Patient-Representing Population Perceptions of GPT-Generated vs. Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment

Thomas Huang, Conrad Safranek, Vimig Socrates, David Chartash, Donald Wright, Monisha Dilip, Rohit B. Sangal, Richard Andrew Taylor

Submitted to: Journal of Medical Internet Research
on: May 08, 2024

Table of Contents

Patient-Representing Population Perceptions of GPT-Generated vs. Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment

Thomas Huang[1, 2] BS; Conrad Safranek[1, 2] BS; Vimig Socrates[2, 3] MS; David Chartash[2, 4] PhD; Donald Wright[1, 2] MD, MHS; Monisha Dilip[1] MD; Rohit B. Sangal[1] MD, MBA; Richard Andrew Taylor[1, 2] MD, MHS

[1]Department of Emergency Medicine Yale School of Medicine New Haven US
[2]Department for Biomedical Informatics and Data Science Yale School of Medicine New Haven US
[3]Program of Computational Biology and Bioinformatics Yale University New Haven US
[4]School of Medicine, University College Dublin National University of Ireland Dublin IE

Corresponding Author:
Richard Andrew Taylor MD, MHS
Department of Emergency Medicine
Yale School of Medicine
333 Cedar St
New Haven
US

Abstract

Background: Discharge instructions are a key form of documentation and patient communication in the time of transition from the Emergency Department (ED) to home. Discharge instructions are time-consuming and often under-prioritized, especially in the ED, leading to discharge delays and patient instructions that are either impersonal. Generative artificial intelligence and large language models (LLMs) offer promising methods of creating high-quality and personalized discharge instructions, however there exists a gap in understanding patient perspectives of LLM-generated discharge instructions.

Objective: We aimed to assess the use of LLMs such as ChatGPT in synthesizing accurate and patient-accessible discharge instructions from the ED.

Methods: We synthesized 5 unique, fictional ED encounters meant to emulate real ED encounters that included a diverse set of clinician H&P notes and nursing notes. These were passed to GPT-4 in Azure OpenAI service to generate corresponding LLM-generated discharge instructions. Standard discharge instructions were also generated for each of the 5 unique ED encounters. All GPT-generated and standard discharge instructions were then formatted into standardized after-visit summary documents. These after-visit summaries containing either GPT-generated discharge instructions or standard discharge instructions were given to Amazon MTurk respondents subjects representing patient populations through Amazon MTurk Survey Distribution. Discharge instructions were assessed based upon metrics of interpretability of significance, understandability, and satisfaction.

Results: Our findings revealed 155 survey respondents assigned favorable ratings more frequently to GPT-generated discharge instructions along the metrics of interpretability of significance in discharge instruction subsections regarding diagnosis, procedures, treatment, post-ED medications or any changes to medications, and return precautions (GPT/Standard respectively: 89.2%/79.5%, 86.7%/65.8%, 74.7%/61.6%, 63.9%/49.3%, 86.7%/68.5%). Survey Respondents found GPT-generated instructions more understandable when rating procedures, treatment, post-ED medications or medication changes, post-ED follow-up, and return precautions (80.7%/61.6%, 85.5%/68.5%, 68.7%/57.5%, 86.7%/76.7%, 85.5%/76.7%). Satisfaction with GPT-generated discharge instruction subsections were most favorable in procedures, treatment, post-ED medications or medication changes, and return precautions (75.9%/54.8%, 85.5%/68.5%, 62.7%/53.4%, 83.1%/71.2%). Kruskal-Wallis analysis of Likert-responses between GPT-generated and standard discharge instructions did not conclude significant differences within any specific metric and discharge instruction subsection.

Conclusions: This study demonstrates the potential for LLMs such as ChatGPT to act as a method of augmenting current documentation workflows in the ED to reduce documentation burden of physicians. The ability for LLMs to provide tailored instructions for patients by improving readability and making instructions more applicable to patients could possibly improve upon the methods of communication that currently exist.

(JMIR Preprints 08/05/2024:60336)
DOI: https://doi.org/10.2196/preprints.60336

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Thomas Huang BS[1,2], Conrad Safranek BS[1,2], Vimig Socrates MS[2,3], David Chartash PhD[2,4], Donald Wright MD MHS[1,2], Monisha Dilip MD[1], Rohit B Sangal MD, MBA[1], R Andrew Taylor MD MHS[1,2]

[1] Department of Emergency Medicine, Yale School of Medicine, New Haven, CT
[2] Department for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT
[3] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT
[4] School of Medicine, University College Dublin - National University of Ireland, Dublin, Co. Dublin, Republic of Ireland

Corresponding Author:
Richard Andrew Taylor, MD, MHS
richard.taylor@yale.edu
464 Congress Ave. Suite 260
New Haven, CT 06519
203-909-3012

Patient-Representing Population Perceptions of GPT-Generated vs. Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment

Background: Discharge instructions are a key form of documentation and patient communication in the time of transition from the Emergency Department (ED) to home. Discharge instructions are time-consuming and often under-prioritized, especially in the ED, leading to discharge delays and patient instructions that are either impersonal. Generative artificial intelligence and large language models (LLMs) offer promising methods of creating high-quality and personalized discharge instructions, however there exists a gap in understanding patient perspectives of LLM-generated discharge instructions.

Objectives: We aimed to assess the use of LLMs such as ChatGPT in synthesizing accurate and patient-accessible discharge instructions from the ED.

Methods: We synthesized 5 unique, fictional ED encounters meant to emulate real ED encounters that included a diverse set of clinician H&P notes and nursing notes. These were passed to GPT-4 in Azure OpenAI service to generate corresponding LLM-generated discharge instructions. Standard discharge instructions were also generated for each of the 5 unique ED encounters. All GPT-generated and standard discharge instructions were then formatted into standardized after-visit summary documents. These after-visit summaries containing either GPT-generated discharge instructions or standard discharge instructions were given to Amazon MTurk respondents subjects representing patient populations through Amazon MTurk Survey Distribution. Discharge instructions were assessed based upon metrics of interpretability of significance, understandability, and satisfaction.

Results: Our findings revealed 155 survey respondents assigned favorable ratings more frequently to GPT-generated discharge instructions along the metrics of interpretability of significance in discharge instruction subsections regarding diagnosis, procedures, treatment, post-ED medications or any changes to medications, and return precautions (GPT/Standard respectively: 89.2%/79.5%, 86.7%/65.8%, 74.7%/61.6%, 63.9%/49.3%, 86.7%/68.5%). Survey Respondents found GPT-generated instructions more understandable when rating procedures, treatment, post-ED medications

or medication changes, post-ED follow-up, and return precautions (80.7%/61.6%, 85.5%/68.5%, 68.7%/57.5%, 86.7%/76.7%, 85.5%/76.7%). Satisfaction with GPT-generated discharge instruction subsections were most favorable in procedures, treatment, post-ED medications or medication changes, and return precautions (75.9%/54.8%, 85.5%/68.5%, 62.7%/53.4%, 83.1%/71.2%). Kruskal-Wallis analysis of Likert-responses between GPT-generated and standard discharge instructions did not conclude significant differences within any specific metric and discharge instruction subsection.

Conclusions: This study demonstrates the potential for LLMs such as ChatGPT to act as a method of augmenting current documentation workflows in the ED to reduce documentation burden of physicians. The ability for LLMs to provide tailored instructions for patients by improving readability and making instructions more applicable to patients could possibly improve upon the methods of communication that currently exist.

Keywords: Artificial Intelligence, ChatGPT, Discharge Instructions, Emergency Medicine, Surveys and Questionaries

Introduction

Discharge instructions serve as an essential bridge between hospital treatment and at-home recovery. Particularly in the Emergency Department (ED), where clinical team members have limited time to review the events of the ED encounter and follow-up recommendations with their patients, discharge instructions are critical to communicating information to patients and increasing adherence to follow-up care. Specifically, discharge instructions serve to inform the patient about key details such as their diagnosis, evaluations performed, preliminary diagnostic test results, medications, treatment plans, follow-up care, and reasons to return to the ED.[1,2] Improved patient understanding of discharge instructions and self-efficacy following discharge have been linked to improved health outcomes and decreased readmission rates.[3–5] Despite their important role, many patients leave the hospital with discharge instructions that are jargon-filled and difficult to navigate, with studies showing that up to 88% of discharge instructions are unreadable to the population served.[6–8] General guidelines suggest that all health-related information provided by a physician to patients should be readable at a sixth-grade reading level.[2,9] Furthermore, the lack of personalization in many current discharge instructions misses an opportunity to better engage patients with understanding their care and diagnosis, which has been shown to be key to increasing patient health outcomes.[10,11] This communication gap drives non-adherence and higher rates of patient readmission.[12,13] Addressing this challenge of bridging physician and patient discharge communication is essential for improving patient health outcomes and the overall effectiveness of care delivery.

The recent advancement of Large Language Models (LLMs), such as ChatGPT, offers a potential solution to the longstanding issue of inaccessible medical communication and time-demanding nature of providing care in the ED.[14,15] Research on LLMs has demonstrated their broad capabilities on medical question-answering, including passing the United States Medical Licensing Examinations.[16] These models are able to generate medical text that is accurate, informative, more readable, and even more empathetic.[17–21] By leveraging LLMs, there is the potential to create discharge instructions that are not only tailored to individual patients, but also presented in a format that is engaging and easy to understand.

Recent research has underscored the feasibility of using LLMs for patient discharge instructions. Translation of 50 patient discharge summaries from the medical sublanguage to regular English by GPT-4 demonstrated significant improvement in objective readability scores.[21] This study also showed these LLM-generated summaries had a generally acceptable level of accuracy and completeness, per physician assessment. Other research has extended to specialties such as neurology and radiology, demonstrating that LLMs can digest complex medical text and produce summaries that are patient-centric and sufficiently comprehensive that physicians are comfortable releasing them to patients.[22–24]

While this existing LLM research has laid a robust foundation, it has so far lacked a key evaluative component: integration of patient perspectives in a randomized unbiased comparison of LLM-generated text against the current standard of care. In addition, current investigations of leveraging LLMs to generate discharge instructions or summaries in the ED remain limited. We aim to address this gap by surveying patient-representing population using a randomized blinded approach to compare LLM-generated fully formatted discharge instructions to the standard discharge instructions.

Methods

Study Population and Setting

The study population was a convenience sample of adult subjects (≥ 18 years of age) taken from Amazon MTurk from 12/31/23 - 3/31/24, residing in the United States, with an approval rate (an indicator in Amazon MTurk of work quality) of >90%, who were also categorized as Amazon MTurk Masters.[25] Research has shown that the majority of patients do not have medical training and have low health literacy and difficulty understanding medical jargon.[26,27] Thus, exclusion criteria for this population included respondents having any significant healthcare training background to best represent the general patient population. Any Amazon MTurk survey respondent was not allowed to repeat the survey therefore all responses are from unique respondents, reducing bias from individual responders and improving data quality. This study was approved by the institutional review board as an exemption (HIC #2000036301).

Synthesizing Fictional ED encounters

A set of synthetic fictional ED notes were generated that represent five separate independent ED encounters (Appendix A1). These notes were diverse in patients' age, sex, ethnicity, and clinical presentation. All ED notes were authored by an emergency medicine attending physician (DW) to emulate realistic ED settings and were rendered in multiple document styles. The emergency medicine physician note included a brief chief complaint statement, history of present illness, past medical history, physical exam, and assessment and plan. The nursing note contained brief summaries of the presentation and care performed.

ChatGPT Prompt Development

Initial prompts were developed based upon Society for Academic Emergency Medicine and Joint Commission guidelines for discharge instructions.[1,2] Prompts were applied to the five synthetic clinical note sets used for the final surveys. (Final GPT prompt available in Appendix A2) Slight adjustments were made iteratively to improve the GPT-generated discharge instruction output including changes the organization provided in the prompt, additional directive phrasing, and evaluation of output. Final prompts added to the synthetic clinical notes described above were deployed via a pipeline to automatically query and retrieve responses from OpenAI's GPT-4 for the final five clinical scenarios. Other strategies described in previous prompt engineering work were also included in this iterative process to improve GPT-4 recognition of the note and generative responses.[28]

Discharge Instructions Development

Each clinical scenario represented by the synthetic clinical note was passed to GPT-4 with the final prompt generated by the methodology described above. Each of the five clinical scenarios was associated with two sets of discharge instructions for a total of 10 different variants of discharge instructions. These discharge instructions consist of one GPT-4 generated discharge instruction and one standard discharge instruction a patient would expect to receive from the ED without GPT-4 generated text. The standard discharge instructions (provided by RS), are the standardized Epic After Visit Summary with complaint-specific discharge instructions from one of two institutionally contracted patient education resources (Elsevier Patient Education, UpToDate). These discharge instructions were then populated, through Epic's "Playground" environment, into a generated ED encounter so that an entire discharge instruction could be generated with the same overall formatting. The standard discharge instructions did not include physician-generated free text. All specific identifiers, such as phone numbers, insurance names, and trademarks, were censored out of the final generated discharge instructions.

Development, Recruitment, and Distribution

Surveys were created in Qualtrics to be distributed through the distribution service within Amazon MTurk, a reliable and widely used survey distribution service. (Qualtrics Survey available in Appendix A3) These surveys were iteratively reviewed by ED physicians, residents, and medical students to ensure readability and ease of use. The survey was an open survey format, available to all Amazon MTurk Masters that also had an approval rating of >90%. The survey was listed on Amazon MTurk and was only visible to MTurkers that met the requirements. No other forms of advertising occurred other than the listing on Amazon MTurk.

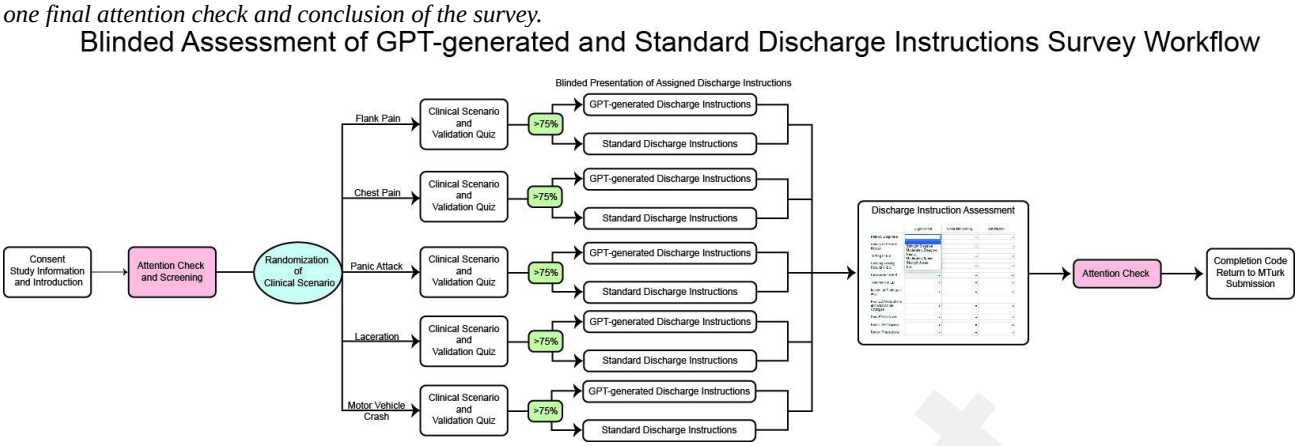
Survey Administration

The survey was distributed on Amazon MTurk over a 3-month period, during which respondents were given a link to the complete survey built on Qualtrics. The survey structure included an initial explanation of what the survey responder would be presented with one clinical scenario described by a note that contains the pertinent information to the ED visit. The survey flow saw each respondent blindly randomized to one of ten possible workflows. The respondent was initially presented with consent, initial screening questions, and attention checks that must be passed for survey results to be accepted.[29,30] Respondents were then randomized to one of five random clinical scenarios where they responded to an additional quality control quiz including 4 multiple choice questions regarding the clinical scenario. (Figure 1)

The survey respondent was then tasked to fill out a matrix of Likert-scale questions in regards to the discharge instructions they read in the context of the clinical scenario (Clinical Scenarios and Discharge Instructions available in Appendix A1). The y-axis included all components of the discharge instruction: Diagnosis, History of the Current Problem/Illness, Testing you received in the ED, Any Pending test results, Procedures, Treatment, Incidental Findings, Post-ED Medications or any changes to Medications, Post-ED self-care, Post-ED Follow up, and Return Precautions. The x-axis of the matrix asked the survey respondent to rate on a traditional Likert scale on the three metrics of interpretability of significant findings or information, ease of understanding, and satisfaction with respect to each of the discharge instruction subsections (Appendix A4).

Figure 1. The Qualtrics survey design for discharge instruction randomized blind assessment by Amazon MTurk respondents. Survey respondents first passed an initial consent documentation, screener for healthcare or medical background, and a series of attention checks. Respondents were then randomized to one possible clinical scenario of five, then randomized to view either the GPT-generated or standard version of discharge instructions. This respondent was then tasked to answer Likert-scale questions regarding the three metrics: interpretability of significance, understandability, and satisfaction in regards to each discharge instructions subsection before

one final attention check and conclusion of the survey.



Attention checks and quality checks were included throughout the survey to ensure attention and increase likelihood of high-quality data. Attention checks, such as asking the patient to answer simple but targeted questions with expected correct answers, are a method of ensuring survey respondents are reading questions and responding to the best of their ability.[31] Manual review of each round of distribution was performed to maintain data quality by assessing percentage of attention checks passed per survey respondent and validating respondent understanding of the clinical scenario through a passing score (>75%) on corresponding quiz. Respondents were allowed to go back to change their answers when necessary. Survey respondents' results were only retained for the final dataset if they passed the content validation test (>75% score) and failed no more than one of four attention checks to ensure a high-quality dataset.

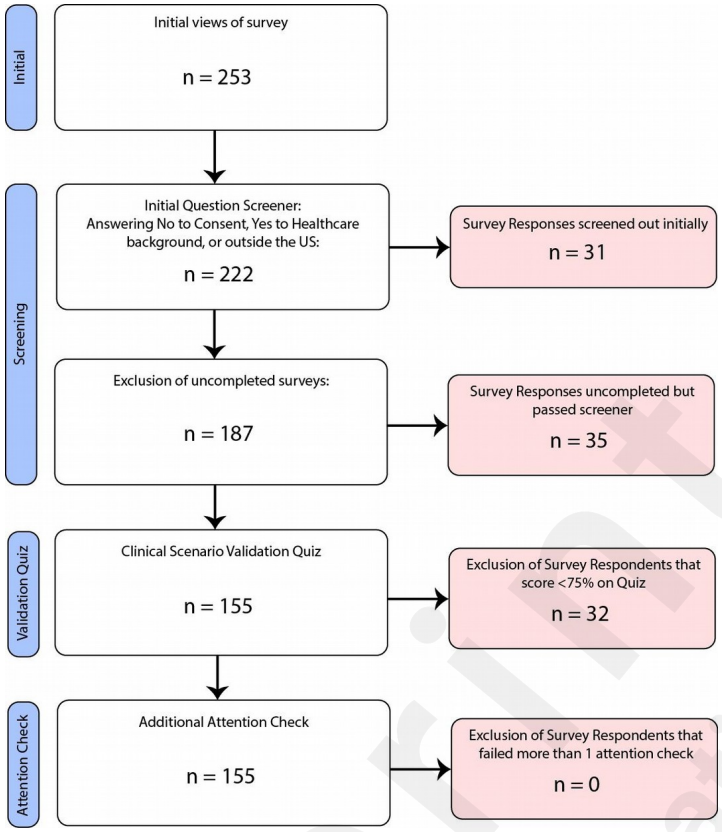
Data Analysis

Survey responses were extracted from Qualtrics and analyzed using Python v3.11.5. Frequency counts of Likert responses to each metric and discharge instruction subsection were organized and visualized in clustered stacked multi-bar charts. The Kruskal-Wallis test was used to analyze for significant differences between Likert responses between GPT-generated and standard discharge instructions.

Results

From a total of 253 Amazon MTurk Survey unique initial Amazon MTurk views, 222 (88%) passed the initial screening criteria, 187 (84%) respondents successfully completed the survey, and 155 (83%) of those respondents passed all validation and attention checks and were eligible for the final cohort. (Figure 2)

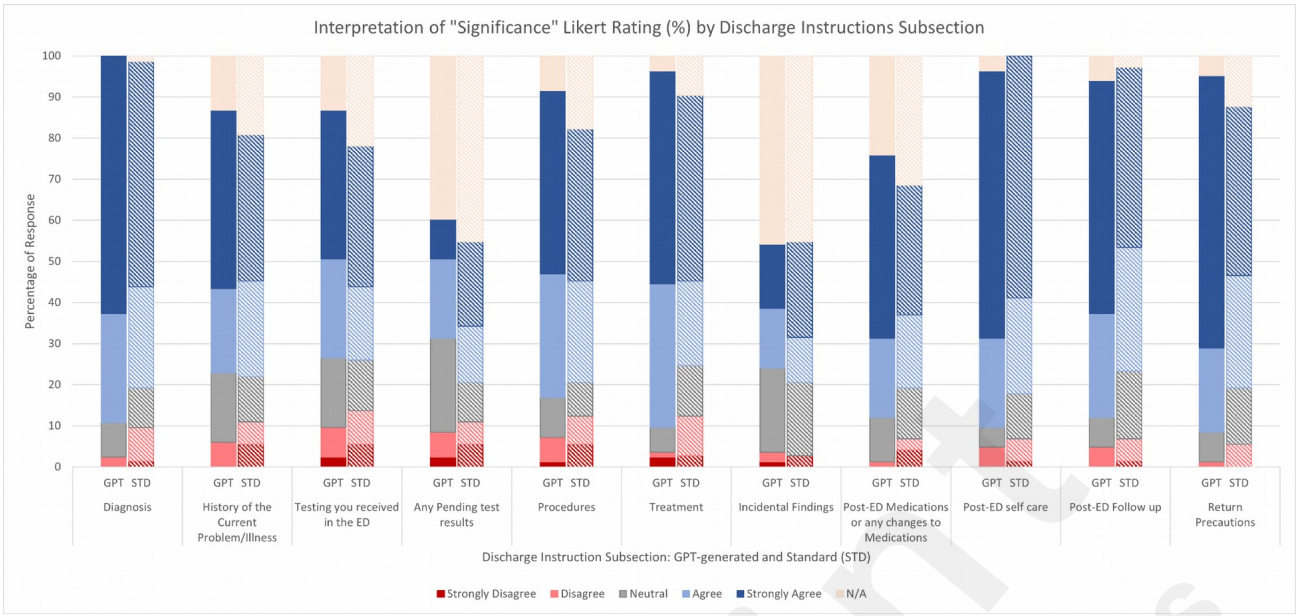
Figure 2. The stepwise method towards exclusion of survey respondents. From an initial 253 views of the survey, 222 were not screened out by the initial screen of consent, not having a healthcare background, and residing in the US. Following the initial screener, 35 did not finish the entire survey. Surveys also had their validation quiz in relation to the clinical scenario, and attention checks graded. Of 187 completed surveys, 155 passed the validation quiz, and all respondents that passed the validation quiz also successfully answered all other attention checks correctly.



Domain 1: Interpretability of Significance

To assess interpretability of the significance of the discharge instructions, each survey respondent was tasked to rate their agreement with the statement: “The information in the discharge instruction (subsection) effectively explains the interpretability of significance of the findings in a way that’s personalized to me (the hypothetical recipient) and is easy to follow.” They were asked to individually assess interpretability of significance in regards to each important subsection of the discharge instructions. Of note, the frequency of agree and strongly agree selected by survey respondents were greater across all GPT-generated discharge instructions subsections in regards to interpretability of significance except for pending test results and incidental findings.

Figure 3. This is a clustered stacked multi-bar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions in relation to the prompt: “The information in the discharge instruction (subsection) effectively explains the significance of the findings in a way that’s personalized to me (the hypothetical recipient) and is easy to follow.”



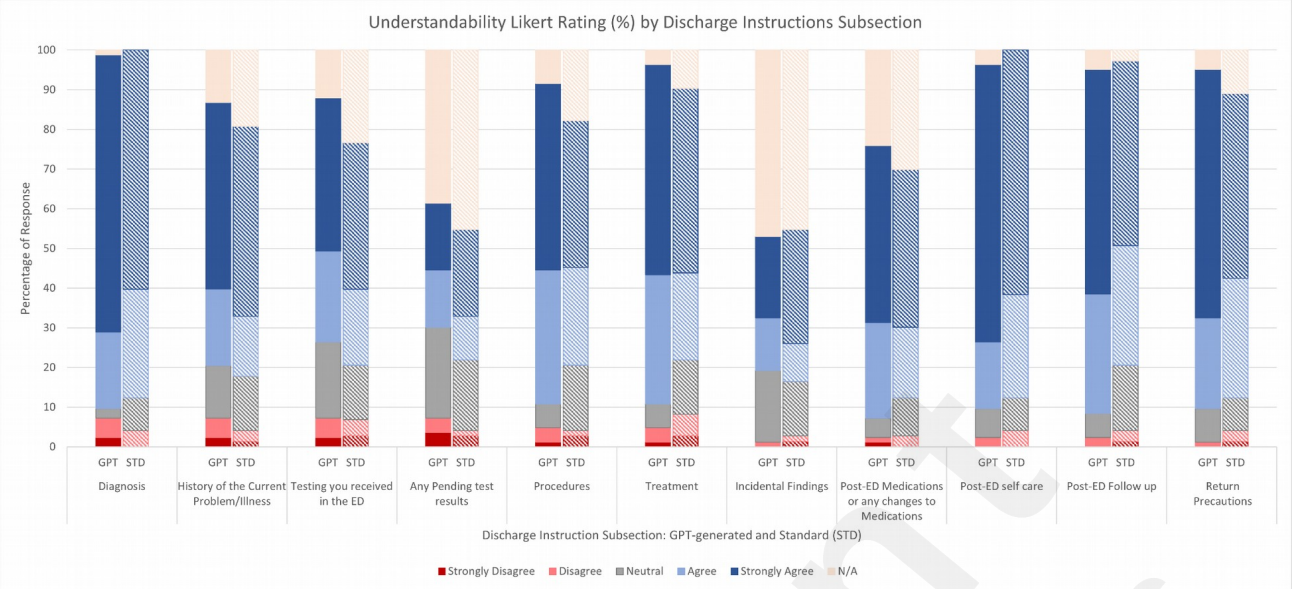
Regarding “pending test results”, 28.9% of respondents rated favorably for their ability to interpret significant information in GPT-generated discharge instructions and 34.25% favorably in standard discharge instructions. For incidental findings, 30.1% and 34.2% of respondents rated favorably for GPT-generated and standard discharge instructions, respectively.
These two sections saw the greatest percentage of N/A responses as well, with 39%/45% in pending testing results and 45%/45% for incidental findings, respectively for GPT-generated and standard discharge instructions. (Figure 3) Multiplot of all Likert response frequencies across all metrics and discharge instructions subsections is available in Appendix A4.

All other subsections of the GPT-generated discharge instructions were scored more favorably in terms of the interpretability of significance metric, with the most notable difference coming from interpreting the significance of Diagnosis (89.2% GPT vs. 79.5% standard), procedures (74.7% GPT vs. 61.6% standard), treatment (86.7% GPT vs. 65.8% standard), post-ED medications or any changes to medications (63.9% GPT vs. 49.3% standard), and return precautions (86.7% GPT vs. 68.5% standard). However, these differences in ratings between GPT-generated and standard discharge instructions were not statistically significant. (Table 1) Differences in average Likert score as graded on a numerical scale for the interpretability of significance is available in Appendix A5.

Domain 2: Understanding

Understanding was assessed by their agreement with the phrase: “The information in the discharge instruction (subsection) is written such that it is presented in a clear and straightforward manner that is easily comprehensible.” Although not statistically significantly different, there are some differences to note between the discharge instructions authored by GPT compared to the standard discharge instructions from the ED.

Figure 4. This is a clustered stacked multi-bar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions to assess understanding in relation to the prompt: The information in the discharge instruction (subsection) is written such that it is presented in a clear and straightforward manner that is easily comprehensible.

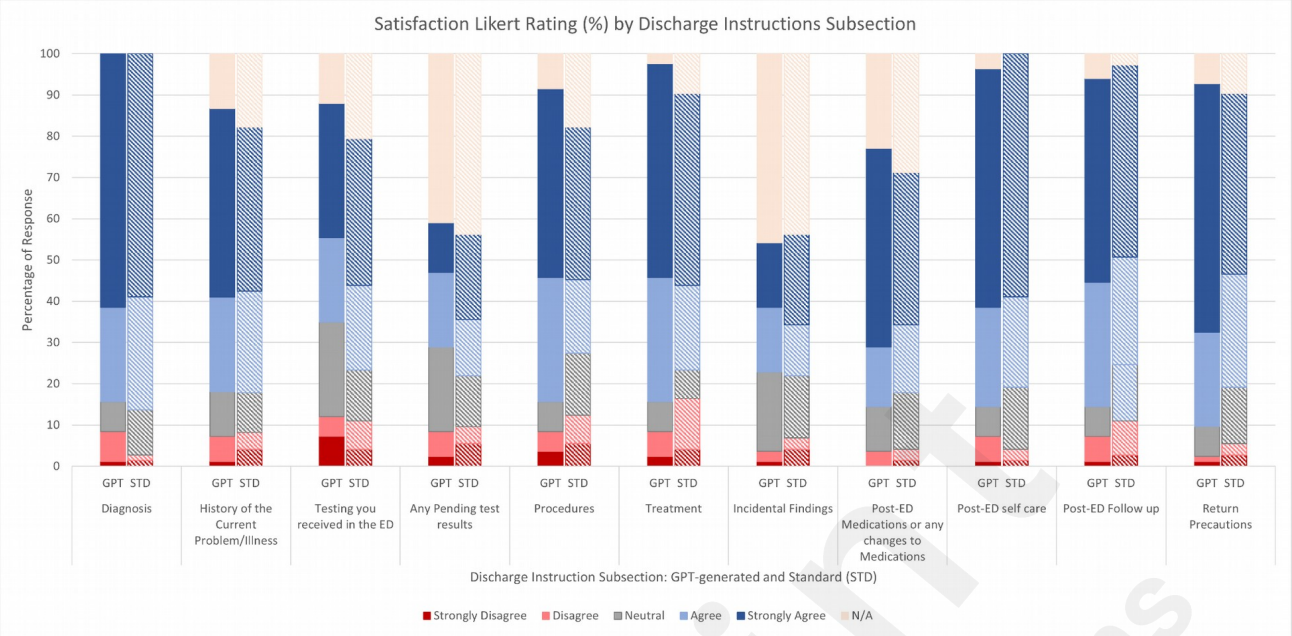


While survey respondents recorded a greater interpretation of significance from diagnosis, their understanding was very similar between GPT (89.2% agree or greatly agree) and standard discharge instructions (87.7%) However, the other subcategories that saw greater percentage of survey respondents rating their understanding of the information in specific discharge instruction subsections favorably came from procedures (80.7% GPT vs. 61.6% standard), treatment (85.5% GPT vs.68.5% standard), post-ED medications or medication changes (68.7% GPT vs. 57.5% standard), post-ED follow-up (86.7% GPT vs. 76.7% standard), and return precautions (85.5% GPT vs. 76.7% standard). (Figure 4) It is important to note, however, that these differences did not achieve statistically significant differences. (Table 1) Differences in average Likert score as graded on a numerical scale for understandability is available in Appendix A6.

Domain 3: Satisfaction

Satisfaction was assessed based on their agreement with the phrase: The information in the discharge instructions (subsection) fulfill your personal expectations of the quality you would expect to receive in an ED setting. Within satisfaction, the GPT-generated discharge instructions once again saw greater frequency of agree and strongly agree across the majority of discharge instruction subsections, signaling positive sentiment towards GPT-generated discharge instructions in comparison to the standard discharge instructions. (Figure 5) A greater percentage of survey respondents rated their satisfaction with GPT-generated discharge instruction subsections most favorably in procedures (75.9% GPT vs.54.8% standard), treatment (85.5% GPT vs.68.5% standard), post-ED medications or medication changes (62.7% GPT vs. 53.4% standard), and return precautions (83.1% GPT vs. 71.2% standard).

Figure 5. This is a clustered stacked multi-bar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions to assess satisfaction of the survey respondent in reference to the statement: The information in the discharge instructions (subsection) fulfill your personal expectations of the quality you would expect to receive in an ED setting.



However, there was no statistically significant difference between the ratings regarding GPT-generated and standard discharge instructions. (Table 1) The sections in which survey respondents were slightly more dissatisfied with GPT-generated discharge instructions were testing received in the ED (53.0% GPT vs. 56.2% standard, rated satisfied or very satisfied), pending test results (30.1% GPT vs. 34.3% standard), and incidental findings (31.3% GPT vs. 34.3% standard). Differences in average Likert score as graded on a numerical scale for satisfaction is available in Appendix A7.

Comparative Analysis

We conducted a Kruskal-Wallis test to compare the Likert results across GPT and standard Discharge Instruction cohorts. Due to the non-parametric, ordinal nature of the data, the Wallis test was best suited to better understand the results.

Table 1. The results of Kruskal-Wallis comparative analysis between GPT and standard discharge instructions across all three metrics of interpretability of significance, understandability, and satisfaction on a traditional Likert scale.

	Significance		Understandability		Satisfaction	
DC Instructions Subsection	Statistic	*P*	Statistic	*P*	Statistic	*P*
Diagnosis	1.620061	0.444844	2.055773	0.357762	0.294928	0.862893
History of the Current Problem/Illness	4.243612	0.119815	3.719087	0.155744	1.421541	0.491266
Testing you received in the ED	1.893903	0.387922	2.885329	0.236297	3.411629	0.181624
Any Pending test results	2.926186	0.231519	2.575909	0.275834	1.928699	0.381231
Procedures	0.319409	0.852396	0.233069	0.889999	0.011596	0.994219

Treatment	0.861691	0.649959	0.07175	0.964761	0.466844	0.791819
Incidental Findings	1.155608	0.561129	2.190465	0.334462	0.843157	0.65601
Post-ED Medications or any changes to Medications	1.335953	0.512745	0.572786	0.750968	0.497961	0.779595
Post-ED self care	3.419674	0.180895	4.535689	0.103535	2.486425	0.288456
Post-ED Follow up	5.860095	0.053395	4.219115	0.121292	2.079271	0.353584
Return Precautions	2.649938	0.265811	0.996082	0.60772	2.885199	0.236313

Notably, no specific subsection of the discharge instructions was rated by respondents to statistically significantly differ between the GPT-generated and standard discharge instructions on a standard Likert scale. (Table 1) The most significant difference was observed in Post-ED Follow up, regarding interpretability of significance and understandability. In both incidences, GPT-generated discharge instructions saw greater frequency of agree and strongly agree, noting greater positive sentiment for the GPT-generated discharge instructions compared to the standard, albeit not to a significant measure.

Discussion

Principal Results

We present a novel assessment of a general patient population's perception of discharge instructions, a key method of communication between patients and physicians. Although the GPT-generated discharge instructions received higher ratings across all three metrics of interpretability of significant information, understandability, and satisfaction in several key subsections of the discharge instructions, the standard discharge instructions were not found to be statistically significantly different from the discharge instructions authored by a large language model (GPT-4).

The similar and even slightly-higher ratings for GPT-generated discharge instructions suggest potential for LLMs to be able to serve as possible adjuncts or interventions in improving discharge instructions or throughput in the ED. The ability of LLMs to generate not only accurate, but detailed and personalized discharge instructions that may improve patients' abilities to interpret the significance of complex medical information through simplification can not only play an important role in reducing documentation burden for the physicians, but also possibly ease patient transitions out of the hospital through augmented and possibly improved forms of documentation and medical information.[32,33]

Despite the lack of statistically significant superiority in the performance of LLM-generated discharge instructions compared to standard methods, the findings suggest that the former are not inferior either. Several subsections of the GPT-generated discharge instructions were even rated favorably in comparison to the standard discharge instructions. Notably, information regarding procedures and treatments received in the ED, as well as important follow-up information including

post-ED medication or medication changes and return precautions had a greater frequency of favorable ratings across all three metrics, interpretability of significance, understandability, and satisfaction, as compared to Standard discharge instructions. This equivalence opens up practical applications for LLMs in the discharge process, particularly in terms of efficiency and reducing clinician workload. EDs often have bottlenecks in patient flow that may benefit from LLM integration to expedite the discharge process. Since LLM-generated instructions can be produced rapidly and tailored to individual patient profiles, they have the potential to decrease the time clinicians spend writing these instructions, there may be value in reducing the discharge-to-departure time.[33,34] Additionally, this may free up clinician time for other high value tasks. For example, incidental findings are often embedded within radiology reports and need to be put into context for patients. LLMs had trending higher scores for incidental findings, post-ED care and follow up which often have to be customized for each patient. Furthermore, the adaptability of LLMs could lead to more dynamic discharge instructions. For instance, electronic discharge instructions could be automatically updated as soon as a clinician finalizes their notes, ensuring that patients receive the most current and relevant information without delay.

Comparison to Prior Work

Prior methods of generating fictional cases, specifically in orthopedics, and synthesizing discharge documentation by ChatGPT in comparison to a standard showed they were comparable in quality as assessed by expert panels.[35] Our study found similar results, with patient-representing survey respondents finding that discharge instructions generated using ED synthetic notes were of similar quality to standard ED discharge instructions. Barak-Corren et al. evaluated pediatric EM attending perceptions regarding summaries regarding completeness, accuracy, efficiency, readability, and overall satisfaction.[14] Our findings represented similar results from the patient perspective, with survey respondents assessing GPT-generated discharge instructions with greater positive sentiment than negative sentiment.

Limitations

This study poses several limitations. The subjects recruited in this study were compensated to answer to their best ability in these surveys. They were asked to take on the role of a patient in an ED, given information regarding a clinical scenario, and to answer as a patient receiving the corresponding discharge instructions. Survey respondents were asked to complete rigorous attention checks and quality control metrics to ensure they properly understood the clinical scenario, but it is possible that survey respondents did not accurately assess the discharge instructions following the attention checks and other quality control metrics. The task presented to survey respondents was not simple and was a complex task.

The generated discharge instructions were improved through an iterative prompt engineering process that may have assessed and improved quality specifically for the ED setting. Due to the use of synthetic notes in this pipeline, the generalizability of these methods may be limited based upon specialty of discharge instructions as well as important variation of discharge instructions from provider to provider. The discharge instructions were also generated from documentation such as the provider H&P, which may not consistently be documentation that is fully finished by the time discharge instructions are written. In addition, the prompt development occurred over the same set of notes and clinical scenarios that were used for the final discharge instructions. These limitations may act as barriers to implementation of the proposed pipeline in current workflows in the ED.

In addition, it is important to note that Amazon MTurk survey respondents are not the same

subpopulation presenting to EDs, only a convenience sample representative of the general population. However, the demographics of Amazon MTurk survey respondents are similar to that of patients presenting to EDs in Connecticut, which is considered to be one of the most representative states of the US.[36] Among Amazon MTurkers, 57% identify as female, with 29.7% between the ages of 18 and 29 y.o, 36.8% between 30-39 y.o., 16.8% 40-49 y.o., 10.7% 50-59 y.o., and 6% 60-69 y.o. This is quite similar to ED usage in Connecticut (CT), with 58% of ED presentations being female, about 41% of ED presenters being between 18-44 y.o. and about 20% in the range of 45-64. Survey respondents slightly over-represent White non-Hispanic by approximately 79.9% compared to ED presentation in CT of 66% white non-hispanic.[37]

Future Directions

Our study explored the initial stages of patient-representing perceptions of using LLM-generated discharge instructions in the ED setting. Due to initial constraints on protected health information, synthetic clinical scenarios, clinician notes, and nursing notes were used in developing Discharge Instructions in this survey format. Future work can focus on the integration of LLMs into existing EHR infrastructure to leverage real patient notes with HIPAA-compliant methods of ChatGPT such as through Doximity GPT or the Azure OpenAI system.

The current scope of our work only looked into five different presenting chief concerns due to the synthetic note constraint. Next steps could look towards expanding the scope of the use of LLMs in discharge instructions, such as integrating multicultural and multilingual personalization into discharge instructions as well as expanding the amount of teaching and patient education embedded into the discharge instructions.[38,39] The iterative process of GPT-4-based discharge instruction development can also be improved with embedding methods of patient feedback, such as structured interviews, as demonstrated in prior research of other forms of discharge instructions could further help develop improved generation of tailored discharge instructions.[40]

Conclusions

The results of this study indicate the promising capabilities for generative LLMs such as GPT-4 in improving methods of healthcare communication. Future research in the use of LLMs in ED workflows can utilize this in real-world applications to assess the perceptions of LLM-generated medical documentation by healthcare staff as well as a possible survey distribution within EDs and real clinical settings that our survey distribution aimed to imitate.

Acknowledgements

This publication was made possible by the Yale School of Medicine Fellowship for Medical Student Research.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

ChatGPT: Chat Generative Pre-Trained Transformer
ED: Emergency Department
EM: Emergency Medicine
GPT-4: Generative Pre-Trained Transformer 4
LLM: Large Language Model

Multimedia Appendix

A1. Repository of Clinical Scenario Clinician and Nursing Notes, GPT-generated and Standard Discharge Instructions
Discharge Instructions Survey Notes

A2. Final ChatGPT Prompts used to generate the final set of discharge instructions.

I am an ED physician and need help writing discharge instructions for my patient, Mr. John Smith. Our EHR system automatically generates basic discharge instructions that cover the standard pertinent material (visit date, medication list, medication changes, procedures performed, incidental findings, f/u instructions etc.); However, I need help writing the "manual" supplemental field, which is the personalized part of the discharge instructions highlighting the most important instructions and information regarding my patients diagnosis and necessary ongoing treatment/follow-up, particularly information that may not be present in the automatically generated discharge instructions.

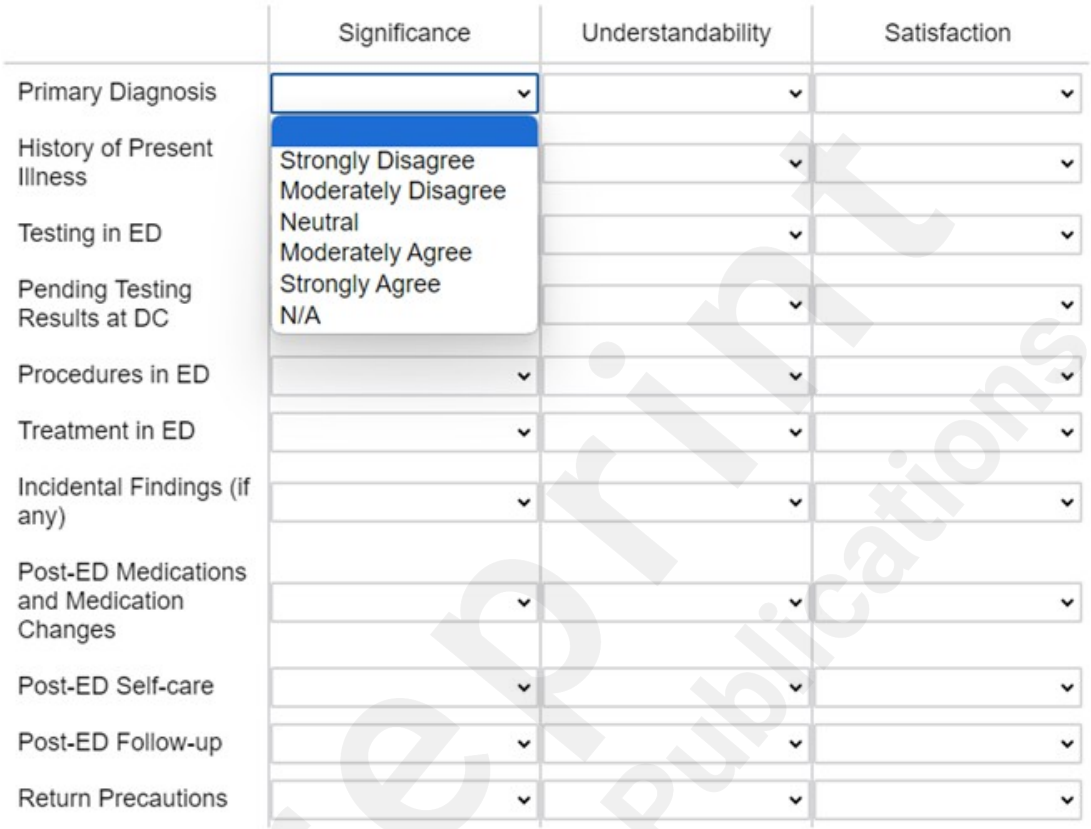
Based on the encounter notes below, please write the manual personalized discharge instruction section to supplement the automatically generated discharge note. Please make sure it is personalized to the patient, interesting to read, and covers pertinent information very succinctly. No need for too much preamble please jump straight into the most pertinent information. Please provide some personalized patient education for any key diagnoses or incidental findings. Make the note readable to a patient with high school education. However, remember you are a physician and keep a formal and serious but informative tone. Be sure to include the following subsections of the discharge instruction, IF the information is available from the encounter notes below:

1. Primary Diagnosis
2. History of Present Illness
3. Testing in the ED (if any)
4. Pending Testing Results upon Discharge (if any)
5. Procedures received in the emergency department (if any)
6. Treatment received in emergency department (if any)
7. Incidental Findings (if any)
8. Post-emergency department medications and medication changes (if any)
These changes include new medications, changes in dosage, or discontinuation
9. Post-emergency department follow-up Instructions/Advice
a. Further investigation if necessary
b. Planned investigations, whether and where these investigations will occur
c. Contact or make appointment with PCP if applicable
10. Return Precautions
a. What red flags to return to the ED for
b. Specify time frame in which patient should return to the ED

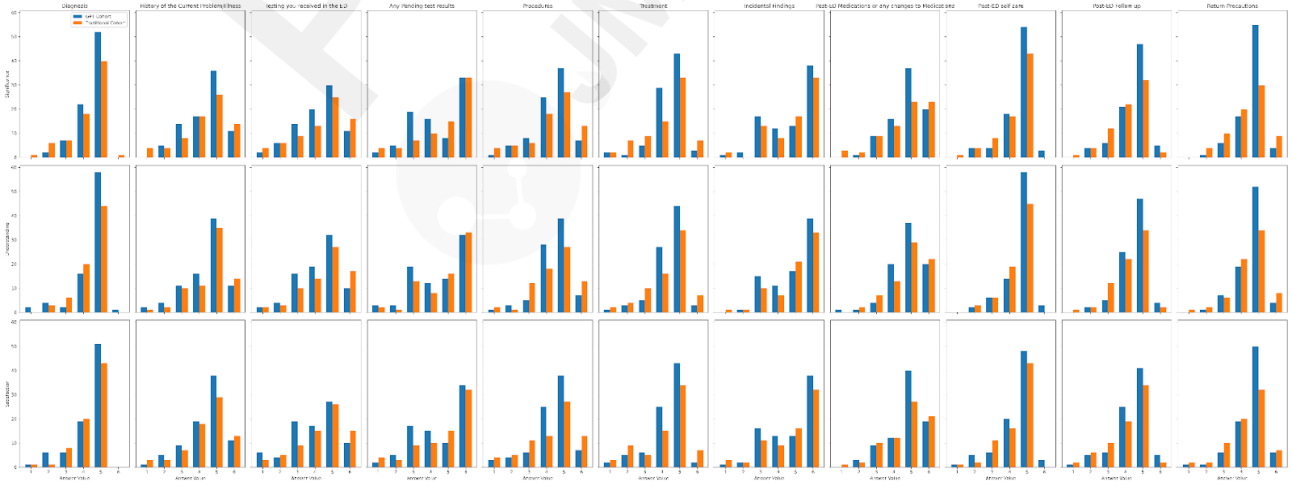
Most importantly, I feel like my patients never finish reading all the important info in the full discharge instructions… To combat this, please make the manual discharge note portion concise and to the point. Use subsections, headers, and bulleted lists for ease of reading. Make sure to highlight key information that my patient might gloss over in the automatically generated discharge instructions. Remember to keep it professional.

A3. Qualtrics Survey Link to the Discharge Instructions Survey Distributed on Amazon MTurk: https://yalesurvey.ca1.qualtrics.com/jfe/form/SV_e9VD90YRhaTKLlk

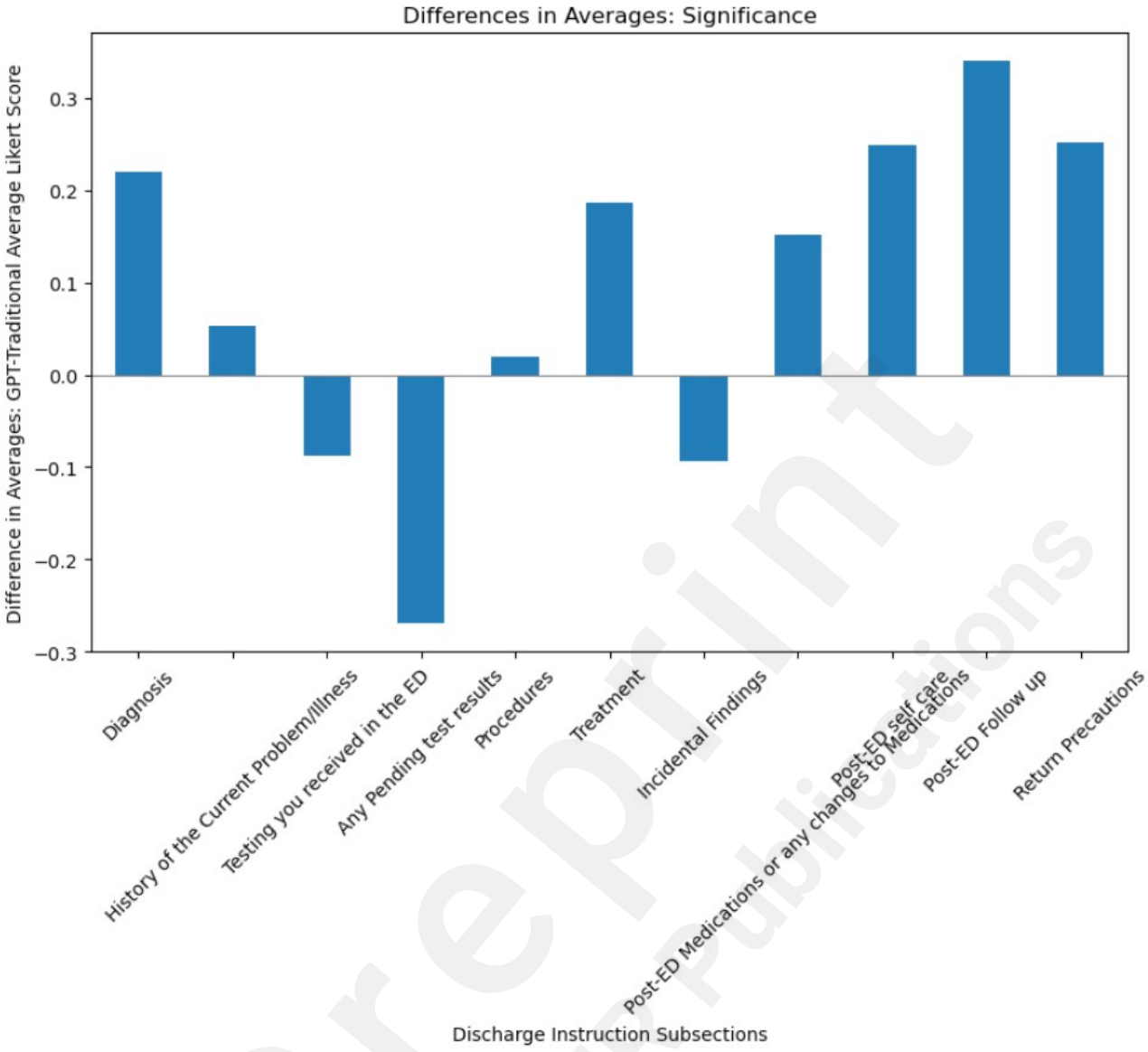
A4. Likert-Style Matrix Survey Respondents were Tasked to Answer in Regards to the Metrics: Interpretability of Significance, Understandability, and Satisfaction for Discharge Instruction Subsection.



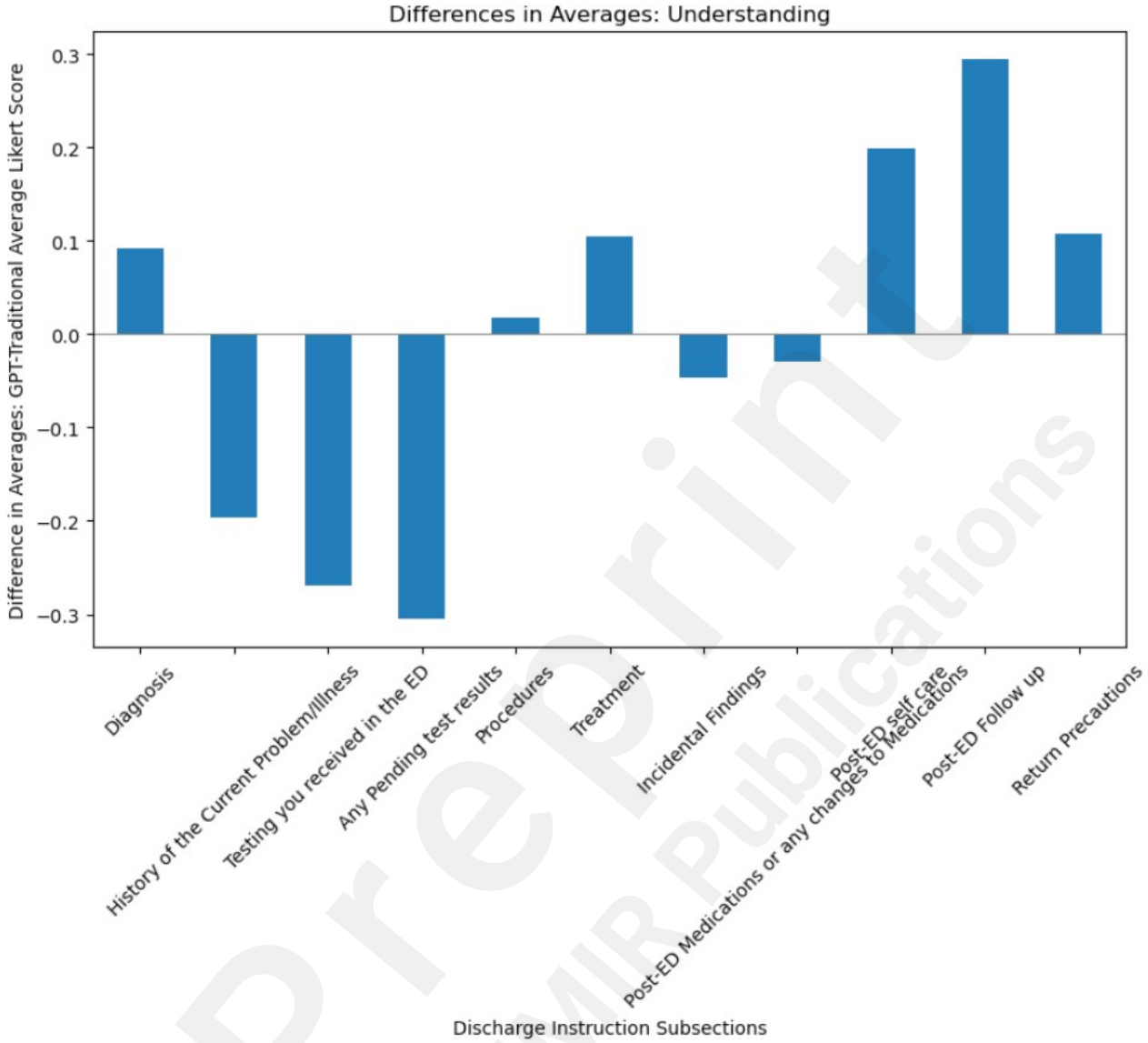
A5. Multiplot of Likert response frequencies (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree, 6 = N/A) by metric (Interpretation of Significance, Ease of Understanding, and Satisfaction) and discharge instruction subset.



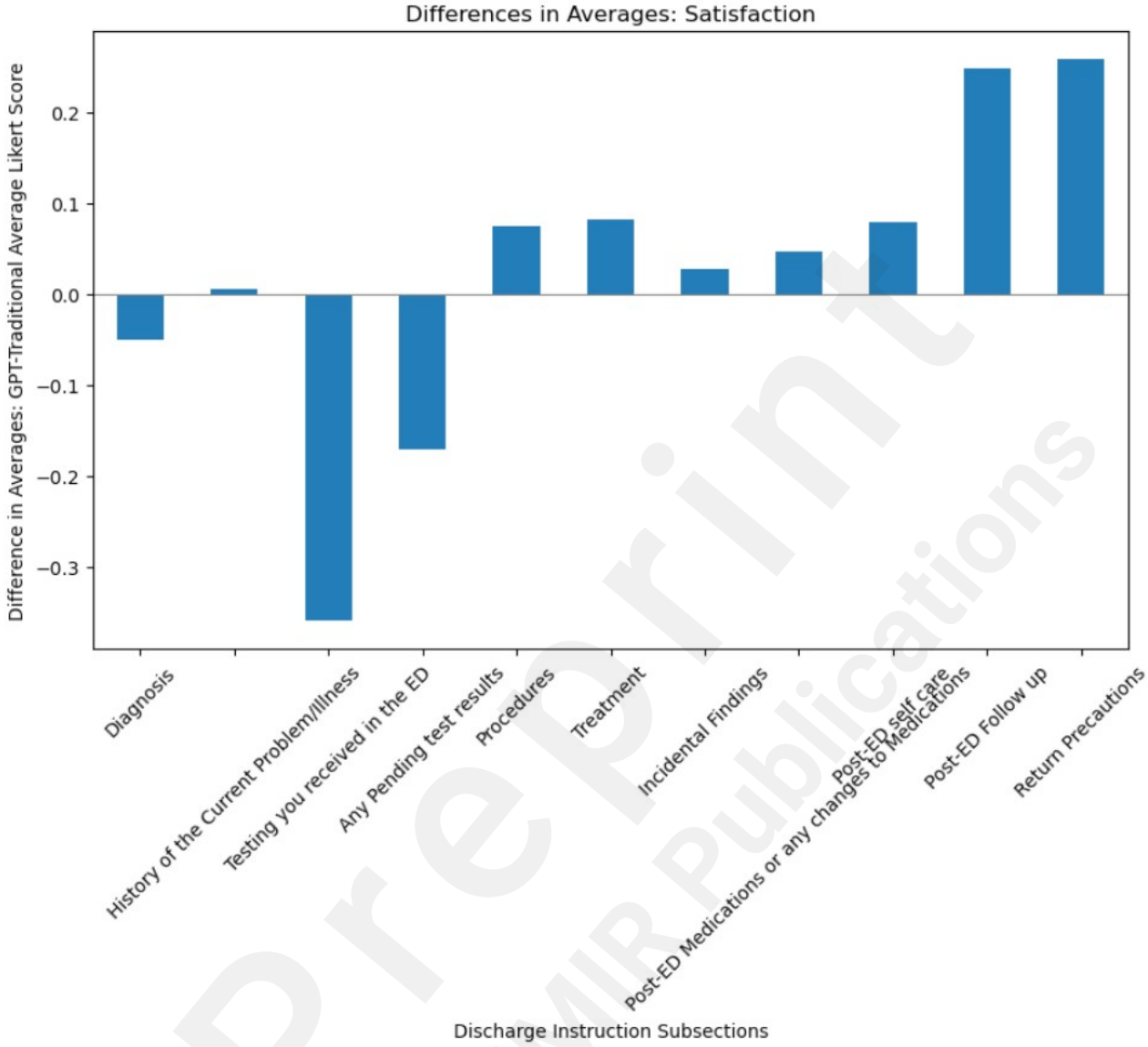
A6. Differences in averages of Likert scores between GPT and standard discharge instructions for evaluation of Interpretability of Significance by discharge instruction subsection.



A7. Differences in averages of Likert scores between GPT and standard discharge instructions for evaluation of Ease of Understanding by discharge instruction subsection.



A8. Differences in averages of Likert scores between GPT and standard discharge instructions for evaluation of Satisfaction by discharge instruction subsection.



References

1. Joint Commission on Accreditation of Healthcare Organizations. *Hospital Accreditation Standards: Standards, Intents : HAS*. The Commission; 2006.

2. Discharge instructions. Default. Accessed April 18, 2024. https://www.saem.org/about-saem/academies-interest-groups-affiliates2/cdem/for-students/online-education/m3-curriculum/disposition/discharge-instructions

3. Engel KG, Heisler M, Smith DM, Robinson CH, Forman JH, Ubel PA. Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand? *Ann Emerg Med*. 2009;53(4):454-461.e15.

4. Buckley BA, McCarthy DM, Forth VE, et al. Patient input into the development and enhancement of ED discharge instructions: a focus group study. *J Emerg Nurs*. 2013;39(6):553-561.

5. DeSai C, Janowiak K, Secheli B, et al. Empowering patients: simplifying discharge instructions. *BMJ Open Qual*. 2021;10(3). doi:10.1136/bmjoq-2021-001419

6. Burns ST, Amobi N, Chen JV, O'Brien M, Haber LA. Readability of Patient Discharge Instructions. *J Gen Intern Med*. 2022;37(7):1797-1798.

7. Choudhry AJ, Baghdadi YMK, Wagie AE, et al. Readability of discharge summaries: with what level of information are we dismissing our patients? *Am J Surg*. 2016;211(3):631-636.

8. Unaka NI, Statile A, Haney J, Beck AF, Brady PW, Jerardi KE. Assessment of readability, understandability, and completeness of pediatric hospital medicine discharge instructions. *J Hosp Med*. 2017;12(2):98-101.

9. Clarke C, Friedman SM, Shi K, Arenovich T, Monzon J, Culligan C. Emergency department discharge instructions comprehension and compliance study. *CJEM*. 2005;7(1):5-11.

10. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. *J Gen Intern Med*. 2012;27(5):520-526.

11. Hibbard JH, Mahoney ER, Stock R, Tusler M. Do increases in patient activation result in improved self-management behaviors? *Health Serv Res*. 2007;42(4):1443-1463.

12. Becker C, Zumbrunn S, Beck K, et al. Interventions to Improve Communication at Hospital Discharge and Rates of Readmission: A Systematic Review and Meta-analysis. *JAMA Netw Open*. 2021;4(8):e2119346.

13. Doyle SK, Rippey JC, Jacques A, et al. Effect of personalised, mobile-accessible discharge instructions for patients leaving the emergency department: A randomised controlled trial. *Emerg Med Australas*. 2020;32(6):967-973.

14. Barak-Corren Y, Wolf R, Rozenblum R, et al. Harnessing the Power of Generative AI for Clinical Summaries: Perspectives From Emergency Physicians. *Ann Emerg Med*. Published online March 12, 2024. doi:10.1016/j.annemergmed.2024.01.039

15. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595.

16. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. 2023;9:e45312.

17. Nayak A, Alkaitis MS, Nayak K, Nikolov M, Weinfurt KP, Schulman K. Comparison of History of Present Illness Summaries Generated by a Chatbot and Senior Internal Medicine Residents. *JAMA Intern Med*. 2023;183(9):1026-1027.

18. Eppler MB, Ganjavi C, Knudsen JE, et al. Bridging the Gap Between Urological Research and Patient Understanding: The Role of Large Language Models in Automated Generation of Layperson's Summaries. *Urol Pract*. 2023;10(5):436-443.

19. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596.

20. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):16492.

21. Zaretsky J, Kim JM, Baskharoun S, et al. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw Open*. 2024;7(3):e240357.

22. Hartman VC, Bapat SS, Weiner MG, Navi BB, Sholle ET, Campion TR Jr. A method to automate the discharge summary hospital course for neurology patients. *J Am Med Inform Assoc*. 2023;30(12):1995-2003.

23. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. *Radiology*. 2023;309(2):e232561.

24. Singh S, Djalilian A, Ali MJ. ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes. *Semin Ophthalmol*. 2023;38(5):503-507.

25. Blokdyk G. *Amazon Mechanical Turk: Projects-Ready*. Createspace Independent Publishing Platform; 2018.

26. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC Health Serv Res*. 2022;22(1):1148.

27. Safeer RS, Keenan J. Health literacy: the gap between physicians and patients. *Am Fam Physician*. 2005;72(3):463-468.

28. Basics of prompting. Accessed February 29, 2024. https://www.promptingguide.ai/introduction/basics

29. Berinsky AJ, Margolis MF, Sances MW. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *Am J Pol Sci*. 2014;58(3):739-753.

30. Mellis AM, Bickel WK. Mechanical Turk data collection in addiction research: utility, concerns and best practices. *Addiction*. 2020;115(10):1960-1968.

31. Hauser DJ, Schwarz N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res Methods*. 2016;48(1):400-407.

32. Abdullahi T, Singh R, Eickhoff C. Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models. *JMIR Med Educ*. 2024;10:e51391.

33. Nguyen J, Pepping CA. The application of ChatGPT in healthcare progress notes: A commentary from a clinical and research perspective. *Clin Transl Med*. 2023;13(7):e1324.

34. Shahab O, El Kurdi B, Shaukat A, Nadkarni G, Soroush A. Large language models: a primer and gastroenterology applications. *Therap Adv Gastroenterol*. 2024;17:17562848241227031.

35. Rosenberg GS, Magnéli M, Barle N, et al. ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases. *Acta Orthop*. 2024;95:152-156.

36. Kolko J. “normal America” is not A small town of white people. FiveThirtyEight. Published April 28, 2016. Accessed April 18, 2024. https://fivethirtyeight.com/features/normal-america-is-not-a-small-town-of-white-people/

37. Moss A. Demographics of people on Amazon mechanical Turk. CloudResearch. Published June 12, 2020. Accessed April 18, 2024. https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/

38. Bang Y, Cahyawijaya S, Lee N, et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv [csCL]*. Published online February 8, 2023. http://arxiv.org/abs/2302.04023

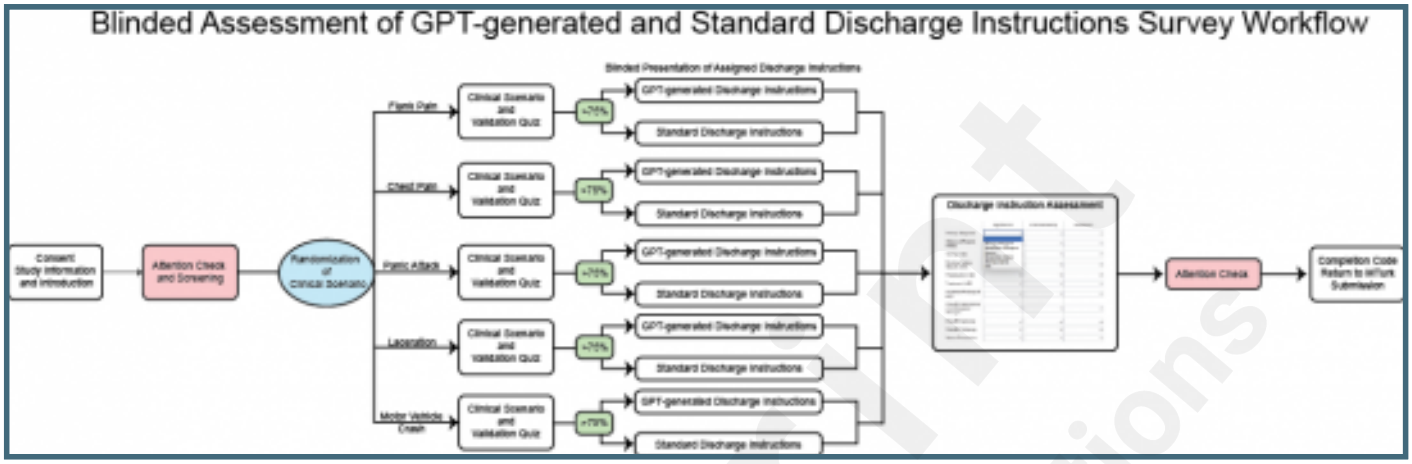
39. Nazir A, Wang Z. A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. *Meta Radiol*. 2023;1(2). doi:10.1016/j.metrad.2023.100022

40. Horstman MJ, Mills WL, Herman LI, et al. Patient experience with discharge instructions in postdischarge recovery: a qualitative study. *BMJ Open*. 2017;7(2):e014842.

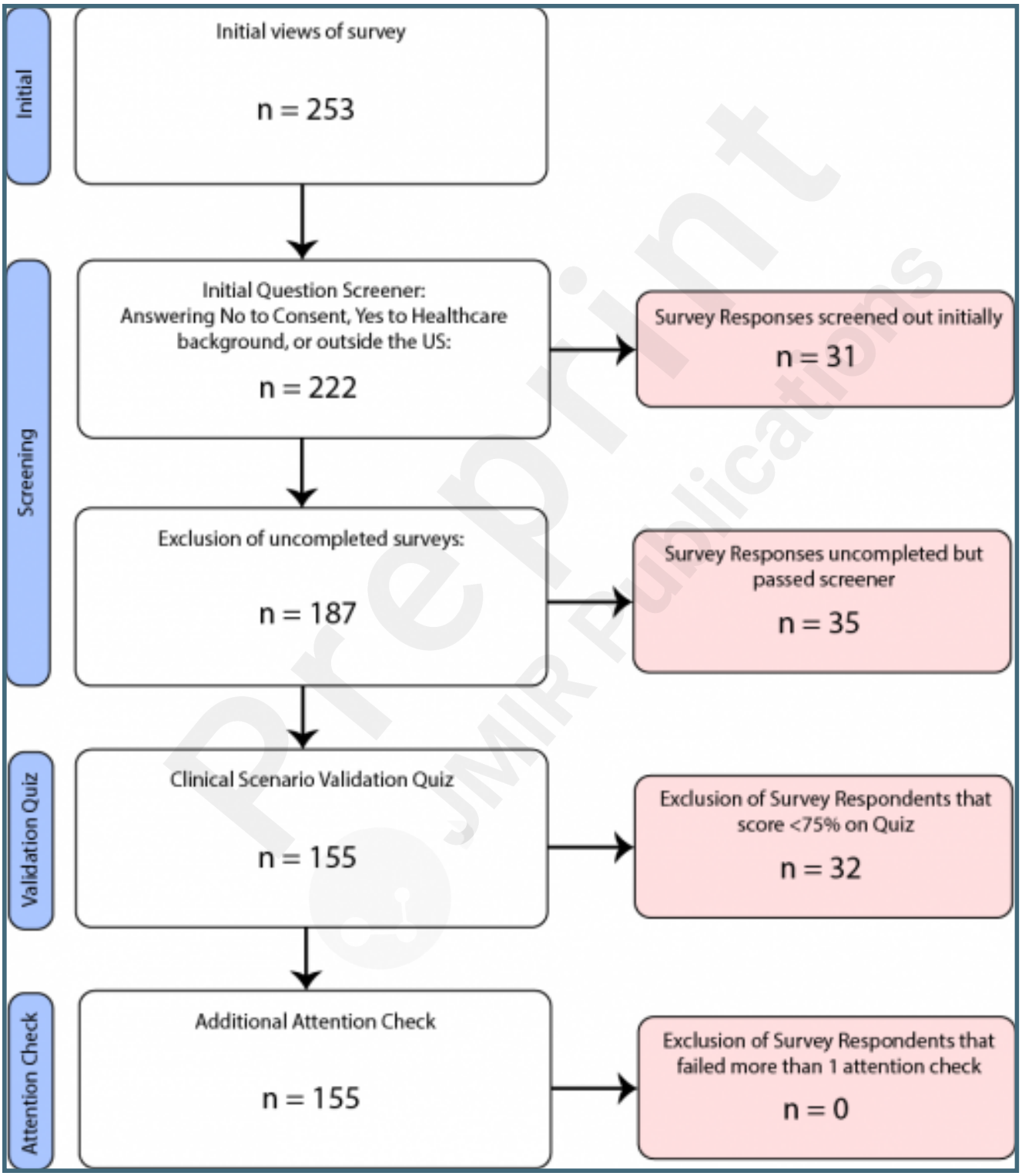
Supplementary Files

Figures

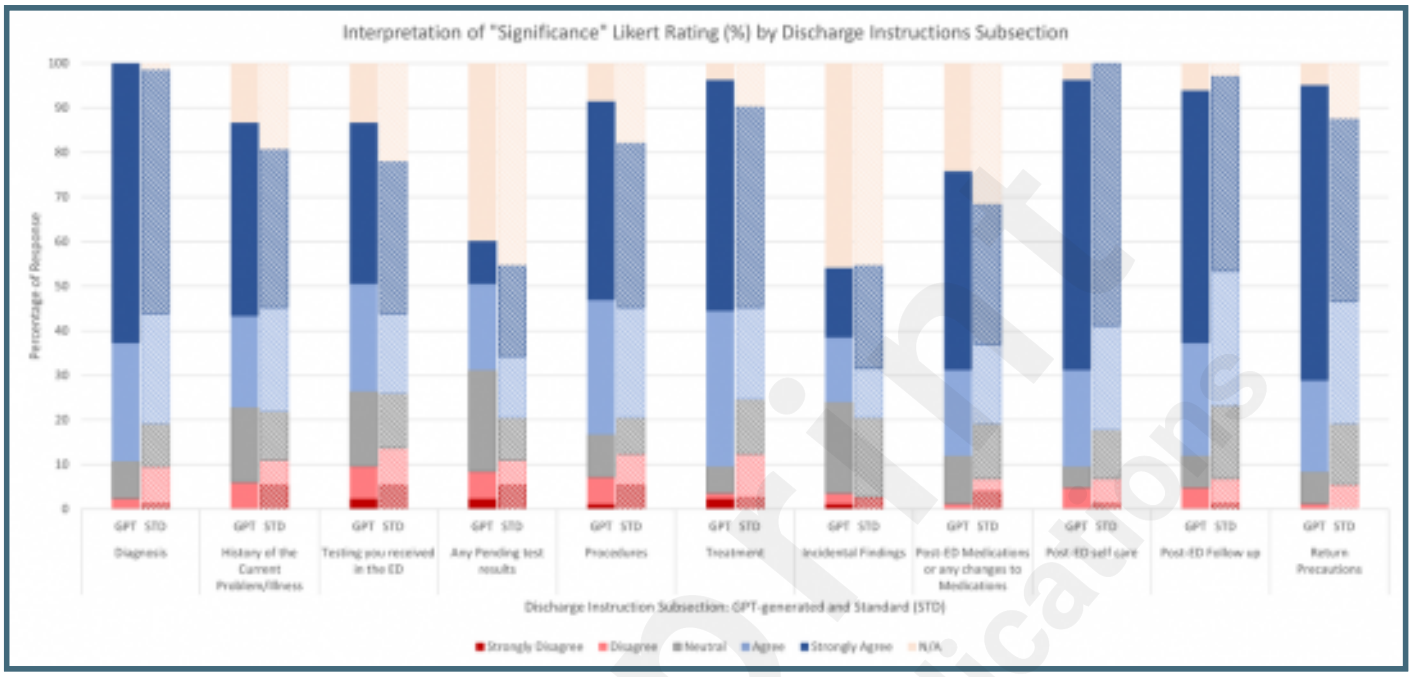
The Qualtrics survey design for discharge instruction randomized blind assessment by Amazon MTurk respondents. Survey respondents first passed an initial consent documentation, screener for healthcare or medical background, and a series of attention checks. Respondents were then randomized to one possible clinical scenario of five, then randomized to view either the GPT-generated or standard version of discharge instructions. This respondent was then tasked to answer Likert-scale questions regarding the three metrics: interpretability of significance, understandability, and satisfaction in regards to each discharge instructions subsection before one final attention check and conclusion of the survey.



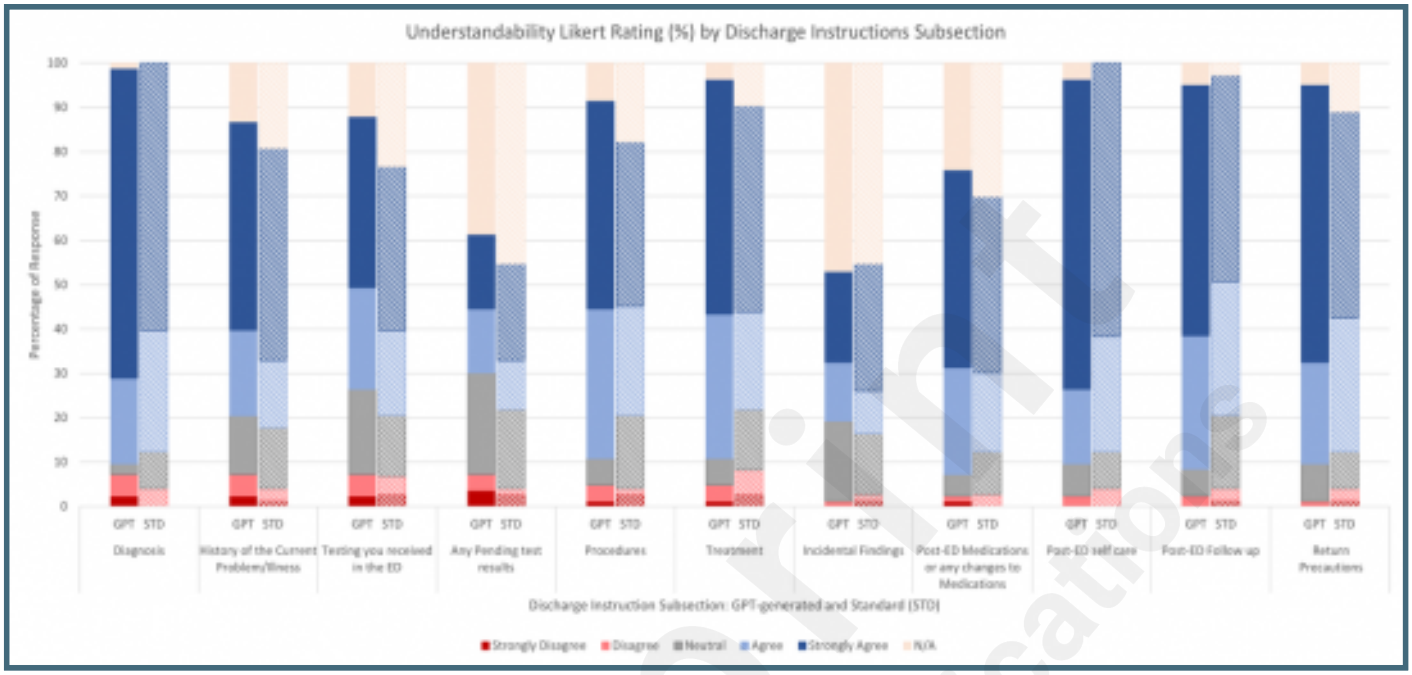
The stepwise method towards exclusion of survey respondents. From an initial 253 views of the survey, 222 were not screened out by the initial screen of consent, not having a healthcare background, and residing in the US. Following the initial screener, 35 did not finish the entire survey. Surveys also had their validation quiz in relation to the clinical scenario, and attention checks graded. Of 187 completed surveys, 155 passed the validation quiz, and all respondents that passed the validation quiz also successfully answered all other attention checks correctly.



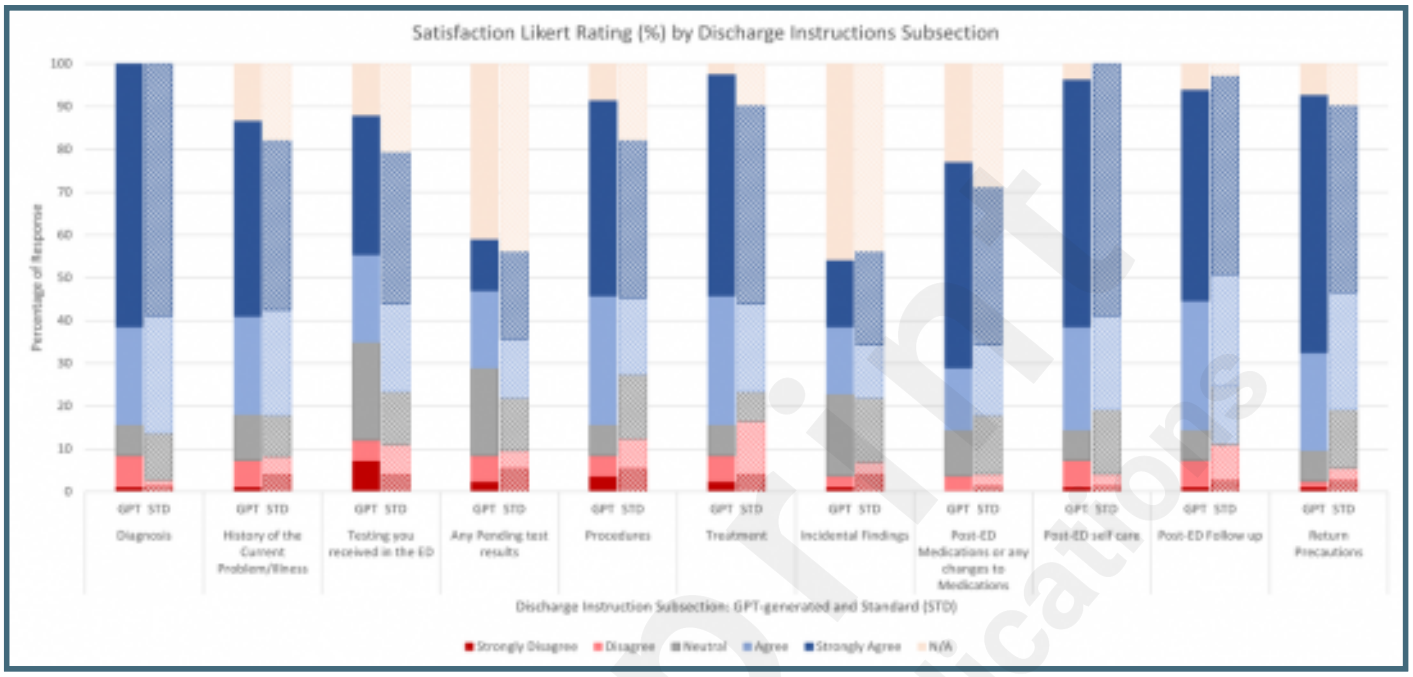
This is a clustered stacked multi-bar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions in relation to the prompt: “The information in the discharge instruction (subsection) effectively explains the significance of the findings in a way that’s personalized to me (the hypothetical recipient) and is easy to follow".



This is a clustered stacked multi-bar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions to assess understanding in relation to the prompt: The information in the discharge instruction (subsection) is written such that it is presented in a clear and straightforward manner that is easily comprehensible.



This is a clustered stacked multi-bar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions to assess satisfaction of the survey respondent in reference to the statement: The information in the discharge instructions (subsection) fulfill your personal expectations of the quality you would expect to receive in an ED setting.



Multimedia Appendixes

Likert-Style Matrix Survey Respondents were Tasked to Answer in Regards to the Metrics: Interpretability of Significance, Understandability, and Satisfaction for Discharge Instruction Subsection.
URL: http://asset.jmir.pub/assets/6d1c7d4f48b200771a9ef5cf587ab2b6.png

Multiplot of Likert response frequencies (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree, 6 = N/A) by metric (Interpretation of Significance, Ease of Understanding, and Satisfaction) and discharge instruction subset.
URL: http://asset.jmir.pub/assets/ab9da696907700c3b1b7d0f9c3de95c7.png

Differences in averages of Likert scores between GPT and standard discharge instructions for evaluation of Interpretability of Significance by discharge instruction subsection.
URL: http://asset.jmir.pub/assets/77faf433fb689abbf8c60f226e779dad.png

Differences in averages of Likert scores between GPT and standard discharge instructions for evaluation of Ease of Understanding by discharge instruction subsection.
URL: http://asset.jmir.pub/assets/9b71f1c01250229a29874f160bc4e5f1.png

Differences in averages of Likert scores between GPT and standard discharge instructions for evaluation of Satisfaction by discharge instruction subsection.
URL: http://asset.jmir.pub/assets/fe5069528473dcb65aed83a9954de617.png