American College of Neuropsychopharmacology

www.nature.com/dpn

METHODOLOGY OPEN

Language models in digital psychiatry: challenges with simplification of healthcare materials

Ankit Aich[1,2], Tingting Liu[1], Salvatore Giorgi[1], Kelsey Isman[1], Ruhshana Bobojonova[1], Lyle Ungar[2,3] and Brenda Curtis[1,3]✉

Linguistic hurdles in healthcare, such as complex language, significantly affect patient outcomes, including satisfaction with interaction, comprehension of healthcare materials, and engagement with the healthcare system. Reducing these hurdles has been a focus in healthcare delivery, as they significantly hinder patient engagement and adherence to treatments. The growing use of large language models (LLMs) in healthcare opens the possibility to reduce these linguistic hurdles. This study evaluates the ability of five prominent LLMs—GPT-3.5, GPT-4, GPT-4o, LLaMA-3, and Mistral—to simplify healthcare information to the standard recommended by the American Journal of Medicine. Our results indicate that while LLMs can approximate targeted reading levels, their outputs are inconsistent, with significant variability in reading level and deviation from the topic, making them unsuitable for deployment in healthcare settings.

NPP – Digital Psychiatry and Neuroscience; https://doi.org/10.1038/s44277-025-00029-w

LAY SUMMARY

In this work we aim to see if public-facing healthcare materials can be simplified using Large Language Models (LLMs). Currently, the American Journal of Medicine recommends that healthcare materials be provided to people at a reading level of 6. In this work we take five state of the art LLMs viz. GPT-3.5, GPT-4, GPT-4o, LLaMA-3, and Mistral-7b and experiment with prompt engineering to see if these models can simplify healthcare materials from different sources such as academic venues, CDC and WHO releases or public releases from bodies like Mayo Clinic. We find significant variability, shown through large standard-deviations in the performance of LLMs. This work paves the pathway to develop and nurture better simplification and summarization pipelines in healthcare.

INTRODUCTION

Linguistic hurdles, such as the complex language of healthcare materials, hinder accessibility and understanding and are well-documented challenges in health informatics and digital medicine. These hurdles have profound and far-reaching effects, influencing patient satisfaction, adherence to psychiatric treatment regimens, and overall health outcomes. Addressing these linguistic hurdles is particularly critical in psychiatry, where clear communication can foster therapeutic engagement and improve treatment adherence [1, 2]. Significant research has addressed these hurdles to enhance patient health outcomes [2].

Language-related obstacles encompass spoken and written communication, including critical elements of linguistic style such as reading level. Addressing these factors is essential for fostering effective patient engagement in healthcare settings and to expand access to digital health materials in psychiatry. In the context of psychiatry, where clear and empathetic communication is vital for adherence to treatment regimens and therapeutic engagement, simplified language can significantly enhance patient outcomes. Because of this, medical organizations have recommended health materials and patient interactions be provided with less linguistic complexity for better understanding and overall improvement in patient health outcomes. The American Journal of Medicine recommends health materials be provided at a reading level of sixth grade for best outcomes in patient health [3]. Research shows that readers overwhelmingly prefer more straightforward writing, which helps them process more information and enhances understanding [4–6]. Patients also report better care when there is a stylistic match in how information is communicated [7].

In general, simple reading material facilitates language learning, aids individuals with cognitive impairments such as aphasia or dyslexia, and enhances patient understanding in multiple settings, leading to improved adherence to treatments across multiple medical domains, from infectious diseases and immunology [8] to general medicine [9, 10] to psychiatry [11], substance use [12]. It has been shown that patient-clinician communication needs improvement through interventions that address patient needs for communication style [13]. Additionally, simplicity of language has been shown to improve health outcomes by making content more relatable and engaging for people from different backgrounds [14]. Implementing simplified and understandable healthcare material promotes communication in multiple areas and better comprehension of the materials [15], fostering greater

[1]National Institute on Drug Abuse, Baltimore 21224 MD, USA. [2]University of Pennsylvania, Philadelphia 10587 PA, USA. [3]These authors contributed equally: Lyle Ungar, Brenda Curtis. ✉email: brenda.curtis@nih.gov

Received: 6 January 2025 Revised: 25 April 2025 Accepted: 28 April 2025
Published online: 22 May 2025

engagement and adherence to psychiatric care. Incorporating simplification into materials has also positively impacted fields beyond healthcare, from sociology to economics [16, 17].

Large language models (LLMs) are increasingly being used across the spectrum of mental health care [18, 19], ranging from Bipolar Disorder [20, 21], Schizophrenia [21, 22], depression [23], suicidal ideation [24], stress [25], and mixed mental health tasks [26]. However, their potential to improve patient engagement and adherence to treatments remains limited by their current inability to consistently simplify complex medical information into accessible formats. In the current landscape of LLM applications in digital health, where they interact with a patient, they have been largely confined to psychiatric and mental health related settings. As such in this section, we focus on the various ways in which LLMs have been used in psychiatric care. They are also used [1] to talk to humans and generate human-like texts [27], but challenges in readability and language alignment hinder their effectiveness in fostering clear communication and engagement in general-purpose digital healthcare settings.

Despite the increasing use of artificial intelligence (AI), generative AI, and natural language processing (NLP) in healthcare, there is a notable lack of effort aimed at creating content that effectively supports patient health outcomes. In this paper, we show how public-facing, informational healthcare materials are far more complicated than the recommended level and that simplification is a significant challenge even with the latest technology. Prior studies in text simplification using natural language processing (NLP) methods often suffer from severe limitations. For instance [28] conducted simplification experiments on a mere four texts, rendering their findings difficult to generalize. Similarly, computationally intensive approaches, such as those proposed by [29], predominantly focus on simplifying medical journals, neglecting alternative sources like online blogs which are more representative of public information consumption patterns.

As exemplified by [30], general-purpose neural text simplification efforts rely on older non-transformer models and machine translation systems, largely downplaying the advancements offered by state-of-the-art LLMs. While some research has aimed at enhancing the quality of information retention in medical simplification tools [31], these efforts rarely address the capabilities and limitations of contemporary LLMs. Other related works have explored adjacent areas such as summarizing electronic health records (EHRs) [32], facilitating doctor-patient interactions [33], identifying named entities in healthcare texts [34], or using more computationally intensive methods such as entity linking [35], but these do not directly assess public-facing medical informational materials using LLMs in the broader medical domain.

While past studies have explored text simplification this paper explores the capacity of cutting-edge large language models (LLMs) to adhere to instructions concerning control over reading levels, addressing the critical issue of style and language alignment in public-facing healthcare material. With LLMs now being ubiquitous in multiple areas, this investigation is essential. We conducted experiments with five prominent LLMs, assessing their competence in simplifying medical information from various sources, including government publications, online clinics, and academic journals. Our results indicate that while LLMs demonstrate some ability to adjust style, they exhibit significant inconsistencies in reading levels, characterized by wide standard deviations. These findings raise serious concerns about the suitability of LLM deployment in healthcare-AI contexts, including patient interactions in digital healthcare.

METHODS

In this section we describe our methods comprising of data curation, prompt selection, and finally healthcare topic simplification. Our main objective is to evaluate the effectiveness of large language models (LLMs) in simplifying complex healthcare content to the recommended reading-level of 6. Before doing this, we first need to identify an effective prompt-strategy. Specifically, we aim to identify the most effective prompting strategy for optimizing performance on the simplification task. To do this, we first conduct a prompt selection phase, described in subsection "Testing the Prompting Process", testing different prompt formulations with GPT-4. We keep all model parameters fixed and vary only the prompt text. Once we identify the best-performing prompt, we apply it uniformly across five models, GPT-3.5, GPT-4, GPT-4o, LLaMA-3, and Mistral-7B, to carry out the simplification task. The data used for the simplification task is described in the Data subsection below.

Data

We selected five healthcare topics of significant public health importance —Attention Deficit Hyperactivity Disorder (ADHD), Vaccinations and Immunizations, Influenza (Flu), Substance Use, and Human Immunodeficiency Virus (HIV)—drawing upon existing literature to underscore their societal relevance [36–49].

For each topic, we curated five representative documents (for example a full blog post from the CDC) from authoritative sources, including government health agencies such as the Centers for Disease Control and Prevention (CDC)[2] and the World Health Organization (WHO)[3], leading medical institutions like the Mayo Clinic[4], academic journals, and high-visibility online media platforms. These serve as the foundation for subsequent text simplification tasks.

Measuring reading level

To evaluate readability, we use the Flesch-Kincaid (FK) reading level, a widely adopted metric for assessing text complexity [5]. FK scores correspond roughly with U.S. grade levels, ranging from 1 (very easy) to 20 (very difficult). However, it is possible to have texts with an FK score of more than 20, where such texts usually contain non-English characters, such as symbols or mathematical notation, or contain extremely difficult English. Texts with scores around 6 are considered accessible for middle-school students and are widely recommended for healthcare communication [3]. We compute FK scores using the Python library *Py-Readability-Metrics*[5]. Table 1 shows the initial FK reading levels of the collected texts prior to simplification.

Testing the prompting process

Before we perform our health information simplification task, we run a smaller experiment to find the best prompting method, since LLM output is highly dependent on the input prompt (i.e., text that contains instructions on how the LLM should behave). Prompts often contain examples of the desired behavior. Such prompts are referred to as *k-shot* prompts, where *k* is the number of examples appended to the instruction. This initial experiment is done to choose the *k* which attains the best simplified output (i.e., the FK score closest to the specified reading level of 6).

Since our main objective is a simplification task, we pick five questions from grade-school reading materials[6] have GPT-4 answer these questions to get a baseline reading level. Next, we prompt GPT-4 to answer the same questions at four grade levels (1, 3, 5, and 7). GPT-4 was chosen to test our prompting process since it was the largest commercially available model with the least variance in performance at the time of conducting these experiments (November 2024). We test three different prompting paradigms: $k = 0$ (prompting only by specifying the required FK grade level), $k = 1$ (specifying the required FK grade level with one example), and $k = 2$ (specifying the required FK grade level with two examples). We prompted GPT-4 by asking it the following questions and asking it to generate an output at the specified grade level:

[1] https://woebothealth.com/

[2] https://cdc.gov

[3] https://who.int

[4] https://mayoclinic.org

[5] https://pypi.org/project/py-readability-metrics/

[6] Sourced from K-5 Learning https://www.k5learning.com/

Table 1. Initial Flesch-Kincaid Reading Level of Articles for Each Topic.

Topic	Post 1	Post 2	Post 3	Post 4	Post 5
ADHD	11.4	18.9	11.9	15.5	15.8
Flu	16.5	18.3	18.3	16.9	16.4
HIV	13.0	12.1	14.1	13.1	14.2
Substance Use	18.0	27.7	15.3	15.4	16.6
Vaccine	14.3	11.3	11.2	13.9	17.3

Table 2. Prompt Selection Experiment.

Grade Level	$k = 0$	$k = 1$	$k = 2$
1	**2.2**	8.9	2.9
3	**3.8**	5.4	5.3
5	**6.1**	7.5	7.2
7	**7.5**	9.9	10.1

Numbers denote the best FK levels for the required grade level for GPT-4.
Numbers closest to the grade level are the best.
The best performing values are shown in bold font.

- Explain parts of a plant
- What is the water cycle?
- What is pollution?
- What does the sun do?
- Explain the parts of a human body.

Table 2 shows the performance difference for GPT-4 among the three prompting method experiments. We find that prompting only by specifying the FK grade level worked better for all tasks than providing examples using one or two-shot learning. We use this best-performing prompt for our healthcare experiment.

Healthcare topics simplification task

This section discusses the healthcare-topics simplification task for the LLMs. As discussed in the introduction, complex language in healthcare information disbursal remains a problem. We perform a text simplification task to mitigate this problem and understand the performance of LLMs in this context.

To perform the healthcare simplification task, we use the best performing prompt (from the previous section) and pass it as an input to the five LLMs: GPT-3.5-turbo, GPT-4, GPT-4o, LLaMA-3, Mistral-Instruct-7B. These language models were chosen for a mix of their large parameter size (Mistral-7B has 7B parameters GPT-4 has 1.8 T parameters), commercial availability (GPT-4 is closed source, whereas Llama3 is open source), and popular-usage in generative AI applications [50]. A sample input is shown below.

> "You are an intelligent and adaptable writing assistant. Given the following article, simplify this to a Flesch-Kincaid grade level of 6. You must also use at least 200 words[7] in your generated answer. **article text verbatim**"

We then measure the FK reading level of the generated outputs (described in "Methods") using the PyPi FK Python library.

RESULTS

We find that prompting only by specifying the FK grade level worked better for all tasks than providing examples or using one or two-shot learning. We use the best-performing prompt for our healthcare experiment. Table 2 shows the performance difference for GPT-4 among the three prompting methods.

Table 1 shows the initial FK reading level for all 25 articles selected. As shown, the lowest reading level is close to an FK score of 11. This score is considered to be at high school reading level and is higher than what is recommended. However, most articles have a reading level of 15+, which is much higher than recommended for healthcare information.

For each of our five topics, we present results in Fig. 1 (vaccine), 2 (ADHD), 3 (substance use), 4 (HIV), 5 (flu). In each figure, the X-axis shows the models and original posts, and the Y-axis shows the FK reading levels. Each bar cluster on the X-axis comprises five bar graphs, each representing the FK level of that post. In the bar cluster, every individual bar corresponds to a post with increasing order from left to right. Specifically, the left most bar in each bar cluster is post 1, the next is post 2, and the right most bar in each bar cluster is post 5. Each graph also has the original FK level of the posts (denoted in gray), the target FK level of six (shown in a dotted line), and the outputs of five LLMs denoted by Blue for GPT-3.5, Green for GPT-4, Yellow for GPT-4o, Red for LLaMA-3, and Purple for Mistral-7B. We will discuss these graphs one by one.

As depicted in Fig. 1 (vaccines), none of the models reach the target FK reading level. While GPT-4o comes very close in two cases for posts 1 and 2, it fails to be consistent and the FK level increases for posts 3,4,5. We also see that, in general, when the baseline information has a higher initial FK level, the models struggle more. In Fig. 1 (ADHD), as shown, for the first post, all models come very close to the FK level required. In the fourth post, *LLaMA-3* reaches below the FK level in the score. However, its result is not in English (it provides a long string of text characters comprising of punctuation marks and other symbols). This artificially deflates the score but does not provide simplified information. In Fig. 1 (substance use), the models are more consistent, coming close to the required level in posts 1, 4, and 5. For HIV and flu not only are the models unable to reach the required grade level, but they also complicate the information.

We also see information in the generated text which is not related to the topic being discussed. Especially in LLaMA-3, the reading level generated is inconsistent and ranges from 0.49 to 36 when asked to generate at a sixth-grade reading level, doing worse than other models. We similarly see that the Mistral model fails to ever reach a sixth-grade level as required. This shows how some current LLMs cannot simplify or distill healthcare information at the required level of simplicity.

DISCUSSION

This study explores the feasibility of deploying large language models (LLMs) in digital healthcare by evaluating their ability to generate tailored text that meets specific stylistic requirements. The investigation focused on controlling output through reading level (to ensure grade-specific accessibility). Using prominent models such as GPT-3.5, GPT-4, GPT-4o, LLaMA-3, and Mistral, we examined how well these systems can adapt their language to match a required level of simplicity, denoted by FK reading level, to promote effective communication. This work provides insight into their potential role in improving healthcare communication, such as in areas of digital psychiatry, general medicine, substance use or immunology, by assessing these models' strengths and

[7] We specify 200 words because FK calculation requires at least 100 words

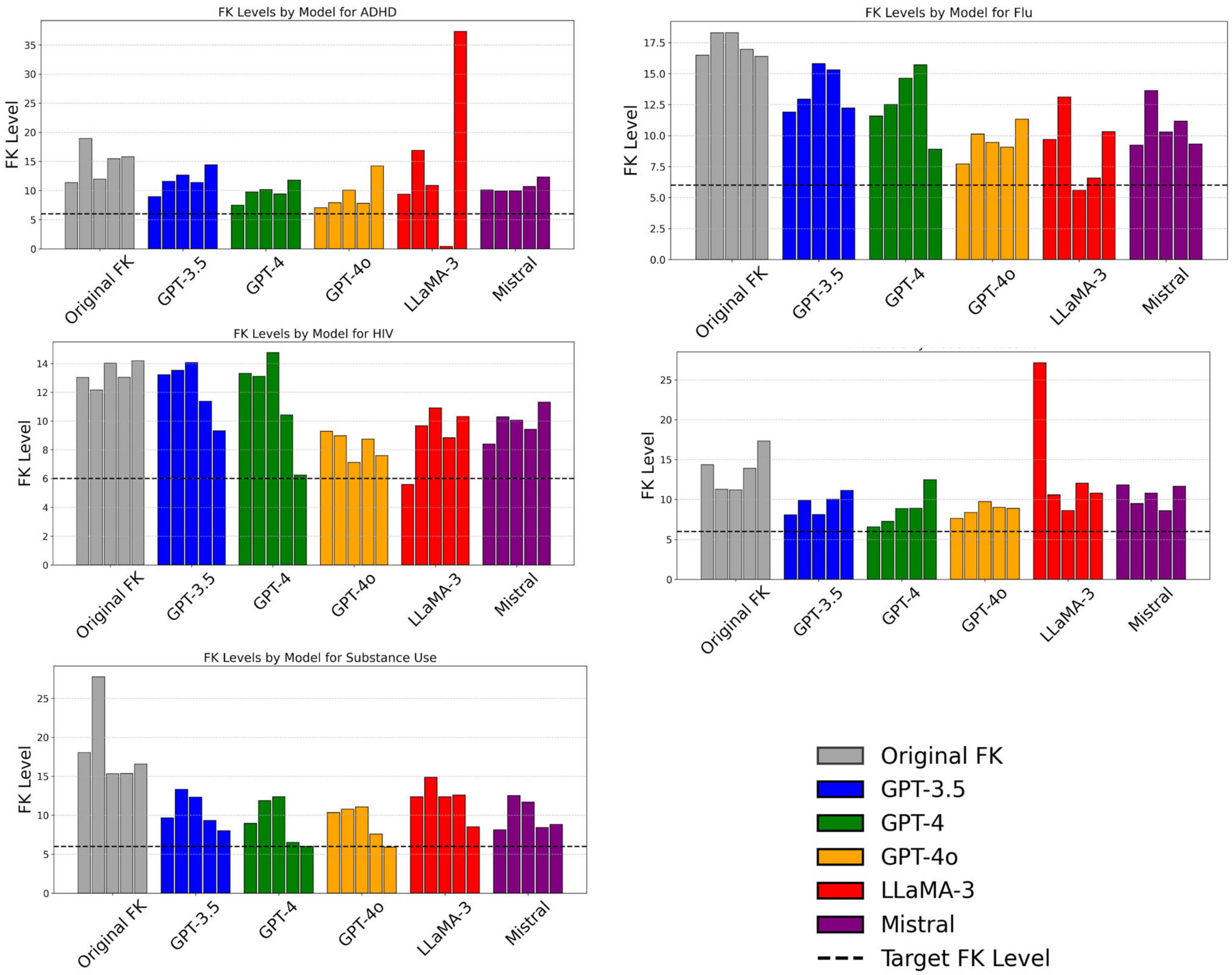

Fig. 1 Flesch-Kincaid reading level across each topic. The dashed line indicates the ideal reading level recommended by the American Journal of Medicine [3]. The gray bars represent the reading levels of the source materials. Each bar corresponds to a specific post within the respective domain, with the bar's position across models showing the same post's performance across different models. For each bar cluster the bars' position corresponds to the post number increasing left to right. Specifically, the left-most bar in a bar cluster is the FK level of post 1 and the right-most bar in a bar cluster is the FK level of post 5. Models that perform better will show reading levels closer to the dashed line than gray bars. In the above 1a - relates to vaccinations, 1b - relates to ADHD, 1c - relates to substance use, 1 d - relates to HIV, and 1e - relates to Flu.

limitations in meeting linguistic needs, which is crucial for patient engagement. We also note how bodies like the WHO and CDC do not post their findings at a required sixth-grade reading level, despite available medical recommendations and humans being able to edit text at this level. This furthers the importance of our study. Our findings highlight the opportunities and current challenges in applying LLMs to real-world, diverse healthcare settings.

Adapting the reading level of healthcare communication is essential to promote patient understanding and participation. It is widely recommended that health materials be written at a reading level of sixth to seventh grade [3] to accommodate diverse literacy needs. This study found that while some LLMs demonstrated the capability to simplify language to lower reading levels (e.g., GPT-4o achieved a minimum mean reading level of 3.2) they often struggled to achieve a low reading level over multiple iterations, as evidenced by high variability in performance (LLaMA-3 had a maximum standard deviation $\sigma = 27.6$).

Smaller models like LLaMA-3, in particular, exhibited significant challenges, frequently generating overly complex or dissimilar language output. In general, we find that the relatively smaller LLMs, such as LLaMA-3 and Mistral, perform worse by failing to maintain consistency (defined above) and, in many cases, failing to achieve the target. The larger state-of-the-art models in the GPT family manage to stay on the topic of the question, but still fall short of achieving the required simplification target, as shown in Fig. 1 (combined graph). It is important to clarify that our observations regarding whether a model *stays on topic* were made qualitatively through manual review of the generated outputs. Specifically, we evaluated each model's output along two dimensions: *topical relevance* and *accuracy*. We define *topical relevance* as the degree to which a response remains focused on the subject matter of the original healthcare prompt, avoiding unrelated or tangential content. *Accuracy* refers to the factual correctness and contextual appropriateness of the information provided.

While we did not quantify these patterns, the consistency of differences across multiple examples—particularly between larger and smaller models—prompted the inclusion of a qualitative assessment in the Discussion. These limitations reveal a critical gap in the ability of current large language models (LLMs) to consistently produce information in a style appropriate for healthcare, where patient comprehension directly influences treatment adherence, satisfaction, and outcomes. Addressing this challenge is essential to expanding the role of LLMs in generating effective, accessible healthcare communication. Simplifying psychiatric healthcare materials may enhance patient engagement, reduce obstacles to understanding treatment plans, and improve adherence to therapies. While prior research has explored summarizing complex texts [51–54], extracting relevant information to aid understanding [55], and supporting mental health assessments [56, 57], our findings underscore the limitations that persist in automating these simplifications reliably. This is especially relevant in mental health, infection, and immunology contexts [13], where clarity is paramount.

Adherence to digital healthcare systems is heavily influenced by the clarity and empathy of communication, especially in linguistically sensitive environments [11, 58–61]. Although LLMs show potential in simplifying healthcare content, their inconsistency in maintaining readability across outputs suggests the need for further refinements before they can reliably support patient communication. Our study highlights the difficulty of aligning automated simplifications with human-centric communication goals. Standardizing healthcare materials to simpler readability levels—such as through Flesch-Kincaid metrics—could help achieve more consistent, patient-friendly outputs. Ultimately, even though we did not directly assess adherence, improving simplification methods could foster greater comprehension and engagement, which are well-established foundations for informed decision-making and adherence to care.

This paper has several limitations. First, we conducted experiments using only five language models. Given the rapidly evolving landscape of LLMs, we focused on five widely used models that are readily accessible and possess sufficient parameter sizes to handle simplification tasks. Additionally, our healthcare simplification experiments were limited to a dataset of 25 articles. Future research could expand upon our methods by utilizing larger datasets to enhance the generalizability of the results. Future work should focus on refining LLMs by incorporating domain-specific datasets and prompts tailored for psychiatric communication, enabling more reliable and simplified patient interactions. In addition, researchers should explore integrating these models into broader digital healthcare ecosystems, such as telehealth platforms and wearable devices, to deliver more accessible and patient-centered care. Collaboration with clinical experts and patient groups will also be essential to ensure that AI tools address real-world challenges effectively and ethically. Addressing the readability challenges identified in this study can transform digital healthcare by improving patient understanding, engagement, and treatment adherence. Furthermore, developing advanced LLMs could pave the way for more all-encompassing and impactful healthcare solutions. While a huge portion of the current landscape of LLM in healthcare research has focused on psychiatry and mental health solutions, we believe stronger testing and collaboration will open them up to be used in a plethora of digital-health settings.

Citation diversity statement. The authors have attested that they made efforts to be mindful of diversity in selecting the citations used in this article.

DATA AVAILABILITY

Data for all articles and code can be made available by emailing the first author of the manuscript.

REFERENCES

1. Lopez Vera A, Thomas K, Trinh C, Nausheen F. A case study of the impact of language concordance on patient care, satisfaction, and comfort with sharing sensitive information during medical care. J Immigr Minor Health. 2023;25:1261–9. https://doi.org/10.1007/s10903-023-01463-8.
2. Alshamsi H, Almutairi A, Al Mashrafi S, Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman Med J. 2020;35:e122.
3. Hutchinson N, Baird G, Garg M. Examining the reading level of internet medical information for common internal medicine diagnoses. Am J Med. 2016;129:637–9. https://doi.org/10.1016/j.amjmed.2016.01.008.
4. Shulman HC, Markowitz DM, Rogers T. Reading dies in complexity: online news consumers prefer simple writing. Sci Adv. 2024;10:1–8.
5. Owusu-Acheaw M. Reading habits among students and its effect on academic performance: a study of students of koforidua polytechnic. Library Philosophy Pract. 2014.
6. Cimmiyotti CB. Impact of reading ability on academic performance at the primary level. Graduate Master's Theses, Capstones, and Culminating Projects, Dominican University of California. 2013.
7. Squires A. Strategies for overcoming language barriers in healthcare. Nurs Manag. 2018;49:20–27.
8. Trepka MJ, Gong Z, Ward MK, Fennie KP, Sheehan DM, Jean-Gilles M, et al. Using causal bayesian networks to assess the role of patient-centered care and psychosocial factors on durable HIV viral suppression. AIDS Behav. 2024;28:1–18.
9. Shardlow M. A survey of automated text simplification. Int J Adv Computer Sci Applications, Spec Issue Nat Lang Process. 2014;5:58–70.
10. Elkefi S, Asan O. The impact of patient-centered care on cancer patients' QOC, self-efficacy, and trust towards doctors: analysis of a national survey. J Patient Exp. 2023;10:237437352311515.
11. Ruffalo ML, Kottapalli M, Anbukkarasu P. Empathy in the care of individuals with schizophrenia: a vital element of treatment. Am J Psychother. 2023;77:30–4. https://doi.org/10.1176/appi.psychotherapy.20230022.
12. Park S(Ethan), Mosley J, Grogan C, Pollack H, Humphreys K, D'Aunno T, et al. Patient-centered care's relationship with substance use disorder treatment utilization. J Subst Abuse Treat. 2020;118:108125.
13. Kuchinad K, Park J, Han D, Saha S, Moore R, Beach MC. Which clinician responses to emotion are associated with more positive patient experiences of communication? Patient Educ Couns. 2024;124:108241.
14. Ladson-Billings G. Toward a theory of culturally relevant pedagogy. Am Educ Res J. 1995;32:465–91.
15. Milner H. Culturally relevant pedagogy in a diverse urban classroom. Urban Rev. 2011;43:66–89.
16. Hu H, Gangning Y, Xueli X, Guo L, Huang J. Cultural diversity and innovation: an empirical study from dialect. Technol Soc. 2022;69:101939.
17. Falck O, Heblich S, Lameli A, Suedekum J. Dialects, cultural identity, and economic exchange. J Urban Econ. 2010;72:225–39. https://doi.org/10.1016/j.jue.2012.05.007.
18. Obradovich N, Khalsa S, Khan W, Suh J, Perlis R, Ajilore O, et al. Opportunities and risks of large language models in psychiatry. NPP - Digital Psychiatry Neurosci. 2024;2:8 https://doi.org/10.1038/s44277-024-00010-z.
19. Smith A, Hames J, Joiner T. Status update: maladaptive facebook usage predicts increases in body dissatisfaction and bulimic symptoms. J Affect Disord. 2013;149:235–40. https://doi.org/10.1016/j.jad.2013.01.032.
20. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. 2024;49:1412–6.
21. Aich A, Quynh A, Badal V, Pinkham A, Harvey P, Depp C, et al. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In: Findings of the association for computational linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022 pp 2871-87.
22. Sharma B, Puri H, Rawat D Digital psychiatry - curbing depression using therapy chatbot and depression analysis. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018 pp 627-31.
23. Choi S, Moreira R, Costa A, Regatieri C, dos Santos V Development of a Chatbot to identify depression through a questionnaire. J Psychiatry Psychiatr Disord. 2021. https://doi.org/10.1101/2021.10.28.21265110.
24. Harmer B, Lee S, Duong TviH, Saadabadi A. Suicidal ideation. StatPearls. 2022.
25. Turcan E, McKeown K Dreaddit: a Reddit dataset for stress analysis in social media. In: Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019). Hong Kong: Association for Computational Linguistics; 2019. pp 97–107.
26. Aich A, Parde N Are you really okay? A transfer learning-based approach for identification of underlying mental illnesses. Proc Eighth Workshop Computational Linguist Clin Psychol. 2022, 89-104. https://doi.org/10.18653/v1/2022.clpsych-1.8.
27. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated

conversational agent (woebot): a randomized controlled trial. JMIR Ment Health. 2017;4:e19.

28. Leroy G, Kauchak D, Haeger D, Spegman D. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. JAMIA Open. 2022;5:044 https://doi.org/10.1093/jamiaopen/ooac044.
29. Lu J, Li J, Wallace B, He Y, Pergola G NapSS: paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In: Vlachos A, Augenstein I (eds) Findings of the association for computational linguistics: EACL 2023. Dubrovnik, Croatia: Association for Computational Linguistics; 2023. pp 1079-91.
30. Bercken L, Sips R-J, Lofi C Evaluating neural text simplification in the medical domain. In: WWW '19: The World Wide Web Conference. 2019. pp 3286-92.
31. Kauchak D, Apricio J, Leroy G. Improving the quality of suggestions for medical text simplification tools. AMIA Annu Symp Proc. 2022;2022:284–92.
32. Laxmisan A, McCoy A, Wright A, Sittig D. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. Appl Clin Inform. 2012;3:80–93.
33. Irfan AA, Khatim NA, Arief MM Using LLM for real-time transcription and summarization of doctor-patient interactions into ePuskesmas in indonesia. 2024. https://doi.org/10.48550/arXiv.2409.17054.
34. Sasikala D, Sudarshan R, Sivasathya S Harnessing LLMs for medical insights:NER extraction from summarized medical text. In: 2024 15th international conference on computing communication and networking technologies (ICCCNT). 2024. pp 1–6.
35. Borchert F, Llorca I, Schapranow M-P Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. Database. 2024, 2024. https://doi.org/10.1093/database/baae067.
36. Mace J, HaileMariam A, Zhu J, Howell N. Involuntary remembering and ADHD: Do individuals with ADHD symptoms experience high volumes of involuntary memories in everyday life? Br J Psychol. 2024;116:216–32. https://doi.org/10.1111/bjop.12749.
37. He L, Zhao Y, Gong J-X, Zhao L, Ma Z-R, Xiong Q-W, et al. Contrasting presentations of children with ADHD and subthreshold ADHD. Pediatric Res. 2024. https://doi.org/10.1038/s41390-024-03502-y.
38. Li T, Van Rooij D, Roth Mota N, Buitelaar JK, ENIGMA ADHD Working G, Hoogman M, et al. Characterizing neuroanatomic heterogeneity in people with and without ADHD based on subcortical brain volumes. J Child Psychol Psychiatry. 2021;62:1140–9.
39. Paduano G, Sansone V, Pelullo C, Angelillo S, Gallè F, Giuseppe G. Recommended vaccinations during adolescence: parents' knowledge and behaviors. Vaccines. 2024;12:1342.
40. Karczmarz S, Owczarczyk A, Zuk P, Środziński L, Prusaczyk A, Bogdan M. Vaccination as an element of health security management - analysis of factors of patient attitudes towards COVID-19 preventive vaccinations based on the opinions of healthcare experts. J Educ Health Sport. 2024;70:55682.
41. Schrenker S, Erpenbeck L. Current vaccination and immunization strategies in dermatology. Dermatologie. 2024;75:889–901. https://doi.org/10.1007/s00105-024-05400-0.
42. Kislaya I, Torres R, Gomes L, Melo A, Machado A, Henriques C, et al. End of season 2022/2023 quadrivalent influenza vaccine effectiveness in preventing influenza in primary care in portugal. Hum Vaccin Immunother. 2023;19:2263219.
43. Hsu P-C, Lin Y-T, Kao K-C, Peng CK, Sheu CC, Liang SJ, et al. Risk factors for prolonged mechanical ventilation in critically ill patients with influenza-related acute respiratory distress syndrome. (2023) https://doi.org/10.21203/rs.3.rs-3446973/v1.
44. Liu Y, Xu J, Wei C, Xu Y, Lyu C, Sun M, et al. Detection of H1N1 influenza virus in the bile of a severe influenza mouse model. Influenza Other Respir Viruses. 2024;18:70012 https://doi.org/10.1111/irv.70012.
45. Gonzalez A, Fardelmann K. Opioid and substance use disorders. pp 323–33. https://doi.org/10.1007/978-3-031-62756-9_40.
46. Burnell K, Andrade F, Hoyle R. Exposure to peers' online postings about substances and adolescents' substance use: a longitudinal study. J Child Family Stud. 2024;33:3854–67.
47. Hardy W, The American Academy of HIV Medicine. Fundamentals of HIV medicine 2023. https://doi.org/10.1093/med/9780197679098.001.0001.
48. Davtyan M, Kacanek D, Berman C, Chadwick E, Smith R, Salomon L, et al. Factors associated with internalized HIV-related stigma among biological mothers living with HIV enrolled in a US cohort study. AIDS Care. 2023;36:1–7.
49. Kim SM, Choi Y, Kang S, study K. Smoothed quantile residual life regression analysis with application to the korea HIV/AIDS cohort study. BMC Med Res Methodol. 2024;24:44 https://doi.org/10.1186/s12874-024-02159-9.
50. Giorgi S, Liu T, Aich A, Isman KJ, Sherman G, Fried Z, et al. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In: Al-Onaizan Y, Bansal M, Chen Y-N (eds) Findings of the association for computational linguistics: EMNLP 2024. Miami, Florida, USA: Association for Computational Linguistics; 2024. pp 7174-88.
51. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30:1–9.
52. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digit Med. 2023;6:158 https://doi.org/10.1038/s41746-023-00896-7.
53. Tariq A, Urooj A, Trivedi S, Fathizadeh S, Ramasamy G, Tan N, et al. Patient centric summarization of radiology findings using large language models. 2024. https://doi.org/10.1101/2024.02.01.24302145.
54. Song Y, Tian Y, Wang N, Xia F. Summarizing medical conversations via identifying important utterances. J Fluoresc. 2020;30:717–29.
55. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. 2022;29:1208–16. https://doi.org/10.1093/jamia/ocac040.
56. Altamimi I, Altamimi A, Alhumimidi A, Temsah M-H. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus. 2023;15:40351 https://doi.org/10.7759/cureus.40351.
57. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digital Health. 2022;1:e0000168.
58. Nembhard I, David G, Ezzeddine I, Betts D, Radin J. A systematic review of research on empathy in healthcare. Health Serv Res. 2022;58:250–63. https://doi.org/10.1111/1475-6773.14016.
59. Yang C-P, Hargreaves W, Bostrom A. Association of empathy of nursing staff with reduction of seclusion and restraint in psychiatric inpatient care. Psychiatr Serv. 2014;65:251–4.
60. Pascucci M, Capobianco F, La Montagna M, Stella E, Ventriglio A, Rubini G, et al. Mental health and empathy: do nursing students have better attitudes to psychiatric patients? Eur Psychiatry. 2016;33:S520.
61. Román-Sánchez D, Paramio-Cuevas J, Paloma-Castro O, Palazón FJL, Lepiani-Díaz I, De la Fuente J, et al. Empathy, burnout, and attitudes towards mental illness among spanish mental health nurses. Int J Environ Res Public Health. 2022;19:692.

ACKNOWLEDGEMENTS

This research was made possible under the Intramural Research Program at the National Institutes of Health and the National Institute on Drug Abuse.

AUTHOR CONTRIBUTIONS

Aich Conceptualization, Methodology, Writing - Original Draft Preparation. **Liu** Formal Analysis - Writing - Original Draft Preparation. **Giorgi** Writing - Original Draft Preparation. **Isman** - Data Curation. **Bobojonova** - Investigation, Writing - Original Draft Preparation. **Ungar** - Supervision of entire research. **Curtis** Supervision of entire research and funding acquisition.

FUNDING

This research was funded by the Intramural Research Program at the National Institute on Drug Abuse.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44277-025-00029-w.

Correspondence and requests for materials should be addressed to Brenda Curtis.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.