Patient Safety Risks from AI Scribes: Signals from End-User Feedback

Jessica Dai JESSICADAI@BERKELEY.EDU
University of California, Berkeley

Anwen Huang ANWEN@BERKELEY.EDU
University of California, Berkeley

Catherine Nasrallah CATHY.NASRALLAH@UCSF.EDU
University of California, San Francisco

Rhiannon Croci RHIANNON.CROCI@UCSF.EDU
University of California, San Francisco

Hossein Soleimani HOSSEIN.SOLEIMANI@UCSF.EDU
University of California, San Francisco

Sarah J. Pollet SARAH.POLLET@UCSF.EDU
University of California, San Francisco

Julia Adler-Milstein JULIA.ADLER-MILSTEIN@UCSF.EDU
University of California, San Francisco

Sara G. Murray SARA.MURRAY@UCSF.EDU
University of California, San Francisco

Jinoos Yazdany JINOOS.YAZDANY@UCSF.EDU
University of California, San Francisco

Irene Y. Chen IYCHEN@BERKELEY.EDU
University of California, San Francisco and University of California, Berkeley

Abstract

AI scribes are transforming clinical documentation at scale. However, their real-world performance remains understudied, especially regarding their impacts on patient safety. To this end, we initiate a mixed-methods study of patient safety issues raised in feedback submitted by AI scribe users (healthcare providers) in a large U.S. hospital system. Both quantitative and qualitative analysis suggest that AI scribes may induce various patient safety risks due to errors in transcription, most significantly regarding medication and treatment; however, further study is needed to contextualize the absolute degree of risk.

Keywords: End-user feedback, AI scribes, patient safety, AI monitoring

Data and Code Availability The data analyzed in this paper is collected as part of ongoing AI monitoring efforts within a large U.S. hospital system; in particular, we study qualitative (text) feedback about a deployed AI scribe product (see Section 1.1). This data is not available to researchers outside of our institution. Our code will not be shared; our quantitative methods are standard and publicly available.

Institutional Review Board (IRB) This studies qualifies as exempt and has undergone limited review by The University of California, San Francisco Institutional Review Board; IRB #25-43915, reference #435961.

1. Introduction

Medical scribing has long been touted as a canonical application for automation by artificial intelligence (AI) systems. Recent advances in AI development have led to a wide range of "AI Scribe" products that claim to realize this vision, and which are now enter-

ing widespread use across the healthcare system (see, e.g., Tierney et al. (2025) for a survey).[1]

Given the nascency of these deployments—and ambiguity about about whether AI scribe products are subject to regulation as medical devices (FDA, 2022)—it is critical to understand the impact of AI Scribe products when applied to real patient encounters. However, while several studies have examined physicians' perspectives (e.g., Tierney et al. (2025); Shah et al. (2025); Duggan et al. (2025)), real-world impacts on patient safety are relatively understudied, with most works utilizing simulated ambulatory encounters (e.g., Anderson et al.; Hose et al. (2025)).

In this work, we seek to understand risks to patient safety by studying feedback submitted by end-users of an AI scribe product deployed in a large U.S. hospital system. Recent work has highlighted end-user feedback as a data source for post-deployment evaluation (Dai et al., 2025); in this work, we leverage feedback both on a per-encounter basis and from a survey of providers about their overall experiences.

Our long-term goal is to design and implement an automated system that can identify safety signals in real time as they arise. This initial work is an exploratory analysis that seeks to understand the extent to which end-user feedback identifies patient safety problems, and to identify considerations for future research and development of such an automated system. To these ends, we take a mixed-methods approach. We begin with standard *quantitative* methods to analyze per-encounter feedback (Section 2), which suggest the presence of errors with clinically-significant impact on patient safety. We thus leverage *qualitative* analysis of complementary survey data to understand these issues in more detail (Section 3).

Overall, we find that per-encounter feedback delivered at point-of-care indicates a clear presence of potential patient safety concerns, especially regarding medication and treatment. This is corroborated by concerns raised in survey responses. However, further study is needed to understand the absolute degree of risk, especially in light of some positive feedback about safety-relevant attributes in survey responses.

1.1. System setup and data sources

The AI scribe products offered by Abridge and a handful of additional vendors were made available to ambulatory providers on an opt-in basis. In this work, we study two main streams of feedback data. The first, analyzed in Section 2, is from ambulatory encounters using Abridge between June 2024 and June 2025 (inclusive); this feedback is delivered at point-of-care within the scribe app, and takes the form of open-ended text about notes generated for a specific encounter. The second, analyzed in Section 3, contains free-text comments from a provider survey sent to all users of AI scribes (including non-Abridge vendors). While the two data sources therefore do not cover identical products and users, our goal in this work is not to make comparative claims across vendors but rather to consider per-encounter clinician feedback as a modality.

2. Quantitative analysis

In this section, we study the per-encounter feedback data as described in Section 1.1. Since this work is exploratory, we focus on simple, out-of-the-box approaches rather than methodological development.

Our main approach is topic modeling, using Sentence-BERT embeddings (Reimers and Gurevych, 2019) with BERTopic (Grootendorst, 2022); to ensure deterministic outputs, we fix a random seed for the latter's underlying clustering model.

In order to focus on patient safety issues, we take a two-phase approach to analyzing the full set of feedback data. In Step 1, we apply the Sentence-BERT/BERTopic pipeline to the entire feedback dataset. The outputs of this initial step included some clusters about non-safety issues (e.g., formatting problems, such as conversions between bullet points and paragraphs, or feature requests). Thus, in Step 2, we remove feedback corresponding to those non-safety clusters, and reapply the clustering pipeline on the remaining "filtered" entries.

Finally, we are also interested in the extent to which modern LLMs can assist in the analysis of this unstructured text feedback. To that end, we label the clusters found by BERTopic both manually and with an LLM summarizer. As a baseline, we also prompt an LLM to generate overall summaries based on the whole collection of feedback.

In total, the data consists of 23,779 unique encounters over 145 total physicians. Of these entries, 365 (1.53% of all encounters) contain text feedback, provided by 29 unique physicians. Of these, 173 entries (47.4% of text feedback) were identified as safety-related in Step 1.

1. Some examples of companies that provide AI scribe products include Abridge, Ambience, DeepScribe, Freed, Nabla, ScribePT, Suki, Tali, and so on.

Cluster	Freq.	Manual Summary	LLM Summary
HPI errors	41 (23.7%)	Incorrect details, specifically in HPI: misplaced diagnostic information, incorrect medication, missing details	Misplaced exam findings, incomplete structure, and blurred lines between HPI and other sections Critical information is often omitted or fabricated within the HPI
Medication errors	34 (19.7%)	Errors in medication dosage and spelling	Incorrect medication names, dosages, taper instructions, or titration schedules Frustration over locked After Visit Summary (AVS) sections that prevent corrections
Speaker details and attribution	28 (16.2%)	Incorrect details from appointment: speaker misattribution, missing symptom discussion, exam results	Statements are often misattributed between patients, caregivers, or providers Details (comorbid conditions, social history, family dynamics) are omitted/inaccurately captured
Sleep-specific	27 (15.6%)	Feedback specific to sleep medicine setting: requests for information like sleep/wake times to be transcribed	Sleep schedules, lifestyle factors, and compliance details are missing in summaries
Pronouns and personalization	17 (9.8%)	Mostly positive feedback; requests to use patient names, correct name misspellings	Misuse of pronouns and failure to use preferred or personalized names Request more conversational language
Surgery-specific	15 (8.7%)	Errors in capturing surgical/treatment discussions, such as surgery risks	Risks, benefits, and alternative surgical options are poorly documented/omitted Surgical plans include fabricated/inaccurate information

Table 1: *Quantitative results: Summaries of BERTopic clusters produced on safety-filtered feedback. LLM summaries have been edited for length; see Table 5 in Appendix A for full LLM outputs.*

Results: Patient safety. Our initial findings suggest the presence of a variety of safety issues. In Table 1, we show the safety-related problems identified by our two-phase clustering approach, with clusters annotated both manually and by GPT-4o.

One prominent safety issue among the feedback submitted at point-of-care is incorrect and missing diagnosis and treatment information. In the History of Present Illness (HPI) section, the scribe is reported to incorrectly record or omit patients' existing conditions as well as to hallucinate conditions the patient does not have; this comprises 23.7% of all safety-related feedback. Another prominent issue is incorrectly recorded medication names, dosages, and instructions (19.7%); for example, the scribe is reported to provide insufficiently detailed instructions on how to taper medication.

The remaining clusters highlight more specific concerns—for instance, speaker misattribution leading to transcription errors specific to the patient's condition (16.2%); missing details in sleep-relevant encounters (15.6%); the omission of surgical risks and benefits in the After-Visit Summary (AVS; 8.7%).

Results: Methodological insights. While simple, our methods in this study highlight three takeaways for future methodological development.

First, heterogeneity in per-user behavior is considerable, and shapes the substantive content of feedback overall. For instance, the presence of a sleep-specific cluster appears to reflect one individual sleep medicine specialist who left a high volume of feedback; it is difficult to conclude that the existence of this cluster implies that the scribe struggles especially with sleep medicine contexts, only that such problems were present. Figure 1 illustrates the wide range of feedback volume and rates across the 29 clinicians who submitted text feedback; meanwhile, the majority of physicians (116 of 145) submitted no feedback at all. We provide additional figures and discussion in A.2, including (negative) results for a heuristic attempt at handling outlier users.

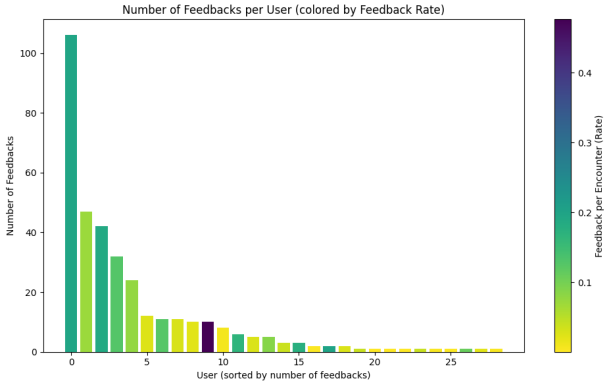

Figure 1: *Heterogeneity in feedback volume (bar chart y-axis) and rates (coloring) per physician. Each bar corresponds to one unique clinician.*

Second, the two-step filtering process, while coarse, is an important step given the nature of observed feedback data. We provide details on this filtering in Appendix A.2. In the first iteration of clustering on unfiltered data, 104 out of the 365 feedback entries are marked as noise; however, when re-clustering the filtered data in Step 2, only 1 out of 173 entries is marked as noise, suggesting that many of the entries initially marked as "noise" in Step 1 were in fact describing safety-relevant topics. Moreover, while the results of clustering in each step were similar thematically, the Step 2 results were clearer; for instance, a more distinct separation is made between errors in AVS (typically corresponding to medication issues), and errors in the HPI section. The full list of clusters and their respective summaries from the unfiltered dataset can be found in Table 3.

Finally, LLMs are reasonably capable of providing summaries of short, free-text pieces of feedback, especially after an initial clustering step that ensures some degree of topical coherence. We show and discuss these results in Appendix A.4.

3. Qualitative analysis

While encounter-level feedback can provide information about specific examples of scribe errors (or successes), physicians' reflections after repeated usage of scribe products can provide information about their broader impressions of AI scribes overall. To this end, clinicians were also given an open-ended survey regarding their usage of AI scribe products. A total of 428 physician responses regarding the use of the AI scribe were analyzed, from 279 unique physicians. Of these physicians, 50 were Abridge users, while the remainder used different AI scribe vendors.[2]

Responses to the open-ended questions were analyzed qualitatively using a multi-stage thematic analysis that incorporated both deductive and inductive approaches (Miles and Huberman, 1994). First, the data were de-identified and reviewed for clarity, relevance, and completeness to inform the development of a preliminary coding framework. Two team members then independently summarized each response and applied a set of inductive codes. Discrepancies were discussed and resolved through consensus meetings with a third coder. Finally, using a grounded approach (Braun and Clarke, 2006), an experienced qualitative researcher reviewed the coded data to generate overarching themes and subthemes.

A total of 428 physician responses regarding the use of the AI scribe were analyzed. Thematic analysis revealed several safety-related themes including accuracy, quality and data completeness. Each subtheme included both perceived strengths and limitations of AI scribe integration in clinical practice (Table 2).

Accuracy. Physicians acknowledged the AI scribe's potential to improve visit documentation and enhance provider efficiency, allowing them to focus more on the patient rather than on documenting patient-reported information. They noted that the AI scribe minimized the likelihood of mistakes and typos that can occur when physicians simultaneously take notes and listen to patients. The AI scribe was recognized in some survey responses for generating accurate clinical documentation, patient information, and AVS, requiring only minor editing.

However, other feedback expressed notable concerns about the accuracy of the data generated by the AI scribe. Some reported incorrect documentation of immunizations; others described clinical hallucinations, where the AI scribe fabricated or misrepresented information regarding diagnoses, physical examination findings, symptoms, dates, medication recommendations, and billing details. Furthermore, the lack of appropriate clinical reasoning in some outputs was highlighted by a few physicians as a critical limitation that could negatively impact clinical outcomes.

Quality. In terms of quality of AI-generated documentation, physicians appreciated the scribe's ability

2. The content of feedback in this survey was very similar across users of different vendors.

Themes	Positives	Negatives
Accuracy	Error reduction; Accurate output; Focus on medical decision-making; Data synthesis	Incorrect immunization documentation; Clinical hallucinations regarding diagnoses, exam, and medications; Lack of appropriate clinical reasoning
Quality	Patient-friendly language; Minimal editing	Extensive editing required; Tense inconsistency; Name confusion and misgendering; Speaker recognition errors when multiple people attend visit
Data completeness	Comprehensive capture of data	Missing data for visit components such as history or exam; Incomplete review of systems

Table 2: *Qualitative results: Themes and sub-themes related to patient safety, with positives and negatives.*

to generate patient-friendly language, which may enhance the clarity and understanding of clinical notes for patients. In addition, a few physicians noted that the AI-generated notes generally required minimal editing, contributing to more efficient workflows.

Conversely, several quality-related concerns that could impact patient safety were highlighted by participants. Some physicians described extensive editing as a burden on both their workload and efficiency of work, which was often necessary to correct errors or improve the notes' readability. Inconsistencies in verb tense affected overall clarity. Physicians also commented on the scribe's confusion with patient names and pronouns, as well as voice recognition inaccuracies that led to incorrect documentation.

Data completeness. Some physicians reported that the AI scribe was effective in ensuring comprehensive capture of clinical data, noting that thorough documentation can support more accurate clinical assessments and continuity of care.

However, notable gaps in data completeness were also reported by several physicians. Incomplete documentation of the review of systems was a frequent challenge, with missing or partially recorded symptoms. Some participants observed missing critical information such as patient names, medical history, and specific test result terminology, which raise concerns about the potential for overlooked clinical details and could compromise the quality of clinical records.

4. Discussion

This work describes initial findings from an analysis of provider feedback about their use of AI scribe products. To the best of our knowledge, our work is the first to study patient safety risks when AI scribes are used in real-world (non-simulated) encounters; though preliminary, our results suggest the importance of monitoring for, and further study of, patient safety issues in real-world deployments.

A secondary contribution is in identifying per-encounter feedback provided at point-of-care as a potentially-fruitful source of data for conducting such evaluations. However, over half of this feedback was non-safety related; thus, more scaffolding for feedback to focus on safety risks could improve data quality. Moreover, as the qualitative results highlight, there is substantial heterogeneity across physicians in their usage and perceptions of the AI scribe. Thus, it is natural to expect that point-of-care feedback also reflects some variation in per-physician behavior (e.g., where some physicians may be more predisposed to submit feedback overall).

In light of this, we emphasize that these initial results should be thought of as signals that safety issues exist in real-world deployments, rather than definitive claims about absolute degrees of risk, especially given the small sample size. However, a better understanding of true prevalence and severity of risk is necessary. Our approaches in this manuscript suggest that future methodological work to this end must explicitly handle the nature of one-sided clinician feedback. This might include, e.g., seeking additional sources of data to supplement problems initially raised in clinician feedback, and/or measuring and handling behavioral heterogeneity across clinicians.

While our data is specific to the deployment of AI scribe products within our institution, we believe that the general principles behind the work apply more broadly: "unknown unknowns" may persist in any AI deployment, and a key challenge of long-run AI adoption in clinical settings is in developing more mature mechanisms for identifying such issues as they arise. We see this work as one step towards this end.

References

Taylor N Anderson, Vishnu Mohan, David A Dorr, Raj M Ratwani, Joshua M Biro, and Jeffrey A Gold. Evaluating the quality and safety of ambient digital scribe platforms using simulated ambulatory encounters. *Available at SSRN 5255300*.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

Jessica Dai, Inioluwa Deborah Raji, Benjamin Recht, and Irene Y Chen. Aggregated individual reporting for post-deployment evaluation. *arXiv preprint arXiv:2506.18133*, 2025.

Matthew J Duggan, Julietta Gervase, Anna Schoenbaum, William Hanson, John T Howell, Michael Sheinberg, and Kevin B Johnson. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. *JAMA Network Open*, 8(2):e2460637–e2460637, 2025.

FDA. *Clinical Decision Support Software - Guidance for Industry and Food and Drug Administration Staff*. U.S. Food and Drug Administration, September 2022. URL https://www.fda.gov/media/109618/download.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Bat-Zion Hose, Jessica L Handley, Joshua Biro, Sahithi Reddy, Seth Krevat, Aaron Zachary Hettinger, and Raj M Ratwani. Development of a preliminary patient safety classification system for generative ai. *BMJ Quality & Safety*, 34(2):130–132, 2025.

Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook*. sage, 1994.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

Shreya J Shah, Trevor Crowell, Yejin Jeong, Anna Devon-Sand, Margaret Smith, Betsy Yang, Stephen P Ma, April S Liang, Clarissa Delahaie, Caroline Hsia, et al. Physician perspectives on ambient ai scribes. *JAMA Network Open*, 8(3):e251904–e251904, 2025.

Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Sarah B Wilson Hannay, Kate Castilla, Cindy S Lau, Patricia Kipnis, Vincent Liu, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. *NEJM Catalyst Innovations in Care Delivery*, 6(5):CAT–25, 2025.

Appendix A. Supplemental materials for quantitative analysis

A.1. Details on filtering step

The initial outputs of our clustering (SentenceBERT+BERTopic) on the full set of data are shown in Table 3. We examined these clusters manually and determined that feedback belonging to clusters -1, 3, 6, 7, 10, and 11, should be removed; these clusters were either marked as noise by BERTopic, uninformative, largely positive, or irrelevant to patient safety concerns.

Cluster	Count	Manual Summary
-1	104	Mix of many issues, including misspellings, missing information, misplaced information, hallucinations, as well as positive feedback
0	34	Incorrect medication plans, dosages, and names
1	29	HPI issues: formatting, misplaced information, incorrect information
2	29	Issues with not using patient names, pronouns, misspellings of physician names
3	28	Positive feedback on summary
4	27	Feedback specific to sleep medicine (likely submitted by one physician)
5	27	Incorrect appointment information: misattributing speakers, missing points of discussion
6	22	Mostly positive feedback on correctly transcribing Restless Legs Syndrome
7	17	One word feedback: "comprehensive"
8	15	Missing surgical/medical discussion details, such as discussions about surgery risks
9	12	Missed diagnoses and medications in the HPI section
10	11	Feedback specific to treating Obstructive Sleep Apnea (OSA)
11	10	Miscellaneous, sentence fragments ("Hi", "Also", "could not yet", etc.)

Table 3: *Summaries of BERTopic clusters (pre-filtering)*

A.2. Details on handling physician heterogeneity

Additional figures illustrating heterogeneity. In Figures 2 and 3, we provide additional illustrations of per-clinician heterogeneity. Figure 2 provides an alternative visualization of the coloring data from Figure 1; on the other hand, 3 shows the types of feedback left by users over the course of all encounters with the scribe. In this set of data, the provision of feedback does not appear necessarily to be strongly correlated with any part of the sequence of encounters (e.g., neither only at the begining of encounters nor only at the end of all encounters).

Condensing feedback volume for outlier physicians. Finally, to attempt to handle outlier physicians who submitted a large quantity of feedback, we implemented a feedback capping procedure prior to the second-round clustering. After removing non-safety clusters Step 1, each physician was limited to a maximum of 10 feedback entries. For physicians with more than 10 entries, we computed sentence embeddings of each of their feedback entries, computed the centroid of these embeddings, and then kept the 10 feedback entries with embeddings that were closest in distance (Euclidean) to the centroid. In total, 5 physicians had their feedback entries capped. We then performed the second round of clustering on this capped data, the results of which are shown in Table 4 below. However, as the contents of Table 4 suggest, this procedure did not meaningfully improve the coherence of the final clusters, nor did it dispel the cluster of sleep-specific feedback.

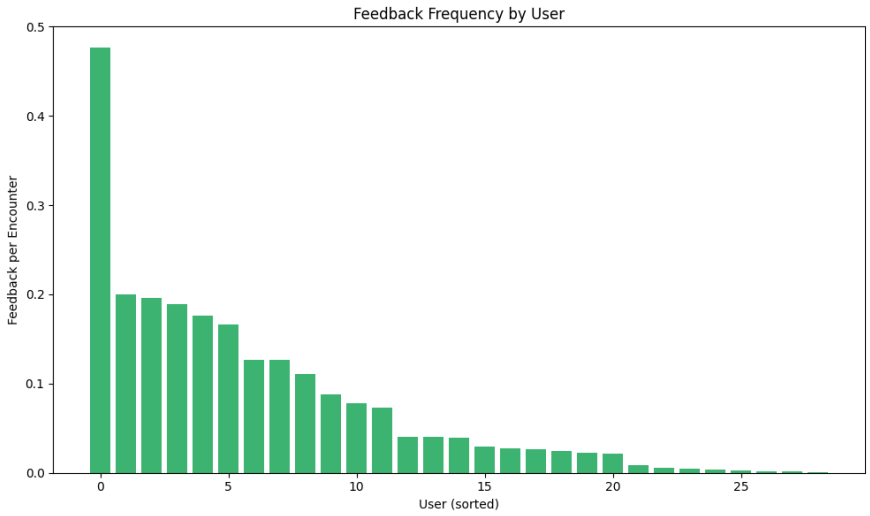

Figure 2: *Feedback rates (i.e. fractions of encounters in which feedback was submitted), per user. Alternative visualization of coloring data from Figure 1.*

Cluster	Count	Manual Summary
-1	17	Issues documenting surgical discussions, inaccurate HPI and A/P
0	28	Mispelling of names, preferences for wording style, inaccurate transcriptions
1	22	HPI issues: formatting, misplaced information, incorrect information
2	21	Incorrect medication names, dosages, and plans
3	13	Omitting details relevant to sleep disorders

Table 4: *Summaries of BERTopic clusters with capped feedback (post-filtering)*

A.3. Details on LLM per-cluster summary

To produce LLM-generated per-cluster summary labels, we concatenated all feedback entries for each cluster into one text chunk, formatted as `Document: [feedback list] | Topic: [cluster label]`.

We then prompted our internal GPT-4o model with the following prompt, passing in the above list.

```
You are a precise analyst. Summarize recurring themes per cluster. Here is a
list structured as Document: Feedback | Topic label. The topic labels are the
different clusters.

[feedback and cluster label list]

Task: Extract 1-2 headline themes for each topic label group (including -1). No
more than 2 themes for each. For each theme, write a short name (2-4 words) and
one-sentence description.
```

The summaries in Table 1 were edited for space; the full outputs are included in Table 5.

Figure 3: *Encounters at which feedback is provided, per user. Blue dots were marked as non-safety related; red were safety-relevant. Gray dots are encounters where no text feedback was provided.*

Cluster	LLM Summary
HPI errors	Disorganized HPI Content: Feedback highlights misplaced physical exam findings, incomplete structure, and blurred lines between HPI and other sections like assessment and plan. Missing or Fabricated Details: Critical information is often omitted or fabricated within the HPI, leading to inaccuracies in patient documentation.
Prescription errors	Medication Documentation Errors: Frequent errors include incorrect medication names, dosages, taper instructions, or titration schedules that require manual correction. Locked Editing Restrictions: Providers express frustration over locked After Visit Summary (AVS) sections that prevent timely updates to medication-related inaccuracies.
Speaker details and attribution	Confused Attribution: Statements are often misattributed between patients, caregivers, or providers, causing inaccuracies in the documented history. Missed Contextual Information: Important details such as comorbid conditions, social history, or family dynamics are frequently omitted or inaccurately captured.
Sleep-specific	Sleep and Lifestyle Omissions: Sleep schedules, lifestyle factors, and compliance details are often missing in summaries, especially for sleep-related visits. Conciseness vs. Detail: Summaries occasionally lack sufficient detail or are overly verbose, requiring edits for clarity and balance.
Pronouns and personalization	Pronoun and Name Usage: Misuse of pronouns and failure to use preferred or personalized names reduces clarity and patient-centeredness in documentation. Improving Narrative Context: Providers request more conversational language and inclusion of social elements like school or lifestyle details for better narrative quality.
Surgery-specific	Incomplete Surgical Details: Risks, benefits, and alternative surgical options are often poorly documented or omitted, leaving gaps in patient care notes. Fabricated or Hallucinated Information: Surgical plans sometimes include fabricated or inaccurate details, requiring significant corrections during review.

Table 5: *Summaries of BERTopic clusters produced on safety-filtered feedback (as shown in Table 1), with full LLM outputs.*

A.4. Details on LLM-overall summary

To produce the overall (non-clustered) LLM summary, we concatenated all feedback entries into one text chunk. We then prompted GPT-4o model with the following:

```
You are a precise analyst.  Summarize recurring themes.  Here are many short
feedback notes:

[list of all feedback]

Task:  Extract 5-10 headline themes across the notes.  No more than 10 themes.  For
each theme, write a short name (2-4 words) and one-sentence description.
```

The output of using this prompt on the unfiltered per-encounter data is shown in Table 6, and on the filtered per-encounter data in Table 7. While the latter is remarkably consistent with Table 1, it appears that the additional coherence induced by filtering on clusters was necessary, as Table 6 is less well-structured.

Topic	Feedback Issues
Medication and Dosing	Medication Errors and Corrections: Frequent issues include incorrect medication names, dosages, titration schedules, or tapering instructions that require manual correction. Locked Sections: Providers report frustration with locked After Visit Summary (AVS) sections, preventing edits to correct medication errors or add necessary details.
HPI Structure and Accuracy	Inaccurate HPI Details: Feedback highlights missing, fabricated, or disorganized HPI content, such as omitted diagnoses, misplaced physical exam findings, or incorrect statements. Improper Section Mixing: Physical exam findings, assessment details, and plans are often erroneously placed in the HPI, leading to confusion and poor organization.
Pronouns and Naming	Pronoun and Name Errors: Misuse of pronouns and incorrect or inconsistent spelling of patient, provider, and medication names are frequent issues that require manual corrections. Clarity in Attribution: Confusion arises from unclear attribution of statements or recommendations, particularly between patient comments and provider instructions.
Summary Quality	Good Summaries with Exceptions: Summaries are generally well-written, but occasional issues include omitted details, overly wordy sections, and disorganized information requiring edits. Interpreter and Translation Challenges: Errors in interpreter-provided information or translated text sometimes lead to inaccurate summaries, especially in multilingual visits.
Sleep and Lifestyle Details	Missing Sleep Schedules: Providers frequently note missing or incomplete documentation of sleep/wake schedules, which are critical for sleep-related visits. Detailed Lifestyle History: Important lifestyle factors, such as caffeine use, alcohol consumption, or positional aids, are often omitted or insufficiently detailed.
Patient History and Context	Missed Historical Context: Critical elements of patient history, such as family dynamics, comorbidities, or social factors, are often overlooked or inaccurately reported. Confusion in Statements: Misattributions of statements between patients, caregivers, and providers frequently lead to errors in the documented history.
Diagnoses and Conditions	Incorrect Diagnoses: Errors include misclassified, omitted, or fabricated diagnoses, as well as incorrect terminology and inconsistent documentation of conditions. Combined or Repetitive Information: Issues arise when related diagnoses, symptoms, or treatments are separated into different sections or repeated unnecessarily.
Comprehensive and Concise Notes	Balancing Detail and Brevity: Feedback highlights the importance of balancing concise documentation with sufficient detail for complex cases or decision-making. Categorization and Organization: Providers prefer clear categorization of content, such as bulleted lists, for easier readability and integration with templates.
Surgical Discussions	Incomplete Surgical Documentation: Risks, benefits, and alternative surgical options are often poorly transcribed or omitted, leaving critical gaps in treatment discussions. Hallucinated or Fabricated Details: Misdocumentation of surgical plans or fabricated information about procedures creates inaccuracies that require extensive corrections.
Assessment and Plan Consistency	Discrepancies Between Sections: Issues arise when information in the assessment and plan contradicts or fails to match the HPI, leading to confusion and miscommunication. Omitted Key Details: Significant omissions in the assessment and plan include prior medication history, referrals, and critical diagnoses discussed during the visit.
Compliance and Follow-Up	Compliance Documentation: Providers request consistent inclusion of compliance statements, such as benefits of CPAP usage, and follow-up instructions for better patient care. OSA and Related Details: Sleep apnea-related discussions often lack sufficient detail, such as treatment options, compliance benefits, or follow-up plans.
General Feedback	Minimal Edits Needed: Notes are occasionally praised for requiring minimal edits, reflecting accurate transcription and logical organization. Positive Overall Impression: Feedback often highlights strong performance in capturing complex histories and providing useful summaries with few errors.

Table 6: *Output of GPT-4o when tasked to identify relevant topics from entire set of (unfiltered) feedback.*

Topic	Feedback Issues
HPI Structure and Accuracy	Disorganized HPI Content: Feedback highlights misplaced physical exam findings, incomplete structure, and blurred lines between HPI and other sections like assessment and plan. Missing or Fabricated Details: Critical information is often omitted or fabricated within the HPI, leading to inaccuracies in patient documentation.
Medication and Dosing	Medication Documentation Errors: Frequent errors include incorrect medication names, dosages, taper instructions, or titration schedules that require manual correction. Locked Editing Restrictions: Providers express frustration over locked After Visit Summary (AVS) sections that prevent timely updates to medication-related inaccuracies.
Patient Context and Attribution	Confused Attribution: Statements are often misattributed between patients, caregivers, or providers, causing inaccuracies in the documented history. Missed Contextual Information: Important details such as comorbid conditions, social history, or family dynamics are frequently omitted or inaccurately captured.
Summary Completeness	Sleep and Lifestyle Omissions: Sleep schedules, lifestyle factors, and compliance details are often missing in summaries, especially for sleep-related visits. Conciseness vs. Detail: Summaries occasionally lack sufficient detail or are overly verbose, requiring edits for clarity and balance.
Personalization and Narrative Style	Pronoun and Name Usage: Misuse of pronouns and failure to use preferred or personalized names reduces clarity and patient-centeredness in documentation. Improving Narrative Context: Providers request more conversational language and inclusion of social elements like school or lifestyle details for better narrative quality.
Surgical Documentation	Incomplete Surgical Details: Risks, benefits, and alternative surgical options are often poorly documented or omitted, leaving gaps in patient care notes. Fabricated or Hallucinated Information: Surgical plans sometimes include fabricated or inaccurate details, requiring significant corrections during review.

Table 7: *Output of GPT-4o when tasked to identify relevant topics from feedback filtered to include only safety-relevant concerns.*