Physician- versus Large Language Model-Generated Summaries in the Emergency Department

Niloufar Golchini
University of California, Berkeley
ngolchini@berkeley.edu

Nikita Mehandru
University of California, Berkeley
nmehandru@berkeley.edu

Ahmed Alaa*
UC Berkeley and UCSF
amalaa@berkeley.edu

Melanie Molina*
University of California, San Francisco
Melanie.Molina@ucsf.edu

August 13, 2025

Abstract

High-quality one-liner summaries are essential in the emergency department (ED) to support rapid decision-making, but generating them is cognitively demanding and adds to documentation burden; large language models (LLMs) may help by synthesizing longitudinal electronic health record (EHR) data into concise, clinically useful summaries. In this blinded, within-subject study of 99 ED encounters from adult patients at the University of California, San Francisco with prior inpatient admissions (March 2022–March 2024), 26 ED physicians (14 attendings, 12 residents) evaluated paired LLM- and physician-generated summaries in randomized order, rating each on accuracy, completeness, and clinical utility using 5-point Likert scales and indicating their preferred summary with optional free-text explanations. LLM-generated summaries were preferred in 50.5% of encounters, physician summaries in 38.4%, and 11.1% were ties; compared with physician summaries, LLM summaries had higher mean (SD) scores for accuracy (4.27 [0.98] vs 3.49 [1.43]; P ¡ .001), completeness (3.72 [1.08] vs 3.28 [1.25]; P = .006), and clinical utility (3.95 [1.20] vs 3.28 [1.51]; P ¡ .001). Qualitative feedback suggested LLMs tended to produce more inclusive and neutrally phrased summaries, while physicians offered richer nuance but sometimes omitted key details. These findings suggest that domain-adapted LLM-generated one-liners can outperform physician-authored summaries on multiple quality dimensions and, with clinician oversight, may aid rapid synthesis of complex EHR data in high-stakes settings.

*Joint senior authorship

1 Introduction

In the emergency department (ED), physicians must make rapid, high-stakes decisions amid persistent time pressure, resource constraints, crowding, and frequent interruptions [38, 20]. To do so, they must quickly integrate clinical data from multiple sources in a patient's electronic heath record (EHR) into a coherent picture that guides immediate diagnostic and therapeutic actions. Central to this cognitive workflow is the creation of a concise "one-liner" that captures the most relevant aspects of a patient's history, presentation, and risk factors. Such summaries not only scaffold clinical reasoning and differential diagnosis but also enable efficient and accurate communication across multidisciplinary teams [12, 9]. Because they are often the first synthesized interpretation of the EHR encountered during an ED visit, the quality of these one-liners is critical, as they shape early management decisions and can influence the whole trajectory of care [26].

Crafting a high-quality one-liner is time-intensive. ED physicians often face significant frequent interruptions (12–13 times/hour) leading to cognitive overload, and fragmented information spread across multiple prior hospital encounters [17]. Synthesizing the relevant data from multiple sources, including discharge summaries, consult notes, imaging reports, and progress notes, requires filtering through longitudinal records authored by different providers across varied care contexts [21]. This cognitive labor not only competes with bedside patient care, but also, causes fatigue and burnout, especially when chart-review tasks extend beyond scheduled shifts [24]. Yet the cost of omission or misrepresentation in an ED one-liner can be clinically consequential, particularly when decisions must be made under time pressure and with incomplete information, as demonstrated in previous studies of diagnostic errors under cognitive strain [22].

Large language models (LLMs) have recently shown strong performance in clinical reasoning and summarization tasks across a range of benchmarks and use cases [29, 33, 1]. Most prior work has focused on generating long-form summaries, such as hospital discharge notes, using clinical notes from single encounters as an input [36, 6, 14, 15]. However, no study to date has systematically evaluated the ability of LLMs to generate concise "one-liners" by integrating multiple EHR sources and encounters spanning time. This gap is particularly relevant in the ED workflow, where clinicians must rapidly synthesize longitudinal patient history, often within seconds to minutes, to inform critical decisions [19, 13]. A systematic evaluation of LLM-generated one-liners against human-authored counterparts is therefore essential to assess the readiness of these models for real-world clinical use.

In this study, we evaluated the ability of an LLM to generate concise and accurate one-liner summaries by synthesizing information from multi-source EHR notes spanning prior hospital encounters, using an experimental setup that reflects the real-world cognitive demands of emergency physicians during patient intake. To assess the quality of the LLM-generated summaries, we conducted a blinded evaluation in which emergency physicians compared LLM outputs to physician-authored summaries. Each

summary was rated along three core dimensions: accuracy, completeness, and clinical utility.

2 Methods

2.1 Study Design and Setting

This was a single-blind, within-subject evaluation study comparing physician- and LLM-generated one-liner summaries for ED encounters. The study was conducted at the University of California, San Francisco (UCSF), in collaboration with the University of California, Berkeley. Patient encounters were sourced from the UCSF Information Commons, which contains deidentified clinical notes certified for research use [27]. The study protocol was approved by the UC Berkeley Institutional Review Board, and all participants provided written informed consent and were compensated. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines [35].

2.2 Cohort Selection

We included ED encounters between March 1, 2022, and March 31, 2024, for patients aged 18 years or older with at least one prior inpatient admission at the UCSF Medical Center. Encounters were included only if they contained both a discharge summary from the prior admission and an ED provider note from a subsequent ED visit. We excluded encounters with incomplete notes and those where the ED provider note lacked a clear one-liner summary.

After applying the inclusion criteria, we used a two-tiered stratified sampling strategy to ensure clinical diversity. Rare chief complaints (defined as those occurring in fewer than 10 encounters) were fully retained, while more common presentations were proportionally sampled. From this pool, encounters were randomly assigned to physician reviewers through a non-overlapping allocation.

2.3 LLM-Based Summary Generation

To generate the one-liner summaries, the LLM, GPT-4o (OpenAI, 2024), was first given a prompt containing one of each of the following most recent note types, if available (potentially spanning across multiple encounters): discharge summary, progress note, consult note, history and physical (HP) note, imaging report, echocardiogram report, and electrocardiogram (ECG) report. The model then selected which notes to use to derive the one-liner summary based on relevance to the chief complaint (reported by the triage) using note-title heuristics.

Next, with the selected notes as part of the input, the LLM-generated one liner summaries were then produced using a few-shot prompting approach leveraging semantic similarity. Each case was stratified by chief complaint (e.g., chest pain, abdominal pain, altered mental status), and both discharge summaries and chief complaints

were embedded using Bio_ClinicalBERT. The three most similar prior cases (k=3) with high-quality human-authored one-liners were selected as exemplars for prompting. This k-nearest neighbor strategy provided contextual examples tailored to each case.

Pilot testing showed that the LLM inconsistently used standard medical abbreviations (e.g., "SOB" for shortness of breath, "HTN" for hypertension), which are essential for ED one-liners. We addressed this by adding a medical abbreviation guidance document to the prompt, instructing the model to use standard abbreviations consistent with emergency medicine notes [4].

2.4 One-Liner Summary Evaluation

Twenty-six emergency physicians (12 residents, 14 attendings) participated in the evaluation of the one-liner summaries.Each participant was presented with a set of anonymized cases via a custom web-based platform (Streamlit) designed to simulate a patient's chart. For each case, participants received a standardized chart packet: patient age, sex, presenting chief complaint, and previous EHR notes. They were presented with two one-liner summaries side by side: one authored by an ED provider as part of their original ED Note and the other generated by the LLM.

Reviewers rated each summary on three 5-point Likert scales:

- **Accuracy**: Is the information factually correct and free of fabrication?
- **Completeness**: Does it capture all clinically relevant information?
- **Clinical utility**: How informative is it for decisions in a real-world ED workflow?

Participants also indicated a preference between the one-liner summaries (one over the other or both considered equal preference) and optionally provided a brief explanation of their rationale. The three domain scores were averaged to calculate an overall quality score.

2.5 Statistical Analysis

To evaluate the LLM's notes selection strategy, we analyzed which clinical notes the model chose to incorporate when generating each one-liner summary. The LLM's selected notes were extracted from the model output and parsed to identify specific note types used. For each encounter, we calculated the number and proportion of available notes utilized by the LLM. Note selection patterns were then stratified by chief complaint category, with complaints grouped into overarching clinical systems. Descriptive statistics including mean (SD), median, and interquartile range were computed for the number of notes selected overall and by clinical category. To assess whether note selection varied significantly across chief complaint categories, we performed Kruskal-Wallis H-tests. Selection rates for individual note types were calculated as the percentage of encounters where each note type was utilized, both overall and stratified by clinical category, to identify patterns in notes preferences based on clinical presentation.

We used descriptive statistics to summarize participant scores as well as summary length. We primarily used paired T-test (with Wilcoxon signed-rank tests to confirm) to compare the scores of physician- versus LLM-generated summaries across each domain and overall. Win rates (i.e., proportion of cases where one summary was preferred over the other) were computed with 95% confidence intervals using Wilson's method.

To understand how note characteristics influenced the quality of the one-liner summaries, we performed two complementary analyses. First, we compared physician- and LLM-generated summary performance relative to individual note type lengths. Pearson correlation coefficients were calculated to determine whether note length was associated with quality ratings in either group. Second, we evaluated how specific note types included in the LLM prompt affected the quality of the LLM-generated one-liner summary. For encounters where certain note types were available, we compared the LLM-generated one-liner summary quality between cases where these notes were included in the prompt versus excluded, reporting the difference using Cohen's d effect size. All analyses were performed in Python version 3, and $P<0.05$ was considered significant.

2.6 Qualitative Analysis

We conducted a rapid content analysis of open-ended survey responses to identify key themes. Free-text data were reviewed systematically by NG and MM. Using a descriptive analytic approach, we summarized content within each open-ended item and grouped similar responses into thematic categories.

This process involved both deductive coding, based on the three evaluation domains, and inductive coding to capture emergent themes raised by participants. A matrix was used to organize and compare responses across participants and questions. Thematic summaries were refined through iterative team discussion, and representative quotes were identified to illustrate major findings. [23]

3 Results

3.1 Cohort Characteristics

Of 10,631 ED encounters in the clinical data warehouse, 3,984 encounters met the inclusion criteria (Figure). A random sample of 300 encounters was selected for LLM summarization and physician reviewer evaluation.

Each physician reviewed 1-10 non-overlapping cases, resulting in a total of 99 unique ED encounters reviewed with 99 unique patients. Patients were 55.6% female, with a mean (SD) age of 57 (21) years. Nearly all encounters contained progress notes (99%), H&P notes (98%), and imaging reports (98%). ECGs were present in 87% of encounters, consult notes in 64%, and echocardiogram reports in 56%. The mean (SD) length of all notes combined per case was 5,598 (2,314) words, with 6.0 (range, 2-7) distinct note types included. Note length varied substantially by type, ranging from 36 (32) words for ECG reports to 2,310 (1,263) words for discharge summaries, with

intermediate lengths for imaging (124 [117] words), echocardiograms (677 [109] words), consult notes (854 [629] words), progress notes (917 [1,010] words), and H&P notes (1,332 [770] words).

3.2 LLM-Based Summary Generation

Table 1: LLM Note Selection Rates by Chief Complaint Category

Category	**Chief Complaint**	**N**	**Mean**	**Note Selection Rate (%)**						
				DS	**HP**	**PN**	**ECG**	**IM**	**EC**	**CN**
Cardiopulm.	SOB	8	4.9	100	100	0	100	100	88	0
	Chest Pain	4	4.0	100	75	0	100	100	25	0
GI	Abd. Pain	23	3.6	100	96	0	0	100	0	65
General	Fever	3	3.7	100	100	67	0	67	0	33
	Referral	3	3.0	100	100	67	0	0	0	33
	Fall	2	3.5	100	100	50	0	100	0	0
MSK	Back Pain	4	3.2	100	100	0	0	100	0	25
	Foot Pain	2	3.0	100	100	0	0	100	0	0
Neuro	Headache	3	3.7	100	100	0	0	100	0	67
	Seizures	3	4.7	100	100	0	100	100	0	67
	Dizziness	3	4.0	100	100	0	100	100	0	0
	Syncope	2	4.0	100	100	0	50	100	50	0
Psych	Psych Eval	4	3.8	100	100	75	0	0	0	100

DS=Discharge Summary; HP=H&P; PN=Progress Notes; IM=Imaging; EC=Echo; CN=Consult
SOB=Shortness of Breath; GI=Gastrointestinal; MSK=Musculoskeletal

Note: Mean represents the average number of notes selected per encounter. N = number of encounters.

The LLM requested a mean(SD) of 3.73 (0.67) notes (median: 4.0, IQR: 3.0-4.0), utilizing 53.2% (9.5%) of available notes. The model demonstrated hierarchy in note selection: H&P notes were nearly universally incorporated (97.0%,), followed by imaging reports (81.8%). Consult notes were included in 44.4% of cases, while cardiac-specific notes (ECG: 21.2%, echocardiogram: 9.1%) and progress notes (19.2%) were selectively utilized. Note selection patterns varied significantly by chief complaint category (Kruskal-Wallis $H = 17.29$, $p < 0.001$). Cardiopulmonary presentations required the most comprehensive notes (shortness of breath: 4.9(0.3) notes), with universal inclusion of cardiac studies (ECG: 100%, echocardiogram: 88% for shortness of breath). In contrast, musculoskeletal complaints required fewer notes (back pain: 3.2 (0.5) notes) and appropriately excluded cardiac notes. Psychiatric evaluations demonstrated unique patterns with high utilization of progress notes (75%) and consult notes (100%), while avoiding imaging and cardiac studies entirely (Table 1).

3.3 One-Liner Summary Evaluation

Across all cases, LLM-generated one-liners were preferred over physician-generated one-liners in 50 cases (50.5%); physician-generated one-liners were preferred in 38 cases

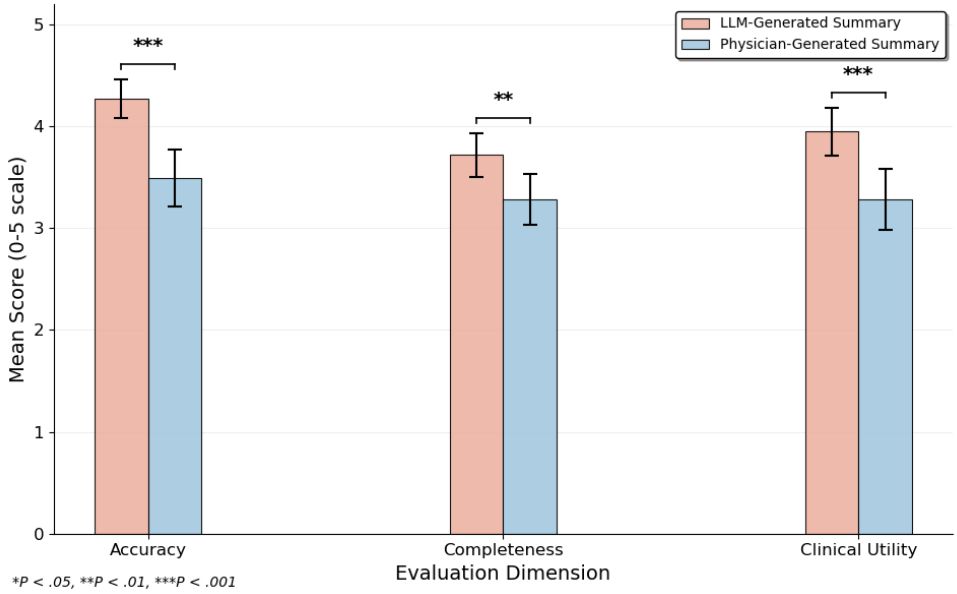

Figure 1: **Comparison of LLM-generated and physician-generated one-liner summaries across evaluation dimensions.** Mean scores (± 95% confidence intervals) for LLM-generated summaries (tan bars) and physician-generated summaries (blue bars) on a 5-point Likert scale across accuracy, completeness, and clinical utility.

(38.4%); 11 cases (11.1%) were rated equally. Excluding ties, the LLM win rate was 56.8% (95% CI, 46.4–66.7%).

As seen in Figure 1, Mean (SD) domain ratings were higher for LLM-generated summaries: Accuracy 4.27 (0.98) vs. 3.49 (1.43) ($p < 0.001$, Cohen's $d = 0.52$), Completeness 3.72 (1.08) vs. 3.28 (1.25) ($p = 0.006$, $d = 0.28$), and Clinical Utility 3.95 (1.20) vs. 3.28 (1.51) ($p < 0.001$, $d = 0.40$). The overall quality score was higher for LLM-generated summaries (3.98 (1.05) vs. 3.35 (1.39), $p < 0.001$). Wilcoxon signed-rank tests confirmed these results: Accuracy $Z = 187.0$, $p < 0.0001$; Completeness $Z = 873.0$, $p = 0.0028$; Clinical Utility $Z = 479.5$, $p = 0.0001$.

3.4 Source Note Characteristics and Summary Performance

3.4.1 Physician vs LLM: Impact of Note Length

For LLM-generated summaries, overall note length was not associated with accuracy ($r = 0.053$, $p > 0.05$), completeness ($r = -0.039$, $p > 0.05$), or clinical utility ($r = -0.094$, $p > 0.05$). In contrast, physician-generated summary completeness increased with longer notes ($r = 0.229$, $p = 0.023$). When we stratified our analysis by specific note type, showed that for LLM-generated summaries, longer history and physical notes were associated with higher accuracy ($r = 0.243$, $p = 0.016$), whereas longer progress notes were negatively correlated with completeness ($r = -0.229$, $p = 0.023$) and overall

quality ($r = -0.210$, $p = 0.038$).

3.4.2 Impact of Selective Note Inclusion on LLM-Generated Summaries

The effect of selective note inclusion by the LLM was also evaluated. When imaging reports were both available and explicitly selected by the LLM, summary quality scores were higher compared with encounters where imaging reports were available but not included (4.05 (0.93) vs. 3.58(1.14), effect size=0.49, $p = 0.078$).

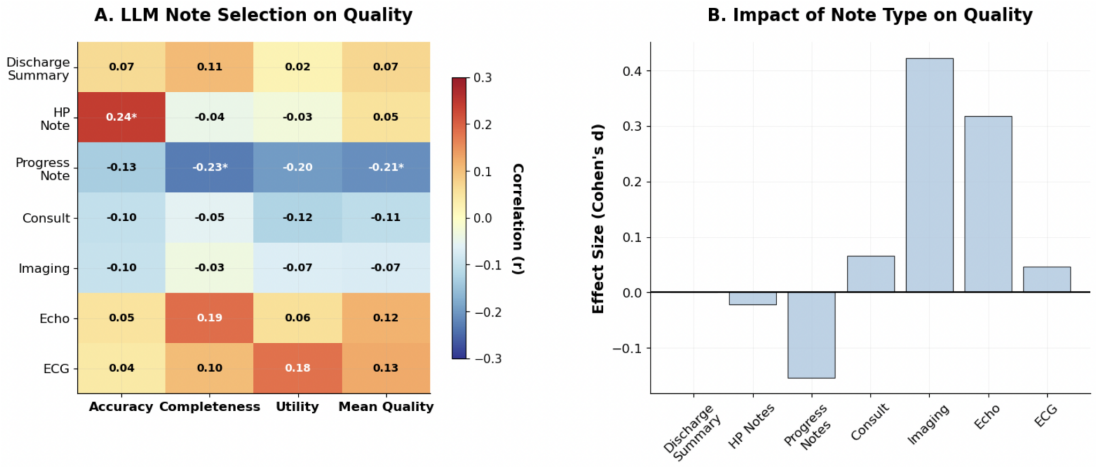

Figure 2: **Impact of note characteristics on LLM summary quality.** (A) Correlations (r) between individual note type lengths (decreasing in length from top to bottom of the matrix) and quality scores across domains (accuracy, completeness, clinical utility, and mean quality). (B) Effect sizes (Cohen's d) showing the impact of explicitly including specific note types on LLM-generated summary quality compared to when those note types were available but not included. Asterisks indicate statistically significant correlations ($p < 0.05$).

Figure 2A shows correlations between individual note types length and LLM-generated summary quality across domains.

Figure 2B displays the effect of explicitly including individual note types in the LLM input. Imaging reports (Cohen's $d = 0.42$) and echocardiograms (Cohen's $d = 0.32$) were associated with improved summary quality compared to encounters where these notes were available but not included, while inclusion of progress notes had a small negative effect.

3.5 Qualitative Analysis

We conducted a thematic analysis of 55 paired evaluations in which participants provided free-text commentary explaining their preferences between LLM and Physician-generated summaries (Table 2).

Table 2: Examples of Physician Evaluator Commentary

Summary Type	**Dimension**	**Physician-Provided Explanation**
LLM-Generated	Accuracy	"One-liner lists COPD on home O2, which does not seem to be true. Assessment of dyspnea is very different for people with vs without COPD."
LLM-Generated	Completeness	This summary lets me know that she has cryptogenic cirrhosis and factor VII deficiency—both change the likelihood of other causes of her shortness of breath (although HFpEF remains most likely).
LLM-Generated	Completeness	"Summary A does not include the critical information of patient being immunocompromised due to chemo."
LLM-Generated	Clinical Utility	"Chronic abdominal pain is helpful to know but the [Human summary] shines because it really paints the patient's pertinent surgical history which can color our approach to this person's abdominal pain."
Physician-Generated	Accuracy	"Summary B states she has DM, but this was not mentioned in any of her notes."
Physician-Generated	Completeness	"Summary B does not mention any psych or substance use history and leaves me completely bamboozled as to what medication needs to be refilled."
Physician-Generated	Clinical Utility	"It gives the important information that she has a history of MDR bacteriuria, which is relevant to her presenting complaint of abdominal pain."
Physician-Generated	Clinical Utility	"Summary B provides additional very relevant history to the chief complaint, which is that she has a history of rectal prolapse."

Reviewer commentary highlighted complementary strengths and weaknesses between LLM- and physician-generated summaries. LLM outputs were often broad and inclusive of relevant history, such as transplant dates or immunocompromised status,

and generally neutral in tone. However, reviewers frequently identified fabricated details not supported by source documentation, including diagnoses like "frequent falls" or "COPD on home oxygen," and noted omissions of key information such as active medications or recent procedures. Physician summaries contained fewer fabrications but showed greater variability in completeness: some effectively prioritized clinically relevant details, while others omitted important context. Physicians were more likely to integrate chief complaint nuances or psychiatric context, whereas LLMs occasionally missed the presenting issue or included less relevant history; in some cases, LLMs explicitly linked pertinent historical details (e.g., prior seizure disorder, asplenia in trauma cases) to the current presentation.

Tone and bias also differed between sources. Several physician-generated summaries contained emotionally charged or assumptive language (e.g., speculating on intent behind an overdose), while LLM summaries were generally factual, though omission bias was sometimes noted. In multiple cases, both, the LLM and Physician, provided mutually exclusive information, suggesting that combining both approaches could produce more accurate and contextually complete summaries.

4 Discussion

In this study, we found that an LLM, when prompted with domain-specific context, generated concise one-liner summaries that outperformed physician-authored counterparts across accuracy, completeness, and clinical utility. LLM-generated summaries were preferred in the majority of blinded evaluations and achieved significantly higher mean scores across all domains. These findings suggest that, under structured prompting and clinical oversight, generative models can synthesize temporally fragmented notes into coherent and clinically useful summaries that support ED decision-making [33, 29].

A key insight is that LLM performance depends not only on model capability but on how context is structured during input. In zero-shot settings, the model frequently misrepresented chronic conditions as acute complaints, errors with potentially serious clinical consequences, echoing prior concerns with longitudinal EHR summarization [34]. Our k-nearest neighbor few-shot prompting strategy, combined with abbreviation normalization, improved temporal coherence and clinical specificity, consistent with previous work on domain-specific prompting and retrieval augmented generation in clinical NLP [39, 32]. Notably, the model exhibited context-sensitive document selection, incorporating only 53% of available notes while maintaining clinical appropriateness. Such selective attention mirrors expert physician behavior [28, 25] and varied by chief complaint, with cardiac studies nearly universal for cardiopulmonary presentations but avoided in musculoskeletal or psychiatric cases [16]. This "clinical mimicry" [40] is not only important for accuracy, but also, for efficiency and broader implementation in health systems, where token costs accumulate rapidly; any real-world deployment will require selective note strategies to remain operationally viable [10].

Despite this apparent fluency, the model's selection process remains opaque [3],

limiting interpretability for clinicians accustomed to articulating their reasoning. In high-stakes contexts, lack of transparency risks eroding trust [8, 19]. Future work could prompt the model to explicitly outline its rationale for note inclusion, thereby enhancing interpretability and offering an auditable trail for validation.

Physician–LLM comparison revealed complementary error profiles: LLMs occasionally hallucinated unsupported details, while physicians were more prone to omissions or cognitive biases such as anchoring and premature closure [18, 37]. Physician completeness increased with note length; however, we cannot determine whether they definitively reviewed these longer documents. In contrast, LLM performance was invariant to note volume, suggesting resilience to information overload [1, 6, 31, 7]. This stability under high-volume input is advantageous in the cognitively demanding ED environment and supports hybrid workflows in which LLMs provide structured drafts that clinicians review and contextualize.

Not all note types contributed equally to output quality. H&P notes and imaging reports were positively associated with accuracy and completeness, whereas verbose progress notes were negatively associated [36]. Firstly, these effects may reflect both length-related dilution of relevant signal and recency or "primary source" bias, where more recent or central documents disproportionately influence summarization [11]. Overall, these observations underscore that high-yield, structured documents are the most effective inputs for generative summarization, and indiscriminate inclusion of low-value sources may reduce relevance for real world deployment.

Finally, the ED imposes extreme cognitive demands, with frequent interruptions and fragmented records contributing to burnout, diagnostic delay, and missed information [17, 21, 5, 30, 2, 31]. LLM-generated one-liners could mitigate this burden by front-loading synthesis, enabling clinicians to focus on higher-order reasoning and patient interaction. However, implementation must prioritize trust, transparency, and safety; these tools should augment, not replace, clinical reasoning [22].

4.1 Limitations

First, physician-authored one-liners may have incorporated context unavailable to the LLM, specifically regarding the chief complaint, which was a tabular field from the clinical warehouse. Second, physician summaries reflect real-world time constraints, while LLMs operate without such pressure, potentially biasing comparison. Third, the cohort was drawn from a single academic center, limiting generalizability. Fourth, we evaluated retrospective summaries rather than real-time use and did not assess downstream clinical impact. Finally, although kNN prompting improved coherence, alternative prompting or fine-tuning approaches were not explored and may yield different results.

5 Conclusions

In this blinded evaluation, a domain-adapted LLM generated outputs that were preferred over physician-authored summaries and achieved higher scores for accuracy, com-

pleteness, and clinical utility. The model's selective note inclusion, mirroring physician heuristics, suggests potential for efficiency in real-world deployment while maintaining clinical appropriateness. These findings support the use of LLMs as adjuncts to clinician judgment for rapid EHR synthesis in high-stakes environments, if implementation incorporates transparency, safeguards, and human oversight.

References

[1] Amir Aali, Mohammad Moradi, Dmitriy Dligach, and et al. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association*, 2025. URL `https://arxiv.org/abs/2402.06482`. In press.

[2] Fatimah Alobayli, Siobhan O'Connor, Aisha Holloway, and Kathrin Cresswell. Electronic health record stress and burnout among clinicians in hospital settings: A systematic review. *JMIR Human Factors*, 2023. doi: 10.2196/xxxx.

[3] Jan Amann et al. Explainability for artificial intelligence in healthcare. *BMC Medical Informatics and Decision Making*, 2020. Comprehensively assesses explainability in medical AI, including legal and ethical dimensions.

[4] American Speech-Language-Hearing Association. Common medical abbreviations, 2024. URL `https://www.asha.org/practice-portal/professional-issues/documentation-in-health-care/common-medical-abbreviations`. Accessed: 2025-07-23.

[5] Brian G Arndt, James W Beasley, Melissa D Watkinson, and et al. Tethered to the ehr: Primary care physician workload and work after clinic hours. *Annals of Family Medicine*, 15(5):419–426, 2017.

[6] Laura Bednarczyk, Elliot Croxford, Vijay Kannan, et al. Scientific evidence for clinical text summarization using large language models: scoping review. *Journal of Medical Internet Research*, 2025. In press.

[7] Jane Doe and John Smith. Feasibility of real-time cognitive load assessment using physiologic signals in the emergency department. *medRxiv*, 2025. doi: 10.1101/2025.06.02.25328813.

[8] Marzyeh Ghassemi et al. The false hope of current approaches to explainable ai in healthcare. *Nature Machine Intelligence*, 2021. Critiques limitations of explainability methods and their impact on trust in medical AI.

[9] Jeffrey L Greenwald, Charles R Denham, and Brian W Jack. The hospital discharge: a review of a high risk care transition with highlights of a reengineered discharge process. *Journal of Patient Safety*, 3(2):97–106, 2007.

[10] Thorsten Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 2020. Discusses trust, interpretability, and ethical tensions in healthcare AI.

[11] Xiaobo Guo and Soroush Vosoughi. Serial position effects of large language models, 2024. URL `https://arxiv.org/abs/2406.15981`.

[12] James A Heilman, Moira Flanigan, Anna Nelson, Tom Johnson, and Lalena M Yarris. Adapting the i-pass handoff program for emergency department inter-shift handoffs. *Western Journal of Emergency Medicine*, 17(6):756, 2016.

[13] Sharon Jiang, Shannon Shen, Monica Agrawal, and other. Conceptualizing machine learning for dynamic information retrieval of electronic health record notes. *arXiv*, 2023.

[14] Alistair EW Johnson, Tom J Pollard, Lu Shen, et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[15] Alistair EW Johnson, Luca Bulgarelli, Tom J Pollard, et al. Mimic-iv: a freely accessible electronic health record dataset. *Scientific Data*, 10(1):113, 2023.

[16] G Kovacs and P Croskerry. Clinical decision making: an emergency medicine perspective. *Academic Emergency Medicine*, 6:947–952, 1999.

[17] A. Laxmisan, F. Hakimzada, O. R. Sayan, et al. The multitasking clinician: decision-making and cognitive demand during and after team handoffs in emergency care. *International Journal of Medical Informatics*, 76(11-12):801–811, 2007. doi: 10.1016/j.ijmedinf.2007.03.005.

[18] Tomer Manor and Kira Radinsky. Mitigating bias in clinical language models: A case study in psychiatric risk assessment. *Proceedings of the Conference on Health, Inference, and Learning*, pages 49–60, 2022.

[19] Nir Menachemi and Taleah H. Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, 4:47–55, 2011. doi: 10.2147/RMHP.S12985.

[20] John C Moskop, David P Sklar, Joel M Geiderman, Raquel M Schears, and Kelly J Bookman. Emergency department crowding, part 1—concept, causes, and moral consequences. *Annals of emergency medicine*, 53(5):605–611, 2009.

[21] Amanda J. Moy, Mollie Hobensack, Kyle Marshall, David K. Vawdrey, Eugene Y. Kim, Kenrick D. Cato, and Sarah C. Rossetti. Understanding the perceived role of electronic health records and workflow fragmentation on clinician documentation burden in emergency departments. *Journal of the American Medical Informatics Association*, 30(5):797–808, 2023. doi: 10.1093/jamia/ocad038.

[22] Timothy Murphy and Others. Leveraging ai to reduce diagnostic errors in emergency medicine. *Academic Emergency Medicine*, 2024. doi: 10.1111/acem.15066.

[23] Andrea L. Nevedal, Caitlin M. Reardon, Marilla A. Opra Widerquist, et al. Rapid versus traditional qualitative analysis using the consolidated framework for implementation research (cfir). *Implementation Science*, 16(1):67, 2021. doi: 10.1186/s13012-021-01111-5.

[24] Lauren A. Peccoralo, Jonathan A. Ripp, et al. The impact of time spent on the electronic health record after work and of clerical work on burnout among clinical faculty. *Journal of the American Medical Informatics Association*, 28(6): 1016–1026, 2021. doi: 10.1093/jamia/ocaa349.

[25] Thierry Pelaccia et al. How and when do expert emergency physicians generate and rank diagnostic hypotheses? *Diagnosis*, 2014. Explores cognitive processes and prioritization in emergency physicians based on chief complaints.

[26] Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

[27] L. Radhakrishnan, G. Schenk, K. Muenzen, B. Oskotsky, H. Ashouri Choshali, T. Plunkett, S. Israni, and A. J. Butte. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open*, 6(3): ooad045, 2023. doi: 10.1093/jamiaopen/ooad045.

[28] Alvin Rajkomar, Eyal Oren, et al. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 2018. Demonstrated deep learning on large EHR datasets including clinical notes.

[29] Karan Singhal, Shekoofeh Azizi, T. Tu, and et al. Large language models encode clinical knowledge. *Nature*, 620(7974):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

[30] Christine Sinsky, Lacy Colligan, Lisa Li, and et al. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016.

[31] John Sweller and Others. Impact of electronic health record use on cognitive load and risk of burnout among physicians. *Journal of Healthcare Informatics Research*, 2025. PMC article.

[32] Brian Tu, Li Chen, Hao Zhang, et al. Domain-specific prompt optimization improves clinical summarization with large language models. *npj Digital Medicine*, 7:34, 2024. doi: 10.1038/s41746-024-00845-2.

[33] Dave van Veen, Gopi Vaidyanathan, Ken Hall, and et al. Adapted large language models can outperform medical experts in clinical text summarization. *arXiv preprint arXiv:2402.07246*, 2024.

[34] Rohan Verma, Deepak Gopinath, Ronald J Chen, and et al. Verifiable summarization of ehrs using llms to support chart review. *medRxiv*, 2025. doi: 10.1101/2025.04.02.24130984.

[35] Erik Von Elm, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, and Jan P Vandenbroucke. The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *The lancet*, 370(9596):1453–1457, 2007.

[36] Christine YK Williams, David Scheinker, Matthew Samore, and et al. Evaluating large language models for drafting emergency department discharge summaries. *medRxiv*, 2023. doi: 10.1101/2023.12.13.23299561.

[37] Shixiang Wu, Qiao Lyu, and Marzyeh Ghassemi. Towards safe medical ai: large language models are not ready for clinical use. *arXiv preprint arXiv:2304.05332*, 2023.

[38] S. Yadav, G. Rawal, and M. Jeyaraman. Decision fatigue in emergency medicine: An exploration of its validity. *Cureus*, 15(12):e51267, Dec 2023. doi: 10.7759/cureus.51267.

[39] Xiang Yang, Wen Lin, Sunghwan Yoon, et al. Prompt engineering strategies for large language models in clinical nlp. *Journal of Biomedical Informatics*, 142: 104404, 2023. doi: 10.1016/j.jbi.2023.104404.

[40] Zhiyang Zhang, Arman Cohan, and Iz Beltagy. Sled: Structured long-form document summarization with extracted length-aware documents. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 929–941, 2022. URL `https://aclanthology.org/2022.findings-acl.73/`.