# Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the Emergency Department

Christopher Y.K. Williams (MB BChir)[1*], Brenda Y. Miao (BA)[1], Atul J. Butte (MD, PhD)[1]

[1]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA

*Corresponding author:

Dr Christopher Y.K. Williams

Postdoctoral Scholar; Bakar Computational Health Sciences Institute, UCSF

cykw2@doctors.org.uk

Word count: 2773 words

## Abstract

The release of GPT-3.5-turbo (ChatGPT) and other large language models (LLMs) has the potential to transform healthcare. However, existing research evaluating LLM performance on real-world clinical notes is limited. Here, we conduct a highly-powered study to determine whether GPT-3.5-turbo can provide clinical recommendations for three tasks (admission status, radiological investigation(s) request status, and antibiotic prescription status) using clinical notes from the Emergency Department. We randomly select 10,000 Emergency Department visits to evaluate the accuracy of zero-shot, GPT-3.5-turbo-generated clinical recommendations across four different prompting strategies. We find that GPT-3.5-turbo performs poorly compared to a resident physician, with accuracy scores 24% lower on average. GPT-3.5-turbo tended to be overly cautious in its recommendations, with high sensitivity at the cost of specificity. Our findings demonstrate that, while early evaluations of the clinical use of LLMs are promising, LLM performance must be significantly improved before their deployment as decision support systems for clinical recommendations and other complex tasks.

## Introduction

Since its November 2022 launch, the Chat Generative Pre-Trained Transformer (ChatGPT; GPT-3.5-turbo) has captured widespread public attention, with media reports suggesting over 100 million monthly active users just two months after launch.[1] Along with its successor, GPT-4, these large language models (LLMs) use a chat-based interface to respond to complex queries and solve problems.[2,3] Although trained as general-purpose models, researchers have begun evaluating the performance of GPT-3.5-turbo and GPT-4 on clinically-relevant tasks. For instance, GPT-3.5-turbo was found to provide largely appropriate responses when asked to give simple cardiovascular disease prevention recommendations.[4] Meanwhile, GPT-3.5-turbo responses to patients' health questions on a public social media forum were both preferred, and rated as having higher empathy, compared to physician responses.[5]

While there are a growing number of studies that explore the uses of the GPT models across a range of clinical tasks, the majority do not use real-world clinical notes. They instead apply these models to answer questions from medical examinations such as the USMLE, solve publicly available clinical diagnostic challenges such as the *New England Journal of Medicine* (NEJM) clinicopathologic conferences, or evaluate performance on existing clinical benchmarks.[3,6–9] This is due to the challenges associated with disclosing protected health information (PHI) with LLM providers such as OpenAI in a Health Insurance Portability and Accountability Act (HIPAA) compliant manner, where business associate agreements must be in place to allow secure processing of PHI content.[10] This is a notable hurdle given the inherent differences between curated medical datasets, such as the USMLE question bank, and real-world clinical notes. In addition, this issue is particularly problematic when you

consider that the GPT models have likely been trained on data obtained from open sources on the Internet and therefore their evaluation on existing publicly available benchmarks or tasks may be confounded by data leakage.[11]

As the availability and accessibility of these models increases, it is now critically important to better understand the potential uses and limitations of LLMs applied to actual clinical notes. In our previous work, we showed that GPT-3.5-turbo could accurately identify the higher acuity patient when provided only the clinical histories of pairs of patients presenting to the Emergency Department.[12] This was despite a lack of additional training or fine-tuning, known as zero-shot learning.[13] Elsewhere, Kanjee and colleagues evaluated the diagnostic ability of GPT-4 across 70 cases from the NEJM clinicopathologic conferences, obtaining a correct diagnosis in its differential in 64% of cases and as its top diagnosis in 39%.[7] However, the ability of these general-purpose large language models to assimilate clinical information from de-identified clinical notes and return clinical recommendations is still unclear.

In this study, we sought to evaluate the zero-shot performance of GPT-3.5-turbo when prompted to provide clinical recommendations for patients evaluated in the Emergency Department. We focus on three recommendations in particular: 1) Should the patient be admitted to hospital; 2) Should the patient have radiological investigations requested; and 3) Should the patient receive antibiotics? We first evaluate performance on balanced (i.e equal numbers of positive and negative outcomes) datasets, to examine the sensitivity and specificity of GPT recommendations, before determining overall model accuracy on an unbalanced dataset that reflects real-world distributions of patients presenting to the Emergency Department.

## Results

From a total of 251,401 adult Emergency Department visits, we first created balanced samples of 10,000 ED visits for each of the three tasks (Figure 1). Using only the information provided in the *Presenting History* and *Physical Examination* sections of patients' first ED physician note, we queried GPT-3.5-turbo to determine whether 1) the patient should be admitted to hospital, 2) the patient requires radiological investigation(s), and 3) the patient requires antibiotics, comparing its output to the ground-truth outcome extracted from the electronic health record.

Across all three clinical recommendation tasks, overall GPT-3.5-turbo performance was poor (Table 1). The initial prompt of 'Please return whether the patient should be admitted to hospital / requires radiological investigation / requires antibiotics' (Prompt A) led to high sensitivity and low specificity performance. For this prompt, GPT-3.5-turbo recommendations had a high true positive rate but similarly high false positive rate, with GPT-3.5-turbo recommending admission / radiological investigation / antibiotic prescription for the majority of cases. Altering the prompt to 'only suggest … if absolutely required' (Prompt B) only marginally improved specificity. The greatest performance was achieved by removing restrictions on the verbosity of GPT-3.5-turbo response (Prompt C) and adding the 'Let's think step by step' chain-of-thought prompting (Prompt D). These prompts generated the highest specificity in GPT-3.5-turbo recommendations with limited effect on sensitivity.

To compare this performance with that of a resident physician, we took a balanced n = 200 subsample for manual annotation and compared performance between physician and machine across each of the four prompt iterations (Table 2). Notably, physician sensitivity was below

that of GPT-3.5-turbo responses (0.73 vs [range: 0.93-1.00], 0.76 vs [range: 0.93-0.96] and 0.64 vs [range: 0.89-0.93] for admission, radiological investigation, and antibiotic prescription tasks, respectively), but specificity was significantly higher than GPT-3.5-turbo (0.74 vs [range: 0.07-0.40], 0.79 vs [range: 0.09-0.17] and 0.78 vs [range: 0.26-0.37]).

We next sought to test the performance of GPT-3.5-turbo in a more representative setting using an unbalanced, n = 1000 sample of ED visits that reflects the real-world distribution of admission, radiological investigation, and antibiotic prescription rates at our institution (Table 3). We found that the accuracy of resident physician recommendations, when evaluated against the ground-truth outcomes extracted from the electronic health record, was significantly higher than GPT-3.5-turbo recommendations: 0.83 for physician vs [range: 0.29-0.53 for GPT-3.5-turbo], 0.79 vs [range: 0.68-0.71] and 0.78 vs [range: 0.35-0.43] for admission, radiological investigation, and antibiotic prescription tasks, respectively (Figure 2; Table 3).

Lastly, in our sensitivity analyses conducted on a balanced, n = 200 subsample for each task, results were largely similar regardless of the written order of labels in the original prompt (e.g ‘0: Patient should be admitted to hospital. 1: Patient should not be admitted to hospital.’ vs ‘1: Patient should be admitted to hospital. 0: Patient should not be admitted to hospital.’) (Table S3). Reversing the order of labels in the original prompt led to almost identical results for all tasks except the antibiotic prescription task, where specificity was improved for Prompts 2-4, but at the cost of sensitivity.

## Discussion

This study represents an early, highly powered evaluation of the potential uses and limitations of GPT-3.5-turbo for generating clinical recommendations based on real-world clinical text. Across three different clinical recommendation tasks, we found that GPT-3.5-turbo performed poorly, with high sensitivity but low specificity across tasks. Model performance was marginally improved with iterations of prompt engineering, including the addition of zero-shot chain-of-thought prompting.[14] On evaluation of an unbalanced (n = 1000) sample reflective of the real-world distribution of clinical recommendations, GPT-3.5-turbo performance was significantly worse than that of a resident physician, with 24% lower accuracy averaged across tasks.

Our results suggest that GPT-3.5-turbo is overly cautious in its clinical recommendations – it exhibits a tendency to recommend intervention for each of the three tasks and this leads to a notable number of false positive suggestions. Such a finding is problematic given the need to both prioritize hospital resource availability and reduce overall healthcare costs.[15,16] This is also true at the patient level, where there is an increasing appreciation that excessive investigation and/or treatment may cause patients harm.[16] It is unclear, however, what is the best balance of sensitivity/specificity to strive for amongst clinical large language models – it is likely that this balance will differ based on the particular task. The increase in GPT-3.5-turbo specificity, at the expense of sensitivity, across our iterations of prompt engineering suggests that improvements could be made bespoke to the task, though the extent to which prompt engineering alone may improve performance is unclear.

Across all three tasks, overall performance remained notably below that of a human physician. This may reflect the inherent complexity of clinical decision making, where

clinical recommendations may be influenced not only by the patient's intrinsic clinical status, but also by patient preference, current resource availability and other external factors.

Before large language models can be integrated within the clinical environment, it is important to fully understand both their capabilities and limitations. Otherwise, there is a risk of unintended harmful consequences, especially if models have been deployed at scale.[17,18] Current research deploying large language models, particularly the current state-of-the-art GPT models, on real-world clinical text is limited. Recent work from our group has demonstrated accurate performance of GPT-3.5-turbo in both assessing patient clinical acuity in the Emergency Department and extracting detailed oncologic history and treatment plans from medical oncology notes.[19] Elsewhere, GPT-3.5-turbo has been used to convert radiology reports into plain language, to classify whether statements of clinical recommendations in scientific literature constitute health advice, and to accurately classify five diseases from discharge summaries in the MIMIC-III dataset.[20–22] Much of the current literature focuses on the strengths of large language models such as GPT-3.5-turbo and GPT-4.[3,9,12,19] However, it is equally important to identify areas of medicine in which LLMs do not perform well. For example, in one evaluation of GPT-4's ability to diagnose dementia from a set of structured features, GPT-4 did not surpass the performance of traditional AI tools, while fewer than 20% of GPT-3.5-turbo and GPT-4 responses submitted to a clinical informatics consult service were found to be concordant with existing reports.[23,24] While early signs of the utility of large language models in medicine are promising, our findings suggest that there remains significant room for improvement, especially in more challenging tasks such as complex clinical decision making.

This study has several limitations. Firstly, it is possible that, for each task, not all the information which led to the real-life clinical recommendation extracted from the electronic health record was present in the *Presenting History* and *Physical Examination* sections of the ED physician note. For instance, radiological investigations requested following the Emergency Medicine physician review may lead to unexpected and/or incidental findings which were not detected during the initial review and may warrant admission or antibiotic prescription. However, even with this limitation, physician classification performance remained a very respectable 78-83% accuracy across the three tasks, suggesting it is challenging, but not impossible, to make accurate clinical recommendations based on the available clinical text. Secondly, we only trialled three iterations of prompt engineering, in addition to our initial prompt, and this was done in a zero-shot manner. Further attempts to refine the provided prompt, or incorporate few-shot examples for in-context learning, may improve model performance.[13,25–27] Lastly, this study did not evaluate the performance of the recently released, more advanced GPT-4 model. It is possible that GPT-4 performance may surpass that of GPT-3.5-turbo in these more complex reasoning tasks, though the ability to test this at a similar scale is limited by the increased costs associated with GPT-4 usage across a sample of this size. Similarly, evaluation of the performance of other natural language processing models, such as a fine-tuned BioClinicalBERT model or bag-of-word-based and other simpler techniques, has not been performed.[28] It is possible that these more traditional NLP models, which are typically trained or fine-tuned on a large training set of data, may outperform the zero-shot performance of GPT-like large language models.[21]

## Methods

The UCSF Information Commons contains deidentified structured clinical data as well as deidentified clinical text notes, deidentified and externally certified as previously described.[29] The UCSF Institutional Review Board determined that this use of the deidentified data within the UCSF Information Commons is not human participants research and therefore was exempt from further approval and informed consent.

We identified all adult visits to the University of California San Francisco (UCSF) Emergency Department (ED) from 2012 to 2023 with an ED Physician note present within Information Commons (Figure 1). Regular expressions were used to extract the *Presenting History* (consisting of 'Chief Complaint', 'History of Presenting Illness' and 'Review of Systems') and *Physical Examination* sections from each note (Supplementary File 1).

We sought to evaluate GPT-3.5-turbo performance on three binary clinical recommendation tasks, corresponding to the following outcomes: 1) Admission status – whether the patient should be admitted from ED to hospital. 2) Radiological investigation(s) request status – whether an X-ray, US scan, CT scan, or MRI scan should be requested during the ED visit. 3) Antibiotic prescription status – whether antibiotics should be ordered during the ED visit.

For each of the three outcomes, we randomly selected a balanced sample of 10,000 ED visits to evaluate GPT-3.5-turbo performance (Figure 1). Using its secure, HIPAA-compliant Application Programming Interface (API) through Microsoft Azure, we provided GPT-3.5-turbo (model *gpt-3.5-turbo-0301*) the *Presenting History* and *Physical Examination* sections of the ED Physician's note for each ED visit and queried it to determine if 1) the patient should be admitted to hospital, 2) the patient requires radiological investigation, and 3) the

patient should be prescribed antibiotics. GPT-3.5-turbo performance was evaluated against the ground-truth outcome extracted from the electronic health record. Separately, a resident blinded to both the GPT-3.5-turbo labels and ground-truth labels reviewed a balanced n = 200 subsample for each of the three tasks to allow a comparison of human and machine performance. The following evaluation metrics were calculated: true positive rate, true negative rate, false positive rate, false negative rate, sensitivity and specificity.

We subsequently experimented with three iterations of prompt engineering (Table S1, Supplementary File 1) to test if modifications to the initial prompt could improve GPT-3.5-turbo performance. Chain-of-thought (CoT) prompting is a method found to improve the ability of large language models to perform complex reasoning by decomposing multi-step problems into a series of intermediate steps.[25] This can be done in a zero-shot manner (zero-shot-CoT), with large language models shown to be decent zero-shot reasoners by adding a simple prompt, 'Let's think step by step' to facilitate step-by-step reasoning before answering each question.[14] Alternatively, few-shot chain-of-thought prompting can be used, with additional examples of prompt and answer pairs either manually (manual CoT) or computationally (e.g auto-CoT) provided and concatenated with the prompt of interest.[25,26] Current understanding of the impact of zero-shot-CoT, manual CoT, and auto-CoT prompt engineering techniques applied to clinical text is limited. In this work, we sought to focus on zero-shot-CoT and investigate the effect of adding 'Let's think step by step' to the prompt on model performance.

Our *initial prompt* (Prompt A) simply asked GPT-3.5-turbo to return whether the patient should be e.g. admitted to hospital, without any additional explanation. We additionally attempted to engineer prompts to a) reduce the high false positive rate of GPT-3.5-turbo

recommendations (Prompt B) and b) examine whether zero-shot chain-of-thought prompting could improve GPT-3.5-turbo performance (Prompts C and D). Attempting to reduce the high GPT-3.5-turbo false positive rate, Prompt B was constructed by adding an additional sentence to Prompt A: 'Only suggest **clinical recommendation** if absolutely required'. This modification was kept for Prompts C and D, which were constructed to examine chain-of-thought prompting. Because chain-of-thought prompting is most effective when the LLM provides reasoning in its output, we removed the instruction 'Please do not return any additional explanation' from Prompts C and D, and added the chain-of-thought prompt 'Let's think step by step' to Prompt D, increasing GPT-3.5-turbo response verbosity (Table S2, Supplementary File 1). Prompt C therefore served as a baseline for comparison of GPT-3.5-turbo performance when it is permitted to return additional explanation (in addition to its outcome recommendation), allowing comparisons with both Prompt A (where no additional explanations were allowed in the prompt) and Prompt D (where the effect of chain-of-thought prompting was examined).

To evaluate the performance of GPT-3.5-turbo in a real-world setting, we constructed a random, unbalanced sample of 1000 ED visits where the distribution of patient outcomes (i.e. admission status, radiological investigation(s) request status and antibiotic prescription status) mirrored the distributions of patients presenting to ED from our main cohort. The *Presenting History* and *Physical Examination* sections of the ED Physician's note for each ED visit were again passed to the GPT-3.5-turbo API in an identical manner to the balanced datasets, while a resident physician manually labelled the entire sample to allow human vs machine comparison. Classification accuracy was calculated in addition to the aforementioned evaluation metrics utilised for the balanced datasets to provide a summative evaluation metric for this real-world simulated task.

*Sensitivity analysis*

Due to the stochastic nature of large language models, it is possible that the order of labels reported in the original prompt may affect the subsequent labels returned. To test this, we conducted a sensitivity analysis on a balanced n = 200 subsample for each outcome where the positive outcome was referenced before the negative outcome in the initial prompt (e.g. '1: Patient should be admitted to hospital' precedes '0: Patient should not be admitted to hospital' in the GPT-3.5-turbo prompt).

## Figures

**Figure 1.** Flowchart of included Emergency Department visits and construction of both balanced (n = 10,000 samples) and unbalanced (n = 1000 sample reflecting the real-world distribution of patients presenting to the Emergency Department) datasets for the following outcomes: 1) Admission status, 2) Radiological investigation(s) status, and 3) Antibiotic prescription status.

**Figure 2.** Evaluation of physician and GPT-3.5-turbo accuracy across four iterations of prompt engineering [Prompt A-D] evaluated on an unbalanced n = 1000 sample reflective of the real-world distribution of clinical recommendations among patients presenting to ED, for the following three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics.

## Tables

**Table 1.** GPT-3.5-turbo performance across four iterations of prompt engineering (Prompt A-D) evaluated on a balanced n = 10,000 sample for three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics.

**Table 2.** Comparison of physician and GPT-3.5-turbo performance across four iterations of prompt engineering [Prompt A-D] evaluated on a balanced n = 200 subsample for three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics. *Physicians were provided the same prompt text as in Prompt A.

**Table 3.** Comparison of physician and GPT-3.5-turbo performance across four iterations of prompt engineering [Prompt A-D] evaluated on an unbalanced n = 1000 sample reflective of the real-world distribution of clinical recommendations among patients presenting to ED, for the following three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics. *Physicians were provided the same prompt text as in Prompt A.

## Tables

| **Task** | | **True positives, n (%)** | **False positives, n (%)** | **True negatives, n (%)** | **False Negatives, n (%)** | **Sensitivity** | **Specificity** |
|---|---|---|---|---|---|---|---|
| 1) Admission status | Prompt A | 4994 (49.9) | 4639 (46.4) | 361 (3.6) | 6 (0.1) | **1.00** | 0.07 |
| | Prompt B | 4904 (49) | 3527 (35.3) | 1473 (14.7) | 96 (1) | 0.98 | 0.29 |
| | Prompt C | 4683 (46.8) | 3255 (32.6) | 1745 (17.5) | 317 (3.2) | 0.94 | 0.35 |
| | Prompt D | 4617 (46.2) | 3165 (31.7) | 1835 (18.4) | 383 (3.8) | 0.92 | **0.37** |
| 2) Radiological investigation(s) request status | Prompt A | 4922 (49.2) | 4361 (43.6) | 639 (6.4) | 78 (0.8) | **0.98** | 0.13 |
| | Prompt B | 4805 (48.1) | 3906 (39.1) | 1094 (10.9) | 195 (2) | 0.96 | 0.22 |
| | Prompt C | 4792 (47.9) | 3855 (38.6) | 1145 (11.5) | 208 (2.1) | 0.96 | **0.23** |
| | Prompt D | 4819 (48.2) | 3991 (39.9) | 1009 (10.1) | 181 (1.8) | 0.96 | 0.20 |
| 3) Antibiotic prescription status | Prompt A | 4812 (48.1) | 3955 (39.6) | 1045 (10.5) | 188 (1.9) | **0.96** | 0.21 |
| | Prompt B | 4690 (46.9) | 3687 (36.9) | 1313 (13.1) | 310 (3.1) | 0.94 | 0.26 |
| | Prompt C | 4658 (46.6) | 3639 (36.4) | 1361 (13.6) | 342 (3.4) | 0.93 | **0.27** |
| | Prompt D | 4544 (45.4) | 3379 (33.8) | 1621 (16.2) | 456 (4.6) | 0.91 | 0.32 |

**Table 1.** GPT-3.5-turbo performance across four iterations of prompt engineering (Prompt A-D) evaluated on a balanced n = 10,000 sample for three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics.

| **Task** | | **True positives, n (%)** | **False positives, n (%)** | **True negatives, n (%)** | **False Negatives, n (%)** | **Sensitivity** | **Specificity** |
|---|---|---|---|---|---|---|---|
| 1) Admission status | *Physician* | *73 (36.5)* | *26 (13)* | *74 (37)* | *27 (13.5)* | *0.73* | ***0.74*** |
| | Prompt A | 100 (50) | 93 (46.5) | 7 (3.5) | 0 (0) | **1.00** | 0.07 |
| | Prompt B | 98 (49) | 67 (33.5) | 33 (16.5) | 2 (1) | 0.98 | 0.33 |
| | Prompt C | 95 (47.5) | 61 (30.5) | 39 (19.5) | 5 (2.5) | 0.95 | 0.39 |
| | Prompt D | 93 (46.5) | 60 (30) | 40 (20) | 7 (3.5) | 0.93 | 0.40 |
| 2) Radiological investigation(s) request status | *Physician* | *76 (38)* | *21 (10.5)* | *79 (39.5)* | *24 (12)* | *0.76* | ***0.79*** |
| | Prompt A | 96 (48) | 91 (45.5) | 9 (4.5) | 4 (2) | **0.96** | 0.09 |
| | Prompt B | 93 (46.5) | 83 (41.5) | 17 (8.5) | 7 (3.5) | 0.93 | 0.17 |
| | Prompt C | 95 (47.5) | 83 (41.5) | 17 (8.5) | 5 (2.5) | 0.95 | 0.17 |
| | Prompt D | 95 (47.5) | 84 (42) | 16 (8) | 5 (2.5) | 0.95 | 0.16 |
| 3) Antibiotic prescription status | *Physician* | *64 (32)* | *22 (11)* | *78 (39)* | *36 (18)* | *0.64* | ***0.78*** |
| | Prompt A | 93 (46.5) | 74 (37) | 26 (13) | 7 (3.5) | **0.93** | 0.26 |
| | Prompt B | 91 (45.5) | 71 (35.5) | 29 (14.5) | 9 (4.5) | 0.91 | 0.29 |
| | Prompt C | 92 (46) | 68 (34) | 32 (16) | 8 (4) | 0.92 | 0.32 |
| | Prompt D | 89 (44.5) | 63 (31.5) | 37 (18.5) | 11 (5.5) | 0.89 | 0.37 |

**Table 2.** Comparison of physician and GPT-3.5-turbo performance across four iterations of prompt engineering [Prompt A-D] evaluated on a balanced n = 200 subsample for three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics. *Physicians were provided the same prompt text as in Prompt A.

| **Task** | | **True positives, n (%)** | **False positives, n (%)** | **True negatives, n (%)** | **False Negatives, n (%)** | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|---|---|---|---|---|
| 1) Admission status | *Physician* | *151 (15.1)* | *79 (7.9)* | *683 (68.3)* | *87 (8.7)* | *0.63* | ***0.90*** | ***0.83*** |
| | Prompt A | 237 (23.7) | 714 (71.4) | 48 (4.8) | 1 (0.1) | **1.00** | 0.06 | 0.29 |
| | Prompt B | 234 (23.4) | 514 (51.4) | 248 (24.8) | 4 (0.4) | 0.98 | 0.33 | 0.48 |
| | Prompt C | 232 (23.2) | 475 (47.5) | 287 (28.7) | 6 (0.6) | 0.97 | 0.38 | 0.52 |
| | Prompt D | 226 (22.6) | 463 (46.3) | 299 (29.9) | 12 (1.2) | 0.95 | 0.39 | 0.53 |
| 2) Radiological investigation(s) request status | *Physician* | *527 (52.7)* | *109 (10.9)* | *261 (26.1)* | *103 (10.3)* | *0.84* | ***0.71*** | ***0.79*** |
| | Prompt A | 619 (61.9) | 314 (31.4) | 56 (5.6) | 11 (1.1) | **0.98** | 0.15 | 0.68 |
| | Prompt B | 604 (60.4) | 274 (27.4) | 96 (9.6) | 26 (2.6) | 0.96 | 0.26 | 0.70 |
| | Prompt C | 604 (60.4) | 268 (26.8) | 102 (10.2) | 26 (2.6) | 0.96 | 0.28 | 0.71 |
| | Prompt D | 608 (60.8) | 276 (27.6) | 94 (9.4) | 22 (2.2) | 0.97 | 0.25 | 0.70 |
| 3) Antibiotic prescription status | *Physician* | *96 (9.6)* | *142 (14.2)* | *686 (68.6)* | *76 (7.6)* | *0.56* | ***0.83*** | ***0.78*** |
| | Prompt A | 162 (16.2) | 642 (64.2) | 186 (18.6) | 10 (1) | **0.94** | 0.22 | 0.35 |
| | Prompt B | 159 (15.9) | 594 (59.4) | 234 (23.4) | 13 (1.3) | 0.92 | 0.28 | 0.39 |
| | Prompt C | 158 (15.8) | 596 (59.6) | 232 (23.2) | 14 (1.4) | 0.92 | 0.28 | 0.39 |
| | Prompt D | 155 (15.5) | 552 (55.2) | 276 (27.6) | 17 (1.7) | 0.90 | 0.33 | 0.43 |

**Table 3.** Comparison of physician and GPT-3.5-turbo performance across four iterations of prompt engineering [Prompt A-D] evaluated on an unbalanced n = 1000 sample reflective of the real-world distribution of clinical recommendations among patients presenting to ED, for the following three clinical recommendation tasks: 1) Should the patient be admitted to hospital; 2) Does the patient require radiological investigation; and 3) Does the patient require antibiotics.

## References

1. Hu K, Hu K. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Published February 2, 2023. Accessed August 7, 2023.

2. GPT-4. Accessed August 7, 2023. https://openai.com/gpt-4

3. OpenAI. GPT-4 Technical Report. Published online March 27, 2023. doi:10.48550/arXiv.2303.08774

4. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*. 2023;329(10):842-844. doi:10.1001/jama.2023.1044

5. 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. doi:10.1001/jamainternmed.2023.1838

6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

7. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA*. 2023;330(1):78-80. doi:10.1001/jama.2023.8288

8. Singhal K, Tu T, Gottweis J, et al. Towards Expert-Level Medical Question Answering with Large Language Models. Published online May 16, 2023. doi:10.48550/arXiv.2305.09617

9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. Published online April 12, 2023. doi:10.48550/arXiv.2303.13375

10. Kanter GP, Packel EA. Health Care Privacy Risks of AI Chatbots. *JAMA*. 2023;330(4):311-312. doi:10.1001/jama.2023.9618

11. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

12. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Butte AJ. Assessing clinical acuity in the Emergency Department using the GPT-3.5 Artificial Intelligence Model. Published online August 13, 2023:2023.08.09.23293795. doi:10.1101/2023.08.09.23293795

13. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv*. 2023;55(9):195:1-195:35. doi:10.1145/3560815

14. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. Published online January 29, 2023. doi:10.48550/arXiv.2205.11916

15. Barasa EW, Molyneux S, English M, Cleary S. Setting healthcare priorities in hospitals: a review of empirical studies. *Health Policy Plan*. 2015;30(3):386-396. doi:10.1093/heapol/czu010

16. Latifi N, Redberg RF, Grady D. The Next Frontier of Less Is More—From Description to Implementation. *JAMA Intern Med*. 2022;182(2):103-105. doi:10.1001/jamainternmed.2021.6908

17. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med*. 2021;181(8):1065-1070. doi:10.1001/jamainternmed.2021.2626

18. Habib AR, Lin AL, Grant RW. The Epic Sepsis Model Falls Short—The Importance of External Validation. *JAMA Intern Med*. 2021;181(8):1040-1041. doi:10.1001/jamainternmed.2021.3333

19. Sushil M, Kennedy VE, Miao BY, Mandair D, Zack T, Butte AJ. Extracting detailed oncologic history and treatment plan from medical oncology notes with large language models. Published online August 7, 2023. doi:10.48550/arXiv.2308.03853

20. 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. Published online March 28, 2023. doi:10.48550/arXiv.2303.09038

21. Chen S, Li Y, Lu S, et al. Evaluation of ChatGPT Family of Models for Biomedical Reasoning and Classification. Published online April 5, 2023. doi:10.48550/arXiv.2304.02496

22. Zhang J, Sun K, Jagadeesh A, et al. The Potential and Pitfalls of using a Large Language Model such as ChatGPT or GPT-4 as a Clinical Assistant. Published online July 16, 2023. doi:10.48550/arXiv.2307.08152

23. Wang Z, Li R, Dong B, et al. Can LLMs like GPT-4 outperform traditional AI tools in dementia diagnosis? Maybe, but not today. Published online June 2, 2023. doi:10.48550/arXiv.2306.01499

24. Dash D, Thapa R, Banda JM, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. Published online April 30, 2023. doi:10.48550/arXiv.2304.13714

25. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv.org. Published January 28, 2022. Accessed August 7, 2023. https://arxiv.org/abs/2201.11903v6

26. Zhang Z, Zhang A, Li M, Smola A. Automatic Chain of Thought Prompting in Large Language Models. Published online October 7, 2022. doi:10.48550/arXiv.2210.03493

27. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. Published online July 22, 2020. doi:10.48550/arXiv.2005.14165

28. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. arXiv.org. Published April 6, 2019. Accessed May 13, 2023. https://arxiv.org/abs/1904.03323v3

29. Radhakrishnan L, Schenk G, Muenzen K, et al. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open*. 2023;6(3):ooad045. doi:10.1093/jamiaopen/ooad045

## Acknowledgements

The authors acknowledge the use of the UCSF Information Commons computational research platform, developed and supported by UCSF Bakar Computational Health Sciences Institute. The authors also thank the UCSF AI Tiger Team, Academic Research Services, Research Information Technology, and the Chancellor's Task Force for Generative AI for their software development, analytical and technical support related to the use of Versa API gateway (the UCSF secure implementation of large language models and generative AI via API gateway), Versa chat (the chat user interface), and related data asset and services.

## Data availability

The code accompanying this manuscript is available at https://github.com/cykwilliams/GPT-3.5-Clinical-Recommendations-in-Emergency-Department/.

## Conflicts of Interest

CYKW has no conflicts of interest to disclose. BYM is a paid consultant for SandboxAQ. AJB is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. AJB receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. AJB's research has been funded by NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. None of these entities had any bearing on the design of this study or the writing of the manuscript.

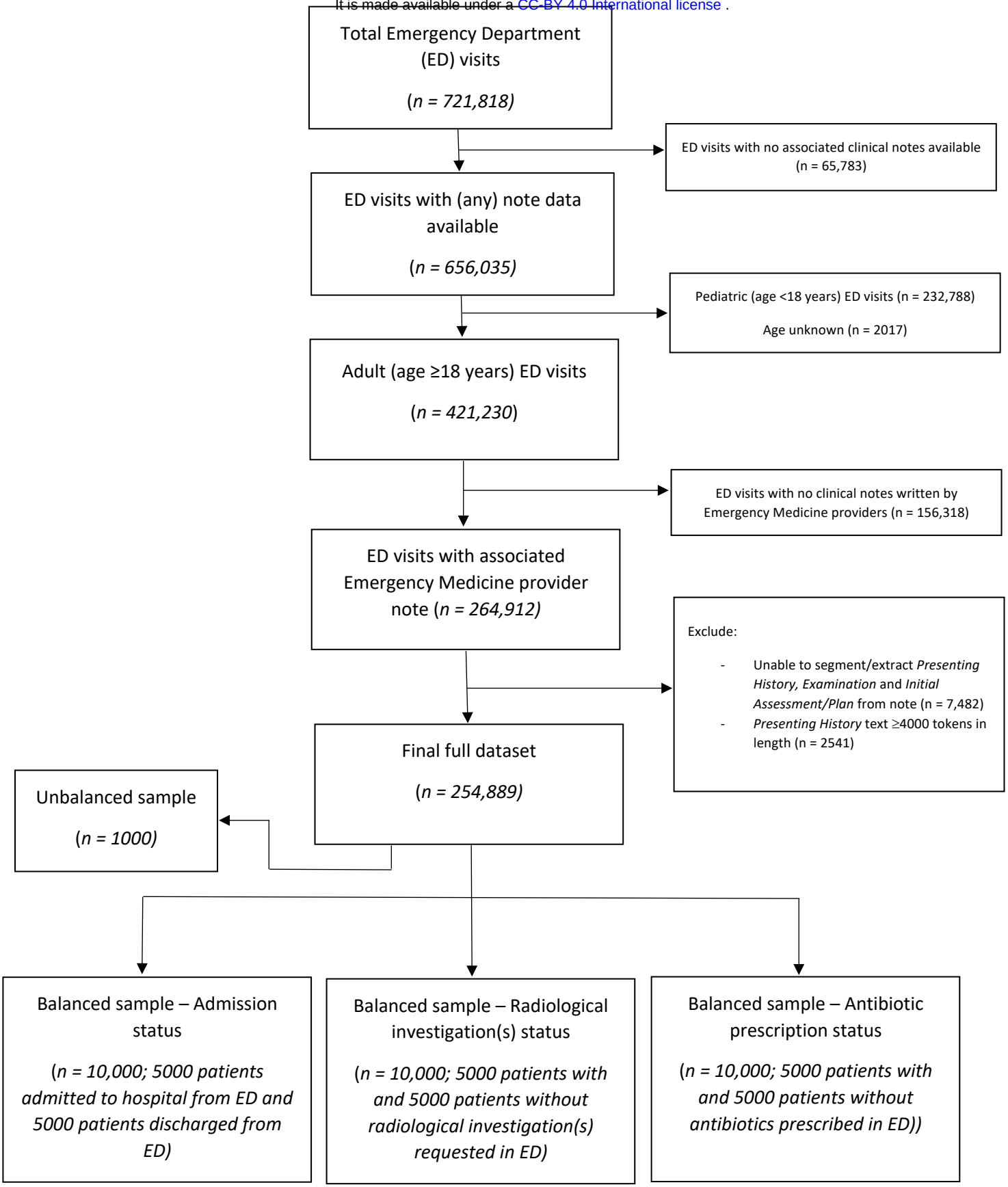
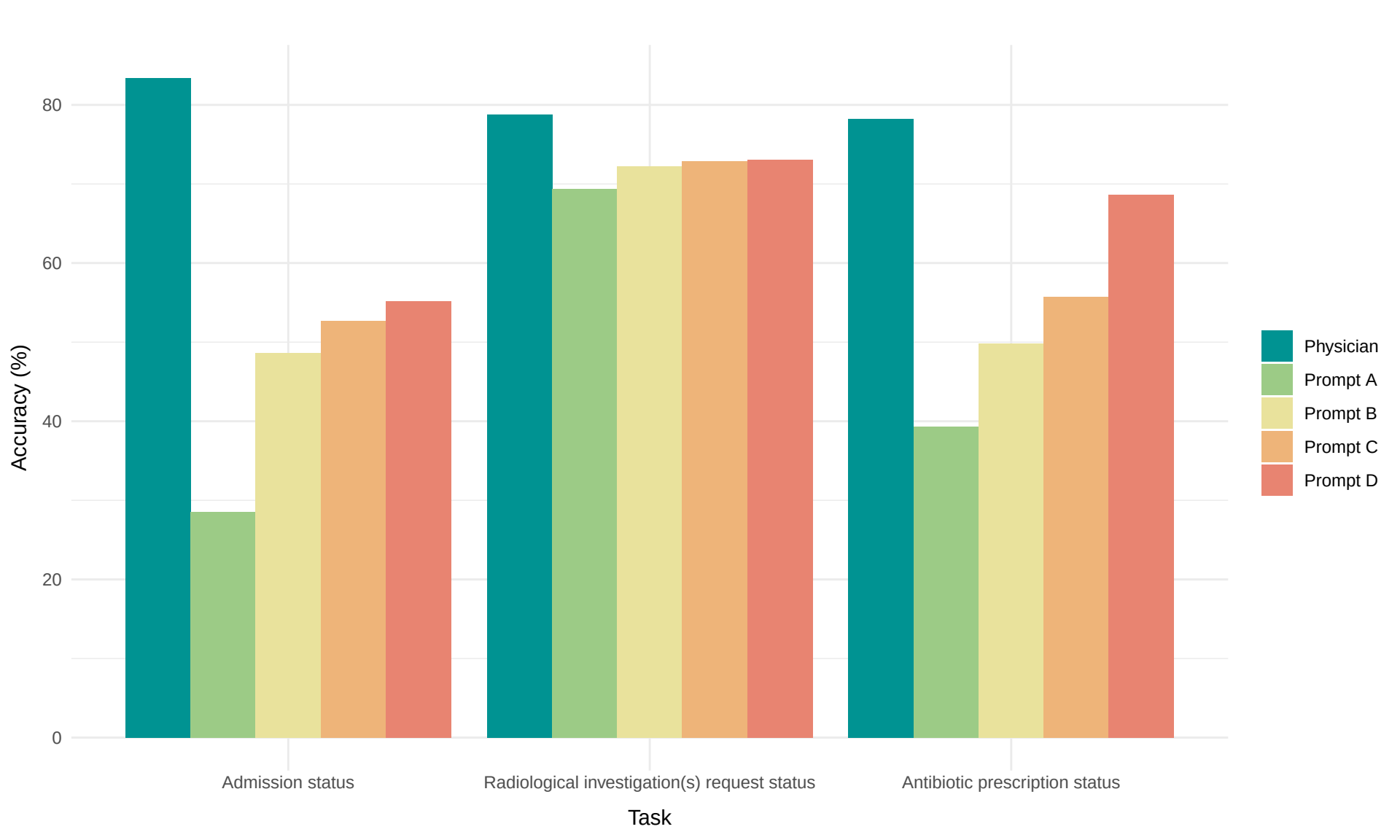


**Figure 1**. Flowchart of included Emergency Department visits and construction of both balanced (n = 10,000 samples) and unbalanced (n = 1000 sample reflecting the real-world distribution of patients presenting to the Emergency Department) datasets for the following outcomes: 1) Admission status, 2) Radiological investigation(s) status, and 3) Antibiotic prescription status


80
60
40
20
0
Accuracy (%)
Admission status
Radiological investigation(s) request status
Antibiotic prescription status
Task
Physician
Prompt A
Prompt B
Prompt C
Prompt D