

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Evaluating Large Language Models for Drafting Emergency Department Discharge Summaries

Christopher Y.K. Williams^{1*}, Jaskaran Bains², Tianyu Tang², Kishan Patel², Alexa N. Lucas²,
Fiona Chen², Brenda Y. Miao¹, Atul J. Butte¹, Aaron E. Kornblith^{1,2}

¹Bakar Computational Health Sciences Institute; University of California, San Francisco
²Department of Emergency Medicine; University of California, San Francisco

*Corresponding author:
Dr Christopher Y.K. Williams
Postdoctoral Scholar; Bakar Computational Health Sciences Institute, UCSF
cykw2@doctors.org.uk

Word count: 2951 words

30

Abstract

31 **Importance:** Large language models (LLMs) possess a range of capabilities which may be
32 applied to the clinical domain, including text summarization. As ambient artificial intelligence
33 scribes and other LLM-based tools begin to be deployed within healthcare settings, rigorous
34 evaluations of the accuracy of these technologies are urgently needed.

35 **Objective:** To investigate the performance of GPT-4 and GPT-3.5-turbo in generating
36 Emergency Department (ED) discharge summaries and evaluate the prevalence and type of
37 errors across each section of the discharge summary.

38 **Design:** Cross-sectional study.

39 **Setting:** University of California, San Francisco ED.

40 **Participants:** We identified all adult ED visits from 2012 to 2023 with an ED clinician note. We
41 randomly selected a sample of 100 ED visits for GPT-summarization.

42 **Exposure:** We investigate the potential of two state-of-the-art LLMs, GPT-4 and GPT-3.5-turbo,
43 to summarize the full ED clinician note into a discharge summary.

44 **Main Outcomes and Measures:** GPT-3.5-turbo and GPT-4-generated discharge summaries
45 were evaluated by two independent Emergency Medicine physician reviewers across three
46 evaluation criteria: 1) Inaccuracy of GPT-summarized information; 2) Hallucination of
47 information; 3) Omission of relevant clinical information. On identifying each error, reviewers
48 were additionally asked to provide a brief explanation for their reasoning, which was manually
49 classified into subgroups of errors.

50 **Results:** From 202,059 eligible ED visits, we randomly sampled 100 for GPT-generated
51 summarization and then expert-driven evaluation. In total, 33% of summaries generated by GPT-
52 4 and 10% of those generated by GPT-3.5-turbo were entirely error-free across all evaluated
53 domains. Summaries generated by GPT-4 were mostly accurate, with inaccuracies found in only
54 10% of cases, however, 42% of the summaries exhibited hallucinations and 47% omitted
55 clinically relevant information. Inaccuracies and hallucinations were most commonly found in
56 the Plan sections of GPT-generated summaries, while clinical omissions were concentrated in
57 text describing patients' Physical Examination findings or History of Presenting Complaint.

58 **Conclusions and Relevance:** In this cross-sectional study of 100 ED encounters, we found that
59 LLMs could generate accurate discharge summaries, but were liable to hallucination and
60 omission of clinically relevant information. A comprehensive understanding of the location and
61 type of errors found in GPT-generated clinical text is important to facilitate clinician review of
62 such content and prevent patient harm.

63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84

Introduction

Clinical documentation is an essential part of high-quality patient care.^{1,2} However, in recent years there has been an increase in the complexity of clinical documentation as a result of the transition from paper-based to electronic health records (EHRs).³ This has had downstream effects on the amount of time physicians spend on the EHR, with recent studies suggesting that every hour of direct clinical time spent with patients is associated with 2 extra hours of EHR documentation.^{4,5} This concerning increase in EHR burden is a significant contributing factor to the rising prevalence of physician burnout, which may lead to a reduction in the overall quality of patient care.⁶⁻⁹

A foundational element of clinical documentation is the patient discharge or encounter summary, created following both Emergency Department (ED) visits and inpatient hospital admissions. Discharge summaries serve as a critical method of patient information transfer and provide instructions for the ongoing management of patients' illness.¹⁰⁻¹² However, the process of writing discharge summaries is time-consuming and, consequently, these summaries are often not completed in a timely manner or finished at all.^{12,13} This is problematic given that the timeliness and availability of discharge summaries is associated with patients' readmission rates, with the absence of a discharge summary associated with a 79% increased rate of 7-day readmission and 37% increased rate of readmission within 28 days.¹³ The AHRQ identifies the lack of adequate post-discharge summarization and communication as primary reasons for ED discharge failures.¹⁴

85 The recent introduction of large language models (LLMs) such as ChatGPT has led to renewed
86 focus on the use of natural language processing (NLP) to improve both quality and efficiency in
87 healthcare.¹⁵ LLMs possess a range of capabilities which may be applied to the clinical domain,
88 one of which is text summarization. Previous reports have evaluated the potential use of LLMs in
89 summarizing scientific literature, radiology reports, patient problem lists and doctor-patient
90 conversations, with varying success.^{16,17} However, there has been limited research on the ability
91 of LLMs to summarize information from a patient's hospital encounter into a discharge
92 summary.¹⁸ As ambient AI scribes and other LLM-based tools begin to be deployed within
93 healthcare settings,¹⁹ rigorous evaluations of the accuracy of these technologies are urgently
94 needed.

95

96 In this study, we investigate the performance of two state-of-the-art LLMs, GPT-4 and GPT-3.5-
97 turbo, in generating ED discharge summaries and evaluate the prevalence and type of errors
98 across each section of the discharge summary.

99

Methods

100 The UCSF Information Commons contains deidentified structured clinical data as well as
101 deidentified clinical text notes, with externally certified deidentification as previously
102 described.²⁰ The UCSF Institutional Review Board determined that this use of deidentified data
103 within the UCSF Information Commons is not human participants research and, therefore, was
104 exempt from further approval and informed consent. This study was completed according to a
105 prospectively developed protocol (Supplementary File 1).

106

107 We identified all adult patients discharged from the University of California, San Francisco
108 (UCSF) ED from 2012 to 2023 with an ED clinician note present within Information Commons
109 (Figure 1). If more than one Emergency Medicine (EM) clinician note was available for a
110 particular ED visit, the earliest note was selected as subsequent notes were often attending
111 attestation notes. In the case of multiple notes with the same chart time, the longest note (by
112 character count) was selected. Clinical notes were minimally preprocessed – only line breaks and
113 extra spaces were removed. Software packages incorporating a series of regular expressions were
114 created and used to examine the structure of notes, confirming the presence/absence of the
115 following note headers: ‘Chief Complaint’ (274,983/278,629 notes); ‘Review of Systems’
116 (263,219/278,629 notes); ‘Physical Exam’ (276,834/278,629 notes); ‘ED Course’
117 (245,900/278,629 notes); and ‘Initial Assessment’ (139,838/278,629 notes). Notes which did not
118 contain appropriate history, physical examination and assessment/plan sections were excluded.
119 Each note was tokenized using the OpenAI Tiktoken tokenizer.²¹ Notes containing ≥ 3500 tokens
120 were excluded to allow sufficient tokens for the GPT-3.5-turbo API response to be completed
121 within the model’s 4096 token context window, which was the shortest context window of the

122 models used. Patients who were admitted to hospital from the ED were identified from the
123 structured electronic health record and excluded so that only patients discharged from the ED
124 were included in our cohort.

125
126 Next, we randomly selected two $n = 100$ samples to be used as the *development* and *test* sets. All
127 prompt engineering and resident annotator training was conducted on the *development* set, while
128 evaluation was conducted on the held-out *test* set. Using the secure, HIPAA-compliant, UCSF
129 Versa Application Programming Interface (API) on Microsoft Azure, we prompted both GPT-
130 3.5-turbo (model = 'gpt-3.5-turbo-0613', role = 'user', temperature = 0; all other settings at
131 default values) and GPT-4 (model = 'gpt-4-0613', role = 'user', temperature = 0; all other
132 settings at default values) to summarize the full ED clinician note into a discharge summary. The
133 following prompt was used, followed by the corresponding note for each patient, denoted by
134 triple quotation marks: "*You are an Emergency Department physician. Below is the History and*
135 *Physical Examination note for a patient presenting to the Emergency Department who was*
136 *subsequently discharged. Write a discharge summary for the patient based on this note. Do not*
137 *include any additional information not present in the note.* \n\n "" Note text "" "

138
139 The GPT-3.5-turbo and GPT-4 generated discharge summaries were evaluated by two
140 independent EM resident reviewers (from AL, FC, KB, KP, TT) in accordance with the protocol.
141 Initial rates of inter-reviewer agreement were over 90% (Table S1). Disagreements were
142 resolved by consensus and, if required, by an attending EM physician reviewer (AK). We
143 selected three evaluation criteria for review: 1) Inaccuracy of GPT-summarized information; 2)
144 Hallucination of information; 3) Omission of relevant clinical information. An inaccuracy refers

145 to information that is not factual and/or is contradicted by the original ED clinician note.
146 Hallucination refers to the fabrication of information in the discharge summary that is not present
147 in the original ED clinician note. Omissions refer to information from the ED clinician note that
148 the reviewer deemed relevant for inclusion in the discharge summary but was not included. The
149 following aspects of a patient's ED visit were evaluated for the presence of inaccuracies,
150 hallucinations, and omissions: Presenting complaint; History of presenting complaint; Past
151 medical history; Allergies/contraindications; Review of systems; Positive examination findings;
152 Laboratory test results; Radiological investigations; Plan; Other notable events during ED stay (if
153 any). On identifying each error, reviewers were additionally asked to provide a brief explanation
154 for their reasoning, which was subsequently manually classified into subgroups of errors within
155 each of the above three evaluation criteria.

156

157 *Statistical analysis*

158 For both the GPT-3.5-turbo and GPT-4 discharge summaries, counts of each error (Inaccuracy,
159 Hallucination or Omission) across each section (Presenting complaint; History of presenting
160 complaint; Past medical history; Allergies/contraindications; Review of systems; Positive
161 examination findings; Laboratory test results; Radiological investigations; Plan; Other notable
162 events during ED stay [if any]) relating to the ED visit were collated and reported in a
163 descriptive analysis. The median word count with interquartile range (IQR) for the original EM
164 clinician notes, alongside both the GPT-4-generated and GPT-3.5-turbo generated summaries
165 was calculated. To evaluate discharge summary readability, the average Flesch-Kincaid Reading
166 Ease Score (FRES) and Flesch Kincaid Grade Level (FKGL) was calculated for each GPT
167 model. Median word counts and FRES/FKGL values were compared using the Mann-Whitney U

168 test against the null hypothesis that there is no significant difference between GPT-4-generated
169 and GPT-3.5-turbo-generated discharge summaries. Categorical variables were compared using
170 the Chi-squared test. $P < 0.05$ was significant. Analyses were performed in Python and R.

171

Results

172 From 202,059 eligible ED visits with an EM clinician note, we randomly sampled 100 for GPT-
173 generated summarization and then expert-driven evaluation (Figure 1; Table 1). The average
174 length of the original EM clinician notes summarized by the GPT models was 802.5 words (IQR
175 643.5-1053.25) (Figure S1). GPT-4-generated discharge summaries (median word count = 235
176 words, IQR 205-264) were shorter than those generated by GPT-3.5-turbo (median word count =
177 369.5 words, IQR 307.75-445) (Figure S2; Mann-Whitney U, $p < 0.001$). The average Flesch-
178 Kincaid Grade Level for GPT-4-generated summaries was lower (FKGL = 10.0, IQR 9.5-11.1)
179 than for GPT-3.5-turbo-generated summaries (FKGL = 10.7, IQR 9.7-11.7) (Mann-Whitney U, p
180 = 0.02), indicating greater readability of GPT-4-generated discharge summaries. This was also
181 reflected in the Flesch Reading Ease Scores, with GPT-4 summaries (FRES = 48.6, IQR 41.0-
182 52.0) having a higher score on average than GPT-3.5-turbo summaries (FRES = 46.7, IQR 39.7-
183 49.5), though this did not meet statistical significance (Mann-Whitney U, $p = 0.10$).

184

185 Overall, GPT-4-generated discharge summaries contained fewer errors than GPT-3.5-turbo-
186 generated summaries across all three domains (Figure 2). In total, 33% of summaries generated
187 by GPT-4 and 10% of those generated by GPT-3.5-turbo were entirely error-free across all
188 evaluated domains. Summaries generated by GPT-4 were mostly accurate, with inaccuracies
189 found in only 10% of cases. However, 42% of the summaries exhibited hallucinations and 47%
190 omitted clinically relevant information. This compares to 36% of GPT-3.5-turbo summaries
191 containing an inaccuracy, with 64% and 50% of the predecessor model's summaries containing
192 hallucinations and clinical omissions, respectively. Initial inter-reviewer agreement rates were

193 95.8%, 93.6% and 91.9% for inaccuracy, hallucination and omission errors, respectively, prior to
194 consensus agreement (Table S1).

195
196 Error rate by domain and discharge summary section is shown in Figure 3. The few inaccuracy
197 errors identified in GPT-4-generated discharge summaries predominantly occurred in the *Plan*
198 section of the summary (n = 4). When comparing GPT-3.5-turbo and GPT-4 models, there was a
199 notable improvement in the accuracy of reporting patients' *Past Medical History*, in which 10%
200 of GPT-3.5-turbo summaries contained an error compared to only 1% of GPT-4 summaries.
201 Most hallucination errors, across both GPT-3.5-turbo and GPT-4 models, occurred in either the
202 *Plan* or *Other* sections of the summary, with GPT-4 recording 36% fewer hallucinations in these
203 sections than GPT-3.5-turbo. Omissions were most frequently present in the *Physical*
204 *Examination* section for both GPT-4 (20%) and GPT-3.5-turbo (18%) summaries, followed by
205 the *History of Presenting Complaint* section (10% of GPT-4 summaries vs 17% of GPT-3.5-
206 turbo summaries).

207
208 Finally, we manually categorized free-text reviewer comments detailing the subtype of each
209 error (Table 2 & Figure S3). Among the GPT-4 summaries, inaccuracy errors included
210 inaccurate follow-up details (e.g., reviewer comment: “[*the original note states that the*] patient
211 *had follow-up with GI for colonoscopy.. already scheduled [whereas the GPT summary states*
212 *the patient was advised to obtain this*]”), inaccurately reporting the interim plan as the follow up
213 plan (e.g., reviewer comment: “*the final plan is listed [by GPT-4] as ‘follow-up labs/psych*
214 *recommendations’, but this was the sign-out plan – the final plan was actually: ‘safe for*
215 *discharge’*”) and inaccurate reporting of physical examination findings (e.g., reviewer comment:

216 “[the GPT summary] states HINTS exam was positive, but is in fact negative”). The most
217 commonly identified hallucination error subtype was hallucination of information in the note that
218 had been redacted during the de-identification process (n = 15; e.g., reviewer comment:
219 “redacted portion [of original note] filled in [in GPT summary] as ‘headache’”). The next most
220 common hallucinations related to patients’ follow up, with GPT-4 either providing details of
221 outpatient specialty follow-up that had not been arranged (n = 11; e.g., reviewer comment: “[the
222 GPT summary] hallucinated follow-up with Rheumatology and Neurology, though [there is] no
223 mention of this in [the original] note”), hallucinating ED return precautions (n = 7), and
224 hallucinating follow-up instructions (n = 3; e.g., reviewer comment: “no instructions to continue
225 current meds or avoid morphine were provided in the original note”). Meanwhile, examples of
226 the most common omission errors include GPT-4 omitting certain positive physical examination
227 findings (n = 13; e.g., “[GPT summary] omitted left sided laceration” or “[GPT summary]
228 omitted murmur”), imaging results (n = 8), details of patients’ management in ED (n = 7; mostly
229 relating to specialty consults that had taken place) and symptom(s) reported (n = 7; e.g. “[GPT
230 summary] does not mention Tylenol overdose concern”). The manually categorized reviewer
231 comments for the GPT-3.5-turbo-generated summaries are shown in Supplementary File 2
232 (Table S2 & Figure S4).

233

Discussion

234 In this cross-sectional study of 100 ED encounters, we found that LLMs could generate accurate
235 discharge summaries, but were liable to hallucination and omission of clinically relevant
236 information. Overall, GPT-4-generated summaries contained fewer errors than GPT-3.5-turbo
237 summaries across all three domains, with 10%, 42% and 47% of summaries containing
238 inaccuracies, hallucinations and omissions, respectively. GPT-4-generated summaries were also
239 shorter and more readable than those generated by GPT-3.5-turbo, with an average Flesch-
240 Kincaid Grade Level of 10.

241

242 The improved performance of GPT-4 compared to GPT-3.5-turbo aligns with prior literature
243 which has shown superior GPT-4 performance across both medical and non-medical tasks.²²⁻²⁴
244 Moreover, the fact that GPT-4 summaries contained a lower number of omissions than GPT-3.5-
245 turbo, whilst summarizing the same information in fewer words, suggests increased summary
246 concision that may be welcomed by primary care physicians and others on the receiving end of
247 the transition of care.²⁵

248

249 Although only 33% of summaries generated by GPT-4 were entirely error-free across all
250 domains, a more detailed review of the subtypes of error demonstrated that a majority of
251 hallucinations either related to information redacted in the original note as part of our
252 institution's de-identification process or resulted from GPT-4 hallucinating follow-up
253 instructions and/or return precautions. In the latter instance, such follow-up instructions were
254 often appropriate for the patient's care (as if they were derived from a standard set of precautions
255 associated with the patient's final diagnosis), but because they had not been explicitly mentioned

256 in the original EM provider's note, they were classified as hallucinations in accordance with our
257 pre-specified protocol. After excluding these specific types of errors post-hoc, the proportion of
258 GPT-4 generated summaries considered error-free across all domains increased by 14%,
259 reaching 47% error-free across the three domains.

260
261 Meanwhile, there were notable differences in initial inter-reviewer agreement between error type
262 prior to consensus agreement, with 91.9% agreement on the presence of clinical omissions
263 compared to 95.8% and 93.6% agreement for inaccuracies and hallucinations, respectively. This
264 reflects the subjective nature of classifying clinical omissions, where the inclusion of different
265 clinical details may depend on the preference of the discharging clinician. It is possible that,
266 with either dedicated prompt engineering or the addition of few-shot examples during future
267 prompting, clinician-specific preferences of what information ought to be included in each
268 discharge summary may be incorporated to address this.

269
270 There is a paucity of existing literature examining the performance of LLMs when generating
271 discharge summaries, either in the Emergency Department or inpatient hospital setting. This is
272 concerning given reports of the recent deployment of ambient artificial intelligence (AI) scribes
273 at a large healthcare organisation.¹⁹ In that study, 35 example patient transcripts and encounter
274 summaries generated by the AI scribe were rated using a modified version of the Physician
275 Documentation Quality Instrument, with an average score of 48/50 achieved.^{19,26} However, a
276 quantitative analysis of the number and type of errors present was not reported. Meanwhile, a
277 separate study of neurology inpatient encounters showed that Bidirectional Encoder
278 Representations from Transformers (BERT) and Bidirectional and Auto-Regressive

279 Transformers (BART) models could be used to generate summaries which met the standard of
280 care in 62% of cases, but acknowledged that future work should count the number and type of
281 hallucinations in automated summaries.¹⁸

282
283 Since clinicians will ultimately be responsible for auditing and modifying clinical
284 documentation produced by LLMs, gaining a thorough understanding of potential error sources
285 in this documentation is critically important. Without a thorough understanding of where errors
286 may occur, there's a risk that errors made by LLMs could be overlooked, potentially harming
287 patient care.²⁷ Additionally, the increased workload on clinicians to meticulously audit the
288 discharge summary could lead to worsening burnout, potentially negating the benefits of using
289 this technology. Our findings suggest that the location of errors within a GPT-generated
290 discharge summary may vary based on the type of error: inaccuracies and hallucinations are most
291 commonly found within the *Plan* sections of GPT-generated discharge summaries, while the
292 *Physical Examination* and *History of Presenting Complaint* sections should be checked closely
293 for clinical omissions. Future studies should evaluate the application of LLMs themselves to
294 identify instances of inaccuracy, hallucination and clinical omission errors within LLM-
295 generated clinical documents when compared to the original source documents, allowing
296 clinicians to audit and amend areas that are subject to discordance.

297
298 This study has several limitations. First, in this study only the initial EM clinician note was
299 summarized. While this note typically contains the patient's clinical history, physical
300 examination findings, results of investigations performed and overall plan, other pertinent
301 information that is found in notes from other providers, such as physical or occupational

302 therapist recommendations and specialty consult advice, may not have been included in the
303 discharge summary. Future work should evaluate the performance of LLMs in the more complex
304 task of multi-document summarization before deployment to EDs can be considered. Second,
305 due to the time and labor-intensive process of manual expert review, we included 100 randomly
306 selected ED encounters in our sample, which may limit generalizability across different types of
307 patient demographics and presenting symptoms. Notably, our randomly selected sample
308 predominantly consisted of White, Asian or Black/African American patients, with limited
309 representation of other minority groups. As LLM performance continues to be evaluated across
310 different medical tasks, racial and gender bias assessments of these tools must be performed
311 prior to their integration into clinical care.²⁸ Third, GPT model performance may improve with
312 further iterations of prompt engineering and/or in-context learning. For instance, in comparing
313 GPT-3.5-turbo to GPT-4, there was an enhancement in summarization capabilities across all
314 domains evaluated, including over ED discharge summary length. Fourth, we did not directly
315 compare the GPT-generated discharge summaries with the actual clinician-generated discharge
316 summaries for these encounters. It is possible that important information might have been
317 missing, or inaccurately reported, in the clinician-generated discharge summaries as well.

318

319 **Conclusion**

320 In this cross-sectional study of 100 ED encounters, we found that LLMs could generate accurate
321 discharge summaries, but were liable to hallucination and omission of clinically relevant
322 information. Our results suggest that the location of errors within a GPT-generated discharge
323 summary may vary based on the type of error. A comprehensive understanding of where errors

324 are most likely to occur in GPT-generated clinical text is critically important to facilitate
325 clinician review and revision of such content and prevent patient harm.

326
327

Tables

Variable	Category	Number of patients, n
<i>Sex</i>	Male	44
	Female	56
<i>Race/ethnicity</i>	White	39
	Asian	20
	Black or African American	18
	Latinx	11
	Other	4
	Native Hawaiian or Other Pacific Islander	3
	Southwest Asian and North African	3
	Unknown/Declined	2
<i>Age, median (IQR)</i>	48.1 years (37.4 – 67.9)	
<i>ESI Acuity Level</i>	Urgent	54
	Less Urgent	27
	Emergent	16
	Non-Urgent	2
	Unspecified	1
<i>Discharge disposition</i>	Home or Self Care	95
	Skilled Nursing Facility	2
	Other	3

328 **Table 1.** Patient demographics in n = 100 sample of Emergency Department encounters
 329 randomly selected for GPT-3.5-turbo and GPT-4 discharge summary generation. ED =
 330 Emergency department; ESI = Emergency Severity Index; IQR = interquartile range.

331

Error Type	Error category	Example reviewer comment*	Count
Inaccuracy	Inaccurate follow-up details	“[The original note states that the] patient had follow-up with GI for colonoscopy and EGD and hematology follow-up [was] already scheduled [whereas the GPT summary states the patient was advised to obtain this]”	3
	Inaccurate examination findings	“[The GPT summary] states HINTS exam was positive, but is in fact negative”	2
	Inaccurately reported the interim plan as the follow-up plan	“The final plan is listed as ‘follow-up labs/psych recommendations’, but this was the sign-out plan – the final plan was actually: ‘safe for discharge’”	2
	Inaccurately reported patient’s management in ED	“Written for but did not get acetaminophen in ED”	1
	Inaccurately reported imaging as normal	“CT pelvis was not negative”	1
	Inaccurate social history reported	“States patient is a former smoker, when in fact patient is a former drinker and never smoked”	1
Hallucination	Hallucinated redacted information	“Redacted portion [of original note] filled in [in GPT summary] as ‘headache’”	15
	Hallucinated outpatient follow-up details	“[The GPT summary] hallucinated follow-up with Rheumatology and Neurology, though [there is] no mention of this in [the original] note”	11
	Hallucinated ED return precautions	“No return precautions mentioned in non-redacted portion of [original] note”	7
	Hallucinated medication plan	“[GPT summary] hallucinated plan of continuing medications as prescribed - there was no reference to this in original note”	3
	Hallucinated primary care physician follow-up details	“[GPT summary] hallucinated PCP follow-up”	3
	Hallucinated follow-up instructions	“No specific return precautions (fever, chest pain, SOB) provided in [original] note; no instructions to continue current meds or avoid morphine were provided in the original note”	3
	Hallucinated patient’s management in ED	“Patient did not receive Nitro spray in ED”	3
	Hallucinated cause of symptoms	“Under diagnosis, [GPT summary] states headache is due to post-surgical changes, which was not documented in the initial note”	1
	Hallucinated patient’s diagnosis	“No mention of what final diagnosis was on original note, yet GPT wrote ‘likely	1

		due to cyst or surgery”	
	Hallucinated symptoms	“[GPT summary] hallucinated patient as having carotid tenderness”	1
Clinical Omission	Omission of positive physical examination findings	“Omitted left sided laceration”; “Did not mention contracture”; “Omitted bilateral conjunctival injection”; “Omitted murmur”; “Omitted patient’s somnolence and gait stability”	13
	Omission of imaging performed	“Omitted chest x-ray”; “Omitted MRI results”; “Omission of all radiology results”	8
	Omission of symptom reported	“Does not mention Tylenol overdose concern”; “No mention of watery diarrhea”; “Omitted that bleeding was seen by ED nurse and pressure dressing was applied”	7
	Omission of details of patient’s management in ED	“Omitted orthopedics consult”; “Omitted gynecology consult”; “Omitted reassessment and repeat check of ambulatory saturation”	7
	Omission of pertinent negative physical examination findings	“Omitted patient was afebrile”; “Should include that patient had a benign GU exam”; “Should have included benign abdominal exam”	5
	Omission of details of patient’s medication history	“Omitted that she was on antibiotics”; “Omitted estrogen use”	4
	Omission of details of patient’s Past Medical History	“Omitted history of known pulmonary embolus”; “Did not mention clarification on baseline bradycardia (this is significant abnormality that provider contacted PMD to clarify)”; “Omitted key medical history including patient was unable to walk secondary to dizziness”	4
	Omission of details of patient’s allergies	“No mention of allergies”	2
	Omission of details of patient’s Past Surgical History	“Omitted cholecystectomy surgery”; “Omits history of PEG”	2
	Omission of ECG performed	“Omission of any mention of an electrocardiogram for patient’s tachycardia”	1
	Omission of laboratory tests performed	“Omitted mention of stool studies collected”	1
	Omission of symptom time course	“Omitted timeline of symptoms – improvement, but woke her up second night in a row”	1
	Omission of symptom character	“Does not mention that the reason the patient’s chest pain is different now is that it is now constant”	1
	Omission of suspicious injury	“Omission of suspicious injury report”	1

report		
Omission of pertinent normal laboratory test results	“Omitted negative troponins”	1
Omission of follow-up information	“Discussion of possible bowel regimen not included”	1
Omission that patient declined physical examination	“Refusal of rectal exam”	1
Omission of diagnosis	“Did not include the presumptive diagnosis selection (menstrual cramps) amongst the various differential diagnosis entities”	1
Omission of code stroke activation	“Omits activation of code stroke”	1
Omission of bedside imaging done	“Omits bedside ultrasound (but mentions x-ray)”	1
Omission of urinalysis results	“Omitted positive urine drug screen for cocaine (not extremely relevant)”	1

332 **Table 2.** Manual categorization of expert reviewer comments providing further details for each
333 error subtype among GPT-4-generated discharge summaries compared to the ground-truth,
334 original Emergency Medicine provider note. *Comments reported with minor modifications to
335 syntax for improved readability.

336

Figures

337 **Figure 1.** A) Flowchart of included Emergency Department (ED) visits. B) Study workflow.

338

339 **Figure 2.** Proportion of discharge summaries with 1 or more error identified by clinical

340 reviewers in each of the three domains evaluated: 1) Inaccuracy, 2) Hallucination and 3) Clinical

341 Omission.

342

343 **Figure 3.** Breakdown of errors for each domain (Accuracy, Hallucination and Clinical Omission)

344 by section of discharge summary. PC = Presenting Complaint; HPC = History of Presenting

345 Complaint; PMH = Past Medical History; ROS = Review of Systems; PE = Physical

346 Examination.

347 **Supplementary Figures and Tables**

348 **Figure S1.** Histogram of original Emergency Medicine provider note length among the n = 100
349 sample of Emergency Department encounters randomly selected for GPT-3.5-turbo and GPT-4
350 summarization.

351
352 **Figure S2.** Histogram of word counts of a) GPT-3.5-turbo and b) GPT-4 generated discharge
353 summaries.

354
355 **Figure S3.** Manual categorization of reviewer comments providing further details for each error
356 subtype [a) Inaccuracy, b) Hallucination, and c) Clinical omission] among GPT-4-generated
357 discharge summaries compared to the ground-truth, original Emergency Medicine provider note.

358
359 **Figure S4.** Manual categorization of reviewer comments providing further details for each error
360 subtype [a) Inaccuracy, b) Hallucination, and c) Clinical omission] among GPT-3.5-turbo-
361 generated discharge summaries compared to the ground-truth, original Emergency Medicine
362 provider note.

363
364

365 **Table S1.** Initial inter-reviewer agreement rates by error type, prior to consensus agreement.

366
367 **Table S2.** Manual categorization of expert reviewer comments providing further details for each
368 error subtype among GPT-3.5-turbo-generated discharge summaries compared to the ground-

369 truth, original Emergency Medicine provider note. *Comments reported with minor
370 modifications to syntax for improved readability.
371

372 **References**

- 373 1. Ngo E, Patel N, Chandrasekaran K, Tajik AJ, Paterick TE. The Importance of the Medical
374 Record: A Critical Professional Responsibility. *J Med Pract Manag MPM*. 2016;31(5):305-
375 308.
- 376 2. Ebbers T, Kool RB, Smeele LE, et al. The Impact of Structured and Standardized
377 Documentation on Documentation Quality; a Multicenter, Retrospective Study. *J Med Syst*.
378 2022;46(7):46. doi:10.1007/s10916-022-01837-9
- 379 3. Gesner E, Gazarian P, Dykes P. The Burden and Burnout in Documenting Patient Care: An
380 Integrative Literature Review. *Stud Health Technol Inform*. 2019;264:1194-1198.
381 doi:10.3233/SHTI190415
- 382 4. Mishra P, Kiang JC, Grant RW. Association of Medical Scribes in Primary Care With
383 Physician Workflow and Patient Experience. *JAMA Intern Med*. 2018;178(11):1467-1472.
384 doi:10.1001/jamainternmed.2018.3956
- 385 5. Sinsky C, Colligan L, Li L, et al. Allocation of Physician Time in Ambulatory Practice: A
386 Time and Motion Study in 4 Specialties. *Ann Intern Med*. 2016;165(11):753-760.
387 doi:10.7326/M16-0961
- 388 6. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records
389 and burnout: Time spent on the electronic health record after hours and message volume
390 associated with exhaustion but not with cynicism among primary care clinicians. *J Am Med*
391 *Inform Assoc*. 2020;27(4):531-538. doi:10.1093/jamia/ocz220
- 392 7. Ortega MV, Hidrue MK, Lehrhoff SR, et al. Patterns in Physician Burnout in a Stable-Linked
393 Cohort. *JAMA Netw Open*. 2023;6(10):e2336745. doi:10.1001/jamanetworkopen.2023.36745
- 394 8. Tajirian T, Stergiopoulos V, Strudwick G, et al. The Influence of Electronic Health Record
395 Use on Physician Burnout: Cross-Sectional Survey. *J Med Internet Res*. 2020;22(7):e19274.
396 doi:10.2196/19274
- 397 9. Peccoralo LA, Kaplan CA, Pietrzak RH, Charney DS, Ripp JA. The impact of time spent on
398 the electronic health record after work and of clerical work on burnout among clinical faculty.
399 *J Am Med Inform Assoc*. 2021;28(5):938-947. doi:10.1093/jamia/ocaa349
- 400 10. Taylor DM, Cameron PA. Discharge instructions for emergency department patients:
401 what should we provide? *Emerg Med J*. 2000;17(2):86-90. doi:10.1136/emj.17.2.86
- 402 11. Sorita A, Robelia PM, Kattel SB, et al. The Ideal Hospital Discharge Summary: A Survey
403 of U.S. Physicians. *J Patient Saf*. 2021;17(7):e637-e644.
404 doi:10.1097/PTS.0000000000000421
- 405 12. Robelia PM, Kashiwagi DT, Jenkins SM, Newman JS, Sorita A. Information Transfer
406 and the Hospital Discharge Summary: National Primary Care Provider Perspectives of
407 Challenges and Opportunities. *J Am Board Fam Med JABFM*. 2017;30(6):758-765.
408 doi:10.3122/jabfm.2017.06.170194

- 409 13. Li JYZ, Yong TY, Hakendorf P, Ben-Tovim D, Thompson CH. Timeliness in discharge
410 summary dissemination is associated with patients' clinical outcomes. *J Eval Clin Pract*.
411 2013;19(1):76-79. doi:10.1111/j.1365-2753.2011.01772.x
- 412 14. Improving the Emergency Department Discharge Process: Environmental Scan Report.
- 413 15. Wachter RM, Brynjolfsson E. Will Generative Artificial Intelligence Deliver on Its
414 Promise in Health Care? *JAMA*. 2024;331(1):65-69. doi:10.1001/jama.2023.25054
- 415 16. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence
416 summarization. *Npj Digit Med*. 2023;6(1):1-8. doi:10.1038/s41746-023-00896-7
- 417 17. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can
418 outperform medical experts in clinical text summarization. *Nat Med*. Published online
419 February 27, 2024;1-9. doi:10.1038/s41591-024-02855-5
- 420 18. Hartman VC, Bapat SS, Weiner MG, Navi BB, Sholle ET, Campion TR. A method to
421 automate the discharge summary hospital course for neurology patients. *J Am Med Inform*
422 *Assoc JAMIA*. 2023;30(12):1995-2003. doi:10.1093/jamia/ocad177
- 423 19. Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes to
424 Alleviate the Burden of Clinical Documentation. *Catal Non-Issue Content*.
425 2024;5(1):CAT.23.0404. doi:10.1056/CAT.23.0404
- 426 20. Radhakrishnan L, Schenk G, Muenzen K, et al. A certified de-identification system for
427 all clinical text documents for information extraction at scale. *JAMIA Open*.
428 2023;6(3):ooad045. doi:10.1093/jamiaopen/ooad045
- 429 21. openai/tiktoken. Published online March 23, 2024. Accessed March 23, 2024.
430 <https://github.com/openai/tiktoken>
- 431 22. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Butte AJ. Assessing clinical
432 acuity in the Emergency Department using the GPT-3.5 Artificial Intelligence Model.
433 Published online August 13, 2023:2023.08.09.23293795. doi:10.1101/2023.08.09.23293795
- 434 23. OpenAI. GPT-4 Technical Report. Published online March 27, 2023.
435 doi:10.48550/arXiv.2303.08774
- 436 24. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining
437 of Free-Text CT Reports on Lung Cancer. *Radiology*. 2023;308(3):e231362.
438 doi:10.1148/radiol.231362
- 439 25. Chatterton B, Chen J, Schwarz EB, Karlin J. Primary Care Physicians' Perspectives on
440 High-Quality Discharge Summaries. *J Gen Intern Med*. Published online November 27, 2023.
441 doi:10.1007/s11606-023-08541-5

- 442 26. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing Electronic Note Quality Using
443 the Physician Documentation Quality Instrument (PDQI-9). *Appl Clin Inform.* 2012;3(2):164-
444 174. doi:10.4338/ACI-2011-11-RA-0070
- 445 27. Adler-Milstein J, Redelmeier DA, Wachter RM. The Limits of Clinician Vigilance as an
446 AI Safety Bulwark. *JAMA.* Published online March 14, 2024. doi:10.1001/jama.2024.3620
- 447 28. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate
448 racial and gender biases in health care: a model evaluation study. *Lancet Digit Health.*
449 2024;6(1):e12-e22. doi:10.1016/S2589-7500(23)00225-X

450

451 **Conflicts of Interest**

452 AEK is a co-founder and consultant to CaptureDx. AJB is a co-founder and consultant to
453 Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung,
454 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory
455 panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group,
456 AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in
457 Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google),
458 Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma,
459 Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna
460 Health, Assay Depot, and Vet24seven, and several other non-health related companies and
461 mutual funds; and has received honoraria and travel reimbursement for invited talks from
462 Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens,
463 Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions,
464 medical or disease specific foundations and associations, and health systems. AJB receives
465 royalty payments through Stanford University, for several patents and other disclosures licensed
466 to NuMedii and Personalis. AJB's research has been funded by NIH, Peraton (as the prime on an
467 NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon
468 Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the
469 Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile
470 Diabetes Research Foundation, California Governor's Office of Planning and Research,
471 California Institute for Regenerative Medicine, L'Oreal, and Progenity. None of these entities
472 had any bearing on the design of this study or the writing of the manuscript.

473
474 No other authors have conflicts of interest to disclose.

475
476 **Acknowledgements**

477 Dr Aaron E. Kornblith is supported by Eunice Kennedy Shriver National Institute of Child
478 Health and Human Development of the National Institutes of Health under award number
479 K23HD110716.

480
481 The authors acknowledge the use of the UCSF Information Commons computational research
482 platform, developed and supported by UCSF Bakar Computational Health Sciences Institute.
483 The authors also thank the UCSF AI Tiger Team, Academic Research Services, Research
484 Information Technology, and the Chancellor's Task Force for Generative AI for their software
485 development, analytical and technical support related to the use of Versa API gateway (the
486 UCSF secure implementation of large language models and generative AI via API gateway),
487 Versa chat (the chat user interface), and related data asset and services.

488
489 Dr Christopher Y.K. Williams had full access to all the data in the study and takes responsibility
490 for the integrity of the data and the accuracy of the data analysis.

491
492 **Code availability**
493 The code accompanying this manuscript is available at [https://github.com/cykwilliams/GPT-4-](https://github.com/cykwilliams/GPT-4-Emergency-Department-Discharge-Summary/)
494 [Emergency-Department-Discharge-Summary/](https://github.com/cykwilliams/GPT-4-Emergency-Department-Discharge-Summary/)

Figures

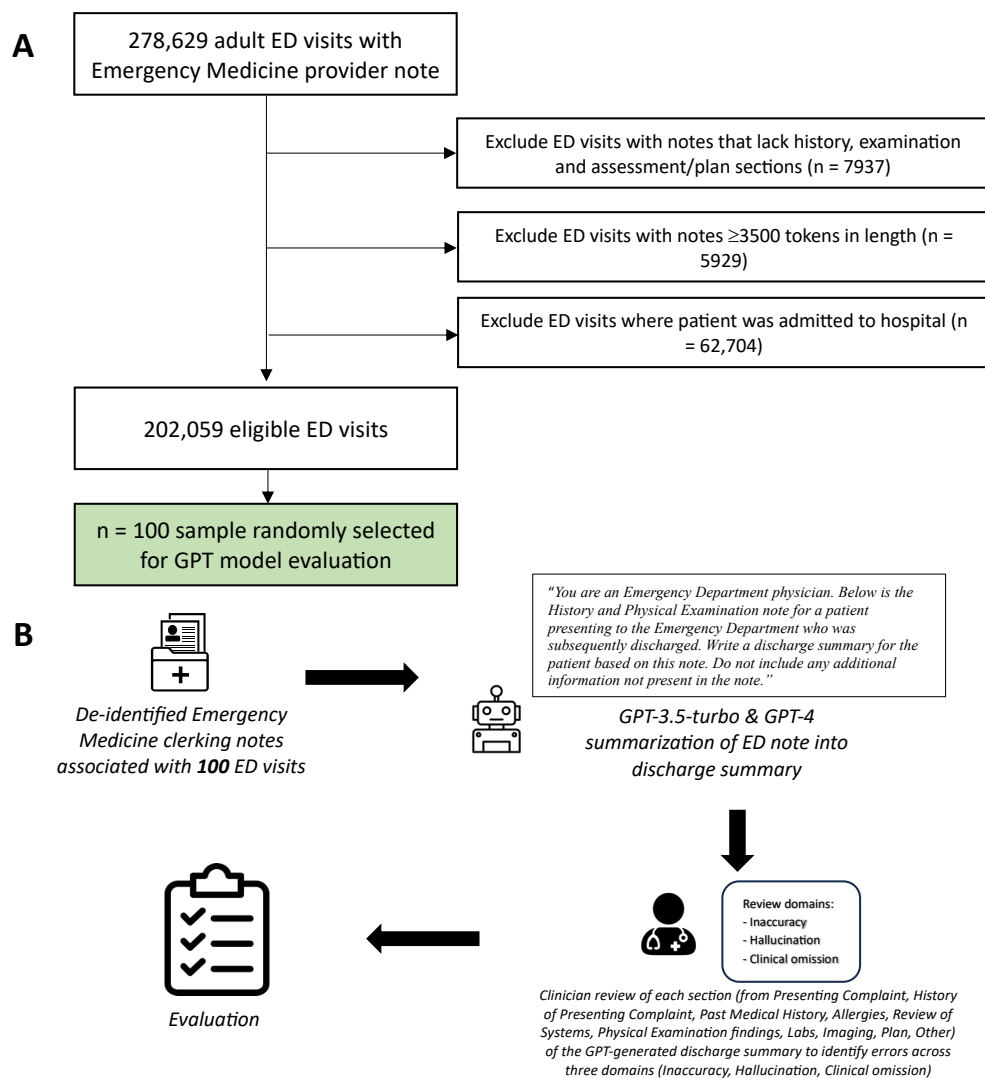


Figure 1. A) Flowchart of included Emergency Department (ED) visits. B) Study workflow.

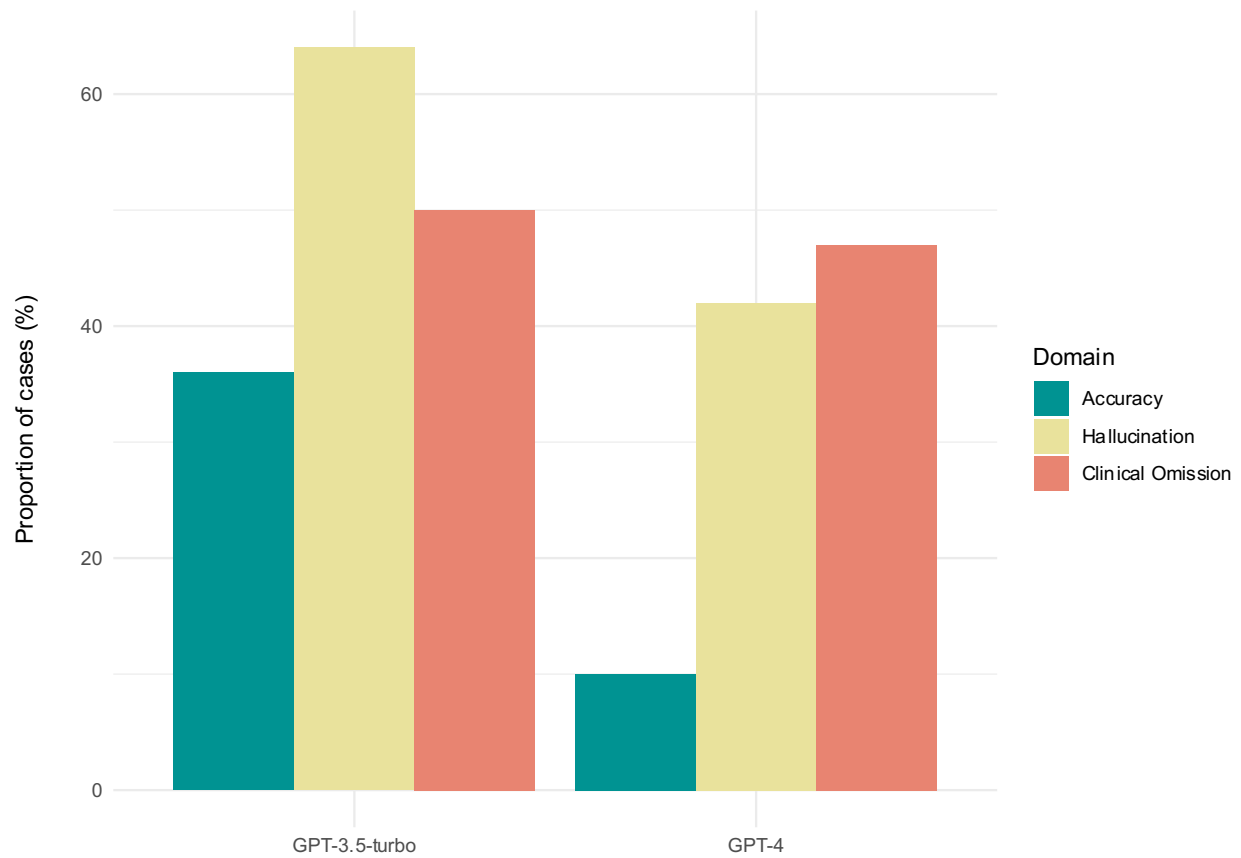


Figure 2. Proportion of discharge summaries with 1 or more error identified by clinical reviewers in each of the three domains evaluated: 1) Inaccuracy, 2) Hallucination and 3) Clinical Omission.

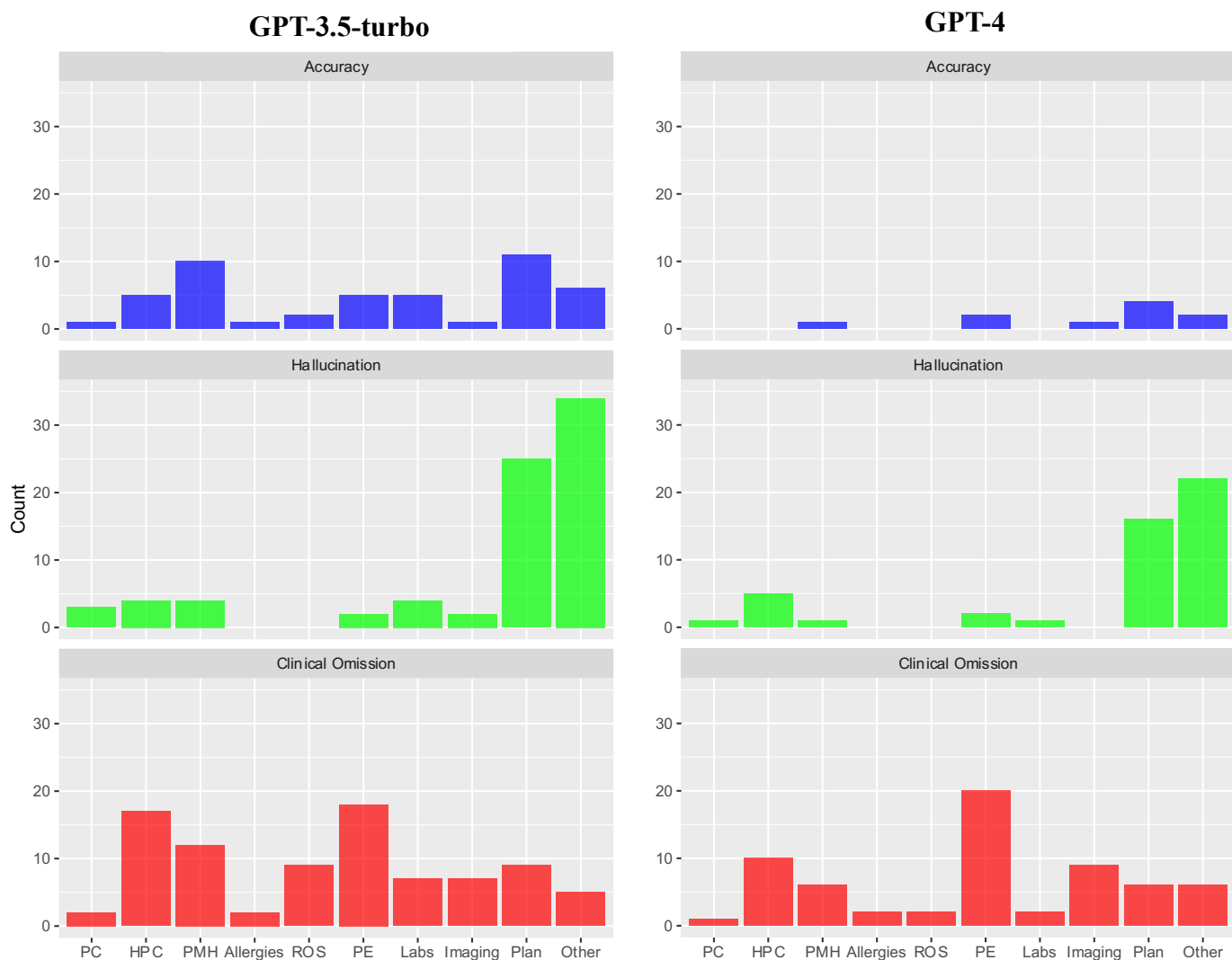


Figure 3. Breakdown of errors for each domain (Accuracy, Hallucination and Clinical Omission) by section of discharge summary. PC = Presenting Complaint; HPC = History of Presenting Complaint; PMH = Past Medical History; ROS = Review of Systems; PE = Physical Examination.