JAMIA Open, 2024, **7(3)**, ooae050
https://doi.org/10.1093/jamiaopen/ooae050
Research and Applications

Research and Applications

Completeness and readability of GPT-4-generated multilingual discharge instructions in the pediatric emergency department

Alex Gimeno, BS[1], Kevin Krause, MS[2], Starina D'Souza, BA[1], Colin G. Walsh, MD, MA*[,1,2,3,4]

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, United States, [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [3]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [4]Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN 37203, United States

*Corresponding author: Colin Walsh, MD, MA, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave, Suite 1475, Nashville, TN 37203, United States (colin.walsh@vumc.org)

Abstract

Objectives: The aim of this study was to assess the completeness and readability of generative pre-trained transformer-4 (GPT-4)-generated discharge instructions at prespecified reading levels for common pediatric emergency room complaints.

Materials and Methods: The outputs for 6 discharge scenarios stratified by reading level (fifth or eighth grade) and language (English, Spanish) were generated fivefold using GPT-4. Specifically, 120 discharge instructions were produced and analyzed (6 scenarios: 60 in English, 60 in Spanish; 60 at a fifth-grade reading level, 60 at an eighth-grade reading level) and compared for completeness and readability (between language, between reading level, and stratified by group and reading level). Completeness was defined as the proportion of literature-derived key points included in discharge instructions. Readability was quantified using Flesch-Kincaid (English) and Fernandez-Huerta (Spanish) readability scores.

Results: English-language GPT-generated discharge instructions contained a significantly higher proportion of must-include discharge instructions than those in Spanish (English: mean (standard error of the mean) = 62% (3%), Spanish: 53% (3%), $P = .02$). In the fifth-grade and eighth-grade level conditions, there was no significant difference between English and Spanish outputs in completeness. Readability did not differ across languages.

Discussion: GPT-4 produced readable discharge instructions in English and Spanish while modulating document reading level. Discharge instructions in English tended to have higher completeness than those in Spanish.

Conclusion: Future research in prompt engineering and GPT-4 performance, both generally and in multiple languages, is needed to reduce potential for health disparities by language and reading level.

Lay Summary

Emergency departments often attempt to increase efficiency with prewritten discharge instructions. However, discharge instructions can be in a language other than the reader's preferred one or can be written at too advanced a level. Therefore, hospital systems may look to artificial intelligence tools such as large language models like generative pre-trained transformer-4 (GPT-4) to automate discharge instruction generation, especially for languages other than English. However, no studies to date examined these models' ability to generate complete, reading-level-appropriate discharge instructions in English and Spanish. Given this, we investigated the completeness and readability of GPT-4-generated discharge instructions in English and Spanish for 6 common pediatric emergency room complaints.

We found that GPT-4-generated discharge instructions were significantly more complete in English than in Spanish. On average, discharge instructions in English contained 62% of must-include discharge items, while instructions in Spanish contained 53% of must-include discharge items. Reading level was not different across languages, and GPT-4 was able to adjust reading level to an eighth-grade, but not fifth-grade, level. These results suggest that before artificial language tools can be used to draft discharge instructions, more work is needed to ensure that these tools are well-validated in their target language and can produce appropriately complete outputs.

Key words: artificial intelligence; pediatric emergency medicine; computer simulations; diversity, equity, inclusion; literacy.

Background and significance

Since the COVID-19 pandemic, emergency departments have been struggling with increased patient volumes and staffing shortages.[1] Prewritten discharge instructions are sometimes recommended to increase emergency department efficiency;[2] however, discharge instructions more generally are often above patients' health literacy levels[3] or not in their preferred language at all.[4,5] In prior studies, over 90% of clinicians indicated that generating language-concordant discharge instructions represented a barrier to care, and 26% of children's hospitals do not translate discharge instructions.[6,7] Additionally, even when language-concordant discharge instructions are provided to patients, Spanish-language instructions are often much less complete than English-language ones.[4]

Received: January 6, 2024; **Revised:** May 16, 2024; **Editorial Decision:** May 17, 2024; **Accepted:** May 28, 2024

This language discordance becomes an issue when one considers that people with limited English proficiency (LEP) experience worse healthcare quality than their English-proficient counterparts,[8,9] especially since Spanish-speaking parents of pediatric patients have indicated they would understand discharge instructions better if information was written in Spanish and that they would prefer to receive instructions in Spanish.[10] Patient-provider language discordance during discharge has also been associated with reduced understanding of medication category and purpose[11] as well as increased rates of readmission.[12,13]

The reasons for the use of English discharge instructions for Spanish-speaking patients are many, including timing issues in coordination of document translation and patient discharge as well as shortages in staffing and funding for document translation.[6] In fact, these issues have led some clinicians to use software such as Google Translate to generate discharge instructions, even though this practice can generate instructions that might lead to harm in more complex cases.[14,15] Further innovation in the process of producing multilingual discharge instructions is therefore needed to increase the efficiency of emergency departments and ameliorate the health inequities that current practices can cause.

Generative large language models (ie, artificial intelligence tools designed to generate text based on user-provided inputs) such as generative pre-trained transformer-4 (GPT-4) may serve as useful tools to help busy clinicians generate discharge instructions for emergency department patients. GPT-4 has shown promise in writing intelligible and coherent discharge summaries and instructions for patients post-surgery,[16,17] modulating the medical complexity of texts,[18] and in answering patient medical questions in Spanish.[19] Therefore, inputting a query requesting discharge instructions in Spanish into a large language model may be an efficient alternative to writing and then translating discharge instructions.

For GPT-4-generated discharge instructions to be useful, they must contain the necessary information to guide care at home and prompt re-presentation to the hospital when necessary. Furthermore, discharge instructions must be at a reading level that patients can understand. To our knowledge, no literature currently exists examining the specific issues of completeness and readability of GPT-4-generated discharge instructions in English and in Spanish. Therefore, the current study's primary objective is to systematically evaluate multilingual GPT-4-generated discharge instructions in terms of completeness and readability in both English and Spanish. A comparison of discharge instructions' completeness and readability by language will help to both evaluate large language models' current potential for the generation of discharge instructions as well as identify language-specific shortcomings that may disproportionally affect Spanish-speaking patients as these models continue to be adapted for use in healthcare settings.

Materials and methods

Generation of GPT-4 output corpus

In June of 2023, GPT-4 was used to create discharge instructions at fifth- and eighth-grade reading levels for 6 common pediatric ED complaints (diarrhea and vomiting, bronchiolitis, fever, abdominal pain, asthma, and minor head injury without concussion), which constitute the chief complaint for approximately two-thirds of pediatric emergency department visits.[20] GPT-4 was chosen over GPT-3.5 due to prior evidence suggesting GPT-4's superior handling of non-English languages.[21] Before analysis was carried out, the input prompt was iterated and refined to specify pediatric (versus general) diagnosis and discharge instructions for parents (versus simply discharge instructions). These changes were made after previous prompts generated discharge instructions for adult (versus pediatric) complaints or were directed to the pediatric patients (as opposed to their parents). The term *emergency department* was not used so as to allow for generalizability of results to those cases where patients are discharged soon after admission (eg, for intravenous rehydration in the setting of severe vomiting). In addition, while no specific guidance regarding length of discharge instructions exists, avoiding excessively lengthy text generally has been used in prior studies to ensure discharge instruction readability[22]; furthermore, the authors wanted to avoid rewarding GPT-4 outputs that contained excessively verbose text. Responses were therefore capped at 500 words, a length that allowed for sufficient word space to produce complete discharge instructions while preventing excessively verbose outputs. The final prompt used was "Create discharge instructions (in under 500 words) for pediatric [diagnosis] at a[n] [5th/8th]-grade reading level in [Spanish/English] for parents." Each instruction case was generated 5 times to appropriately capture the breadth of potential responses from GPT-4. A new session was used for the generation of each output. Therefore, given 6 chief complaints, each with 4 cases (2 languages and 2 grade levels), with each case being generated 5 times, 120 total outputs were generated.

Scoring of outputs

Discharge instructions were analyzed in terms of completeness and readability by 2 individuals, both of whom are clinical medical students and certified English-Spanish medical interpreters. Completeness was assessed based on a previous report of the most important elements of discharge instructions according to emergency room clinicians[23]: given that no universally accepted criteria exist for the specific return criteria that must be included in pediatric discharge instructions,[24] the 5 most important elements of discharge instructions for each diagnosis were extracted from the results of a previous study of ED clinicians (see Table S1).[23] Briefly, this study used a modified Delphi method with 37 clinicians to converge on those discharge instruction features considered most important to include.

To measure completeness, the presence of each discharge instruction criterion in each output was coded as present or absent by the 2 raters. To assess inter-rater reliability, Cohen's kappa was examined across raters. Kappa values were deemed unsatisfactorily low, and so specific discordances between raters were identified (see online repository file *completeness_Scores_Comparisons.xlsx*), and the reason for discordance determined. Then, a third, nonclinical author of the paper was consulted to understand how these discordances would be interpreted by a nonclinical audience (eg, parents) and create final scoring rules for these discordances (see *Resolving Discordances* in the Supplementary Methods).

Quantification of completeness

Completeness for each instruction was calculated by measuring the proportion of most important elements contained within the generated output and was reported as a

percentage. It was also noted that the top 5 items to include in discharge instructions were not unanimously agreed upon by all clinicians for all criteria in the cited Delphi study. While this lack of agreement is expected given normal variations in clinical judgment, this variability in the top 5 items to include could potentially impact the utility of raw completeness scores to determine GPT-4's ability to generate complete discharge instructions. For example, if the agreement amongst experts for the top 5 return criteria for a given chief complaint were only 80%, the average completeness rating of a discharge instruction containing only those top 5 criteria (if rated by those same experts) would be only 80% (see *Proof S1* in the Supplemental Methods for a demonstration of this). Therefore, we endeavored to create a measure comparing the raw completeness of the GPT-4 output to the average expected completeness rating of an "ideal" discharge summary (ie, one containing all top 5 criteria). This was calculated by calculating the ratio of the raw completeness score and the mean percent consensus amongst raters for each chief complaint (see Table S1 for these consensus scores). Completeness scores, therefore, can be used to gain an intuitive sense of GPT-4's ability to incorporate essential discharge information into discharge instructions; completeness ratios, on the other hand, can help readers interpret these completeness scores in the context of normal variation in clinical opinion.

Quantification of readability

Previous research has demonstrated that approximately 80% of ED attendees have a reading level of eighth grade or lower,[25] although more recent research has shown an average reading ability between a seventh- and eighth-grade level in 1 urban ED.[26] Therefore, an eighth-grade reading level was chosen as one of the test cases. Additionally, the American Medical Association has previously recommended that patient education materials be written at a fifth- or sixth-grade reading level;[27] therefore, a fifth-grade case was also included. Readability was quantified using the Fernandez-Huerta readability score for Spanish or the Flesch-Kincaid readability score for English, both of which have previously been used to compare the readability of patient education texts in English and Spanish. For these goal reading levels, a reading score between 70 and 80 corresponds to a reading level of eighth or ninth grade in both these systems, while a score between 90 and 100 corresponds to a reading level of fifth grade (as a higher score indicates higher readability and therefore a lower corresponding grade level).[28]

Statistical analysis

Prior to statistical analysis, both raters read all GPT-4-generated outputs to ensure that none contained instructions that would cause harm.

The predictor variables of interest were output language and reading level case; the outcome variables of interest were completeness, completeness ratio, and Flesch-Kincaid or Fernandez-Huerta reading level. To characterize the general performance of the GPT-4 for the task, the overall completeness, completeness ratio, and reading level of all outputs were calculated. Next, to determine GPT-4's general performance in English versus Spanish, aggregated outputs (ie, all outputs in the fifth-grade and eighth-grade cases combined) in English and in Spanish were compared in terms of completeness (including completeness ratio) and readability. Then, to determine GPT-4's general ability to modulate reading level without omitting important information, aggregated outputs (ie, all outputs in English and in Spanish combined) in the fifth-grade and eighth-grade reading level cases were compared in terms of completeness (including completeness ratio) and readability. Finally, to better understand how specific grade-level cases may contribute to language-specific issues with completeness or readability, completeness (including completeness ratio) and readability were compared across languages for outputs at a specific grade level (eg, for English versus Spanish outputs in the fifth-grade reading level case).

Differences between conditions' completeness and reading levels were analyzed using independent 2-sample *t*-tests, with Levene's test of equal variance used to determine appropriate *t*-value for analysis. All significances reported were 2-sided. Analyses were carried out using SPSS statistics version 29.0.1.0.

Results

General performance and characterization of outputs

Outputs were legible and comprehensible in English and in Spanish (Table 1) and did not contain any outputs giving potentially harmful advice. Kappa values of inter-rater reliability ranged from 0.05 to 0.68 (Table S2). The mean completeness of all outputs was 57 ± 21% (mean ± standard deviation), the mean completeness ratio was 74 ± 28%, and the mean readability score was 70.5 ± 6.6 (corresponding to a seventh-grade reading level). By chief complaint, the mean completeness was 56 ± 23% (mean ± standard deviation) for diarrhea and vomiting, 59 ± 26% for bronchiolitis, 62 ± 14% for fever, 59 ± 22% for abdominal pain, 56 ± 20% for asthma, and 52 ± 18% for minor head injury without concussion. For each case (ie, each set of 5 outputs with the same chief complaint, language, and grade level), the standard deviation of completeness had a range between 9% and 30%. (Please see the repository file *completeness_Scores.xlsx* for the standard deviations for specific cases.)

English versus Spanish outputs: completeness and readability

Overall, English discharge instructions at all reading levels ($n = 60$) were significantly more complete than those in Spanish ($n = 60$). Specifically, in each language, 6 discharge diagnoses were examined, with 5 outputs generated for each at goal reading levels of fifth- and eighth-grade (ie, the fifth-grade and eighth-grade cases, respectively). English discharge instructions had a mean completeness of 62 ± 3% (mean ± standard error of the mean), while those in Spanish had a completeness of 53 ± 3% ($P = .02$). Readability (averaged across all outputs) did not differ significantly between languages. Completeness ratios followed a similar pattern, as English discharge instructions had an average completeness ratio of 80 ± 4% (mean ± standard error of the mean), while those in Spanish had an average completeness ratio of 68 ± 3% ($P = .02$) (Table 2).

Fifth-*grade versus* eighth-*grade outputs: completeness and readability*

No significant difference in completeness or completeness ratio was noted between aggregate (ie, both English and

Table 1. Sample outputs from the minor head injury without concussion prompt.

Grade level	English	Spanish
Fifth grade	Dear Parents or Guardians, Your child has had a minor head injury. Good news is, it's not a concussion, and doctors believe it is OK for your child to go home. Still, it's important to keep an eye on them and help them get better. Here's what you should do: 1) **Rest and relaxation**: Your child needs to rest a lot. They should avoid rough play, running, bike riding, or anything that could hurt their head again. Even activities that need a lot of thinking, like schoolwork, should be reduced for a few days. 2) **Pain relief**: If your child feels some pain, you can give them an over-the-counter medicine like Tylenol (or its generic form, acetaminophen). Don't give them aspirin—it's not safe for kids. Remember to always follow the instructions on the bottle. . .	**Instrucciones para el Alta: Lesión Menor en la Cabeza de su Hijo** Estimado(a) padre/madre o tutor(a): 1) **Vigile los Síntomas**: Su hijo ha sufrido una lesión menor en la cabeza, pero no ha mostrado signos de conmoción cerebral. Aun así, vigile a su hijo por si aparecen nuevos síntomas, como dolor de cabeza fuerte, vómitos, mareos, dificultad para caminar o hablar, o cambios en su comportamiento. 2) **Dolor**: Para el dolor, puede darle a su hijo medicamentos sin receta, como paracetamol (Tylenol). Siga las instrucciones de envase para la dosis adecuada. No le de aspirina a su hijo. . .
Eighth grade	Dear Parent/Guardian, Your child has had a minor head injury. We're glad to tell you that there are no signs of a concussion, but we do need you to monitor them at home for any changes. **Rest and activity** 1) Ensure your child gets plenty of rest in the first 24-48 hours. This helps the healing process. 2) Avoid vigorous physical activities, like running or playing sports, until they feel completely well again. . .	**Instrucciones de Alta para un Lesión Menor de Cabeza Sin Conmoción en Pediatría** Estimados padres, Su hijo ha tenido una lesión menor en la cabeza pero no muestra signos de conmoción cerebral. A continuación, encontrará las instrucciones a seguir en casa para ayudarle a recuperarse. **Cuidado en casa** 1) Descanso. Asegúrese de que su hijo descanse durante las primeras 24-48 horas después del incidente. 2) Supervisión: Preste atención a cualquier cambio en su comportamiento o estado de ánimo. Aunque no tenga conmoción cerebral, aún debe ser observado por posibles síntomas. . .

Table 2. Completeness and readability in all English versus all Spanish outputs.

Output attribute	All English ($n = 60$)	All Spanish ($n = 60$)	Difference
Completeness, mean ± SEM	62 ± 3%	53 ± 3%	$t(118) = 2.33$, $P = .02$
Completeness ratio, mean ± SEM	80 ± 4%	68 ± 3%	$t(188) = 2.38$, $P = .02$
Readability, mean ± SEM	70.73 ± 1.08	70.27 ± 0.55	$t(87) = 0.38$, $P = .70$

Spanish) outputs for the 2 grade levels. Readability scores were significantly higher in the fifth-grade case discharge instructions as an aggregate (mean = 74.4 [corresponding to seventh grade], SEM = 0.70, $n = 60$) compared to the eighth-grade case (mean = 66.6 [eighth to ninth grade], SEM = 0.69, $n = 60$); $P < .001$ (Table 3).

English versus Spanish outputs stratified by grade level case: completeness and readability

In the fifth-grade case, English and Spanish outputs did not differ significantly in completeness or completeness ratio; however, English outputs were significantly more readable than Spanish ones in the fifth-grade case (English: mean = 77.24 [seventh grade], SE = 0.98, $n = 30$; Spanish: mean = 71.53 [seventh grade], SE = 0.66, $n = 30$, $P < .001$). For the eighth-grade case,

English and Spanish outputs did not differ significantly in completeness or completeness ratio.; however, the Spanish eighth-grade cases were significantly more readable than those in English (English: mean = 64.22 [eighth to ninth grade], SE = 0.95, $n = 30$; Spanish: mean = 69.0 [eighth to ninth grade], SE = 0.82, $n = 30$, $P < .001$) (Table 4).

Model omissions

Elements that were missing across all discharge instructions were as follows:

- Diarrhea/vomiting prompt: return criterion of bloody or green vomit (no output mentioned green vomit specifically).
- Fever prompt: no output explicitly mentioned that antipyretics can reduce fever, but do not prevent it from returning.

Discussion

To our knowledge, this study is the first to assess generative large language models' production of discharge instructions in English and Spanish. While GPT-4-generated readable discharge instructions with minimal prompting, outputs included only 57% of literature-derived completeness criteria across multiple discharge scenarios. Outputs were slightly more complete in English than in Spanish (62% versus 53%)., suggesting that the model is less competent at the discharge instruction generation task in Spanish than in English.

Table 3. Completeness and readability in all fifth-grade versus all eighth-grade cases.

Output attribute	All fifth grade ($n = 60$)	All eighth grade ($n = 60$)	Difference
Completeness, mean ± SEM	55 ± 3%	59 ± 2%	$t(118) = -1.06$, $P = .29$
Completeness ratio, mean ± SEM	72 ± 4%	77 ± 3%	$t(118) = -0.98$, $P = .33$
Readability, mean ± SEM	74.4 ± 0.70	66.6 ± 0.69	$t(118) = 7.91$, $P < .001$

Table 4. Completeness and readability in English versus Spanish for fifth-grade and eighth-grade cases.

Output attribute	English ($n = 30$)	Spanish ($n = 30$)	Difference
Completeness (fifth grade), mean ± SEM	60 ± 4%	51 ± 4%	$t(58) = 1.61$, $P = 0.11$
Completeness ratio (fifth grade), mean ± SEM	78 ± 6%	65 ± 6%	$t(58) = 1.66$, $P = 0.10$
Completeness (eighth grade), mean ± SEM	63 ± 3%	55 ± 3%	$t(58) = 1.71$, $P = 0.09$
Completeness ratio (eighth grade), mean ± SEM	82 ± 5%	71 ± 4%	$t(58) = 1.72$, $P = 0.09$
Readability (fifth grade), mean ± SEM	77.24 ± 0.98	71.53 ± 0.66	$t(58) = 4.82$, $P < .001$
Readability (eighth grade), mean ± SEM	64.22 ± 0.95	69.0 ± 0.82	$t(58) = -3.82$, $P < .001$

In addition to GPT-4's overall low completeness in the generation of discharge instructions, the model's inferior performance in Spanish should also be an improvement target for developers and clinicians. As generative language models improve and individuals and organizations start turning to these tools for multi-lingual discharge instruction generation, the importance of these models' multilingual performance will only continue to grow. One method to improve large language models' performance in Spanish could be to increase the size of their Spanish-language medical training corpora: because GPT-4 is a transformer-style model which predicts the next token (eg, a word) in a series of text based on previous training data, differences in the size or quality of English- versus Spanish-language training data can introduce differences in output quality across languages.[29–32]

An important caveat to consider when interpreting the current study's results is the heterogeneity in clinician opinion on the most important items to include in discharge instructions. Unfortunately, there exist no specific guidelines on the return criteria to include in pediatric emergency room discharge instructions, as even the American Academy of Pediatrics Joint Policy Statement regarding guidelines for the care of children in the pediatric ED does not provide guidance of specific return criteria.[24] Indeed, the consensus on information to include in discharge instructions in the Delphi study used to determine return criteria ranged from 51% to 100%.[23] While improvements in pediatric ED guidelines may help resolve some of this ambiguity in the future, completeness figures must until then be interpreted with the knowledge that even an output containing all 5 return top criteria may be graded as under 100% complete by some clinicians.

Completeness scores found in this study (ranging from 52% to 62%) are in line with those in other recent studies. For example, in a study of the quality of Spanish-language GPT-4-generated answers to questions regarding chronic diseases, GPT-4-generated instructions were rated by clinicians as "comprehensive" ≤50% of the time, with the remainder being "Correct but inadequate" or containing information that was a mix of correct or incorrect.[33] In another study examining English-language GPT-4-generated answers to urology questions, physician-rated comprehensiveness was generally high, but 66% of the raters did not endorse high comprehensiveness for the emergency cases specifically.[6] Therefore, GPT-4 performance for the discharge-instruction task seems to be in line with its performance on other clinical tasks.

In terms of readability, GPT-4 was successful in generating outputs at the eighth- but not fifth-grade level; specifically, the fifth-grade cases had a readability at the seventh-grade level, and the eighth-grade cases had a readability at the eighth- to ninth-grade levels. While seventh grade is at a higher level than that recommended by the American Medical Association,[27] a document written at a seventh-grade reading level is likely still accessible to the average emergency department patient,[22] and is more accessible than most current discharge instructions. Older data suggest that ED discharge instructions have been written at approximately an 11th-grade reading level historically,[34] with more recent data from internal medicine services revealing an average discharge instruction reading level at the 10th- to 12th-grade reading level, with only 11.3% of discharge instructions being written below the seventh-grade reading level.[35]

Furthermore, these results demonstrate the partial success of prompt engineering (ie, the practice of designing prompts for large language models to guide the model toward generating the desired output), as specification of reading level did lead GPT-4 to successfully modulate this parameter to some degree. Prior studies have also shown GPT-4's ability to reduce medical consent forms from a college- to an eighth-grade reading level.[36] Interestingly, a separate large analysis of GPT-4-generated prompts indicated that the floor for GPT-4 in terms of reading level was sixth grade, which may explain the model's inability to produce fifth-grade-level outputs.[37] This large study was also published in November 2023, 5 months after data generation for the current study, so GPT-4's reading-level modulatory capabilities may have also improved over this time.

Of note, busy clinicians may not always include reading-level specifications; however, given that GPT-4 is able to intake custom instructions to be applied to all outputs,

hospitals could set up GPT-4 (or other large language models) with a specification toward reading level to remove the need of clinicians to input this every time. Future work examining baseline reading level of GPT-4 outputs without guidance would also be illuminating.

Because previous research has shown success in using user input to direct the performance of transformer-based generative models,[38] the authors also briefly explored prompt engineering as a method to ameliorate these issues. Specifically, GPT-4 was given the same prompts in English and Spanish for the head injury with concussion case but instructed (in English) to include the specific completeness criteria examined in the present study. When this additional step was performed, both the English and Spanish outputs successfully incorporated these completeness criteria. Taken together, these findings suggest 2 parallel yet complementary avenues for further research: characterization of the differences in GPT-4 outputs based on language and reading level and further prompt reengineering to improve generated output. Preliminary research suggests GPT-4 might be used improve its own prompts,[39] which presents another potential avenue of study. Large language models may also be used to preemptively generate discharge instructions, allowing for pilot testing and storage of these instructions until they are needed.

Strengths of the current study include a standardized evaluation of multiple discharge instructions, allowing for a systematic and rigorous evaluation of GPT-4 produced outputs and providing future researchers with a template for similar analyses. This study also provides future researchers with a baseline of GPT-4's performance on the generation of simple discharge instructions, which will be useful when evaluating the model's performance on more complex tasks. Finally, this study highlights the importance for researchers and clinicians to approach generative AI technologies with thoughtfulness and care, as these technologies may inadvertently worsen disparities in researchers' attempts to reduce them.

One limitation of this study is our approach to minimize prompt engineering. While this methodologic choice was intended to mimic a busy clinician quickly querying GPT-4 for discharge instructions, the authors acknowledge the large effects that prompt design can have on GPT-4 outputs[40] and welcome further research into applications of prompt engineering in healthcare document generation. A second limitation relates to the metrics used to analyze reading level that, while validated and widely used for this purpose, rely on such factors as number of syllables and number of words. These tools were chosen because of their prior use in medical contexts and because the Fernandez-Huerta score is a derivative of the Flesch-Kincaid score,[28] which allows for better comparability between the measures; however, they may miss more nuanced complexity in discharge instructions. Other health-literacy-related aspects of GPT-4-generated discharge instructions worth examining could include the use of jargon, the structural organization of discharge instructions, and patient-reported measures of comprehensibility and actionability. Finally, the novelty of the completeness ratio measure is a limitation; while the completeness ratio may prove a useful tool to contextualize completeness scores in the context of heterogenous clinical opinions, further validation of the score will be necessary to allow for full confidence in its use and interpretation. Especially given the ambiguous nature of some of the criteria used in the present study, further validation of essential return criteria will also be important in evaluating future attempts to generate complete discharge instructions.

These limitations provide numerous avenues for future work. Future studies should investigate how the quality of discharge instructions can be further improved using refined prompt tuning, both at the level of the individual user as well as automated prompt-editing innovations. More specifically, user-level interventions such as templates for produced text may improve GPT's ability to produce outputs better in line with clinicians' needs,[41] and clinician input can be used to further evaluate the quality of discharge instructions. It will also be important to use other tools (eg, machine learning approaches that can take into account more complex structural features of texts)[42] to compare reading level in terms of vocabulary and other text features. Finally, the relationship between objectively defined output completeness and physician-determined appropriateness of outputs can be further investigated to better delineate performance standards during the development of GPT-4 tools prior to wide-scale testing.

Future studies can also examine prompt engineering to determine the best way to phrase requests for discharge instructions in both languages. Studies of generalizability across other complex diseases or in new clinical settings are also important steps to better understand these tools in the clinical domain. At present, these tools are not yet sufficiently evaluated to be safe for clinical use without human-in-the-loop clinical editing. Prompts including personalized instructions might improve the likelihood that generated text include points critical for particular patients. Additionally, the generated outputs (in both languages) might serve as useful templates that can be minimally edited to provide useful discharge instructions for Spanish-speaking patients.

Conclusion

Busy clinicians in fast-paced clinical settings might reach for GPT-4 and similar tools to quickly produce clinical documentation including discharge instructions. These tools' ability to produce outputs in numerous languages might serve to improve clinical workflows and reduce disparities at discharge; however, without prompt engineering, GPT-4 does not generate complete discharge instructions as of June 2023, a shortcoming which is more apparent in Spanish than in English. While the current study provides a baseline of GPT-4's performance in basic cases of this task, careful and rigorous further evaluation of such tools both in simulation and in controlled clinical studies are still needed before they can be safely applied in the clinical space for more complex discharges.

Author contributions

Alex Gimeno conceived of and designed the study, generated, and analyzed the data, and helped write the manuscript. Kevin Krause aided with study design, document scoring, and writing the manuscript. Starina D'Souza assisted with document scoring and interpretation of results. Colin G. Walsh aided with study design and writing the manuscript.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

C.G.W. is supported by funding from Wellcome Leap MCPsych; NIMH R01 MH121455; NIMH R01 MH120122; NIMH R01 MH118233; NIH 5RM1HG009034; and BRIDGE2AI (A.P.). K.K. is supported by funding from NLM Training Grant T15 LM007450.

Conflicts of interest

None of the authors have competing interests to declare.

Data availability

The data underlying this article are available in the following GitHub repository: https://github.com/AlexGimeno3/GPT_Corpus/.

References

1. Varner C. Emergency departments are in crisis now and for the foreseeable future. *CMAJ*. 2023;195(24):E851-E852.
2. Hill J, Frederick M, Santen SA, et al. Turning lemons into lemonade: teaching strategies in boarded emergency departments. *AEM Educ Train*. 2023;7(5):e10914.
3. Samuels-Kalow M, Rhodes K, Uspal J, et al. Unmet needs at the time of emergency department discharge. *Acad Emerg Med Off J Soc Acad Emerg Med*. 2016;23(3):279–287.
4. Platter E, Hamline MY, Tancredi DJ, et al. Completeness of written discharge guidance for English- and Spanish-speaking patient families. *Hosp Pediatr*. 2019;9(7):516-522.
5. Isbey S, Badolato G, Kline J. Pediatric emergency department discharge instructions for Spanish-speaking families. *Pediatr Emerg Care*. 2022;38(2):E867-e870.
6. Davis SH, Rosenberg J, Nguyen J, et al. Translating discharge instructions for limited English-proficient families: strategies and barriers. *Hosp Pediatr*. 2019;9(10):779-787.
7. D'Annibale D, Keyt L, Seymann G. *Discharge Instructions for Spanish-Speaking Patients: A House Staff Perspective*; 2023. https://escholarship.org/uc/item/4hp872tn
8. Eneriz-Wiemer M, Sanders LM, Barr DA, et al. Parental limited English proficiency and health outcomes for children with special health care needs: a systematic review. *Acad Pediatr*. 2014;14(2):128-136.
9. Clark JR, Shlobin NA, Batra A, et al. The relationship between limited English proficiency and outcomes in stroke prevention, management, and rehabilitation: a systematic review. *Front Neurol*. 2022;13:790553.
10. Jang M, Plocienniczak MJ, Mehrazarin K, et al. Evaluating the impact of translated written discharge instructions for patients with limited English language proficiency. *Int J Pediatr Otorhinolaryngol*. 2018;111:75-79.
11. Karliner LS, Kim SE, Meltzer DO, et al. Influence of language barriers on outcomes of hospital care for general medicine inpatients. *J Hosp Med*. 2010;5(5):276-282.
12. Lindholm M, Hargraves JL, Ferguson WJ, et al. Professional language interpretation and inpatient length of stay and readmission rates. *J Gen Intern Med*. 2012;27(10):1294-1299.
13. Karliner LS, Pérez-Stable EJ, Gregorich SE. Convenient access to professional interpreters in the hospital decreases readmission rates and estimated hospital expenditures for patients with limited English proficiency. *Med Care*. 2017;55(3):199-206.
14. Kreger V, Aintablian H, Diamond L, et al. 10 Google translate as a tool for emergency department discharge instructions? Not so fast!. *Ann Emerg Med*. 2019;74(4):S5-S6.
15. Khoong EC, Steinbrook E, Brown C, et al. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Intern Med*. 2019;179(4):580-582.
16. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107-e108.
17. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci*. 2023;192(6):3197-3200.
18. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: promising results, limitations, and potential. *Vis Comput Ind Biomed Art*. 2023;6:9. https://doi.org/10.1186/s42492-023-00136-5
19. Yeo YH, Samaan JS, Ng WH, et al. 2023. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis, medRxiv, 2023.05.04.23289482, preprint: not peer reviewed.
20. Massin MM, Montesanti J, Gérard P, et al. Spectrum and frequency of illness presenting to a pediatric emergency department. *Acta Clin Belg*. 2006;61(4):161-165. https://doi.org/10.1179/acb.2006.027
21. Koubaa A. GPT-4 vs GPT-3.5: a concise showdown; 2023:24. https://doi.org/10.20944/PREPRINTS202303.0422.V1
22. DeSai C, Janowiak K, Secheli B, et al. Empowering patients: simplifying discharge instructions. *BMJ Open Qual*. 2021;10(3):e001419.
23. Curran JA, Murphy A, Burns E, et al. Essential content for discharge instructions in pediatric emergency care: a delphi study. *Pediatr Emerg Care*. 2018;34(5):339-343.
24. American Academy of Pediatrics, Committee on Pediatric Emergency Medicine, American College of Emergency Physicians, et al. Joint policy statement—guidelines for care of children in the emergency department. *Pediatrics*. 2009;124(4):1233-1243.
25. Williams DM, Counselman FL, Caggiano CD. Emergency department discharge instructions and patient literacy: a problem of disparity. *Am J Emerg Med*. 1996;14(1):19-22.
26. Sheikh S, Hendry P, Kalynych C, et al. Assessing patient activation and health literacy in the ED. *Am J Emerg Med*. 2016;34(1):93-96.
27. Weiss BD. *Health Literacy: A Manual for Clinicians*. American Medical Association Foundation; 2003.
28. Gorrepati PL, Smith GP. Contrasting readability and availability of Spanish language with English language patient education materials. *Pediatr Dermatol*. 2021;38(Suppl 2):142-143.
29. OpenAI. GPT-4 Technical Report, 15 March 2023.
30. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2017.
31. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. *OpenAI Blog*. 2019;1:9.
32. Baevski A, Edunov S, Liu Y, et al. 2019. Cloze-driven pretraining of self-attention networks, arXiv, arXiv:190307785, preprint: not peer reviewed.
33. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, et al. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. *Digit Health*. 2024;10:20552076231224603. https://doi.org/10.1177/20552076231224603
34. Spandorfer JM, Karras DJ, Hughes LA, et al. Comprehension of discharge instructions by patients in an urban emergency department. *Ann Emerg Med*. 1995;25(1):71-74.
35. Burns ST, Amobi N, Chen JV, et al. Readability of patient discharge instructions. *J Gen Intern Med*. 2022;37(7):1797-1798.
36. Ali R, Connolly ID, Tang OY, et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. *Npj Digit Med*. 2024;7:1–6.
37. Amin KS, Mayes L, Khosla P, et al. 2023. ChatGPT-3.5, ChatGPT-4, Google Bard, and Microsoft Bing to improve health literacy and communication in pediatric populations and beyond, arXiv, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2311.10075

38. Kosonocky CW, Feller AL, Wilke CO, et al. 2023. Prompt engineering for transformer-based chemical similarity search identifies structurally distinct functional analogues, arXiv, preprint: not peer reviewed.
39. Peng B, Li C, He P, et al. 2023. Instruction tuning with GPT-4, arXiv, preprint: not peer reviewed.
40. Yang X, Peynetti E, Meerman V, et al. What GPT knows about who is who. In: *Insights 2022—3rd Workshop Insights Negat Results NLP Proc Workshop*; 2022:75-81.
41. Wang J, Shi E, Yu S, et al. 2023. Prompt engineering for healthcare: methodologies and applications, arXiv, preprint: not peer reviewed.
42. Attia M, Samih Y, Ehara Y. Statistical measures for readability assessment. In: Hämäläinen M, Öhman E, Pirinen F, et al., eds. *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics; 2023:153-161. Accessed February 11, 2024. https://aclanthology.org/2023.nlp4dh-1.19

JAMIA Open, 2024, 7, 1–8
https://doi.org/10.1093/jamiaopen/ooae050
Research and Applications