ORIGINAL ARTICLE

Annals of Agricultural and Environmental Medicine
www.aaem.pl

Effectiveness of ChatGPT in simplifying information for Polish patients on cervical cancer screening using advanced prompting

Magdalena Dębska[1,A-B,D], **Joanna Gotlib-Małkowska**[1,A-F]

[1] *Department of Education and Research in Health Sciences, Medical University, Warsaw, Poland*
A – Research concept and design, B – Collection and/or assembly of data, C – Data analysis and interpretation, D – Writing the article, E – Critical revision of the article, F – Final approval of the article

Dębska M, Gotlib-Małkowska J. Effectiveness of ChatGPT in simplifying information on cervical cancer using advanced prompting. Ann Agric Environ Med. doi: 10.26444/aaem/204249

Abstract

Introduction and Objective. ChatGPT can generate reliable medical information in gynaecology and obstetrics,but the content is often difficult to understand for patients with lower educational levels. The aim of the study is to evaluate the impact of Audience Persona Prompting on the simplification and readability of ChatGPT-generated medical information on cervical cancer screening (MICC_GPT) in Polish.

Materials and Method. 392 MICC_GPT were analyzed, with 196 generated using Zero-Shot Prompting (STANDARD) and 196 generated using Audience Persona Prompting (EASY). The Audience Persona prompts included instructions to simplify the content: 'Explain as if to an average Polish woman with only primary education' (8 years of formal schooling). Readability was assessed using 24 objective linguistic indicators available at Jasnopis.pl. Statistica 13 (StatSoft, Poliand), the Brunner-Munzel test, $p < 0.05$.

Results. The average difficulty level of STANDARD output was 5.32 (at least 15 years of formal education), while EASY output averaged 4.15 (12 years of formal education). Of the 24 indicators, 21 showed statistically significant improvements in the simplification of EASY output ($p < 0.05$). While ChatGPT significantly simplified MICC_GPT, the readability levels remained too high for patients with only primary education.

Conclusions. ChatGPT shows promise in tailoring medical information on cervical cancer (CC) screening for the needs of Polish patients with varying educational backgrounds, with the use of advanced prompt engineering techniques. However, further research is required to refine prompt engineering methods and develop effective strategies for generating information on cervical cancer screening that is accessible to individuals with only primary education.

Key words

readability, ChatGPT, cervical cancer information, simplification of medical information, Audience Persona Prompting

INTRODUCTION

Cervical cancer screening is discussed on social media platforms with varying frequency, often peaking during global health campaigns, such as Cervical Health Awareness Month and World Cancer Day. Platforms like Twitter, Instagram, and TikTok play an increasingly important role in disseminating educational content, survivor narratives, and public health messages, particularly through the use of hashtags and influencer engagement. Although awareness-related content is common, messages that directly encourage screening behaviour remain relatively infrequent. Given the widespread reliance on social media for health information – especially among younger audiences – strategically leveraging these platforms is crucial for enhancing public engagement, combating misinformation, and promoting preventive behaviours related to cervical cancer.

A scoping review conducted by Plackett et al. synthesized existing evidence on social media-based interventions for improving cancer screening and early diagnosis. The review concluded that although such interventions can effectively raise awareness, there is limited evidence regarding their direct impact on actual screening uptake. The authors advocate for more robust methodologies and long-term evaluations to better assess behavioural outcomes [1].

In contrast to Plackett et al., Lyson et al. emphasized that social media engagement does not necessarily translate into guideline-concordant cervical cancer screening behaviours. Using data from the Health Information National Trends Survey, the authors found a significant negative association between social media use and adherence to recommended screening guidelines (OR = 0.57; 95% CI = 0.33–0.96), suggesting that while social media serves as a prevalent source of health information, it may also contribute to confusion or misinformation [2]. Similarly, in the study of Zheng et al., researchers examined the effect of a targeted social media intervention aimed at increasing knowledge about HPV and cervical cancer prevention. Participants received tailored messages over a five-day period. While no significant changes were observed in general knowledge or preventive behaviour, a modest but statistically significant increase in HPV awareness was noted (90% – 94%; $p = 0.003$), highlighting the potential – but also the limitations -of brief social media interventions [3].

These findings collectively underscore the complexity of using social media as a tool for public health promotion and point to the need for well-designed, evidence-based digital interventions to support informed cervical cancer screening practices [1–3].

✉ Address for correspondence: Joanna Gotlib-Małkowska, Department of Education and Research in Health Sciences, Medical University, Warsaw, Poland
E-mail: joanna.gotlib@wum.edu.pl

Received: 02.03.2025; accepted: 18.04.2025; first published: 06.05.2025

Over the past two years, large language models (LLMs), such as ChatGPT (OpenAI), Gemini (Google), and Copilot (Microsoft), have become increasingly popular sources of information for patients seeking medical advice online, in much the same way that such search engines as Google, became ubiquitous a few years ago [4]. Their accessibility and ease of use have contributed to their growing popularity. Despite initial concerns, analyses of the available international scientific literature indicate that medical information generated by LLMs, regardless of medical specialty, is generally reliable and not harmful to patients [5–10]. However, although misinformation is rare, researchers highlight that health-related information produced by LLMs is not always up-to-date or consistent with the latest guidelines and scientific evidence [5–10].

One of the key findings in recent scientific literature is that ChatGPT-generated health information may be inappropriate for the average user as it is too complex and requires a higher level of expertise, preferably in medicine [11]. There are also recent reports indicating ChatGPT's ability to simplify medical information, such as imaging reports, so that it can be understood by patient groups with varying levels of education [12]. However, there is a notable lack of research investigating the level of readability of ChatGPT-generated medical information in Polish, and the ability of ChatGPT to simplify medical information for patients with different literacy levels.

OBJECTIVE

The aim of the study is to evaluate the efficacy of the advanced Audience Persona Prompting in adjusting the readability of ChatGPT-generated medical information on cervical cancer screening (MICC_GPT) in Polish.

MATERIALS AND METHOD

Study design. The study followed the guidelines for good practice in healthcare education research using content generated by artificial intelligence algorithms, including LLMs, as outlined by Sallam at al. [13]. According to Sallam et al., there are nine aspects to consider in the design of a study and the subsequent description of its results: 1) the design of the LLM model used to generate the content, 2) the methods of evaluation of the output data (objective vs. subjective assessment), 3) the exact time and date of the LLM model testing and output generation, 4) transparency of the input data, 5) scope of the input data, 6) randomness of the input data, 7) individual factors affecting the consistency of evaluation of input and output data, 8) the number of queries performed, and 9) prompt design [13].

Advanced prompt engineering. Audience Persona Prompting is an advanced method of prompt creation that involves providing clear and transparent information in the prompt itself that defines the target audience of the ChatGPT-generated information [14]. For the present study, the prompt designed to generate MICC_GPT easy to understand for a person with elementary education included additional instructions: *'Explain as if to an average Polish woman with only primary education (eight years of formal education)'*.

Formulation of cervical cancer screening questions. A detailed list of topics was developed by the authors specifically for the purposes of this study. However, it was created based on a thorough analysis of the scientific literature concerning women's knowledge about cervical cytology, as well as the most frequently asked questions on this subject. The authors conducted a comprehensive analysis of recent international scientific publications on women's knowledge of cervical cancer (CC) screening [15–18], identified five key thematic areas, and created an initial database of 76 questions on CC screening divided into five thematic categories: I) general information about the test (16 items); II) preparation for the test (15 items); III) specimen collection for testing (16 items); IV) interpretation of the test results (16 items); and V) management of an abnormal test result (13 items). This approach ensured that the content was both relevant and grounded in existing research.

Expert review of the usefulness of CC screening questions. Seven of 76 questions were rejected as unsuitable following a preliminary assessment of their usefulness by an expert midwife (MS), a professor with 14 years' experience and substantial academic achievements in midwifery.

Subsequently, a group of 20 licensed expert midwives further assessed the relevance of the remaining questions. On 4 October 2024, the experts completed a review form online (Google Forms, available at https://forms.gle/jMVzoWoF1UdeWsXi8) to anonymously rate the usefulness of the remaining questions. The rating was conducted on a scale of: 1 – 'definitely not useful'; 2 – 'rather not useful'; 3 – 'no opinion'; 4 – 'rather useful'; 5 – 'definitely useful'.

Items that received a minimum of 50% of the 'definitely useful' responses or a minimum of 50% of the combined 'rather useful' and 'definitely useful' responses, were included in the subsequent stage of the study. Twenty questions did not meet the expert criterion for inclusion, leaving 49 items in the final stage of the study (Fig. 1, App. 1).

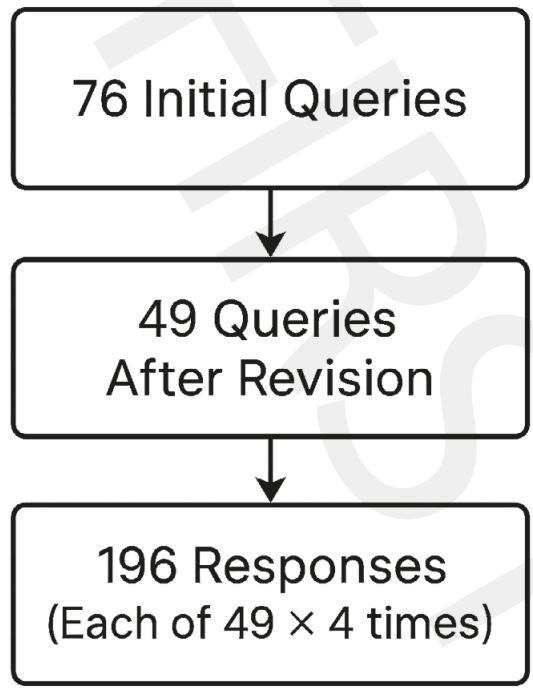

Figure 1. Flowchart of question processing

Appendix 1.
49 items included in the final stage of the study

1. Why is a Pap smear (cervical cytology) performed?
2. What are the types of cervical cytology?
3. Is liquid-based cytology better?
4. Does a Pap smear detect HPV?
5. Does a Pap smear detect sexually-transmitted infections?
6. Does a Pap smear detect infections?
7. When should the first Pap smear be performed?
8. Until what age should Pap smears be performed?
9. How often should a Pap smear be performed?
10. Can a Pap smear be done during pregnancy?
11. Can Pap smears be stopped after menopause?
12. Is it necessary to have Pap smears after a hysterectomy?
13. Do I need a Pap smear if I have only one sexual partner?
14. Do I need a Pap smear if I'm no longer sexually active?
15. Do I still need Pap smears after getting the HPV vaccine?
16. Does a Pap smear require any special preparation?
17. Can a Pap smear be performed before becoming sexually active?
18. Do I need a Pap smear if I have no concerning symptoms?
19. Should vaginal suppositories be used before a Pap smear?
20. What is the best day of the menstrual cycle to have a Pap smear?
21. Why can't a Pap smear be done during menstruation?
22. How long before a Pap smear should I avoid sexual intercourse?
23. Is an infection a contraindication for having a Pap smear?
24. Should I bring my previous Pap smear result to the appointment?
25. Are there any special recommendations for Pap smears during menopause?
26. Can a Pap smear be done after treating an infection?
27. Where can I get a Pap smear?
28. Who can collect a Pap smear?
29. How is a Pap smear collected?
30. Is a Pap smear painful?
31. Why might there be bleeding after a Pap smear?
32. What if cells from the cervical canal cannot be collected?
33. How long does it take to collect a Pap smear?
34. How long does it take to get Pap smear results?
35. What can affect the accuracy of a Pap smear result?
36. Does a normal Pap smear mean I never have to repeat it?
37. What is considered a normal Pap smear result?
38. What does NILM mean in a Pap smear result?
39. What does CIN 1 mean in a Pap smear result?
40. What does CIN 2 mean in a Pap smear result?
41. What does CIN 3 mean in a Pap smear result?
42. What does cervical dysplasia mean?
43. What can cause an abnormal Pap smear result?
44. What should be done after receiving an abnormal Pap smear result?
45. What is the follow-up procedure after an abnormal Pap smear?
46. How soon should an abnormal Pap smear result be discussed with a doctor?
47. If I have an abnormal Pap smear result, should I be tested for HPV?
48. If the result of a Pap smear is abnormal, should it be repeated?
49. Does the HPV vaccine treat abnormal cervical cells?

The process of generation STANDARD and EASY MICC_GPT. The responses to the 49 questions ultimately included in the study were generated by the ChatGPT-4omni Plus chatbot model (model GPT-4), with the personalisation, ChatGPT learning, and customization features disabled. The output was generated by the same researcher (JGM) as in the pilot study. The same computer and IP number were used. To eliminate the potential impact of the ChatGPT model's learning effects, each response was generated in a distinct conversation window. When ChatGPT generated two answers concurrently, both answers were stored in the database for subsequent analysis.

The initial stage of the study involved the zero-shot prompting generation of 196 STANDARD MICC_GPT responses: ChatGPT generated answers to 49 questions. Each answer was generated four times, resulting in a total of 196 responses (49 questions × 4 answers). They were generated on four occasions between 7 – 11 October 2024, starting with Question 1. The detailed results of the analysis and assessment of the content quality of the MICC_GPT items generated in the first stage of the study are described in Michalska et. al. [18].

The second stage of the study generated 196 EASY MICC_GPT responses using Audience Persona Prompting. In addition to the inquiry about CC screening, the prompt designed to generate an MICC_GPT that is easily comprehensible to a person with only primary education, included additional instructions: *'Explain as if to an average Polish woman with only primary education (eight years of formal education)'*. The EASY MICC_GPT output was generated on four occasions between 26 November – 3 December 2024, starting with Question 1.

An Excel spreadsheet was used to record all questions and answers. The data was cleaned of special characters and formatting, such as the symbols '###' and '**' used to mark some of the words selected by ChatGPT (the database is available upon request).

Analysis of the readability of STANDARD and EASY MICC_GPT. The readability analysis of the MICC_GPT output was carried out using the Jasnopis.pl web application available at: https://www.jasnopis.pl/. The general level of readability was measured on a scale from 1: an easy-to-understand text, for an audience with primary education, to 7: very difficult-to-understand text, for an audience with a doctoral degree or expert knowledge in the field.

The detailed evaluation of text readability included the analysis of 24 indicators covering structural and lexical dimensions, also with the use of the Jasnopis.pl web application. The data set comprised 392 responses, evenly distributed between STANDARD and EASY MICC_GPT categories.

Descriptive statistical measures, including the mean, standard deviation, and range (minimum and maximum values), were calculated for each indicator to characterize the distribution and variability inherent in the data. The Brunner-Munzel test with random permutation was used to assess the statistical significance of differences between the two text categories. This method, a robust non-parametric approach, is particularly suitable for comparing two independent groups, as it avoids the assumption of normality.

Effect sizes were determined using the relative effect

metric, which provides a probabilistic interpretation of group differences. Specifically, the relative effect quantifies the probability that a randomly selected value from the EASY group will be less than or equal to a randomly selected value from the STANDARD group.

To further contextualize the results, 95% confidence intervals for the effect sizes were calculated to ensure that the results were both interpretable and robust. All statistical analyses were performed using Jamovi software (version 2.3), with a pre-set significance threshold of $\alpha = 0.05$ (Statistica 13 Software, StatSoft Sp. z o.o. Co., Polish) [19].

RESULTS

Readability level of STANDARD and EASY MICC_GPT. The average difficulty level of the STANDARD MICC_GPT was 5.32, and full understanding of the STANDARD MICC_GPT required the target audience to have completed at least 15 years of formal education (bachelor's or engineering degree).

The average difficulty level of the EASY MICC_GPT was 4.15, and required the target audience to have completed at least 12 years of formal education (high school or equivalent).

Of the 24 readability indicators analyzed, 21 showed statistically significant differences between STANDARD and EASY MICC_GPT ($p<0.001$). The only indicators where the differences were not statistically significant were related to the length of the generated texts, including the number of paragraphs, sentences, and average paragraph length in words ($p>0.05$).

Analysis of the results showed that the EASY MICC_GPT items have a significantly lower linguistic complexity and a lower cognitive load compared to the STANDARD ones. These differences were found across a wide range of structural and lexical indicators, showing coherence of effects across linguistic dimensions. Large effect sizes and narrow confidence intervals confirmed the significance of the findings and demonstrated their methodological soundness and practical relevance. These results suggest that linguistic simplification in EASY texts effectively reduces cognitive load, which may have important implications for the design of patient education materials.

Overall, ChatGPT was able to significantly simplify MICC_GPT items generated using advanced prompt engineering and audience persona prompting. However, the readability level was still too high for MICC_GPT texts generated in Polish to be understood by audiences with elementary education.

DISCUSSION

The present study demonstrates that advanced prompt engineering – specifically Audience Persona Prompting (APP) – can significantly reduce the linguistic complexity of medical information generated by ChatGPT in Polish, making it more accessible to users with lower education levels. While the simplified (EASY MICC_GPT) texts were consistently more readable than their STANDARD counterparts across nearly all linguistic indicators, they still required a minimum of secondary education to be fully understood. This indicates that, although APP enhances readability, further refinements in prompt design are necessary to achieve true accessibility for audiences with only primary education.

These findings align with and extend the results of Haver et al., who used ChatGPT to simplify responses to common breast cancer screening questions [12]. Their study showed an improvement in readability metrics, such as the Flesch Reading Ease score (46 – 70), and a reduction in the average reading level from college-level (13th grade) to 8.9. Yet, only a small fraction (8%) of the responses reached the desired 6th-grade reading level, reflecting the same challenges observed in the current study – namely, that even well-engineered prompts may not consistently produce content suitable for individuals with limited literacy.

The results obyained contribute novel insights by examining the effectiveness of LLMs in a non-English, under-represented language (Polish), and by focusing on preventive healthcare education rather than diagnostic communication. Previous studies, such as those by Schmidt et al. [20] and Lyu et al. [21], primarily investigated the simplification of radiological reports (e.g., MRI or CT scans) in English. While these studies reported that ChatGPT could produce outputs rated as understandable and accurate by both patients and clinicians, they focused on more structured forms of medical documentation rather than general health education materials.

In contrast, the current study evaluated the simplification of complex, unstructured content across five thematic domains related to cervical cancer screening. The significance of the approach lies in the comprehensive linguistic analysis using 24 structural and lexical indicators, which revealed statistically significant differences in 21 of them. These differences were particularly pronounced in metrics related to lexical density, sentence length, word complexity, and syntactic structure – all crucial dimensions affecting the cognitive load of health information.

Moreover, methodology of the current study builds upon previous work by incorporating a population-specific prompt design strategy. Whereas earlier studies, such as those by Grünebaum et al. [11], demonstrated improved readability through general simplification tactics, the use of persona-driven prompts marks an evolution in prompt engineering – tailoring the LLM's outputs to reflect not only linguistic simplicity, but also socio-demographic characteristics of the target audience. This offers a more nuanced and patient-centreed approach to health communication, particularly important in addressing health disparities caused by low health literacy.

The implications of these findings are twofold. First, they reaffirm the growing role of generative AI tools, such as ChatGPT, in producing educational materials that support patient empowerment and informed decision-making. Second, they highlight the limitations of current models in meeting the needs of the most vulnerable populations – those with only basic education or low digital literacy. As noted by Jeblick et al. [22], while ChatGPT can make radiology reports more accessible, approximately half of their simplified outputs contained minor omissions or inaccuracies, emphasizing the importance of expert review and oversight. The results of the current study support this position and further suggest that even when readability improves, clinical validation and user testing are necessary to ensure the material's safety, trustworthiness, and practical utility.

Table 1. Results of Statistical Comparative Analysis of Readability of Linguistic Indicators for Standard and Easy MICC_GPT Outputs Using Jasnopis.pl

Index	Output	Mean	SD	Min.	Max.	Statistic	df	p-value*	Relative effect	95% CI	
										Lower	Upper
Readability	EASY	4.21	0.94	2	7	15.413	385	**<.001**	0.82	0.78	0.86
	STAND	5.32	0.83	3	7						
FOG: headwords	EASY	10.91	3.25	0.40	36.30	10.663	375	**<.001**	0.76	0.71	0.81
	STAND	13.14	2.31	7.20	20.60						
FOG: run-on words	EASY	12.06	3.50	2.00	39.80	12.508	388	**<.001**	0.79	0.74	0.83
	STAND	14.74	2.46	7.79	21.73						
FOG: low-frequency headwords	EASY	9.44	3.27	2.30	35.00	7.921	390	**<.001**	0.71	0.66	0.76
	STAND	10.76	2.08	6.03	16.51						
L-Pisarek: headwords	EASY	10.05	3.16	0.90	39.40	9.188	381	**<.001**	0.73	0.68	0.78
	STAND	11.55	1.86	6.92	17.55						
L-Pisarek: run-on words	EASY	10.93	3.22	1.10	40.80	10.787	390	**<.001**	0.76	0.71	0.81
	STAND	12.86	2.04	7.42	18.51						
L-Pisarek: low-frequency headwords	EASY	8.82	3.04	1.00	38.40	7.415	390	**<.001**	0.70	0.64	0.75
	STAND	9.82	1.73	5.89	14.59						
NL-Pisarek: headwords	EASY	9.99	3.46	1.70	41.00	8.318	389	**<.001**	0.71	0.66	0.77
	STAND	11.44	2.01	6.98	17.78						
NL-Pisarek: run-on words	EASY	10.65	3.53	2.00	38.80	10.836	390	**<.001**	0.76	0.71	0.81
	STAND	12.71	2.16	7.26	18.66						
NL-Pisarek: low-frequency headwords	EASY	9.19	3.35	2.45	36.20	5.472	389	**<.001**	0.65	0.60	0.71
	STAND	10.07	2.00	5.81	15.60						
Number of paragraphs	EASY	4.85	4.60	1	42	1.196	373	0,233	0.54	0.48	0.59
	STAND	5.99	5.87	1	30						
Number of sentences	EASY	9.95	16.83	2	159	-1.748	360	0.081	0.45	0.39	0.51
	STAND	8.61	6.40	3	36						
Number of words	EASY	110.59	56.11	2	286	3.613	377	**<.001**	0.60	0.55	0.66
	STAND	145.39	77.20	41	462						
Number of difficult words	EASY	7.79	4.76	1	34	7.520	382	**<.001**	0.70	0.65	0.75
	STAND	12.23	7.55	1	45						
Average word length [syllables]	EASY	3.56	4.71	2.00	37.60	9.159	306	**<.001**	0.74	0.69	0.80
	STAND	2.42	0.12	2.04	2.83						
Average sentence length [words]	EASY	16.79	6.46	1.80	55.50	4.674	379	**<.001**	0.63	0.58	0.69
	STAND	18.49	4.14	10.20	30.00						
Average paragraph length [words]	EASY	33.17	14.43	1.40	86.00	-0.678	381	0,498	0.48	0.42	0.54
	STAND	33.17	13.89	11.00	101.00						
Percentage of difficult words	EASY	7.09	5.90	0.00	41.40	7.973	362	**<.001**	0.71	0.66	0.76
	STAND	8.41	2.75	0.90	16.50						
Percentage of nouns	EASY	34.52	10.49	1.00	50.50	11.220	387	**<.001**	0.77	0.72	0.81
	STAND	41.70	4.61	28.80	52.10						
Percentage of difficult nouns	EASY	4.82	4.30	0.00	41.00	4.272	389	**<.001**	0.62	0.57	0.68
	STAND	5.33	2.50	0.80	13.90						
Percentage of verbs	EASY	17.38	5.42	0.18	38.70	-10.379	340	**<.001**	0.24	0.19	0.29
	STAND	14.15	3.16	8.20	25.60						
Percentage of difficult verbs	EASY	2.66	5.31	0.00	46.80	-3.508	390	**<.001**	0.40	0.35	0.46
	STAND	1.13	1.06	0.00	5.30						
Percentage of adjectives	EASY	13.77	4.94	0.19	42.20	10.081	388	**<.001**	0.75	0.70	0.80
	STAND	17.37	3.51	8.90	27.10						
Percentage of difficult adjectives	EASY	5.02	4.41	0.00	40.80	9.385	336	**<.001**	0.74	0.69	0.79
	STAND	6.53	2.19	0.00	12.90						

* Brunner-Munzel test with random permutation; CI – Confidence Interval

Table 2. Comparison of Key Findings Across Studies

Study	Language	Medical Domain	Prompt Engineering Used	Improvement in Readability	Limitations Highlighted
Current Study (Dębska et al.)	Polish	Cervical cancer screening	Audience Persona Prompting	YES – statistically significant across 21/24 indicators	Still too complex for primary education; lacks clinical validation
Haver et al. [12]	English	Breast cancer screening	General simplification prompt	YES – improved Flesch score from 46 to 70	Only 8% of texts reached 6th-grade level
Grünebaum et al. [11]	English	Obstetric informed consent	8th-grade readability target	YES – FRE score improved from 30.34 to 67.39	Requires expert oversight to ensure legal validity
Schmidt et al. [20]	English	Radiology (MRI reports)	Simple explanation prompts	YES – Patients rated simplified texts as clearer	Omissions in complex cases noted by clinicians
Lyu et al. [21]	English	Radiology (CT/MRI reports)	Layperson-targeted prompt	YES – 5th-grade level achieved in simplifications	Occasional over-simplification of findings
Jeblick et al. [22]	English	Radiology (general)	Child-level explanation prompt	YES – Radiologists rated simplifications as accurate	51% of outputs had omissions or hallucinations

Taken together, the present study contributes to the emerging consensus that large language models hold great promise for enhancing health communication, but must be guided by precise prompt design and professional oversight. Importantly, this study offers some of the first evidence on this topic in the Polish language, thus filling a critical gap in the international literature and laying the groundwork for future cross-linguistic and cross-cultural comparisons.

A comparative overview of key findings from studies evaluating the effectiveness of large language models (LLMs), such as ChatGPT, in simplifying medical information for patients, are presented in Table 2.

The overview includes six studies representing diverse languages, medical domains, prompt engineering strategies, and levels of content simplification. While all studies demonstrated improvements in readability, each one also highlighted specific limitations, including the continued inaccessibility of content for individuals with low educational attainment, the potential omission of critical medical information, and the ongoing need for expert oversight to ensure clinical accuracy and patient safety.

Studies on the application of large language models (LLMs), such as ChatGPT, ChatGPT-4, Gemini, and Claude, confirm their effectiveness in simplifying complex medical information and tailoring it to the needs of different target groups. Numerous analyses have demonstrated that these models successfully transform radiology reports, MRI findings, and reproductive health documents into more accessible language for non-medical audiences. This improves understanding and increases patient engagement in the treatment process. Notably, although occasional problems with oversimplification or omission of critical details were noted, ChatGPT-4 scored high on accuracy, readability, and completeness. These findings highlight the potential of LLMs for health education and supporting patient-centred care, while emphasizing the need for professional oversight to ensure the accuracy and safe use of these tools in clinical practice.

Limitations of the study. The main limitation of the study is that only ChatGPT-4o chatbot was analyzed, without comparison with other LLM models. Moreover, the focus on the readability index, limits the possibility to evaluate the practical relevance of the model in clinical setting. Another limitation is the lack of qualitative feedback from patients or healthcare professionals. The research does not explore how the simplified texts are perceived in terms of understandability, relevance, or trustworthiness by the target audience. As a result, the actual effectiveness of the materials in real-world communication settings remains unassessed. Additionally, the study does not include an evaluation of the content in terms of clinical accuracy or completeness. This limits the ability to determine whether the simplified texts meet the standards required for safe and effective communication in real-world healthcare settings.

Further directions for on-going research. The aspects discussed in the Limitations section should be incorporated into future research. The inclusion of qualitative feedback from the target audience, as well as assessments of clinical accuracy and content completeness, will be essential to fully evaluate the practical applicability and safety of simplified health information in real-world settings.

Further research should also compare the effectiveness of different LLM models in simplifying medical information. It would also be interesting to analyze the impact of different methods of prompt construction, e.g. testing simple prompts or different prompt lengths, different language styles, including emotional prompts, advanced contextual commands, and iterative prompting, on the simplification of medical information generated by ChatGPT in Polish and their impact on improving patient reception of information.

CONCLUSIONS

Advanced Audience Persona Prompting greatly simplifies the medical information generated by ChatGPT in Polish, improving readability for people with lower education levels. However, even simplified content is still difficult for people with only basic education to fully understand. Therefore, there is a need to further refine prompt design. The results demonstrate the potential of ChatGPT to tailor medical information to different audiences, while ensuring content quality and consistency with current medical knowledge. Further research should compare different methods of prompt design in order to develop more effective strategies to support healthcare communication and education for patients with different levels of education.

REFERENCES

1. Plackett R, Kaushal A, Kassianos AP, et al. Use of Social Media to Promote Cancer Screening and Early Diagnosis: Scoping Review. J Med Internet Res. 2020;22(11):e21582, doi: 10.2196/21582.
2. Lyson HC, Le GM, Zhang J, et al. Social Media as a Tool to Promote Health Awareness: Results from an Online Cervical Cancer Prevention Study. J Canc Educ. 2019; 34: 819–822. https://doi.org/10.1007/s13187-018-1379-8
3. Zheng F, Wang K. The impact of social media on guideline-concordant cervical cancer-screening: insights from a national survey. Public Health. 2023 Oct;223:50–56. doi: 10.1016/j.puhe.2023.07.025.
4. Ayoub NF, Lee YJ, Grimm D, et al. Comparison between ChatGPT and Google search as sources of postoperative patient instructions. JAMA Otolaryngol Head Neck Surg. 2023;149(6):556–558.
5. Burns C, Bakaj A, Berishaj A, et al. Use of generative AI for improving health literacy in reproductive health: case study. JMIR Form Res. 2024;8(1):e59434.
6. Ye Z, Zhang B, Zhang K, et al. An assessment of ChatGPT's responses to frequently asked questions about cervical and breast cancer. BMC Womens Health. 2024;24(1):482.
7. Deng J, Lin Y. The benefits and challenges of ChatGPT: An overview. Front Comput Intell Syst. 2022;2(2):81–83.
8. Medjedovic E, Stanojevic M, Jonuzovic-Prosic S, et al. Artificial intelligence as a new answer to old challenges in maternal-fetal medicine and obstetrics. Technol Health Care. 2024;32(3):1273–1287.
9. Bachmann M, Duta I, Mazey E, et al. Exploring the capabilities of ChatGPT in women's health: obstetrics and gynaecology. NPJ Womens Health. 2024;2(1):26.
10. Eoh KJ, Kwon GY, Lee EJ, et al. Efficacy of large language models and their potential in Obstetrics and Gynecology education. Obstet Gynecol Sci. 2024;67(6):550–556.
11. Grünebaum A, Dudenhausen J, Chervenak FA. Enhancing patient understanding in obstetrics: The role of generative AI in simplifying informed consent for labor induction with oxytocin. J Perinat Med. 2024.
12. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. Radiol Imaging Cancer. 2024;6(2):e230086.
13. Sallam M, Barakat M, Sallam M. A Preliminary Checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence–based models in health care education and practice: development study involving a literature review. Interact J Med Res. 2024;13(1):e54704.
14. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. 2023;2302.11382.
15. Rezaie-Chamani S, Mohammad-Alizadeh-Charandabi S, Kamalifard M. Knowledge, attitudes, and practice about Pap smear among women referring to a public hospital. J Family Reprod Health. 2012;6(4):177–182.
16. Al Ghamdi NH. Knowledge of human papilloma virus (HPV), HPV vaccine, and Pap smear among adult Saudi women. J Family Med Prim Care. 2022;11(6):2989–2999.
17. Deguara M, Calleja N, England K. Cervical cancer and screening: knowledge, awareness and attitudes of women in Malta. J Prev Med Hyg. 2021;61(4):E584–E592.
18. Michalska A, Stefaniak M, Gotlib-Małkowska J. Can ChatGPT provide clear/patient-friendly and reliable information on cervical cancer screening? A study of ChatGPT-generated information in Polish. Med Sci Monit. 2025;in press.
19. Statistica 13 Software, StatSoft, Poland Sp. z o.o. https://www.statsoft.pl/. Available at: https://www.statsoft.pl/. Accessed January 14, 2025.
20. Schmidt S, Zimmerer A, Cucos T, et al. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg. 2024;144(2):611–618. doi:10.1007/s00402-023-05113-4
21. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. 2023;6(1):9. doi:10.1186/s42492-023-00136-5
22. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817–2825. doi:10.1007/s00330-023-10213-1