Journal of the American Medical Informatics Association, 2024, **31(10)**, 2294–2303
https://doi.org/10.1093/jamia/ocae186
Advance access publication 15 July 2024
Research and Applications

Research and Applications

Leveraging artificial intelligence to summarize abstracts in lay language for increasing research accessibility and transparency

Cathy Shyr, PhD[1], Randall W. Grout, MD, MS[2,3], Nan Kennedy, MLAS[4], Yasemin Akdas, PhD[5], Maeve Tischbein, PhD[4], Joshua Milford, BS[4], Jason Tan, MS[4], Kaysi Quarles, BS[4], Terri L. Edwards, RN[4], Laurie L. Novak, PhD[1], Jules White, PhD[6], Consuelo H. Wilkins, MD, MSCI[7], Paul A. Harris, PhD[1,8,9,*]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [2]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN 46202, United States, [3]Regenstrief Institute, Inc, Indianapolis, IN 46202, United States, [4]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [5]Division of Emergency Medicine Research, Vanderbilt University Medical Center, Nashville, TN 37232, United States, [6]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37240, United States, [7]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, United States, [8]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [9]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37240, United States

*Corresponding author: Paul A. Harris, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, TN 37203, United States (paul.a.harris@vumc.org)

Abstract

Objective: Returning aggregate study results is an important ethical responsibility to promote trust and inform decision making, but the practice of providing results to a lay audience is not widely adopted. Barriers include significant cost and time required to develop lay summaries and scarce infrastructure necessary for returning them to the public. Our study aims to generate, evaluate, and implement ChatGPT 4 lay summaries of scientific abstracts on a national clinical study recruitment platform, ResearchMatch, to facilitate timely and cost-effective return of study results at scale.

Materials and Methods: We engineered prompts to summarize abstracts at a literacy level accessible to the public, prioritizing succinctness, clarity, and practical relevance. Researchers and volunteers assessed ChatGPT-generated lay summaries across five dimensions: accuracy, relevance, accessibility, transparency, and harmfulness. We used precision analysis and adaptive random sampling to determine the optimal number of summaries for evaluation, ensuring high statistical precision.

Results: ChatGPT achieved 95.9% (95% CI, 92.1–97.9) accuracy and 96.2% (92.4–98.1) relevance across 192 summary sentences from 33 abstracts based on researcher review. 85.3% (69.9–93.6) of 34 volunteers perceived ChatGPT-generated summaries as more accessible and 73.5% (56.9–85.4) more transparent than the original abstract. None of the summaries were deemed harmful. We expanded ResearchMatch's technical infrastructure to automatically generate and display lay summaries for over 750 published studies that resulted from the platform's recruitment mechanism.

Discussion and Conclusion: Implementing AI-generated lay summaries on ResearchMatch demonstrates the potential of a scalable framework generalizable to broader platforms for enhancing research accessibility and transparency.

Key words: artificial intelligence; large language model; text summarization; ResearchMatch; return of study results.

Introduction

Timely and effective return of aggregate study results plays an essential role in informing the public about clinical findings to promote trust and transparency in the research process.[1] Researchers and US funding agencies, including the National Institutes of Health and the Patient-Centered Outcomes Research Institute, view the return of aggregate results as an important ethical responsibility to demonstrate respect, enhance transparency, and provide insight on findings that may benefit participants' health.[2,3] Similarly, a majority of participants value receiving study results to inform future health decisions and behaviors.[4–7] According to a national return of value survey, receiving study results would make participants more likely to trust researchers and participate in research.[8]

Despite support from both participants and the research community, the practice of providing study results to a lay audience is not widely adopted. Studies showed that fewer than 2% of global clinical trials completed or terminated between 2015 and 2017 returned plain language summaries to their participants.[1] Major barriers include (1) *scarce infrastructure* necessary for implementing lay summaries that can

Received: May 29, 2024; **Revised:** June 26, 2024; **Editorial Decision:** July 3, 2024; **Accepted:** July 4, 2024

effectively reach participants, (2) *high burden in cost and time* to support providing lay summaries at scale, and (3) *researchers' varied proficiency in effectively communicating with a lay audience*.[9–12]

To address these barriers, our study focuses on generating lay summaries using OpenAI's ChatGPT 4[13] and implementing them on a national clinical study recruitment registry, ResearchMatch,[14] to facilitate timely and cost-effective return of aggregate study results at scale. Created in 2009, ResearchMatch is a disease-neutral, web-based recruitment platform that helps match volunteers who wish to participate in clinical research studies with researchers actively searching for volunteers across the US. Aimed at enhancing the recruitment and engagement of diverse populations in clinical studies, ResearchMatch serves as an effective national recruitment platform, having supported over 6600 studies and registered 161 000 volunteers and 14 480 researchers as of April 2024. It prioritizes returning value to volunteers and publicly displays research publications (over 750 as of April 2024) for studies that used the platform's recruitment mechanism. Specifically, its technical infrastructure supports the routine and automated retrieval of bibliometric details, including the title, author(s), abstract, and link to the full article. To generate lay summaries at scale, a cost-effective approach is to leverage advanced large language models (LLMs), which can automatically summarize study abstracts at a literacy level accessible to the general public. Among LLMs, OpenAI's ChatGPT 4 was the most promising for the task because of its ease of integration with ResearchMatch's technical workflow through its application programming interface (API) and impressive performance on text summarization and background explanation tasks, outperforming other LLMs on benchmark datasets.[15–18] ChatGPT 4's ability to generate lay summaries for increasing accessibility and transparency of research studies, however, has not been explored. Previous studies on biomedical text summarization focused on sentence-level simplification rather than summarizing paragraphs or abstracts in plain language.[19–23] Furthermore, they primarily relied on conventional text summarization metrics (eg, ROUGE[24] and BLEU[25]), which are designed to assess lexical overlap and thus are not directly applicable to measuring the quality of lay summaries tailored to non-specialist readers.[26]

To this end, the objective of our study was to generate, evaluate, and implement ChatGPT-generated lay summaries of scientific abstracts on ResearchMatch to return aggregate study results. We engineered prompts for summarizing scientific abstracts at a literacy level accessible to a general audience and rigorously evaluated them for accuracy, relevance, accessibility, transparency, and harmfulness. Implementing LLM-generated lay summaries on a national study recruitment platform like ResearchMatch demonstrates the potential of generalizing this approach to broader platforms, including biomedical publication repositories and clinical trial registries, to return aggregate study results at scale.

Materials and methods

An overview of our study is shown in Figure 1. Our workflow consisted of three key steps: (1) engineer prompts to summarize scientific abstracts in lay language, (2) evaluate ChatGPT 4-generated summaries to select the optimal prompt, and (3) implement summaries generated by the optimal prompt on ResearchMatch for any study that used the platform's recruitment mechanism. For brevity, we henceforth refer to ChatGPT 4 as ChatGPT.

Prompt engineering

We engineered prompts based on five design principles: promote succinctness, reduce technical jargon, increase readability at a literacy level accessible to the public, retain salient information, and emphasize practical importance of the research findings. We formulated these principles based on published best practices and guidance on developing lay summaries of research studies.[27–30] After establishing these design principles, we refined the prompts through collaboration with experienced prompt engineers and researchers specializing in the recruitment and retention of diverse participants for research studies, ensuring our prompts were well founded and effective in engaging a wide audience. Table 1 provides a summary of the prompts used and rationale behind each design principle.

1) *Promote succinctness*: A succinct summary can help readers quickly focus on key elements without being overwhelmed by extensive details. The length of a scientific abstract typically ranges from 200 to 300 words depending on the article type, discipline, and journal. While not a systematic requirement, some journals, including *PLOS* and *BMJ*, ask authors to provide a lay summary ranging from 100 to 200 words. Therefore, we designed two prompts with different word count requirements, that is, A = under 200 words, and B = under 100 words, to generate summaries that were comparable to or shorter than research abstracts and lay summaries required by journals, respectively, in terms of length.
2) *Reduce technical jargon*: While jargon allows researchers to effectively communicate complex and nuanced ideas within their fields, it may alienate lay readers who do not share the same technical background.[31] A 2019 experiment found that the use of technical jargon impaired people's ability to process scientific information and undermined efforts to inform the public.[32] Minimizing technical jargon ensures that the information is accessible to a broader, non-specialist audience. As such, readers from various backgrounds, including journalists, practitioners, and policymakers, can understand and engage with the research findings. Bridging the gap between researchers and the public can encourage informed discussions and potentially enhance trust in the research process.[33]
3) *Increase readability*: Providing summaries at a literacy level accessible to the public ensures that the content is understandable to a broader audience. Following the American Medical Association's recommendation on the readability of educational materials, we instructed ChatGPT to generate lay summaries at a sixth-grade reading level, ensuring that they can be easily understood by the general public.[34]
4) *Retain salient information and 5. Emphasize practical importance*: Scientific abstracts generally consist of four key elements: purpose, methods, results, and conclusion. Therefore, we specifically instructed ChatGPT to highlight these components to help readers gain a holistic understanding of the study, including *why* the study was conducted, *what* was studied, and *how*. In addition, we

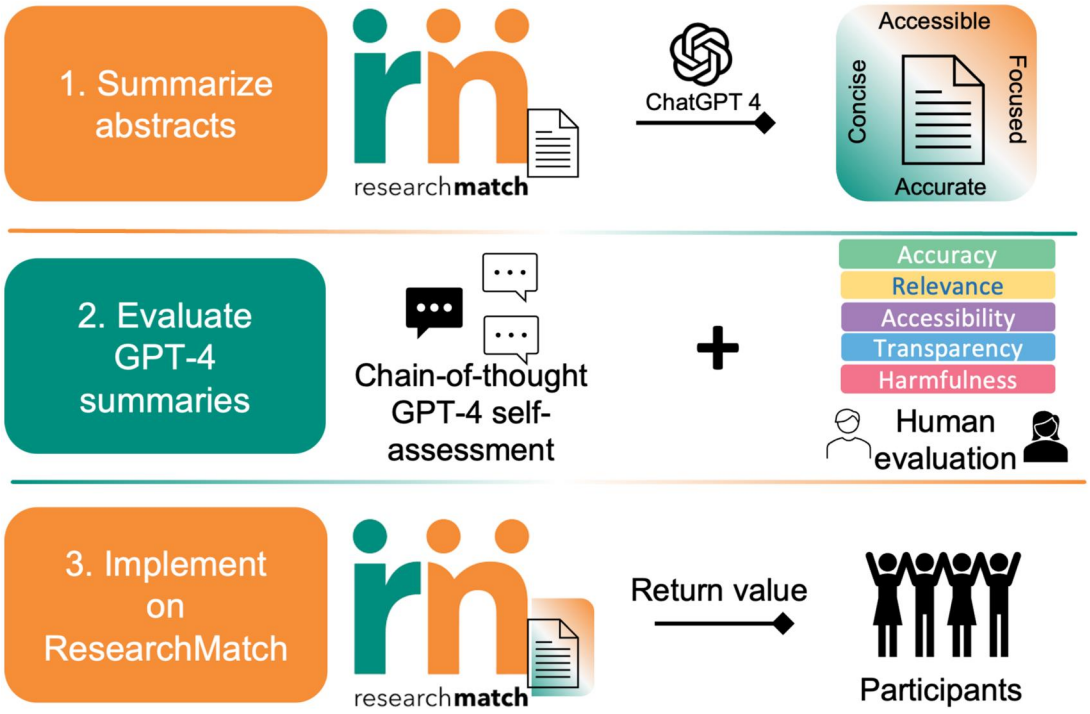

Figure 1. Overview of study.

wanted the model to emphasize the *meaning* of the study to help readers understand its practical relevance, as comprehending the tangible benefits of research findings can help inform evidence-based decision-making and increase public engagement.[4–7]

Based on these design principles, we designed two prompts (Table 1). Prompt A: "Summarize this abstract under 200 words in lay language at a 6th grade reading level, highlighting the study purpose, methods, key findings, and practical importance of these findings for the general public." Prompt B was the same as A except the word limit was reduced to 100.

Lay summary generation

We obtained the list of 657 PubMed IDs associated with published abstracts on ResearchMatch (as of October 2023). The abstracts spanned a wide range of topics, including clinical studies, basic science research, informatics and engineering, among others. We used the National Center for Biotechnology Information's Entrez Programming Utilities API and retrieved the abstracts from the PubMed database. We accessed OpenAI's ChatGPT 4 using its API (model = "gpt-4-0613") on November 30 and December 1, 2023. ChatGPT's temperature parameter, which ranges between 0 and 1, controls the degree of variability in the model's output, with lower values corresponding to more deterministic results. Therefore, we set the temperature to 0, ensuring that the model chose the most likely word or token in its response for reproducibility. We generated two sets of lay summaries, one for Prompt A and another for Prompt B.

Prompt evaluation and selection

Chain-of-thought ChatGPT self-evaluation. After the lay summaries were generated, we prompted ChatGPT to generate a chain-of-thought reasoning process to determine whether each lay summary was supported by the original abstract.[35] The goal was to guide ChatGPT to perform a self-evaluation and detect potential errors prior to human review. A summary of our approach is shown in Figure 2. To implement chain-of-thought prompting, we provided ChatGPT with example sets of summarized abstracts and guided the model to check the accuracy of the lay summary. We provided three example summaries as input: a summary that contradicts the abstract, a summary that contains an unsupported statement, and an accurate summary. Next, we prompted ChatGPT to check the actual lay summary against the original abstract. The final output was a Boolean value (True/False) indicating whether the lay summary was accurate.

Human evaluation. In addition to ChatGPT's self-evaluation, human reviewers evaluated the lay summaries across five dimensions: accuracy, relevance, accessibility, transparency, and harmfulness. We chose these dimensions because they captured key aspects of lay summaries, allowing flexibility in expression, context sensitivity, and focus on specific content for a lay audience. Reviewers also assessed the lay summaries based on four of the five core design principles outlined in Table 1. Specifically, they were asked which of the two summaries (A vs B) was better at reducing technical jargon, easier to read, retaining salient information, and emphasizing the practical importance of the research. We did not ask reviewers to evaluate succinctness, as Summary B was always shorter than Summary A by design (ie, under 100 words vs 200 words). Our evaluation study was approved by the Vanderbilt Institutional Review Board as exempt research (IRB #231679).

Human evaluation—researchers. From November 22, 2023 to January 31, 2024, we used convenience sampling to select five researchers with complementary backgrounds in medicine, biostatistics, informatics, medical anthropology, and psychology to conduct the evaluation study. All

Table 1. Summary of design principles, rationale, and prompts.

Design Principle	Rationale
1. Promote succinctness	Distill the abstract to its most essential elements to ensure that readers can quickly grasp the key points
2. Reduce technical jargon	Bridge the communication gap between researchers and the public to help non-specialist readers better understand and engage with the research findings
3. Increase readability	Ensure the summary is written at a literacy level accessible to the public
4. Retain salient information	Foster a deeper understanding of the research by providing a holistic overview of key information
5. Emphasize practical importance	Create a connection between study findings and real-world applications to increase reader engagement
Prompts (Color-coded by Design Principle)	
A. Summarize this abstract **under 200 words in lay language at a 6th grade reading level**, **highlighting the study purpose, methods, key findings**, and **practical importance of these findings for the general public**.	
B. Summarize this abstract **under 100 words in lay language at a 6th grade reading level**, **highlighting the study purpose, methods, key findings**, and **practical importance of these findings for the general public**.	

researchers have experience in the recruitment and retention of diverse participants for research studies, including community engagement and developing educational resources for diverse populations. Each lay summary was independently evaluated by two researchers, and disagreements were adjudicated and resolved by a third researcher. For each summary sentence, researchers evaluated three dimensions: (1) accuracy (ie, whether the sentence was supported by or could be inferred from the original abstract—Yes/No), (2) relevance (ie, whether the sentence provided relevant information about the study purpose, methods, results, or significance of those results—Yes/No), and (3) harmfulness (ie, whether the sentence contained misinformation that could impact participants negatively, eg, engage in harmful behavior, increase research mistrust—Yes/No).

Human evaluation—volunteers. From January 23 to February 29, 2024, we surveyed volunteers about their perception of the lay summaries. ResearchMatch volunteers over the age of 18 were eligible. Between January 23 and February 19, we contacted eligible volunteers who expressed interest in our study using ResearchMatch's recruitment mechanism. Volunteers were given two weeks to fill out a 10-min secure REDCap[36] survey, and those who completed the survey received a $5 Amazon gift card. In the survey, volunteers were asked to compare ChatGPT-generated lay summaries with the original abstract across three dimensions: (1) accessibility (ie, which one was more understandable—abstract vs summary), (2) transparency (ie, which one communicated the research in a more transparent manner—abstract vs summary), and (3) harmfulness. In addition, volunteers were asked to provide demographic information, including gender, age, race, ethnicity, and highest level of education completed.

Statistical analysis. We used precision analysis and adaptive random sampling to determine the number of abstracts to evaluate. Specifically, we calculated the statistical precision for five measures: (1) accuracy (proportion of accurate sentences), (2) relevance (proportion of relevant sentences), (3) accessibility (proportion of lay summaries that were more

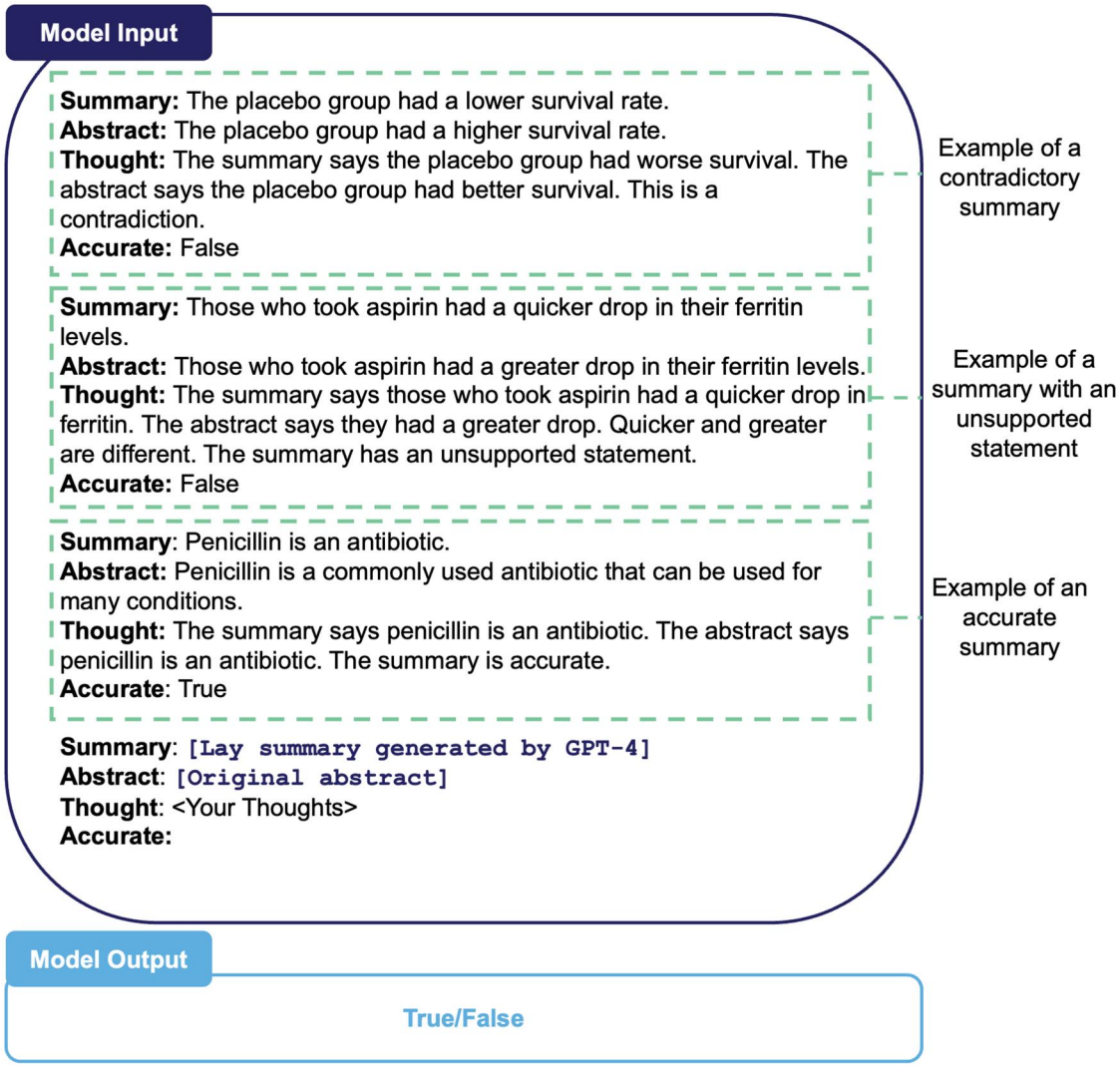

Figure 2. Overview of chain-of-thought prompting.

understandable than the original abstract), (4) transparency (proportion of lay summaries that communicated the research in a more transparent manner), and (5) harmfulness (proportion of sentences or lay summaries that were harmful). We randomly selected one of the 657 abstracts to evaluate until the half width of the 95% confidence interval (CI) around these measures was less than a pre-determined threshold to guarantee a high level of precision. Because all five measures were binary, we calculated the CIs based on a binomial distribution. Our target half-widths were $\leq 5\%$ for accuracy, relevance, and harmfulness, and $\leq 15\%$ for accessibility and transparency. We chose a more stringent half-width (ie, higher precision) for the first three measures because they were crucial in maintaining the integrity of the original abstract. To achieve our target of $\leq 5\%$, assuming a worst-case scenario of 0.5 for the proportion, the estimated number of sentences we needed to evaluate was 381. For accessibility and transparency, on the other hand, we accounted for a higher degree of variability, as both measures were subjective and may vary widely depending on demographic factors (eg, education level). To achieve our target of $\leq 15\%$, assuming 0.5 for the proportion, the estimated number of lay summaries participants needed to evaluate was 39.

After all evaluations were completed, we used one-sample tests of binomial proportion to assess whether accuracy and relevance were $> 90\%$ at a significance threshold of .05. Conversely, we assessed harmfulness for $<10\%$. We assessed whether the majority of volunteers perceived the lay summaries as more accessible and transparent. To choose the final prompt, we used two-sample, one-sided tests for equality of proportions comparing Summary A versus Summary B across the five measures at the .05 level. If both summaries performed similarly, we compared them based on preferences for the four design principles. All statistical analyses were performed using R v4.3.3.[37]

Implementation on ResearchMatch

Implementing ChatGPT-generated lay summaries on ResearchMatch was approved by the Vanderbilt Institutional Review Board (IRB #090207). For full transparency, we informed ResearchMatch liaisons and registered researchers about the planned update in a newsletter prior to implementation. We expanded ResearchMatch's technical pipeline to incorporate automated generation and storage of ChatGPT-generated lay summaries. Specifically, published abstracts were programmatically sent from the ResearchMatch server to a private OpenAI

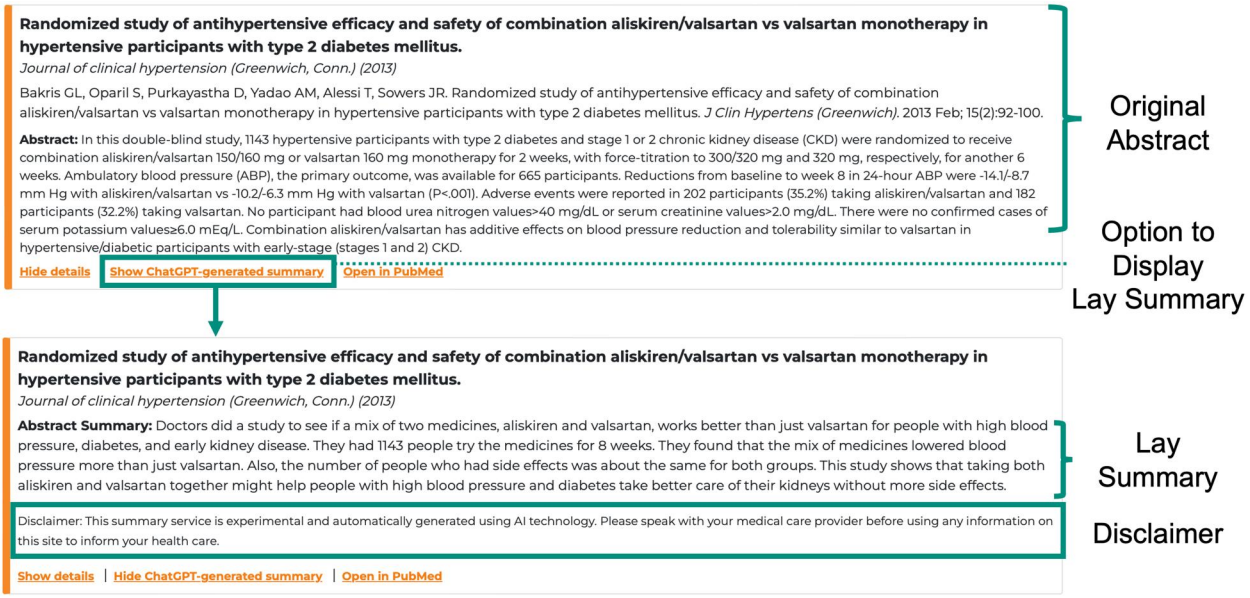

Figure 3. Example of abstract and lay summary on ResearchMatch.

Table 2. Researcher and volunteer evaluation results.

	% of sentences [95% CI]		
Researcher evaluation	**Summary A ($N = 303$ sentences)**	**Summary B ($N = 192$ sentences)**	***P* value[a]**
Accuracy	97.7 [95.3, 98.9]	95.9 [92.1, 97.9]	.20
Relevance	98.2 [96.0, 99.2]	96.2 [92.4, 98.1]	.14
Harmfulness	0 [0, 1.3]	0 [0, 2.0]	NA
	% of preferred summary [95% CI]		
Researcher preference	**Summary A ($N = 33$ summaries)**	**Summary B ($N = 33$ summaries)**	***P* value**
Reduce technical jargon	10.1 [2.88, 26.7]	89.9 [73.3, 97.1]	<.001
Easier to read	8.1 [1.90, 24.2]	91.9 [75.8, 98.1]	<.001
Retain salient information	91.4 [75.2, 97.9]	8.6 [2.13, 24.8]	<.001
Emphasize practical importance	53.5 [35.7, 70.6]	46.6 [29.4, 64.3]	.37
	% of summaries [95% CI]		
Volunteer evaluation	**Summary A ($N = 34$ summaries)**	**Summary B ($N = 34$ summaries)**	***P* value**
Accessibility	82.4 [66.5, 91.7]	85.3 [69.9, 93.6]	.5
Transparency	70.6 [53.8, 83.2]	73.5 [56.9, 85.4]	.5
Harmfulness	0 [6.23e−18, 10]	0 [6.23e−18, 10]	NA
	% of preferred summary [95% CI]		
Volunteer preference	**Summary A ($N = 34$ summaries)**	**Summary B ($N = 34$ summaries)**	***P* value**
Reduce technical jargon	38.2 [22.7, 56.4]	61.8 [43.6, 77.3]	.045
Easier to read	32.4 [18.0, 50.6]	67.6 [49.4, 82.0]	.0038
Retain salient information	82.4 [64.8, 92.6]	17.6 [7.39, 35.2]	<.001
Emphasize practical importance	55.9 [38.1, 72.4]	44.1 [27.6, 61.9]	.23

[a] *P* based on two-sample, one-sided test of equal proportions at the .05 significance level

deployment through its API via an HTTP request. This allowed us to interact with the Chat Completion function, which generated a lay summary of the abstract based on the final prompt. The summary was then returned to ResearchMatch through the HTTP request and stored in a database used to populate the "Study Findings" page. This process is repeated on a daily basis to automatically update the page with new published abstracts and lay summaries.

To inform users about the potential limitations of ChatGPT-generated summaries, we included a disclaimer on ResearchMatch, stating that "this summary service is experimental and automatically generated using AI technology. Please speak with your medical care provider before using any information on this site to inform your health care." In addition, we provided an open channel of communication that allows users to connect with the ResearchMatch team

via email about any platform-related concerns. Figure 3 shows an example of a published abstract[38] and ChatGPT-generated lay summary on ResearchMatch. The implementation of over 750 lay summaries on the platform went live on May 1, 2024.

Results

Researchers' evaluation for accuracy, relevance, and harmfulness

Table 2 provides a summary of the evaluation results. Based on adaptive random sampling, we reached the target precision level for accuracy, relevance, and harmfulness after evaluating 303 sentences from Summary A and 192 from Summary B (ie, 33 total abstracts). The inter-evaluator-agreement F1 scores were 0.97, 0.99, and 0.99 for accuracy, relevance, and harmfulness, respectively, indicating a high degree of between-evaluator agreement. ChatGPT's chain-of-thought self evaluation did not flag any lay summaries as erroneous. According to researcher review, Summary A had an accuracy of 97.7% (95% CI, 95.3–98.9) and Summary B 95.9% (95% CI, 92.1–97.9). Based on one-sample tests of proportion, both were significantly higher than 90% accuracy (P value $< .001$ for Summary A and P value $= .004$ for Summary B). For relevance, Summaries A and B achieved 98.2% (95% CI, 96.0–99.2) and 96.2% (95% CI, 92.4–98.1), respectively, and both were significantly higher than 90% (P value $< .001$ and P value $= .003$, respectively). Researchers did not mark any sentences as harmful for either Summary A or B. There was no statistically significant difference between Summary A and Summary B based on accuracy, relevance, or harmfulness. Researchers did not identify any accurate sentences that collectively misconstrued the original abstract. For the four design principles assessed, researchers felt that Summary B was better at reducing technical jargon (P value $< .001$) and easier to read (P value $< .001$); Summary A, on the other hand, was better at retaining salient information (P value $< .001$).

In free text, researchers provided reasons for marking sentences as inaccurate. Overall, none of the inaccuracies were deemed harmful. These benign inaccuracies can be broadly divided into two categories. First, ChatGPT could make an incorrect extrapolation when information provided in the original abstract was open for interpretation. For example, in the abstract by Yang *et al.*,[39] the authors stated that "smokers … were randomized to either the extinction or control condition." The ChatGPT-generated summary stated that "half the smokers were [assigned to the extinction condition] and half [to the control condition]." In this case, "randomized to either [condition]" in the original abstract does not necessarily mean 1:1 allocation. Second, when simplifying technical jargon, ChatGPT may choose words or phrases that do not completely capture nuances in the original writing. For example, in the study by Stevenson *et al.*,[40] the authors tested differences in audiovisual speech integration among children using the McGurk effect, a perceptual phenomenon where auditory and visual signals create a fused percept. When explaining the McGurk effect in lay language, ChatGPT stated that the effect "shows how our brains mix up what we hear and see." Here, researchers felt that the phrase "mix up" did not completely reflect the "fused" or interactive effect from the original abstract.

Table 3. Summary of self-reported demographic information.

	Volunteers ($N = 34$)
Gender	
Man	9 (26.5%)
Woman	24 (70.6%)
Non-binary	0 (0%)
Transgender	0 (0%)
None of the above	0 (0%)
Prefer not to answer	1 (3%)
Age in years	37 [27]
Race and ethnicity	
American Indian or Alaska Native	0 (0%)
Asian	5 (14.7%)
Black, African American, or African	6 (17.6%)
Hispanic, Latino, or Spanish	2 (5.9%)
Middle Eastern or North African	2 (5.9%)
Native Hawaiian or other Pacific Islander	1 (3%)
White	18 (52.9%)
None of the above	0 (0%)
Prefer not to answer	0 (0%)
Highest level of education attained	
Below Grade 11	0 (0%)
Grade 12 or GED (high school graduate)	4 (11.8%)
1–3 years after high school (some college, Associate's degree, or technical school)	7 (20.6%)
College 4 years or more (College graduate)	15 (44%)
Advanced degree (Master's, Doctorate, etc)	7 (20.6%)
Prefer not to answer	1 (3%)

N is the total sample size. For continuous variables, median [interquartile range] is reported. For categorical variables, frequencies are followed by proportions.

Volunteers' evaluation for accessibility, transparency, and harmfulness

Based on adaptive random sampling, we reached the target precision level for accessibility, transparency, and harmfulness after evaluating 34 lay summaries. A total of 34 volunteers participated in our study; each volunteer evaluated a pair of lay summaries (A and B) for a single abstract. Overall, 82.4% of volunteers (95% CI, 66.5–91.7) perceived Summary A as being more accessible than the original abstract (P value $< .001$) and 85.3% of volunteers (95% CI, 69.9–93.6) Summary B (P value $< .001$). For transparency, 70.6% (95% CI, 53.8–83.2; P value $= .026$) and 73.5% (95% CI, 56.9–85.4; P value $= .01$) of volunteers perceived Summary A and Summary B, respectively, as having communicated the research in a more transparent manner than the original abstract. The volunteers did not flag any summaries as harmful. There was no statistically significant difference between Summary A and Summary B based on accessibility, transparency, or harmfulness. For the four design principles assessed, volunteers felt that Summary B was better at reducing technical jargon (P value $= .045$) and easier to read (P value $= .0038$); Summary A, on the other hand, performed better at retaining salient information (P value $< .001$). Table 3 shows a summary of the volunteers' demographic information. The majority of volunteers (70.6%) were women, white (52.9%), college graduates or above (64.6%), and the median age was 37 years.

Selecting the final prompt

Because both Summary A and B performed similarly in terms of accuracy, relevance, accessibility, transparency, and

harmfulness (ie, no statistical difference), we chose to implement the summary based on researchers' and volunteers' preference on the five design principles. Both summaries performed similarly at emphasizing the practical importance of the research finding. While Summary A did better at retaining salient information from the original abstract, Summary B was more succinct, did better at reducing technical jargon, and was easier to read according to both researchers and volunteers. These results were statistically significant at the .05 level. Therefore, we chose to implement the summaries generated using Prompt B on ResearchMatch.

Discussion

Our study demonstrates the potential of leveraging LLMs like ChatGPT to generate lay summaries of scientific abstracts and implementing them on a national recruitment registry to return aggregate study results at scale. Overall, ChatGPT-generated summaries achieved over 95% accuracy and relevance based on researcher review, and over 80% and 70% of volunteers perceived them as more accessible and transparent than the original abstract, respectively. Though the length of a summary directly influenced detail retention, with the longer summary being better at preserving key information, the shorter summary still maintained a high level of accuracy without leading to harmful misinterpretations.

While previous studies explored the potential of leveraging LLMs to assist in scientific communication,[18,41–43] our work is the first to rigorously assess and implement LLM-generated lay summaries for returning aggregate study results at scale. Partnering with researchers and volunteers to evaluate these summaries across complementary dimensions, including accuracy, relevance, accessibility, transparency, and harmfulness, not only helps ensure the quality and reliability of the lay summaries, but also highlights the potential of LLM-generated lay summaries to promote trust in healthcare research. Implementing lay summaries on a national study recruitment platform like ResearchMatch has important implications for research engagement and participation. Specifically, this may help potential participants better understand the purpose, processes, and benefits of existing studies that have resulted from the platform's recruitment mechanism and lead to increased participation. Our approach offers a scalable framework that can be extended to broader platforms, such as health information systems and clinical trial registries, to enhance participant engagement and foster greater transparency, trust, and inclusivity in healthcare research. Future work includes assessing public uptake and performing real-time evaluation of utility when users click on the AI-generated summaries.

A potential limitation is that the volunteers in our study were recruited from ResearchMatch, which has a self-selected population that is highly educated (>82% of volunteers are college graduates), predominantly White (71%), and majority women (65%). As such, this may limit the generalizability of our findings. Despite this, 47% of the volunteers in this study were from marginalized racial and ethnic groups, and 32% were not college graduates. To diversify our sample, a promising strategy is to leverage adaptive sampling techniques based on demographic characteristics and educational levels to ensure a broader representation in future studies. It is noteworthy that several volunteers who were highly educated (ie, Master's degree or above) expressed that the ChatGPT-generated summaries were too simplistic, stating that they "almost read like a fourth-grade assignment." This sentiment highlights the challenge of creating lay summaries that meet the needs of diverse audiences, as highly-educated individuals who are accustomed to reading technical abstracts may perceive straightforward explanations as overly simplistic. To this end, we provide both the original abstract and lay summary on ResearchMatch. It is important to note that while we selected the final prompt based on preference across the design principles, readers may not weigh each principle equally, as some may value depth over brevity. Another potential limitation is that summaries are currently provided in English. Future work includes translating them into multiple languages to increase accessibility and engagement with non-English speaking populations. Though our study primarily focused on creating clear and engaging summaries for a lay audience, we recognize that research studies can vary widely in their objectives and methodologies. To this end, it may be more straightforward to generate relevant lay summaries for clinical studies compared to basic science research. In addition to the AI-generated summaries, providing additional background information or context can enable the general public to make well-informed conclusions and further engage with the research summaries. Incorporating this into our methodology will be a valuable focus for future work.

Enhancing understanding and transparency is essential for building trust and inclusivity in the research process. While artificial intelligence offers promising opportunities for communicating scientific findings to a wide audience, it is crucial to rigorously evaluate these models to ensure the integrity and reliability of their output.

Author contributions

Conceptualization: Paul A. Harris, Cathy Shyr. Prompt design: Paul A. Harris, Cathy Shyr, Randall W. Grout, Nan Kennedy, Yasemin Akdas, Laurie L. Novak, Jules White. Evaluation study design and execution, including administrative and regulatory support: Paul A. Harris, Cathy Shyr, Randall W. Grout, Nan Kennedy, Yasemin Akdas, Maeve Tischbein, Kaysi Quarles, Terri L. Edwards. Statistical analysis: Cathy Shyr. Acquisition, analysis, or interpretation of data: Paul A. Harris, Cathy Shyr. Implementation: Paul A. Harris, Cathy Shyr, Maeve Tischbein, Joshua Milford, Jason Tan. Critical revision of the manuscript: Paul A. Harris, Cathy Shyr, Randall W. Grout, Nan Kennedy, Yasemin Akdas, Maeve Tischbein, Joshua Milford, Jason Tan, Kaysi Quarles, Terri L. Edwards, Laurie L. Novak, Jules White, Consuelo H. Wilkins Obtained funding: Paul A. Harris, Consuelo H. Wilkins.

Funding

This work was funded by the grant numbers 1U24TR004432-01 from the National Center for Advancing Translational Sciences and 1K99LM014429-01 from the National Library of Medicine. The funders had no role in the design and conduct of the study, collection, management, analysis, and interpretation of the data, preparation, review, or approval of the manuscript, and decision to submit the manuscript for publication.

Conflict of interest

None.

Data availability

Data and code used in this study can be found at https://github.com/cathyshyr/researchmatch_gpt4_lay_summaries.

References

1. Getz K, Farides-Mitchell J. Assessing the adoption of clinical trial results summary disclosure to patients and the public. *Expert Rev Clin Pharmacol*. 2019;12(7):573-578. https://doi.org/10.1080/17512433.2019.1615441
2. Long CR, Purvis RS, Flood-Grady E, et al. Health researchers' experiences, perceptions and barriers related to sharing study results with participants. *Health Res Policy Syst*. 2019;17(1):25. https://doi.org/10.1186/s12961-019-0422-5
3. Patient-Centered Outcomes Research Institute. Returning study results to participants: an important responsibility. Accessed April 22, 2024. https://www.pcori.org/research/about-our-research/returning-study-results-participants-important-responsibility#:~:text=PCORI%20encourages%20researchers%20to%20use,below%20the%208th%20grade%20level
4. Rigby H, Fernandez CV. Providing research results to study participants: support versus practice of researchers presenting at the American Society of Hematology annual meeting. *Blood*. 2005;106(4):1199-1202. https://doi.org/10.1182/blood-2005-02-0556
5. Purvis RS, Abraham TH, Long CR, Stewart MK, Warmack TS, McElfish PA. Qualitative study of participants' perceptions and preferences regarding research dissemination. *AJOB Empir Bioeth*. 2017;8(2):69-74. https://doi.org/10.1080/23294515.2017.1310146
6. Partridge AH, Hackett N, Blood E, et al. Oncology physician and nurse practices and attitudes regarding offering clinical trial results to study participants. *J Natl Cancer Inst*. 2004;96(8):629-632. https://doi.org/10.1093/jnci/djh096
7. McElfish PA, Purvis RS, Scott AJ, Haggard-Duff LK, Riklon S, Long CR. "The results are encouragements to make positive changes to be healthier:" qualitative evaluation of Marshallese participants' perceptions when receiving study results in a randomized control trial. *Contemp Clin Trials Commun*. 2020;17:100543. https://doi.org/10.1016/j.conctc.2020.100543
8. Wilkins CH, Mapes BM, Jerome RN, Villalta-Gil V, Pulley JM, Harris PA. Understanding what information is valued by research participants, and why. *Health Aff (Millwood)*. 2019;38(3):399-407. https://doi.org/10.1377/hlthaff.2018.05046
9. Kuehn BM. Few studies reporting results at US government clinical trials site. *JAMA* 2012;307(7):651-653. https://doi.org/10.1001/jama.2012.127
10. Long CR, Stewart MK, McElfish PA. Health research participants are not receiving research results: a collaborative solution is needed. *Trials*. 2017;18(1):449. https://doi.org/10.1186/s13063-017-2200-4
11. Miller FA, Hayeems RZ, Li L, Bytautas JP. What does 'respect for persons' require? Attitudes and reported practices of genetics researchers in informing research participants about research. *J Med Ethics*. 2012;38(1):48-52. https://doi.org/10.1136/jme.2010.041350
12. Schroter S, Price A, Malički M, Richards T, Clarke M. Frequency and format of clinical trial results dissemination to patients: a survey of authors of trials indexed in PubMed. *BMJ Open*. 2019;9(10):e032701. https://doi.org/10.1136/bmjopen-2019-032701
13. OpenAI. OpenAI GPT-4. Accessed November 30, 2023. https://openai.com/chatgpt
14. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. *Acad Med*. 2012;87(1):66-73. https://doi.org/10.1097/ACM.0b013e31823ab7d2
15. Pu X, Gao M, Wan X. Summarization is (almost) dead. arXiv. 2023, preprint: not peer reviewed.
16. OpenAI. OpenAI application programming interface. Accessed November 30, 2023. https://platform.openai.com/docs/api-reference
17. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2019;21(140):1-67.
18. Guo Y, Qiu W, Leroy G, Wang S, Cohen T. Retrieval augmentation of large language models for lay language generation. *J Biomed Inform*. 2024;149:104580. https://doi.org/10.1016/j.jbi.2023.104580
19. Cai X, Liu S, Han J, Yang L, Liu Z, Liu T. ChestXRayBERT: a pretrained language model for chest radiology report summarization. *IEEE Trans Multimedia*. 2023;25:845-855. https://doi.org/10.1109/TMM.2021.3132724
20. Bui DDA, Del Fiol G, Hurdle JF, Jonnalagadda S. Extractive text summarization system to aid data extraction from full text in systematic review development. *J Biomed Inform*. 2016;64:265-272. https://doi.org/10.1016/j.jbi.2016.10.014
21. Mishra R, Bian J, Fiszman M, et al. Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform*. 2014;52:457-467. https://doi.org/10.1016/j.jbi.2014.06.009
22. Zhang Y, Ding DY, Qian T, Manning CD, Langlotz CP. Learning to summarize radiology findings. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics; 2018:204-213. https://doi.org/10.18653/v1/W18-5623
23. Wang M, Wang M, Yu F, Yang Y, Walker J, Mostafa J. A systematic review of automatic text summarization for biomedical literature and EHRs. *J Am Med Inform Assoc*. 2021;28(10):2287-2297. https://doi.org/10.1093/jamia/ocab143
24. Lin CY. *ROUGE: A Package for Automatic Evaluation of Summaries*. Association for Computational Linguistics; 2004.
25. Papineni K, Bleu ZWJ. A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2002:311-318.
26. Guo Y, August T, Leroy G, Cohen T, Wang LL. APPLS: evaluating evaluation metrics for plain language summarization. arXiv. 2023, preprint: not peer reviewed.
27. Taylor and Francis Author Services. How to write and publish a plan language summary. Accessed November 30, 2023. https://authorservices.taylorandfrancis.com/publishing-your-research/writing-your-paper/how-to-write-a-plain-language-summary/
28. Maurer M, Siegel JE, Firminger KB, Lowers J, Dutta T, Chang JS. Lessons learned from developing plain language summaries of research studies. *Health Lit Res Pract*. 2021;5(2):e155-e161. https://doi.org/10.3928/24748307-20210524-01
29. Wiley Editing Services. Lay Summaries to Communicate Your Research in Plan Language. Accessed November 30, 2023. https://wileyeditingservices.com/en/article-promotion/lay-summaries
30. Zarshenas S, Mosel J, Chui A, et al. Supporting patient and public partners in writing lay summaries of scientific evidence in healthcare: a scoping review protocol. *BMJ Open*. 2022;12(12):e062981. https://doi.org/10.1136/bmjopen-2022-062981
31. Hirst R. Scientific jargon, good and bad. *J Tech Writing Commun*. 2003;33(3):201-229. https://doi.org/10.2190/J8JJ-4YD0-4R00-G5N0
32. Bullock OM, Colón Amill D, Shulman HC, Dixon GN. Jargon as a barrier to effective science communication: evidence from metacognition. *Public Underst Sci*. 2019;28(7):845-853. https://doi.org/10.1177/0963662519865687
33. Hendriks F, Kienhues D. 2. Science understanding between scientific literacy and trust: contributions from psychological and

educational research In: Leßmöllmann A, Dascal M, Gloning T, eds. *Science Communication*. De Gruyter; 2019:29-50. https://doi.org/10.1515/9783110255522-002.
34. Weiss BD. *Health Literacy A Manual for Clinicians*. American Medical Association Foundation; 2003.
35. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. 2022, preprint: not peer reviewed.
36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. https://doi.org/10.1016/j.jbi.2008.08.010
37. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021.
38. Bakris GL, Oparil S, Purkayastha D, Yadao AM, Alessi T, Sowers JR. Randomized study of antihypertensive efficacy and safety of combination aliskiren/valsartan vs valsartan monotherapy in hypertensive participants with type 2 diabetes mellitus. *J Clin Hypertens (Greenwich)*. 2013;15(2):92-100. https://doi.org/10.1111/jch.12032
39. Yang MJ, Brandon KO, Sutton SK, et al. Augmented reality for extinction of cue-provoked urges to smoke: proof of concept. *Psychol Addict Behav*. 2022;36(8):990-998. https://doi.org/10.1037/adb0000868
40. Stevenson RA, Siemann JK, Woynaroski TG, et al. Brief report: arrested development of audiovisual speech perception in autism spectrum disorders. *J Autism Dev Disord*. 2014;44(6):1470-1477. https://doi.org/10.1007/s10803-013-1992-7
41. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res*. 2023;13(4):1148-1154.
42. Biyela S, Dihal K, Gero KI, et al. Generative AI and science communication in the physical sciences. *Nat Rev Phys*. 2024;6(3):162-165. https://doi.org/10.1038/s42254-024-00691-7
43. Schmitz B. Improving accessibility of scientific research by artificial intelligence—an example for lay abstract generation. *Digit Health*. 2023;9:20552076231186245. https://doi.org/10.1177/20552076231186245

Journal of the American Medical Informatics Association, 2024, 31, 2294–2303
https://doi.org/10.1093/jamia/ocae186
Research and Applications