluated LLMs for simplifying formal pathology reports(31). A study converting prostate biopsy and prostatectomy pathology reports into patient-centered summaries using ChatGPT-3.5 showed significantly higher Flesch Reading Ease scores for AI-generated versions compared with original reports, although the associated grade levels remained at or above the high-school range because of unavoidable technical terminology. Together with our finding that prompting can systematically lower grade levels across multiple models for male infertility FAQs, these data support a broader role for LLMs as tools to improve readability.

This study offers several strengths. First, it is the first to evaluate the readability of LLM-generated responses to FAQs related specifically to male infertility, addressing a clinically important yet previously unexamined area of patient education. Second, it provides a robust comparative assessment of readability across five widely used LLMs, offering meaningful insight into model-level differences in communication clarity. Third, by analyzing both prompted and unprompted outputs, the study demonstrates the practical value of prompting in improving readability, generating actionable guidance for optimizing LLM use in clinical and educational contexts. Nonetheless, several limitations should be acknowledged. Although we evaluated the readability of LLM-generated responses, we did not assess their accuracy. Improved readability must be accompanied with careful verification of content accuracy, as both elements are essential when determining an LLM’s suitability for patient education. Second, although prompting improved readability metrics, we did not examine patients’ perceptions of clarity, usefulness, or ease of understanding. Readability indices alone may not fully reflect real-world patient experiences. Future studies should therefore combine readability analysis with assessments of scientific accuracy and adherence with established clinical guidelines in male infertility. Additionally, exploring how LLMs perform in more complex clinical scenarios, such as personalized counseling, risk communication, or shared decision-making, may offer further insight into their potential role in male infertility care. Longitudinal studies evaluating the impact of LLM-assisted education on patient understanding, engagement, and outcomes are also warranted.

Conclusion

This study is the first to assess the readability of LLM-generated responses to patient questions related to male infertility, while also providing a comparative analysis across five commonly used models. DeepSeek produced the most readable non-prompted responses, whereas Grok achieved the highest readability after prompting. Claude demonstrated the lowest readability scores in both settings. Prompting substantially improved readability across all models, highlighting the value of intentional prompt design for patient communication. Because patient education is essential in male infertility care, prompted LLMs hold meaningful potential as supportive tools. However, improvements in readability must be paired with verification of accuracy to ensure information that is both clear and clinically reliable.

Declarations

Disclosures: 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.

Funding: None.

Ethical Committee Approval: Ethical approval was waived as the analysis does not contain

Human subjects data.

Study location: This study was performed at Case Western Reserve University School of Medicine, Cleveland, OH, 44106.

Acknowledgments: None

Clinical Trial Registration: Not relevant.

Level of Evidence: Level IV

References

1. Practice Committee of the American Society for Reproductive Medicine. Definitions of infertility and recurrent pregnancy loss. *Fertil Steril*. 2008 Nov;90(5 Suppl):S60.
2. Agarwal A, Mulgund A, Hamada A, Chyatte MR. A unique view on male infertility around the globe. *Reprod Biol Endocrinol RBE*. 2015 Apr 26;13:37.
3. Leslie SW, Soon-Sutton TL, Khan MA. Male Infertility. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 [cited 2025 Dec 8]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK562258/>
4. Eisenberg ML, Lathi RB, Baker VL, Westphal LM, Milki AA, Nangia AK. Frequency of the male infertility evaluation: data from the national survey of family growth. *J Urol*. 2013 Mar;189(3):1030–4.
5. Agarwal A, Baskaran S, Parekh N, Cho CL, Henkel R, Vij S, et al. Male infertility. *The Lancet*. 2021 Jan 23;397(10271):319–33.
6. Simionescu G, Doroftei B, Maftai R, Obreja BE, Anton E, Grab D, et al. The complex relationship between infertility and psychological distress (Review). *Exp Ther Med*. 2021 Apr;21(4):306.
7. Sahoo S, Das A, Dash R, Behera A, Mishra N, Bal K. The Psychological Impact of Male Infertility: A Narrative Review. *Cureus*. 17(8):e89453.
8. Leung AK, Henry MA, Mehta A. Gaps in male infertility health services research. *Transl Androl Urol*. 2018 Jul;7(Suppl 3):S303-S30309.
9. Robins S, Barr HJ, Idelson R, Lambert S, Zelkowitz P. Online Health Information Regarding Male Infertility: An Evaluation of Readability, Suitability, and Quality. *Interact J Med Res*. 2016 Oct 21;5(4):e6440.
10. Sørensen K, Levin-Zamir D, Duong TV, Okan O, Brasil VV, Nutbeam D. Building health literacy system capacity: a framework for health literate systems. *Health Promot Int*. 2021 Dec 13;36(Supplement_1):i13–23.
11. Ogundiya O, Rahman TJ, Valnarov-Boulter I, Young TM. Looking Back on Digital Medical Education Over the Last 25 Years and Looking to the Future: Narrative Review. *J Med Internet Res*. 2024 Dec 19;26(1):e60312.
12. Frey E, Bonfiglioli C, Brunner M, Frawley J. Parents' Use of Social Media as a Health Information Source for Their Children: A Scoping Review. *Acad Pediatr*. 2022 May 1;22(4):526–39.
13. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. *J Med Internet Res [Internet]*. 2024 Jan 1 [cited 2025 Sep 10];26. Available from: <https://www.sciencedirect.com/org/science/article/pii/S143888712400757X>
14. Hirose T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int J Environ Res Public Health*. 2023 Jan;20(4):3378.
15. Jo E, Epstein DA, Jung H, Kim YH. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems [Internet]*. New York, NY, USA: Association for Computing Machinery; 2023 [cited 2025 Sep 10]. p. 1–16. (CHI '23). Available from: <https://dl.acm.org/doi/10.1145/3544548.3581503>
16. Pal A, Wangmo T, Bharadia T, Ahmed-Richards M, Bhandari MB, Kachhadiya R, et al. Generative AI/LLMs for Plain Language Medical Information for Patients, Caregivers and General Public: Opportunities, Risks and Ethics. *Patient Prefer Adherence*. 2025 Jul 31;19:2227–49.

17. Cevasco KE, Morrison Brown RE, Woldeeslassie R, Kaplan S. Patient Engagement with Conversational Agents in Health Applications 2016–2022: A Systematic Review and Meta-Analysis. *J Med Syst.* 2024;48(1):40.

18. Jung H, Oh J, Stephenson KAJ, Joe AW, Mammo ZN. Prompt engineering with ChatGPT3.5 and GPT4 to improve patient education on retinal diseases. *Can J Ophthalmol.* 2025 Jun 1;60(3):e375–81.

19. Laugesen J, Hassanein K, Yuan Y. The Impact of Internet Health Information on Patient Compliance: A Research Model and an Empirical Study. *J Med Internet Res.* 2015 Jun 11;17(6):e143.

20. Sengupta P, Dutta S, Chakravarthi S, Jegasothy R, Jeganathan R, Pichumani A. Comparative efficacy of ChatGPT 3.5, ChatGPT 4, and other large language models in gynecology and infertility research. *Gynecol Obstet Clin Med.* 2023 Dec 1;3(4):203–6.

21. Bahçeci T, Elmaağaç B, Ceyhan E. Comparative analysis of the effectiveness of microsoft copilot artificial intelligence chatbot and google search in answering patient inquiries about infertility: evaluating readability, understandability, and actionability. *Int J Impot Res.* 2025 Apr 22;1–6.

22. Naghdi M, Cao P, Essers R, Heijligers M, Paulussen ADC, van der Lugt A, et al. Artificial intelligence-simplified information to advance reproductive genetic literacy and health equity. *Hum Reprod.* 2025 Sep 1;40(9):1681–8.

23. Li W, Ur-Rehman A, Ge MW, Shen LT, Peng XY, Zhong K, et al. The potential of large language models in the field of infertility: a systematic review. *J Assist Reprod Genet.* 2025 Nov;42(11):3565–79.

24. 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(2):394–9.

25. Morrison C, Vercnocke J, Moser AM, Cher ML, Lucas S, Cochrane J, et al. Are ChatGPT’s Responses to Urologic Inquiries Readable and Supported by AUA Guidelines? *Int J Urol Nurs.* 2025;19(3):e70023.

26. Thia I, Saluja M. ChatGPT: Is This Patient Education Tool for Urological Malignancies Readable for the General Population? *Res Rep Urol.* 2024 Jan 16;16:31–7.

27. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *J Patient Exp.* 2021 Mar 3;8:2374373521998847.

28. Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis.* 2024 Mar;27(1):159–60.

29. Chiarelli G, Stephens A, Finati M, Cirulli GO, Beatrice E, Filipas DK, et al. Adequacy of prostate cancer prevention and screening recommendations provided by an artificial intelligence-powered large language model. *Int Urol Nephrol.* 2024 Aug;56(8):2589–95.

30. Halawani A, Almeahmadi SG, Alhubaishy BA, Alnefaie ZA, Hasan MN. Empowering patients: how accurate and readable are large language models in renal cancer education. *Front Oncol [Internet].* 2024 Sep 26 [cited 2025 Dec 7];14. Available from: <https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2024.1457516/full>

31. Khanmammadova N, Gevorkyan R, Afyouni AS, Cumpanas AD, Jiang D, Myklak K, et al. Evaluation of prostate cancer pathology reports generated by ChatGPT to enhance patient comprehension. *Sci Rep.* 2025 Oct 2;15:34459.

Table

Table 1: Readability Scores of Large Language Models Responses to Male Infertility FAQs Before and Corresponding Reading Levels Before and After Prompting.

LLM	FRE Score		SMOG Index					
	Non Prompted	Equivalent Reading Level	Prompted	Equivalent Reading Level	Non Prompted	Equivalent Reading Level	Prompted	Equivalent Reading Level
Open AI	31.89 ± 13.94	College level	57.25 ± 7.77	10 th -12 th grade	12.75 ± 1.84	12 th Grade - College	11.63 ± 1.23	11 th -12 th grade
Claude	23.83 ± 12.29	College graduate level (Very Difficult)	51.59 ± 10.24	10 th -12 th grade	16.37 ± 2.47	College (Senior -postgraduate)	13.03 ± 1.85	College (1 st year)
Gemini	38.18 ± 8.83	College level	59.3 ± 7.8	10 th -11 th grade	13.35 ± 1.38	College (Freshman/Sophomore)	11.4 ± 1.09	11 th grade
DeepSeek	46.28 ± 6.86	12 th grade - college level	67 ± 5.57	8 th -9 th grade	11.71 ± 0.76	11 th -12 th grade	10.3 ± 0.67	10 th grade
Grok	29.97 ± 10.01	College level (Very difficult)	69.71 ± 6.09	7 th -8 th grade	14.55 ± 1.86	College (Sophomore)	9.96 ± 1.04	9 th -10 th grade

Figures

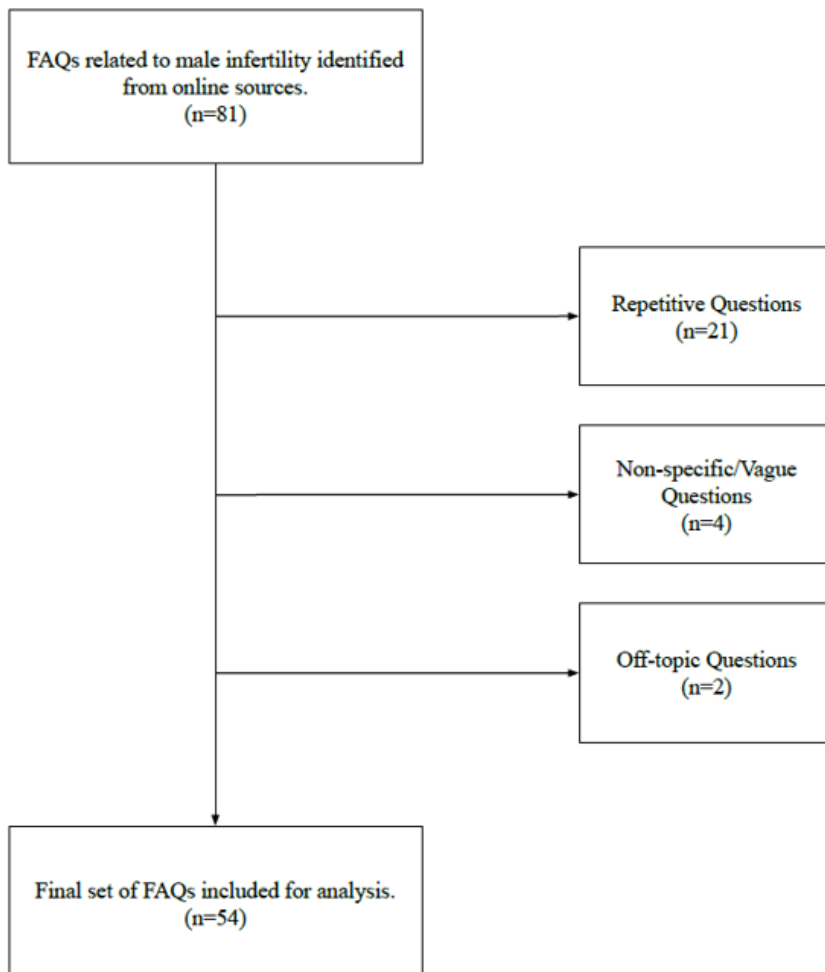


Figure 1

Flowchart of Screening and Selection of Male Infertility FAQs included in the study.

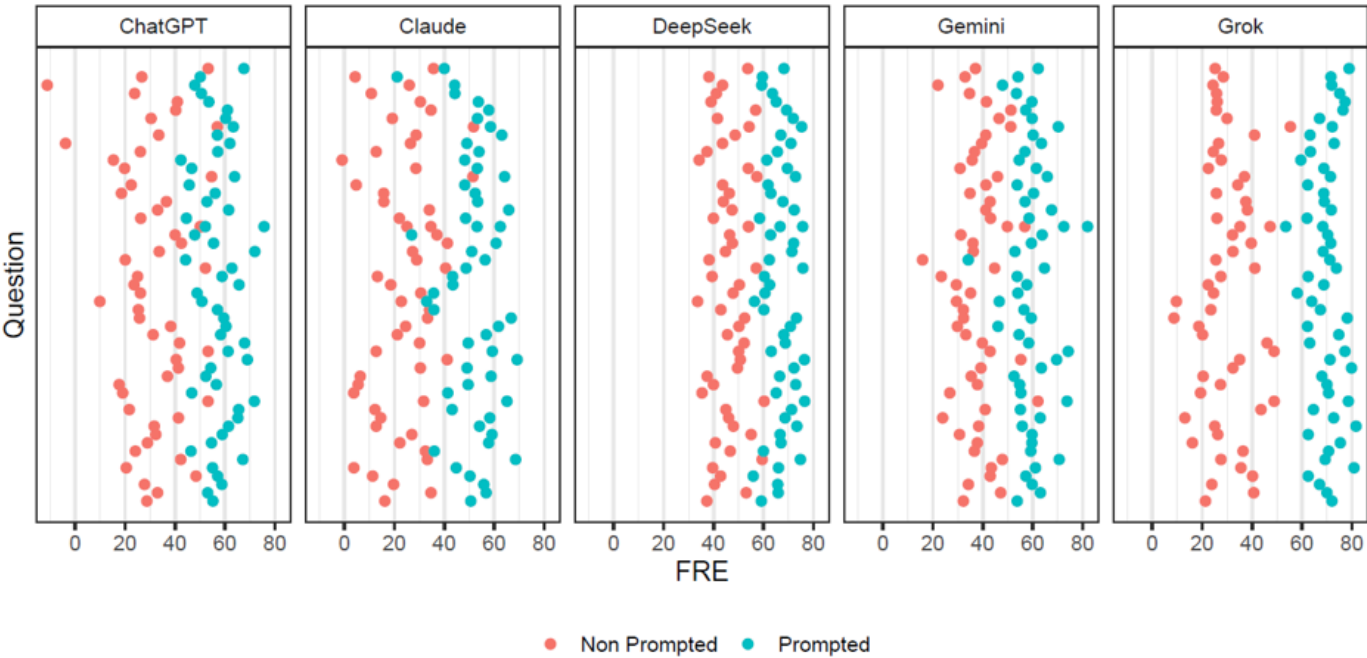


Figure 2

Distribution of Flesch-Kincaid Reading Ease (FRE) Readability Scores for Each Response, Showing Prompted and Non-Prompted Outputs Across the Five Evaluated LLMs.

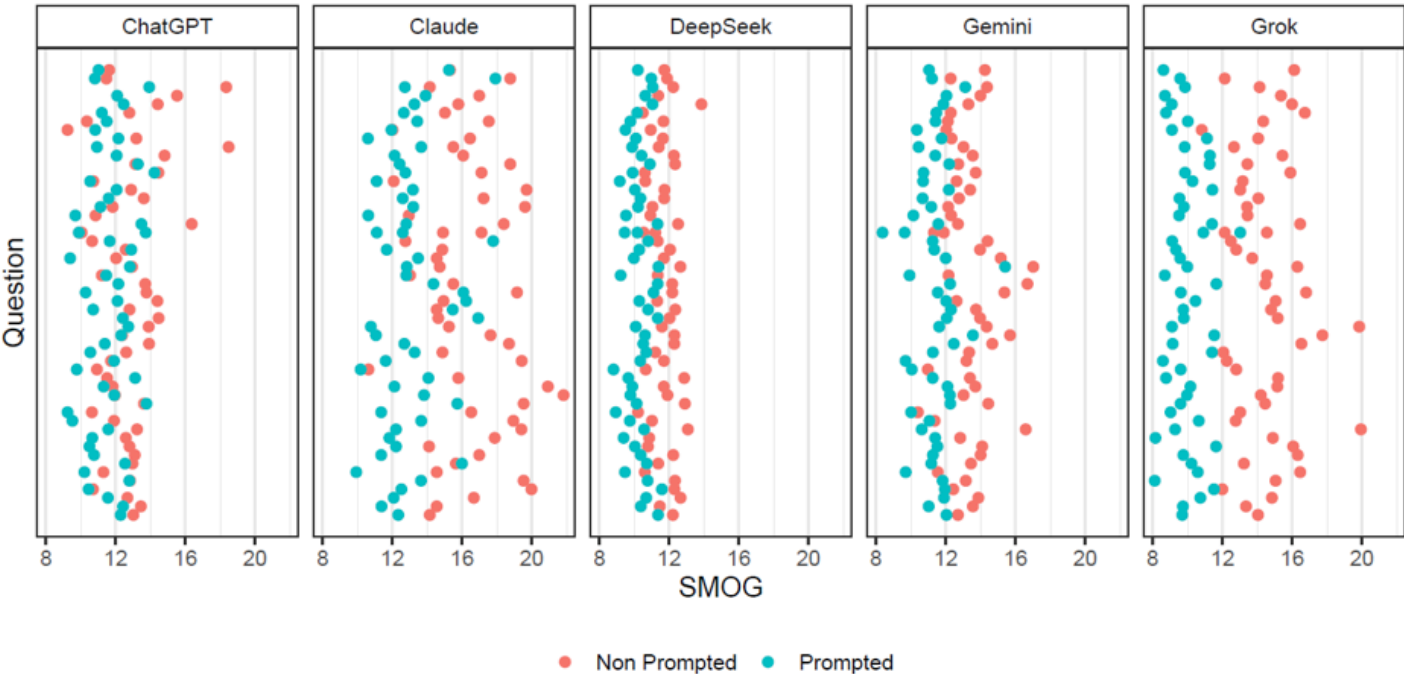


Figure 3

Distribution of Simple Measure of Gobbledygook (SMOG) Index Scores for Each Response, Showing Prompted and Non-Prompted Outputs Across the Five Evaluated LLMs.

Supplementary Files

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

- [Supplement.1Questions.xlsx](#)