UNIVERSITY OF CALIFORNIA,
IRVINE

Human-Centered Evaluation of Ambient Artificial Intelligence Clinical Documentation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Informatics

by

Yawen Guo

Dissertation Committee:
Professor Kai Zheng, Chair
Professor Yunan Chen
Professor Dan M. Cooper
Associate Professor Daniel Epstein

2026

DEDICATION

To

my parents and friends

for being home in different ways

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

This dissertation reflects the support, mentorship, and collaboration of many people and teams. I am fortunate to have learned from and worked with an extraordinary community throughout this journey.

I am especially indebted to my dissertation chair, Professor Kai Zheng, for his steady guidance and unwavering support. His mentorship and expertise were crucial in making this dissertation possible. I am also grateful to my committee members, Professors Yunan Chen, Daniel Epstein, and Dan M. Cooper, for their thoughtful feedback and supervision. Their perspectives strengthened the rigor of this work and expanded my understanding of clinical informatics research.

I sincerely thank my clinical faculty collaborators and coauthors at the University of California, Irvine Health (UCI Health), including Dr. Deepti Pandita, Dr. Emilie Chow, Dr. Steven Tam, and Dr. Danielle Perret, for their partnership and guidance and for helping ensure that this research remained grounded in real clinical workflows. I also thank the clinicians at UCI Health who participated in my studies and generously shared their time and perspectives.

I deeply appreciate my colleagues for their collaboration, discussions, and day-to-day support. In particular, I thank Dr. Lu He and Dr. Zhaoyu (Nick) Su for their practical advice during my early years in the program. I am also grateful to Di Hu, Yiliang Zhou, Dr. Tianchu Lyu, Rachael Zehrung, and the many colleagues with whom I had the privilege of working throughout my doctoral research.

I acknowledge the institutional support that enabled this work, including the University of California, Irvine, the Department of Informatics, and UCI Health, as well as the teams that supported study coordination, data access, and research operations.

The text of Chapter 2 of this dissertation is an adaptation of the material as it appears in Guo et al. (2026), "Evaluating Ambient Artificial Intelligence Documentation: Effects on Work Efficiency, Documentation Burden, and Patient-Centered Care," published in the *Journal of the American Medical Informatics Association*, 33(2), 273–282, used under Oxford University Press's standing author permission for reuse in a thesis or dissertation. The coauthors listed in this publication are Jiayuan Wang, Di Hu, Steven Tam, Charles Gilman, Emilie Chow, Danielle Perret, Deepti Pandita, and Kai Zheng.

The text of Chapter 3 of this dissertation is an adaptation of the preprint, "From Conversation to Chart: An Analysis of Clinician Edits to Ambient AI Draft Notes," by Guo et al. (2026), posted on medRxiv (CC BY 4.0). This work has subsequently been submitted for publication. The coauthors listed in this publication are Di Hu, Yiliang Zhou, Tianchu Lyu, Sairam Sutari, Steven Tam, Emilie Chow, Danielle Perret, Deepti Pandita, and Kai Zheng.

The text of Chapter 4 of this dissertation is an adaptation of the preprint, "What Do Clinicians Edit in Ambient AI-Drafted Clinical Documentation? A Qualitative Content Analysis," by Guo et al. (2026), posted on medRxiv (CC BY 4.0). A subsequent version of this work was published

in the *Journal of the American Medical Informatics Association* (2026, 33(8), 1457–1465, ocag073). The coauthors listed in this publication are Di Hu, Ziqi Yang, Seungjun Kim, Brian Tran, Jamie Lee, Sitha Vallabhaneni, Rachael Zehrung, Sairam Sutari, Steven Tam, Emilie Chow, Danielle Perret, Deepti Pandita, and Kai Zheng.

The text of Chapter 5 of this dissertation is an adaptation of the preprint, "Understanding Clinician Edits to Ambient AI Draft Notes: A Feasibility Analysis Using Large Language Models," by Guo et al. (2026), posted on medRxiv (CC BY 4.0). A subsequent version of this work has been accepted for publication in the proceedings of the *American Medical Informatics Association* (AMIA) 2026 Annual Symposium and is currently in press. The coauthors listed in this publication are Yiliang Zhou, Di Hu, Sairam Sutari, Emilie Chow, Steven Tam, Danielle Perret, Deepti Pandita, and Kai Zheng.

The text of Chapter 6 of this dissertation is an adaptation of the preprint, "Clinicians' Rationale for Editing Ambient AI-drafted Clinical Notes: Persistent Challenges and Implications for Improvement," by Guo et al. (2026), posted on medRxiv (CC BY 4.0). A subsequent version of this work was published in the *Journal of the American Medical Informatics Association*, 33(7), 1345–1353. The coauthors listed in this publication are Di Hu, Ziqi Yang, Emilie Chow, Steven Tam, Danielle Perret, Deepti Pandita, and Kai Zheng.

My deepest gratitude goes to my parents, my grandmother, and my cousin Yixin Yan, whose love and support mean more to me than I can put into words. I am deeply thankful to my childhood friends Ming Wang, Huwan Peng, Yun Liu, and Yiting Sun, who have always been there for me across the years; to Yi Gu, Ryan Cheng, Ruiye Wang, Yueqing Li, Ce Guo, Jingyan Peng, Jixing Sun, and Roy Wang, whose support has felt like family and who have encouraged me through my lowest moments; and to Shiran Wang, Maggie Yin, Jin Seo Kim, Vicky Zhang, Gary Jiang, Hank Chou, Sarina Chen, Yiyang Min, Jae Yun Kim, and Lucas Lin, who filled my PhD years with companionship and happiness. I am also grateful to my fellow PhD students and dear friends Di Hu, Kana Huang, Jun Zhu, Jiayuan Wang, Andrew Wen, Leo Li, Zekai Fan, and Wei Xuan, for the inspiring and thought-provoking conversations, shared experiences, and the opportunity to learn and grow together.

VITA

Yawen Guo

EDUCATION

September 2026
University of California, Irvine — Irvine, CA, USA
Ph.D. in Informatics

December 2020
Carnegie Mellon University — Pittsburgh, PA, USA
Master of Information Systems Management

July 2019
Beihang University — Beijing, China
B.E. in Industrial Engineering and B.S. in Mathematics

FIELD OF STUDY

Informatics

PUBLICATIONS

Guo Y, Zhou Y, Hu D, Sutari S, Chow E, Tam S, Perret D, Pandita D, Zheng K. *Understanding Clinician Edits to Ambient AI Draft Notes: A Feasibility Analysis Using Large Language Models.* AMIA Annu Symp Proc. 2026. Forthcoming.

Walden A*, Yiu A*, Welk G, **Guo Y**, Radom-Aizik S, Schneider M, Reeds A, Shakir A, Cooper DM. *Pathways to Real World Evidence From the Real World of School-Based Physical Fitness Testing (SB-PFT).* Learn Health Syst. 2026. doi:10.1002/lrh2.70108.

Hu D, Flores D, Flores L, Chien R, Lam K, Chow E, **Guo Y**, Tam S, Perret D, Pandita D, Zheng K. *Ambient AI Documentation in Mixed-Language Encounters: A Heuristic Evaluation of Reenacted Mandarin–English and Spanish–English Clinical Conversations.* AMIA Annu Symp Proc. 2026. Forthcoming.

Zhou Y, **Guo Y**, Sutari S, Dhillon J, Beck AL, Chow E, Tam S, Perret D, Pandita D, Sadigh G, McEligot AJ, Zheng K. *Understanding Stigmatizing Language in Clinical Documentation: A Paired Comparison of Ambient AI Drafts and Clinician Finalized Notes.* AMIA Annu Symp Proc. 2026. Forthcoming.

Zhou Y, **Guo Y**, Hu D, Sutari S, Chow E, Tam S, Perret D, Pandita D, Zheng K. *Examine Clinicians' Modification of Hedging Language in Ambient AI Documentation: A Comparative*

Study of AI Drafts and Final Notes. AMIA Annu Symp Proc. 2026. Forthcoming.

Cho HN, **Guo Y**, Sutari S, Chow E, Tam S, Perret D, Pandita D, Zheng K. *Consumer-to-Clinical Language Shifts in Ambient AI Draft Notes and Clinician-Finalized Documentation: A Multi-level Analysis.* AMIA Annu Symp Proc. 2026. Forthcoming.

Tran BD, Hu D, Kim S, **Guo Y**, Mangu R, Reynolds TL, Lafata JE, Tai-Seale M, Zheng K. *Does Recording Hardware Matter for Clinical Speech Recognition? Evaluating ASR Performance Across Consumer Devices.* AMIA Annu Symp Proc. 2026. Forthcoming.

Guo Y, Hu D, Yang Z, Kim S, Tran B, Lee J, Vallabhaneni S, Zehrung R, Sutari S, Tam S, Chow E, Perret D, Pandita D, Zheng K. *What do clinicians edit in ambient AI-drafted clinical documentation? A qualitative content analysis.* J Am Med Inform Assoc. 2026;33(8):1457–1465. doi:10.1093/jamia/ocag073.

Guo Y, Hu D, Yang Z, Chow E, Tam S, Perret D, Pandita D, Zheng K. *Clinicians' rationale for editing ambient AI-drafted clinical notes: persistent challenges and implications for improvement.* J Am Med Inform Assoc. 2026;33(7):1345–1353. doi:10.1093/jamia/ocag059.

Kim S, Zhou Y, **Guo Y**, Xiao C, Zheng K. *Applying natural language processing and large language models to clinical notes for phenotyping and diagnosing rare diseases: a systematic review.* J Am Med Inform Assoc. 2026;33(6):1225–1235. doi:10.1093/jamia/ocag045.

Guo Y, Hu D, Zhou Y, Lyu T, Yang Z, Reynolds TL, Zheng K. *A scoping review of studies on secure messaging through patient portals: persistent challenges and potential solutions.* npj Health Syst. 2026;3:34. doi:10.1038/s44401-026-00091-2.

Guo Y, Wang J, Hu D, Tam S, Gilman C, Chow E, Perret D, Pandita D, Zheng K. *Evaluating ambient artificial intelligence documentation: effects on work efficiency, documentation burden, and patient-centered care.* J Am Med Inform Assoc. 2026;33(2):273–282. doi:10.1093/jamia/ocaf180.

Hu D, **Guo Y**, Zhou Y, Flores L, Zheng K. *A systematic review of early evidence on generative AI for drafting responses to patient messages.* npj Health Syst. 2025;2:27. doi:10.1038/s44401-025-00032-5.

Zhou S, Xu Z, Zhang M, Xu C, **Guo Y**, Zhan Z, Fang Y, Ding S, Wang J, Xu K, Xia L, Yeung J, Zha D, Cai D, Melton GB, Lin M, Zhang R. *Large language models for disease diagnosis: a scoping review.* npj Artif Intell. 2025;1(1):9. doi:10.1038/s44387-025-00011-z.

Thompson HR, Ricks-Oddie JL, Schneider M, Day S, Argenio K, Konty K, Radom-Aizik S, **Guo Y**, Cooper DM. *Data Missingness and Equity Implications in the Nation's Largest Student Fitness Surveillance System: The New York City School Based Physical Fitness Testing Programs, 2006–2020.* J Sch Health. 2025;95(7):498–509. doi:10.1111/josh.70021.

Guo Y, Hu D, Wang J, Zheng K, Perret D, Pandita D, Tam S. *Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload.* Stud Health Technol Inform. 2025;329:653–657. doi:10.3233/SHTI250921.

Hu D, **Guo Y**, Cho HN, Chow E, Mukamel DB, Sorkin D, Reikes A, Perret D, Pandita D, Zheng K. *When AI Writes Back: Ethical Considerations by Physicians on AI-Drafted Patient Message Replies.* AMIA Annu Symp Proc. 2025;2025:481–489. PMID: 41726446.

Guo Y, Hu D, Zhou Y, Lyu T, Zheng K. *Computational Use of Patient–Provider Secure Messaging Data to Achieve Better Clinical Efficiency and Quality of Communication: A Systematic Review.* AMIA Annu Symp Proc. 2025;2025:403–412. PMID: 41726454.

Guo Y, Hu K, Hu D, Zheng K, Cooper DM. *An Interactive Web Application for School-Based Physical Fitness Testing in California: Geospatial Analysis and Custom Mapping.* AMIA Annu Symp Proc. 2024;2024:463–472. PMID: 40417571. **Best Student Paper Finalist.**

Zehrung R*, Hu D*, **Guo Y**, Zheng K, Chen Y. *Investigating the effects of housing instability on depression, anxiety, and mental health treatment in childhood and adolescence.* AMIA Annu Symp Proc. 2024;2024:1303–1312. PMID: 40417547.

Guo Y, Liu X, Susarla A, Padman R. *YouTube Videos for Public Health Literacy? A Machine Learning Pipeline to Curate Covid-19 Videos.* Stud Health Technol Inform. 2024;310:760–764. doi:10.3233/SHTI231067. (MEDINFO 2023.)

Guo Y, Zehrung R, Genuario K, Lu X, Mei Q, Chen Y, Zheng K. *Perspectives on Privacy in the Post-Roe Era: A Mixed-Methods of Machine Learning and Qualitative Analyses of Tweets.* AMIA Annu Symp Proc. 2023;2023:951–960. PMID: 38222378.

Guo Y, Zhu J, Huang Y, He L, He C, Li C, Zheng K. *Public Opinions toward COVID-19 Vaccine Mandates: A Machine Learning-based Analysis of U.S. Tweets.* AMIA Annu Symp Proc. 2022;2022:502–511. PMID: 37128441. **Second Place, Student Paper Competition.**

Guo Y, Liu X, Susarla A, Padman R. *YouTube Video Analytics for Patient Engagement: Evidence from Colonoscopy Preparation Videos.* Workshop on Information Technologies and Systems (WITS). 2020.

ABSTRACT OF THE DISSERTATION

Human-Centered Evaluation of Ambient Artificial Intelligence Clinical Documentation

by
Yawen Guo
Doctor of Philosophy in Informatics
University of California, Irvine, 2026
Professor Kai Zheng, Chair

Ambient artificial intelligence (AI) documentation tools convert patient–clinician conversations into draft clinical notes that clinicians review and modify before signing in the electronic health record (EHR). These systems are increasingly adopted as a response to documentation burden and clinician burnout, but real-world effectiveness depends on the amount of clinician revision required for AI drafts and on persistent barriers that limit uptake and sustained use. This dissertation presents a human-centered, multi-level evaluation of ambient AI documentation conducted within the University of California, Irvine Health (UCI Health).

First, the implementation of ambient AI documentation was evaluated using a mixed-methods approach that paired EHR time-efficiency measures with pre- and post-implementation survey responses from outpatient clinicians at UCI Health across multiple specialties. Results indicated that time spent on documentation decreased, while note length increased during the three-month period following implementation. Survey findings aligned with these trends, reflecting improved perceived efficiency and reduced documentation burden.

Second, clinicians' edits of ambient AI drafts were quantified at scale using AI draft and clinician-finalized note pairs. Across notes that contained one or more ambient AI–generated note sections, 84.4% were revised by clinicians before signing. Editing intensity was highest in

the Assessment and Plan section, while the tendency to accept drafts without edits varied primarily across individual clinicians rather than across specialties.

Third, a qualitative content analysis was conducted using a manually annotated corpus of sentence-level edits derived from draft and final note pairs. The analysis identified recurring patterns in clinicians' revisions to AI drafts, including adding specialty-specific details, calibrating certainty to reflect available evidence and diagnostic framing, and converting transcript-like narratives into professional documentation. Using this annotated dataset, few-shot prompting of large language models (LLMs) was evaluated to assess the feasibility of scalable automated edit measurement across five target categories. Results indicated that LLM-assisted classification was most reliable when the revision could be inferred from the local before-and-after text alone and less consistent when accurate classification required broader clinical context.

Finally, semi-structured interviews with clinicians who use ambient AI in routine practice were conducted to understand the rationale behind editing AI-drafted notes. Clinicians described editing as a deliberate process to ensure accuracy, align drafts with clinical reasoning and documentation standards, and meet billing and medico-legal requirements. Interviews further underscored the need for specialty-aware customization and workflow support for efficient draft review and verification. Recommendations emphasized system-level optimization across model behavior, individual personalization, workflow design, and EHR integration.

Collectively, these findings highlight that effective ambient AI implementation must align with clinical documentation standards, support specialty-specific workflows, and enable clinicians to efficiently verify, refine, and structure notes. By clarifying how ambient AI is used

in routine practice, what revisions clinicians make, and why those revisions occur, this dissertation provides practical evidence and guidance to inform implementation, continuous monitoring, and iterative design improvements, ultimately aiming to reduce documentation burden while maintaining high-quality clinical records.

CHAPTER 1: Introduction

Clinical documentation lies at the core of healthcare communication and clinical decision-making[1,2]. Over the past two decades, the widespread adoption of electronic health record (EHR) systems has transformed how clinical encounters are recorded and shared[3,4]. However, documentation burden has also become a leading source of clinician burnout[5–9]. As clinicians spend increasing amounts of time navigating EHR interfaces and reconciling data, documentation has become a major driver of workload, crowding out time for patient care and reducing job satisfaction [9].

In response to these challenges, efforts have aimed to improve documentation efficiency without compromising the fidelity of clinical communication. This has driven successive generations of digital tools, from early dictation and templating systems to natural language processing (NLP)-based approaches and, more recently, ambient artificial intelligence (AI) scribes [10–12]. Ambient AI scribes leverage automatic speech recognition (ASR) and large language models (LLMs) to draft clinical notes from real-time encounters, aiming to reduce documentation burden and improve workflow efficiency [13–15]. Although early studies report time savings and improved clinician satisfaction, many evaluations continue to emphasize EHR time-efficiency metrics such as reduced after-hours charting, rather than directly assessing the accuracy and completeness of AI-drafted notes [13,14,16–24].

At the same time, ambient AI systems introduce new risks, including factual errors, contextual omissions, and inconsistencies that may affect clinical interpretation [25–27]. The quality of clinical notes directly influences patient safety and care quality. When AI-drafted documentation is incomplete or inaccurate, even small errors can propagate through the EHR and

affect downstream functions such as clinical decision support and billing accuracy. Moreover, if clinicians must spend substantial time revising AI drafts, the promised efficiency gains may be diminished, potentially increasing documentation burden rather than reducing it.

Little is known about how clinicians use these systems in everyday practice, including the extent of revision, the types of changes they make, and the reasons behind those edits. Without granular, content-level evidence, it is difficult to determine how clinician oversight shapes the final record and which system and workflow improvements are needed to support high-quality documentation. Understanding clinician perceptions and editing behavior is therefore essential for responsible adoption and can guide both technical refinement and implementation strategy to reduce documentation burden.

This dissertation addresses these gaps through a series of interrelated studies examining how ambient AI documentation tools are used and adapted in clinical practice, and by developing evaluation methods that reflect multi-level evidence from clinician perceptions, editing behavior, and note content. First, a mixed-methods evaluation combines EHR time-efficiency measures with clinician surveys to characterize perceived benefits, workflow impacts, and priorities for improvement. Second, a large-scale analysis of paired AI-drafted notes and clinician-finalized versions quantifies editing frequency and intensity across clinician groups, specialties, and note sections. Third, a qualitative content analysis examines what clinicians change in AI-drafted notes and characterizes the nature and recurring patterns of those revisions. Fourth, a feasibility study develops and evaluates LLM-based methods to classify edit types as a step toward scalable documentation quality monitoring. Finally, semi-structured interviews investigate the rationale behind edits and elicit recommendations for improvement across model behavior, personalization, workflow design, and EHR integration. Together, these

studies link clinician perceptions, observable editing behavior, and scalable computational assessment to inform responsible design, evaluation, and implementation of ambient AI documentation in healthcare.

Research Questions

The following research questions guide the dissertation and structure the five studies that follow.

1. How do clinicians perceive the impact of ambient AI on documentation efficiency, workload, and patient care?
2. To what extent do clinicians revise ambient AI drafts, and how do editing frequency and intensity vary across individual clinicians, specialties, and note sections?
3. What changes do clinicians make to AI-drafted notes, and do these changes follow systematic patterns?

    3a. How feasible is it to use computational methods (e.g., LLM-based detectors) to measure edit behaviors at scale for ongoing monitoring and improvement?

4. What are clinicians' rationales for revising ambient AI drafts, and what do they believe is needed for effective ambient AI use?

Historical and Technical Overview of Digital Documentation

The passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act and the widespread adoption of EHR systems fundamentally changed how clinical work is documented [3,4]. By the early 2010s, many health systems turned to medical scribes as a strategy

for addressing documentation burden. Human scribes documented encounters in the EHR for clinicians, alleviating documentation burden while introducing new challenges related to workflow coordination and cost [28,29]. Both in-person and remote scribe programs were costly to sustain, variable in quality, and often affected by high turnover. To find more scalable alternatives, healthcare organizations increasingly turned to digital documentation tools. Early approaches included speech-recognition dictation, EHR templates and macros, and NLP-based documentation support systems[10,11,30]. The following sections outline the historical development of these technologies and the limitations that set the stage for today's ambient AI systems.

Voice Dictation Software. As EHR adoption accelerated following the HITECH Act, voice-based dictation became an important digital approach to reducing documentation burden. Speech recognition software, most notably Nuance's Dragon Medical, enabled clinicians to convert spoken language directly into text within the EHR[31]. By the early 2010s, such tools were in use across major health systems, offering a way to capture the narrative of clinical encounters while improving efficiency compared to manual typing [32,33]. However, dictation results required careful review to ensure accuracy, particularly in the presence of accents, background noise, or specialized terminology. Many clinicians found that the process of dictating and editing long notes could feel as laborious as typing[34]. Despite these limitations, speech recognition technologies laid essential groundwork for later innovations, establishing both the technical and cultural foundation for more advanced, context-aware intelligent documentation systems.

EHR Templates and Macro Systems. In parallel with speech recognition, clinicians increasingly relied on EHR templates and macro systems to assist documentation [10]. Major EHR platforms such as Epic and Cerner introduced tools that allowed text expansion, automated data insertion, and reusable templates for common encounter types [35]. Features like "dot phrases" enabled

clinicians to populate large portions of the record, often pulling laboratory results, vital signs, prior assessments, and historical information directly from the EHR [36]. These tools substantially increased documentation efficiency and promoted standardization across encounters. However, the widespread use of copy-forward and default text led many clinical notes to expand with redundant or outdated information, contributing to "note bloat" [37].

Early NLP-based Documentation Assistants. By the late 2010s, many NLP-based approaches were developed to analyze clinical notes and improve documentation specificity or completeness. Computer-assisted clinician documentation tools, for instance, could identify missing details such as laterality, severity, or diagnostic qualifiers, and prompt clinicians to clarify them[38]. These tools functioned as intelligent reviewers in the documentation process, guiding clinicians to produce more accurate and billable notes. Related applications, such as computer-assisted coding, used NLP to extract clinical concepts from notes and suggest billing codes, helping streamline administrative workflows[39]. In addition, voice-enabled virtual assistants began incorporating NLP to support basic EHR workflows [34]. Clinicians could issue spoken commands to open templates, pull up prior results, or navigate within the chart, reducing manual clicks and keystrokes. However, these systems still depended on structured inputs and explicit dictation to capture context, rather than independently summarizing clinical narratives.

Evolution Toward the Ambient Artificial Intelligence Era. In the early 2020s, ambient AI marked a shift from assistive to generative documentation [25-27]. Unlike earlier NLP-based approaches that processed text after clinicians entered it, ambient systems generate draft notes during the encounter. This transition was enabled by several converging advances, including improved medical speech recognition, the availability of large clinical text corpora for training

language models, and lessons learned from earlier generations of computer-assisted documentation and coding tools.

Related Work

As the adoption of ambient AI technologies in clinical practice accelerates, a growing body of research has begun to evaluate their impact on clinician experience, documentation efficiency, and note quality. This related work section is organized into four parts. The first section synthesizes empirical studies examining the deployment of ambient AI scribes across both real-world and simulated clinical environments. The second reviews technical evaluation frameworks and benchmarking efforts for ambient documentation tools. The third expands the technical scope to NLP methods for text comparison and quality assessment, providing conceptual background for the computational approaches later employed in this dissertation. The final section integrates insights across the above domains to highlight persistent challenges and directions for developing reliable, clinically grounded evaluation frameworks for ambient AI systems.

Empirical Evidence on Ambient and Generative AI Documentation in Clinical Practice. Early evaluations of ambient AI demonstrate consistent improvements in documentation efficiency. Pilots in primary care settings have shown reduced time spent writing notes and shorter after-hours charting periods, with specialty clinics such as dermatology, orthopedics, and cardiology reporting similar gains [1,12,16,18,23,40]. Large multispecialty implementations have also noted decreases in "pajama time" and smoother in-visit workflows as clinicians become more accustomed to the technology [41]. However, early studies have also reported longer notes after AI implementation, suggesting that the tool may re-introduce "note bloat" [18,23,24,42]. Surveys

assessing clinician experience generally show improvements in perceived usability, patient engagement, and task load (NASA-TLX) scores following ambient AI deployment. Several studies have also reported reduced burnout and emotional disengagement among frequent users, though others found no statistically significant changes in burnout indicators despite clinicians reporting greater control over documentation time [13,17,19,22,23,42]. Productivity outcomes such as work relative value units (wRVUs) and Current Procedural Terminology (CPT) code submission rates show little to no measurable improvement in fee-for-service payment settings [19]. Cost also varies by vendor and deployment model, and some programs report lower monthly costs than human scribes once onboarding costs are amortized [25]. Patient-reported experiences are generally positive, but evidence to date comes from limited samples, and few studies have linked documentation-related changes to downstream indicators of care quality or patient safety [12,18,19]. These mixed findings likely reflect differences in baseline workload and specialty context, and they also suggest that ambient AI is still in an early stage of deployment with substantial opportunities for usability and workflow optimization.

Evaluation Frameworks and Benchmarks for Ambient Documentation. A small but growing set of benchmark datasets has emerged for end-to-end evaluation of ambient AI documentation systems. ACI-Bench provides over two hundred simulated clinical encounters accompanied by both automated and expert-review metrics [43]. MTS-DIALOG offers a multi-specialty corpus derived from real clinical notes to facilitate research on note generation and factual consistency [44]. Other task-specific datasets such as the emergency department consultation corpus for referral summarization further expand the available benchmarks [45]. However, publicly available datasets largely rely on simulated encounters, while real-world datasets are usually investigated internally within healthcare institutions and are not shared publicly because of privacy constraints [12].

Evaluation strategies typically combine automated text metrics with expert rating instruments. Most human assessments of clinical note quality draw from two principal frameworks: the *Physician Documentation Quality Instrument* (PDQI-9) [46–48] and the *Sheffield Assessment Instrument for Letters* (SAIL) [49]. Automated evaluation commonly uses lexical overlap (e.g., ROUGE), semantic similarity (e.g., BERTScore), learned reference-based metrics (e.g., BLEURT), and medical-domain metrics such as MEDCON (UMLS concept F1) and fact-based measures like Fact-Core and Fact-Full that score extracted clinical facts and modifiers (e.g., negation, status, hedge, laterality, body site) [50–52]. Although these hybrid frameworks provide an important methodological foundation for clinical note quality assessment, prior studies still point to several limitations. Few evaluations link linguistic metrics with clinical correctness or meaningfulness; key assessment dimensions are defined inconsistently across studies, and expert-rating frameworks remain subjective and difficult to apply at scale.

Methods for Text Comparison and Quality Assessment. Many studies quantify textual differences between system output and human reference using token- or character-level comparison methods. Classical measures such as Levenshtein distance, longest common subsequence, and Ratcliff–Obershelp overlap estimate the extent of insertions, deletions, and substitutions between two text versions [53,54]. Cosine similarity or Jaccard overlap on bag-of-words representations provides a vector-space view of lexical agreement [55,56]. In clinical contexts, these measures have been adopted as indicators of editing magnitude or stability among note versions. Evidence from machine-translation post-editing suggests that higher edit distances correlate with lower perceived output quality and greater human revision effort [57,58]. In addition, indicators such as token or sentence counts, section order alignment, duplication ratios, and measures of semantic novelty are also essential to note quality evaluation [46,59–62]. Studies

applying these metrics find that AI-generated documentation can achieve linguistic coherence comparable to human-written notes but still exhibits measurable factual gaps, particularly in medication details, temporality, and attribution [1,63].

Challenges and Opportunities for Standardized Evaluation. Despite rapid progress in ambient AI documentation, evaluation practices remain fragmented and methodologically inconsistent. A key challenge is that automated NLP metrics often do not align with clinically meaningful measures of note quality [25,64]. Bracken et al. observed that while most studies report improvements in documentation efficiency, quality-assessment frameworks vary widely from standardized instruments such as PDQI-9 and SAIL to locally developed Likert-style scales [64]. Recent studies have called for the development of standardized, hybrid evaluation frameworks that integrate NLP metrics with clinically grounded quality rubrics [24,26]. Such frameworks would enable reproducibility while maintaining clinical validity by pairing automated factuality detection with expert review.

Taken together, these challenges point to a lack of clinically interpretable, scalable, and reproducible approaches for evaluating ambient AI documentation in real-world settings. Closing this gap requires moving beyond surface text similarity to understand how clinicians review, modify, and validate AI-drafted notes in practice. This dissertation addresses this need by integrating evidence from clinician perceptions, observed editing behavior, and note content to enable evaluation that is both clinically meaningful and operationally scalable.

CHAPTER 2: Evaluating Ambient AI Documentation: Effects on Work Efficiency, Documentation Burden, and Patient-Centered Care

Abstract

Background and significance

Ambient listening tools powered by generative artificial intelligence (GenAI) offer real-time, scribe-like support that reduces documentation burden and may help alleviate burnout. This study assesses physician-perceived benefits and challenges of ambient AI implementation through surveys and evaluates its effectiveness in clinical workflows using automatically recorded electronic health record (EHR) time-efficiency metrics.

Method and materials

A quality improvement pilot has been underway at UCI Health since December 2023. Epic EHR usage metrics were analyzed to assess changes in note length, documentation time, and same-day encounter closure rates. Matched pre- and post-implementation surveys evaluated physician-perceived changes in documentation burden, clinical efficiency, and care quality. We also examined open-ended survey responses using thematic analysis to supplement quantitative findings.

Results

Analysis of EHR usage data from 167 physicians showed significant reductions in note-writing time, despite an increase in note length. Survey responses (n = 65) also indicated statistically significant improvements across multiple domains. Physicians reported reduced cognitive demand ($p = 0.031$) and documentation effort ($p = 0.014$), alongside perceptions of enhanced clinical efficiency, patient-centered care, and EHR system usability. Thematic analysis confirmed

these quantitative findings and identified opportunities for improvement, including specialty-specific customization and expanded AI functionality.

Discussion and Conclusion

Ambient AI tools demonstrated improved documentation efficiency, perceived care quality, and reduced cognitive workload. Future development should prioritize customization for specialty-specific and individual physician needs, ensure the reliability and accuracy of AI-generated content, and integrate ethical and legal considerations to facilitate safe and scalable implementation in patient-centered care contexts.

Background and Significance

The Health Information Technology for Economic and Clinical Health (HITECH) Act accelerated the widespread adoption of electronic health records (EHRs) and emphasized their meaningful use [3,4]. Although EHRs have benefited many aspects of clinical workflow, their rapid implementation has also significantly increased physicians' documentation burden [5–7]. This added workload is closely associated with clinician dissatisfaction and has emerged as a major contributor to burnout, which affects not only clinician well-being but also the quality and safety of patient care [8,65].

In an effort to reduce this burden, early solutions included digital scribes and voice dictation tools designed to automate routine note-taking [66]. These systems predominantly relied on natural language processing (NLP) techniques for tasks such as clinical entity extraction, text classification, and summarization [67]. While promising, these approaches often required manual correction and lacked the capacity to fully capture clinical nuance and context. More recently, ambient listening technologies have marked a significant advancement over traditional tools [12,41].

By integrating large language models (LLMs) and generative artificial intelligence (GenAI), these systems can passively capture patient-clinician conversations and generate tailored, context-aware clinical notes.

Many large health systems in the United States have begun piloting ambient listening tools, prompting a growing body of research evaluating their impact on documentation efficiency, clinician burnout, patient experience, and adoption factors. Studies such as simulated outpatient consultations, the STREAMLINE pilot, and implementations in dermatology clinics have reported reduced documentation time and EHR interaction, along with improvements in note quality and clinician satisfaction [1,16,18,40]. Paired surveys further demonstrated reductions in task load and burnout, as well as improvements in usability, work-life integration, and patient engagement among physicians [13,19–21]. Additionally, qualitative interviews revealed clinician enthusiasm for ambient AI's potential to enhance workflow and care delivery, while identifying key areas for refinement and better integration into clinical practice [22]. Observational studies incorporating patient feedback have also shown general agreement with the use of ambient AI tools in the clinical setting [23].

Building on this evidence base, our study contributes to the ongoing evaluation of ambient AI as part of a quality improvement pilot project at the University of California, Irvine (UCI) Health system in ambulatory settings. Using matched pre- and post-implementation surveys alongside Epic EHR time data, which automatically capture system usage logs, we assess changes in documentation time and length, as well as physician perceptions of ambient AI scribes' impact on documentation burden, clinical efficiency, and care quality. We hypothesize that integrating ambient AI tools will improve work efficiency, reduce physician workload, and

help mitigate burnout, offering critical insights into the scalability and sustainability of these technologies in routine clinical practice.

Method and Materials

Study setting and pilot recruitment

A pilot study was launched in December 2023 at the University of California, Irvine's healthcare system (UCI Health), a major academic health system with broad regional coverage across Southern California, including locations in Irvine, Newport Beach, Orange, Fountain Valley. In partnership with Epic Systems Corporation, UCI Health recruited physicians from both primary care and specialty care clinics to implement two commercially available ambient listening tools: DAX Copilot (Nuance Communications, Inc., Burlington, MA, USA) and Abridge (Abridge AI, Inc., Pittsburgh, PA, USA) [14,15]. Clinicians accessed these ambient listening tools primarily via the Epic Haiku mobile application or an integrated web browser-based approach. The tools automatically generated draft clinical notes parsed into note sections such as physical exam, assessment, and follow-up plan. Physicians then review, edit, and finalize the notes before submitting them to the EHR system. Patients provided informed consent prior to the use of ambient listening tools during their clinical visits. Clinicians accessed these ambient tools primarily through Epic Haiku (mobile app integration) or via a web browser-based approach (Abridge Integrated), which required additional manual steps, such as switching between browser windows and Epic interfaces. These differences in workflow integration emerged as notable considerations in provider feedback.

The pilot remains ongoing at the time of this analysis. The data analyzed in this study were collected between December 2023 and April 2025. The project met institutional criteria for

quality improvement and was exempt from IRB-mandated informed consent. All reporting followed the Standards for Quality Improvement Reporting Excellence (SQUIRE) guidelines.

EHR Metrics Analysis

Automatically recorded Epic EHR metrics were analyzed from physicians who used ambient listening tools for a minimum of four weeks during the pilot period and generated at least 10 clinical notes. Paired comparisons were conducted to compare pre- and post-intervention means and medians across five Epic metrics: same-day encounter closure rate, average number of characters per note, average time spent per note, total daily documentation time, and time spent documenting per appointment. Metrics were extracted across all types of notes generated within each appointment. It is important to note that Epic time-efficiency metrics may not precisely reflect active writing duration, as they capture general user interaction time with the EHR system, including pauses and idle periods. As such, recorded time may exceed the actual time spent composing documentation.

Baseline values, representing pre-implementation usage, were calculated as the average of each metric during the three months preceding the introduction of ambient listening tools. These were compared with cumulative post-implementation averages at one, two, and three months following the pilot start. Normality of metric distributions was assessed using the Shapiro–Wilk test [68]. Outliers were excluded using the Interquartile Range Method. When normality assumptions were met ($p>0.05$), paired t-tests were applied; otherwise, the non-parametric Wilcoxon signed-rank test was used.

Pre- and Post-Implementation Survey Analysis

Physicians who participated in the ambient listening pilot received pre-implementation and post-implementation surveys approximately four months apart. The timing reflects the survey

administration schedule and does not necessarily correspond to the exact duration of tool usage. The surveys were adapted from the KLAS Arch Collaborative's standardized instrument and included a combination of 5-point Likert-scale items (ranging from strongly disagree to strongly agree), 1–10 scale ratings, and open-ended questions [69]. Survey items assessed perceived changes across domains such as patient safety and care experience, clinical efficiency, EHR training and implementation support, system usability and reliability, and documentation burden.

Only participants who completed both the pre- and post-surveys were included in the paired analysis to enable direct comparison. We analyzed survey responses across three groups: all participants, primary care physicians, and specialty care physicians. Wilcoxon signed-rank tests were used to assess changes in Likert-scale responses and 1–10 scale items [70]. We also used difference-in-differences tests to compare whether the magnitude of change differed between primary care and specialty care providers[71].

To supplement the quantitative results, an exploratory qualitative analysis was conducted on open-ended responses collected at the end of each survey, where respondents could voluntarily provide additional comments. Due to the limited volume of open-ended question responses, thematic coding was performed by a single reviewer[72]. These qualitative findings were intended to provide additional context and insights rather than serve as a matched pre-post comparison.

Results

A total of 319 physicians participated in the ambient listening tool pilot. Participants included those in primary care, various specialty care disciplines, and medical trainees. For the EHR metrics analysis, only participants who met the inclusion criteria, including a minimum duration

of ambient tool usage and sufficient note volume, were included. A subset of enrolled physicians completed both the pre- and post-implementation surveys and were included in the corresponding survey analysis.

EHR Metrics Analysis Results

A total of 167 participants met the inclusion criteria for EHR metrics analysis, defined as using the ambient listening tool for more than four weeks and generating at least 10 clinical notes. After removing outliers and missing data to ensure completeness across all five Epic metrics, final sample sizes varied by metric (Table 1).

Table 1. Changes in EPIC Documentation Metrics Averaged Over a 3-Month Baseline and Cumulative 1–3 Month Post-Pilot Periods

EPIC Signal Metrics	**Avg. chars/note**	**Avg. time/note**	**Avg. time/day**	**Avg. time/ appointment**	**Same-Day Note Closure Rates** [a]
	N = 110	**N = 107**	**N = 112**	**N = 99**	**N = 26**
Baseline (3M) *Mean (SD)*	7522.29 (3069.95)	6.37 (3.30)	55.97 (29.20)	9.91 (5.95)	0.77 (0.30)
1-Month Post *Mean (SD)*	7828.45 (3236.92) **	5.60 (2.94) ***	50.00 (26.86) ***	8.90 (5.49) ***	0.79 (0.32)
2-Month Post *Mean (SD)*	7863.74 (3182.98) ***	5.57 (2.89) ***	49.73 (26.16) ***	8.84 (5.45) ***	0.79 (0.32)
3-Month Post *Mean (SD)*	7862.85 (3154.88) ***	5.58 (2.90) ***	49.13 (26.19) ***	8.82 (5.41) ***	0.78 (0.32)

[a] *Analysis includes only providers with an average of ≥40 notes per week during the study period.*
*Statistical significance: * = p<0.05, ** = p<0.01, *** = p<0.001*

Significant reductions in documentation time were observed following the implementation of the ambient listening tool. The average time spent writing a note decreased from 6.37 minutes at baseline to 5.58 minutes over the 3-month post-pilot period ($p < 0.001$). Similarly, total daily documentation time per provider declined from 55.97 minutes to 49.13 minutes ($p < 0.001$), and documentation time per appointment dropped from 9.91 to 8.82 minutes ($p < 0.001$). In contrast, note length slightly increased post-implementation. The average number of characters per note

rose from 7,522 to 7,862, representing a statistically significant but modest increase ($p < 0.001$). For same-day encounter closure data, we had access to a limited 12-month dataset. To ensure meaningful analysis, we included only physicians who averaged more than 40 encounters per week during the study period. No statistically significant change was observed in same-day encounter closure rates across the baseline and post-implementation periods. The four Epic metrics that demonstrated statistically significant changes are visualized in Figure 1. These results show immediate reductions in documentation burden following pilot implementation, followed by relative stabilization over the subsequent months.

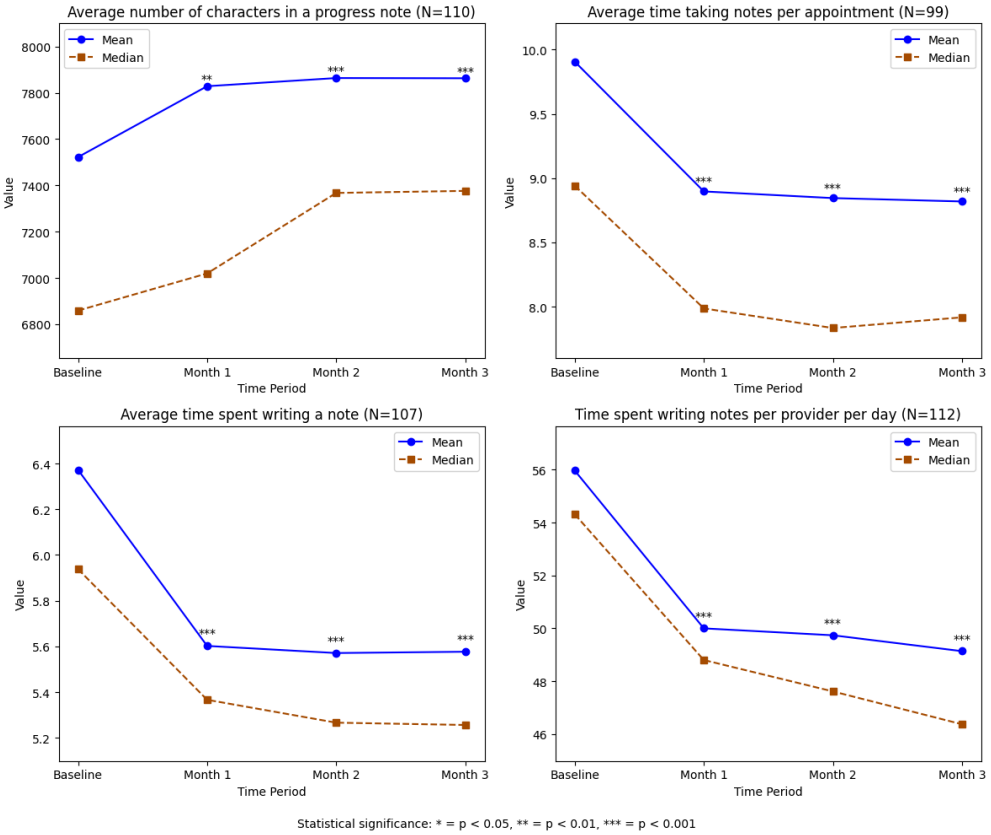

Figure 1. Changes in EPIC Signal Metrics: 3-Month Baseline vs. 3-Month Cumulative Post-Pilot Period

Pre- and Post-implementation Survey Analysis Results

A total of 127 physicians completed the pre-implementation survey, and 133 completed the post-implementation survey. Surveys were anonymous but included a unique identifier that allowed for matching pre- and post-responses. Among the respondents, 65 clinicians completed both surveys and were included in the paired analysis. Table 2 summarizes the clinical characteristics of these 65 participants. Most were attending physicians (n = 44, 68%), followed by physician residents or fellows (n = 16, 25%) and advanced practice providers (n = 5, 8%). Clinical specialties included both primary care (n = 34, 52%) and a range of specialty care fields (n = 31, 48%). Participants reported varied experience with EHRs, with the largest group using EHRs for 6–9 years (n = 24, 37%). Use of ambient listening tools was distributed between Abridge (n = 41, 63%) and Nuance DAX Copilot (n = 24, 37%).

Table 2. Descriptive Characteristics of Physician Survey Participants

	Clinicians, No. (%)	
	Count	**Percentage**
Characteristic Grouping	**N=65**	**%**
Ambient Listening Technology		
Abridge	41	63%
Nuance: DAX Copilot	24	37%
Care Type		
Primary Care[a]	34	52%
Specialty Care[b]	31	48%
Clinical Background		
Advanced practice provider (PA or NP)	5	8%
Physician	44	68%
Physician resident or fellow	16	25%
Years Practicing Medicine		
0–4 years	9	14%
15–24 years	14	22%
25+ years	8	12%

5–14 years	32	49%
Years Using EMR		
Less than 1 year	12	18%
1–2 years	8	12%
3–5 years	20	31%
6–9 years	24	37%
Average Hours of Clinical Practice		
20–39 hours per week	23	35%
40–60 hours per week	29	45%
60+ hours per week	7	11%

[a] *Family medicine, Geriatrics, Internal medicine*
[b] *Allergy and immunology, Dermatology, Emergency medicine, Endocrinology, Gastroenterology, Obstetrics and Gynecology, Hematology/oncology, Hospice/palliative medicine, Infectious disease, Neurology, Neurosurgery, Orthopedics, Otolaryngology, Physical Medicine & Rehabilitation, Psychiatry, Rheumatology*

Below we present paired statistical analyses of pre- and post-implementation surveys assessing the impact of ambient listening tools on physicians' experiences with EHRs. Survey questions included 5-point Likert-scale items, 0–10 numeric rating scales, and percentage-based estimates (Table 3). The results are organized by question domains, emphasizing significant pre- and post-implementation changes and highlighting differences between primary and specialty care providers as identified through difference-in-differences (DiD) analysis.

Patient Safety and Care Experience. Survey results indicated significant improvements post-implementation across multiple aspects of patient safety and care experience (Table 3). Physicians reported an improved ability to give patients their undivided attention during encounters (Pre: 2.70 vs. Post: 3.45; $p < 0.001$), enhanced delivery of patient-centered care (Pre: 3.68 vs. Post: 4.40; $p < 0.001$), and improved perceptions of delivering high-quality care (Pre: 3.98 vs. Post: 4.54; $p < 0.001$). Perceived patient safety also increased significantly (Pre: 3.72 vs. Post: 4.03; $p = 0.027$). Notably, these improvements were consistent across both primary and specialty care providers, with no statistically significant differences in change scores between

groups for most items. The only exception was related to alerts that prevent care-delivery mistakes, where a significant difference in change was observed between primary and specialty providers (DiD $p = 0.040$).

Clinical Efficiency and Workflow. Respondents reported statistically significant improvements in perceived clinical efficiency and workflow following implementation (Table 3). Clinical practice effectiveness was rated higher post-implementation (Pre: 3.66 vs. Post: 4.28; $p < 0.001$), as was overall efficiency (Pre: 3.42 vs. Post: 4.31; $p < 0.001$). Perceptions of specialty-specific workflow support also improved significantly (Pre: 3.75 vs. Post: 4.23; $p = 0.003$), along with perceptions of system response time (Pre: 3.82 vs. Post: 4.22; $p = 0.021$). No statistically significant differences in change scores were observed between primary and specialty care providers for any of the items.

EHR Training and Implementation Support. Participants reported statistically significant improvements in several areas related to EHR training and support (Table 3). Initial training preparedness improved (Pre: 3.52 vs. Post: 3.89; $p = 0.016$), as did the ease of learning the EHR system (Pre: 3.62 vs. Post: 4.23; $p < 0.001$). Participants also reported increased satisfaction with ongoing training and education (Pre: 3.54 vs. Post: 3.88; $p = 0.046$). Perceptions of how well organizations supported EHR implementation improved (Pre: 3.58 vs. Post: 3.94; $p = 0.004$), with a statistically significant difference between primary and specialty care providers in change scores (DiD $p = 0.035$). Additionally, participants felt more confident in their personal mastery of the EHR (Pre: 3.78 vs. Post: 4.05; $p = 0.005$) and perceived improvement in receiving training tailored to their specialty (Pre: 3.32 vs. Post: 3.72; $p = 0.019$).

System Usability and Reliability. Significant improvements were noted in system integration, both within organizations (Pre: 3.98 vs. Post: 4.43; $p = 0.002$) and with external

organizations (Pre: 3.18 vs. Post: 3.73; p = 0.010), the latter being particularly driven by improvements among specialty care providers (p = 0.029). System availability also improved significantly (Pre: 4.09 vs. Post: 4.52; p = 0.005), with a significant difference in change scores between primary and specialty care providers (DiD p = 0.047). No statistically significant change was observed in perceptions of the EHR vendor's overall system quality.

Perceived Documentation Burden. Responses to documentation burden questions demonstrated statistically significant improvements following implementation of the ambient listening tool (Table 3). Among all providers, mental demand decreased (Pre: 6.71 vs. Post: 6.11; p = 0.031), and perceived effort required to write notes declined (Pre: 6.75 vs. Post: 6.02; p = 0.014). Within the primary care group, effort scores also showed a significant reduction (Pre: 7.07 vs. Post: 5.93; p = 0.016), although no statistically significant differences in change scores were observed between primary and specialty care providers (all DiD p-values > 0.1). Temporal demand and charting behavior (i.e., the percentage of notes completed during or immediately after visits) did not show significant changes across any group.

Table 3. Pre- and Post- Perceived Impact Survey Ratings by Care Group

Perceived Impact Questions [a]	**Care Group** [b]	**N**	**Pre** *Mean (SD)*	**Post** *Mean (SD)*	**P-value**	**DiD P**[c]
Patient Safety and Care Experience						
1. Does this EHR include alerts that prevent care-delivery mistakes?	All Providers	59	3.66 (0.84)	3.44 (1.09)	0.157	
	Primary	30	3.83 (0.59)	3.27 (1.05)	0.012*	0.040*
	Specialty	29	3.48 (1.02)	3.62 (1.12)	0.593	
2. Are you able to give patients your undivided attention during the encounter?	All Providers	56	2.70 (1.06)	3.45 (1.14)	0.000***	

	Primary	27	2.74 (1.16)	3.52 (1.09)	0.007**	0.87
	Specialty	29	2.66 (0.97)	3.38 (1.21)	0.002**	
3. Does this EHR help keep your patients safe?	All Providers	64	3.72 (0.79)	4.03 (0.76)	0.027*	
	Primary	34	3.79 (0.59)	4.00 (0.74)	0.22	0.399
	Specialty	30	3.63 (0.96)	4.07 (0.78)	0.057	
4. Does this EHR allow you to deliver patient-centered care?	All Providers	62	3.68 (1.00)	4.40 (0.73)	0.000***	
	Primary	33	3.64 (1.03)	4.52 (0.71)	0.001**	0.325
	Specialty	29	3.72 (1.00)	4.28 (0.75)	0.033*	
5. Does this EHR enable you to deliver high-quality care?	All Providers	65	3.98 (0.94)	4.54 (0.61)	0.000***	
	Primary	34	4.18 (0.80)	4.74 (0.45)	0.001**	.969
	Specialty	31	3.77 (1.06)	4.32 (0.70)	0.025*	
Clinical Efficiency and Workflow						
6. Is your clinical practice enhanced through the use of this EHR?	All Providers	65	3.66 (0.97)	4.28 (0.67)	0.000***	
	Primary	34	3.82 (0.72)	4.44 (0.56)	0.000***	0.986
	Specialty	31	3.48 (1.18)	4.10 (0.75)	0.021*	
7. Does this EHR make you as efficient as possible?	All Providers	65	3.42 (1.14)	4.31 (0.86)	0.000***	
	Primary	34	3.50 (1.02)	4.41 (0.66)	0.000***	0.892
	Specialty	31	3.32 (1.28)	4.19 (1.05)	0.003**	
8. Does this EHR support your specialty-specific workflows?	All Providers	65	3.75 (0.97)	4.23 (0.86)	0.003**	
	Primary	34	4.03 (0.76)	4.50 (0.51)	0.003**	0.965
	Specialty	31	3.45 (1.09)	3.94 (1.06)	0.107	
9. Does this EHR have a fast system response time?	All Providers	65	3.82 (1.09)	4.22 (0.84)	0.021*	
	Primary	34	3.97 (0.97)	4.29 (0.72)	0.081	0.632
	Specialty	31	3.65 (1.20)	4.13 (0.96)	0.113	
EHR Training and Support						

10. Did your initial training prepare you well to use this EHR?	All Providers	62	3.52 (1.07)	3.89 (1.04)	0.016*	
	Primary	32	3.59 (0.98)	4.06 (1.01)	0.007**	0.488
	Specialty	30	3.43 (1.17)	3.70 (1.06)	0.312	
11. Is ongoing EHR training/education helpful and effective?	All Providers	57	3.54 (1.13)	3.88 (0.91)	0.046*	
	Primary	28	3.75 (1.11)	3.79 (1.03)	0.837	0.069
	Specialty	29	3.34 (1.14)	3.97 (0.78)	0.016*	
12. Is this EHR easy to learn?	All Providers	65	3.62 (1.06)	4.23 (0.70)	0.000***	
	Primary	34	3.85 (0.93)	4.32 (0.64)	0.007**	0.276
	Specialty	31	3.35 (1.14)	4.13 (0.76)	0.002**	
13. Has your organization done a good job implementing, training, and supporting the EHR?	All Providers	64	3.58 (0.87)	3.94 (0.77)	0.004**	
	Primary	34	3.56 (0.96)	4.15 (0.78)	0.001**	0.035*
	Specialty	30	3.60 (0.77)	3.70 (0.70)	0.591	
14. Have you personally learned the EHR well enough to be successful?	All Providers	65	3.78 (0.78)	4.05 (0.72)	0.005**	
	Primary	34	3.91 (0.67)	4.24 (0.55)	0.005**	0.467
	Specialty	31	3.65 (0.88)	3.84 (0.82)	0.225	
15. Were you trained on EHR workflows specific to your specialty area?	All Providers	60	3.32 (1.10)	3.72 (1.04)	0.019*	
	Primary	30	3.73 (0.91)	3.97 (0.93)	0.168	0.322
	Specialty	30	2.90 (1.12)	3.47 (1.11)	0.06	
System Usability and Reliability						
16. Does this EHR provide expected integration with outside organizations?	All Providers	49	3.18 (1.07)	3.73 (0.93)	0.010*	
	Primary	23	3.22 (0.90)	3.61 (0.78)	0.147	0.459
	Specialty	26	3.15 (1.22)	3.85 (1.05)	0.029*	
17. Does this EHR provide expected integration within your organization?	All Providers	65	3.98 (0.89)	4.43 (0.73)	0.002*	

	Primary	34	4.21 (0.64)	4.47 (0.71)	0.093	0.183
	Specialty	31	3.74 (1.06)	4.39 (0.76)	0.010*	
18. Is this EHR available when you need it (minimal downtime)?	All Providers	65	4.09 (0.98)	4.52 (0.66)	0.005**	
	Primary	34	4.32 (0.77)	4.47 (0.75)	0.356	0.047*
	Specialty	31	3.84 (1.13)	4.58 (0.56)	0.004*	
19. Has the EHR vendor designed a high-quality system?	All Providers	65	3.85 (0.78)	4.02 (0.54)	0.134	
	Primary	34	3.97 (0.46)	4.09 (0.51)	0.285	0.608
	Specialty	31	3.71 (1.01)	3.94 (0.57)	0.291	
Perceived Documentation Burden						
20. Mental demand (0–10): How demanding is it to write your notes?	All Providers	56	6.71 (1.86)	6.11 (2.20)	0.031*	
	Primary	27	6.81 (1.75)	6.30 (2.00)	0.203	0.740
	Specialty	29	6.62 (1.97)	5.93 (2.40)	0.066	
21. Temporal demand (0–10): How rushed is the pace of your note writing?	All Providers	55	7.07 (1.59)	6.58 (1.90)	0.055	
	Primary	26	6.92 (1.55)	6.35 (1.85)	0.082	0.767
	Specialty	29	7.21 (1.63)	6.79 (1.95)	0.311	
22. Effort (0–10): How much effort is required to write your notes?	All Providers	56	6.75 (1.97)	6.02 (1.91)	0.014*	
	Primary	27	7.07 (2.18)	5.93 (1.90)	0.016*	0.165
	Specialty	29	6.45 (1.72)	6.10 (1.95)	0.413	
23. Ambulatory encounters (%): What percentage of charts do you complete during or right after patient visits?	All Providers	59	53.64 (24.36)	55.75 (28.70)	0.429	
	Primary	33	53.67 (25.64)	55.94 (29.74)	0.652	0.954
	Specialty	26	53.62 (23.14)	55.50 (27.91)	0.628	

[a] *Responses to survey questions 1–19 were recorded on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Questions 20–22 used a 0–10 numerical rating scale, and question 23 captured responses as percentages.*
[b] *Care Group: "Primary" = Primary Care providers; "Specialty" = Specialty Care providers; "Total" = All providers combined.*

[c] *DID = Difference-in-Differences between Primary and Specialty change scores*

Qualitative Analysis of Open-Ended Responses

We included an open-ended question, *"Do you have any other related comments or concerns?"*, to invite participants to share additional thoughts. As responses were optional and the number of comments was limited, an exploratory thematic analysis was conducted by one author (YG) to supplement the quantitative findings. As shown in Table 4, three overarching themes emerged from the qualitative comments: (1) Positive perceptions of ambient AI, including reduced documentation burden, improved efficiency, and enhanced focus on patient interaction; (2) Identified limitations and areas for improvement, such as insufficient customization for individual and specialty-specific needs, workflow disruptions, and content accuracy concerns; and (3) Expectations for future AI integration, highlighting clinicians' interest in expanding AI functionalities beyond documentation to areas such as inbox management, clinical summarization, and billing support.

Table 4. Themes of Identified Benefits, Limitations, and Future Expectations for Ambient AI in Clinical Practice

Themes	**Example Quotation** [a]
Positive Perceptions of Ambient AI	
Streamlined Workflow and Documentation Efficiency	*"I complete 90% of each note during or immediately after the visit, and it only takes 30 minutes to an hour after the clinic to finish all documentation."*
	"Ambient AI charts the visit and brings forward my last note, updates interval history, exam, and plan, which allows me to quickly review and sign off."
	"Since using ambient AI, I've been able to complete charts more efficiently."
	"I use as many personalized templates and speed buttons as possible. This tool has completely changed my ability to document in an efficient manner."
	"Ambient AI has completely changed the way I do notes. Much more efficient."
Reduced Mental Fatigue and Cognitive Load	*"Using ambient AI has lifted a huge mental load. I try to get my notes done quickly and close my encounters the same day."*

	"Love the tool. It's improved note-writing fatigue and burnout."
Enhanced Focus on Patient Interaction	*"This is the right step forward—letting the machine do the work, and the human do the thinking and analyzing."*
	"Using ambient AI for dictation helps me pay more attention to the patients during the visit and write notes more efficiently."
	"I feel it does give me higher satisfaction with my ability to listen and interact with my patients."
Identified Limitations and Desired Improvements for Ambient AI	
Need for Specialty-Specific Customization	*"There is a growing need for AI scribes to be tailored to each medical specialty. In my field, the neurologic exam is critical, but current outputs often fail to capture its complexity."*
	"Ambient AI should support surgical workflows. It currently feels geared toward primary care or emergency visits."
	"We need orthopedic-specific features. The ideal platform should evolve with departments and adapt to diverse types of encounters and documentation styles."
	"In psychiatry, the tool often applies overly medicalized language and misses the nuance of patient interactions. A specialty-specific mode is needed."
Need for Personalization to Individual Styles	*"If we could train the tool to write notes a certain way, it would be much more useful. Right now, editing takes more time than typing from scratch."*
	"It should learn from my preferences and evolve over time. More intelligent customization would reduce editing workload."
	"The system should be 'living software'—something users can refine and adapt to their individual style and needs."
Workflow Disruptions	*"Switching between app, browser, and room computer is inefficient. Seamless EHR integration is essential."*
	"Even minor technical failures are disruptive. We rely on it heavily now, so stability is critical."
	"We shouldn't have to use personal phones. Integration into workstations would improve usability." *"It would be great to use the room's microphone instead of relying on mobile devices."*
Issues with Note Coherence, Accuracy, and Edit Burden	*"It occasionally misplaces physical exam findings in the wrong section, and repeated feedback hasn't led to improvements."* *"The output is sometimes inaccurate or wordy, and I spend time correcting it."*
	"The transcription isn't reliable enough for certain sections, so I've stopped using those features."
Need for Expanded Functional Capabilities	*"There should be support for multiple languages—especially in diverse clinical populations."*
	"The tool is more helpful for new visits than follow-ups. It struggles to build on prior notes."
	"Ambient AI doesn't pull relevant chart data, like labs, demographics, or medical history, which limits its usefulness."
	"It would help if it could generate documentation for things like medical necessity or align with problem-based note structures."
	"Why doesn't the tool recognize basic EHR info like patient age or sex?"
Expectation of AI future Implementation into	

EHR beyond assisting documentation	
AI for EHR Inbox Management	*"AI should assist with inbox management—triaging messages, handling refills, and reducing human error."*
	"Assign AI to handle in-basket tasks and patient messages more effectively."
	"AI inbox support exists but remains limited; I rarely use the suggested responses."
	"More customizable AI tools are needed for in-basket workflows."
AI Summarization of Patient Medical History and Literature	*"AI should help summarize notes from other providers or highlight key findings in the chart."*
	"Integrate AI to summarize labs and imaging findings in the assessment section—flagging important positives and negatives."
	"It would be useful to have AI support in constructing differential diagnoses or treatment plans through literature review."
	"Use AI to generate succinct handoff or discharge summaries."
AI Automation for Administrative and Workflow Tasks	*"AI could assist with billing, medication review, charge entry, and visit timing for time-based billing."*
	"We are overburdened by non-clinical tasks like button-clicking and documentation for charges—AI should help streamline this."
	"Would love to have AI-driven orders and billing assistance."
	"Use AI to draft appeal letters and manage correspondence through the portal."
AI for Chart Structuring and Problem List Optimization	"AI-driven chart review would be beneficial—particularly to extract clinical relevance from lengthy records."
	"It would be helpful if the tool could push relevant findings directly into the problem list."
	"Allow AI to mine notes or scan for HCC codes."

[a]*All quotes have been paraphrased or minimally modified to deidentify respondents and maintain confidentiality.*

Discussion

This study presents one of the first mixed-methods evaluations of ambient AI listening tools implemented across multiple specialties in a real-world clinical setting. By analyzing automatically recorded EHR time-efficiency metrics, pre- and post-implementation survey data, and open-ended feedback, we found that ambient AI use was associated with reduced documentation burden, improved perceptions of clinical efficiency, and enhanced patient-centered care. Consistent findings across both self-reported experiences and

system-generated metrics reinforce the potential of ambient AI to support clinician workflow and alleviate cognitive workload in everyday practice.

Survey results indicated meaningful decreases in mental and physical effort associated with note writing, alongside improved perceptions of system integration and training support. While overall improvements were consistent across provider types, only a few statistically significant differences emerged between primary and specialty care providers. One such difference was in perceptions of implementation and training support, with primary care providers reporting greater improvements. This suggests that onboarding strategies and organizational support may have been better aligned with the workflows of primary care clinicians, highlighting the need for more tailored training approaches to maximize the impact of ambient AI tools across different specialties. Another notable difference was observed in perceptions of alerts that help prevent care-delivery mistakes. This aligns with qualitative feedback emphasizing the need for specialty-specific customization of ambient AI systems. Providers in neurology, psychiatry, orthopedics, and surgery noted that current tools often fail to capture the complexity of their documentation, which remains closely tied to nuanced physical exams and specialty-specific narratives. These findings underscore the importance of developing adaptive, specialty-aware AI solutions that can flexibly accommodate diverse clinical workflows and accurately reflect domain-specific terminology and reasoning [73]. However, this study was conducted within a single academic health system, which may limit generalizability to other care settings. Additionally, potential survey bias and the voluntary nature of participation may have influenced the representativeness of responses. Failure to address these limitations may result in unequal adoption and benefit across specialties, ultimately constraining the broader impact and scalability of ambient AI in clinical practice.

Epic usage data further demonstrated significant reductions in documentation time. These results are consistent with existing studies and further validate the effectiveness of ambient AI in a larger, multispecialty sample [16,24,40]. However, several limitations warrant consideration. First, time spent on notes may not accurately represent active documentation time, as delays related to system performance or clinicians leaving notes open while multitasking may inflate these metrics. Second, ambient AI use was not mandatory for all encounters; even among physicians with access, usage varied depending on personal preference and clinical context. Notably, post-implementation Epic data captured all encounters, regardless of whether the ambient tool was used, which may have diluted the observed effects. Third, no statistically significant change was detected in same-day encounter closure rates. This outcome may be attributed to the constrained analytic sample, limited to physicians with at least 40 weekly encounters within a single institution, and to the high baseline same-day closure rate of approximately 70%, which reduced the potential for improvement. It may also reflect that certain workflow indicators, such as encounter closure, are less sensitive to documentation tools like ambient AI and more dependent on external operational factors, including clinic schedules and staffing patterns [73]. Additionally, some clinicians may not yet be fully familiar with or trained to leverage the full capabilities of the ambient scribe tool, instead using it in its default configuration without customization or optimization, which may have limited its observed impact.

The findings of this study highlight both the potential benefits and current limitations of ambient AI tools in clinical practice. Beyond improvements in documentation burden and efficiency, thematic analysis revealed strong interest in expanding the role of ambient AI to include tasks such as inbox management, billing, chart summarization, and clinical decision-making [74,75]. Participants also emphasized the need for greater personalization, both at

the individual clinician level and across specialties, to ensure the tool aligns with varying documentation styles and workflow demands.

These insights suggest that future research should focus on refining ambient AI systems to better support diverse clinical workflows, enable flexible customization, improve integration with historical medical records, and expand core functionalities to address broader administrative burdens. As ambient AI becomes more deeply integrated into clinical environments, developers must also account for ethical, legal, and regulatory considerations[76–80]. Ensuring transparency, safeguarding patient privacy, and clarifying liability for AI-generated documentation will be essential to building clinician trust and enabling responsible adoption[81]. In parallel, investigating patient perspectives through surveys or interviews will be important to assess perceived communication quality, trust, and acceptance of AI-assisted documentation. Long-term evaluations will be essential to determine whether the benefits of ambient AI, such as improved efficiency and reduced burden, are sustained over time and whether they lead to measurable improvements in clinical outcomes, provider satisfaction, and burnout.

Conclusion

This quality improvement study at UCI Health provides real-world evidence supporting the potential of ambient AI tools to reduce documentation burden, enhance workflow efficiency, and improve patient experiences. By analyzing pre- and post-implementation surveys and EHR usage metrics, our findings demonstrate that ambient AI effectively supports physician productivity and addresses key contributors to burnout. Nevertheless, several areas for further improvement were identified, including specialty-specific customization, integration with historical patient data, and enhanced accuracy of AI-generated content. Future implementation efforts should

emphasize thoughtful scalability, adaptability to diverse clinician needs, and careful consideration of clinical, ethical, and legal requirements. Ongoing evaluation and iterative optimization will be crucial for building clinician trust, promoting equity in AI-supported care, and fully leveraging ambient technologies to facilitate patient-centered healthcare.

CHAPTER 3: From Conversation to Chart: An Analysis of Clinician Edits to Ambient AI Draft Notes

Abstract

Objective

Ambient artificial intelligence (AI) tools are increasingly adopted in clinical practices. This study investigated whether and how clinicians edit AI-generated drafts and the linguistic differences between AI drafts and clinician-finalized notes.

Materials and Methods

This retrospective study analyzed real-world data from ambulatory clinics at a large academic health system spanning two vendor deployments. We quantified clinicians' editing behavior using the Myers diff algorithm to compare AI drafts and final documentation. We then applied statistical and linguistic analysis to study factors associated with the frequency/intensity of editing across note sections, turnaround time, clinician characteristics, and encounter types.

Results

Across 23,760 notes that included one or more ambient AI sections, 84.4% were edited by clinicians before signing off. While rates of unedited notes differed across note sections and care settings, the dominant source of variation was individual clinician practice style rather than specialty-level norms. Notes signed after 24 hours had lower overall edit intensity. The final versions showed small but statistically significant linguistic changes and exhibited slightly higher lexical diversity and modest changes in readability. Editing is most intensive in the assessment and plan section, and varies across specialties.

Conclusion and Discussion

A majority of AI-drafted clinical notes were edited by clinicians, although the editing rate varies across note sections, medical specialties, and individual clinicians. Future research is needed to further analyze this editing behavior to inform improvement in AI-assisted clinical documentation to achieve better documentation quality, efficiency, and clinician satisfaction.

Background and Significance

Clinical documentation is essential to outpatient care but remains a major driver of after-hours "pajama time" and clinician burnout. [17,67,82] Ambient artificial intelligence (AI) documentation systems have emerged to capture visit conversations and draft notes that integrate into the electronic health record (EHR) for clinicians to review. [26,27] Prior studies report that clinicians are generally satisfied with ambient AI tools, citing reduced documentation burden and improvements in workload, work–life integration, and patient engagement. [1,12,19,41,42] Despite this promise, clinicians still spend substantial time reviewing, refining, and signing off notes, and analysis on time savings alone does not reveal how clinicians adopt and modify AI drafts in routine practice. Most evaluations of ambient AI documentation to date rely on before-and-after implementation comparisons, assessing time-based workflow efficiency outcomes such as time spent on note writing. [16,18,40] Early evidence suggests that following deployment, clinicians spend less time on documentation while AI-assisted notes become longer, raising concerns about note bloat, in which increased length may dilute clinically important content and increase downstream review burden. [42,83]

In contrast, routine editing behavior in real-world practice remains poorly characterized. It is still unclear whether clinicians sign ambient AI-drafted notes without changes, how they revise

drafts, and how editing behaviors vary across encounter characteristics. Empirical evidence is also limited in how ambient AI relates to note completion timeliness within everyday ambulatory workflows. Understanding these patterns is essential because documentation practices and information needs differ substantially across clinical settings such as specialties and visit types, and ambient AI documentation should support workflow-specific needs without shifting effort into additional review and correction work. [7,66,67]

To address this gap, we analyzed a large dataset of paired ambient AI-drafted note and clinician-finalized text from a large academic health system, where two commercially available ambient AI documentation systems are integrated into ambulatory EHR workflows. Our analyses addressed three questions: (1) how often AI drafts are submitted without edits and whether no-edit behavior reflects individual clinician writing style versus specialty norms, as well as how it varies by note section, encounter type, and specialty; (2) among edited note sections, how clinician-edited and AI-generated texts differ in linguistic characteristics, including length, lexical diversity, and readability, and how the magnitude of edits varies by note sections and care settings; and (3) how turnaround time relates to edit intensity, and which note sections, encounter types, and specialties are associated with more editing and longer turnaround. Together, these analyses provide a detailed picture of how clinicians interact with ambient AI documentation in routine outpatient practice.

Methods

Setting, Data Source, and Unit of Analysis. A pilot deployment of ambient AI documentation was initiated at University of California, Irvine Medical Center (UCI Health) in 2023. UCI

Health is an academic health system with ambulatory clinics across multiple sites in Southern California. Participating clinicians from selected primary care and specialty clinics used two commercially available ambient AI tools integrated with the Epic EHR. Vendors were referred to as Vendor A and Vendor B to align with institutional data governance and reporting constraints. Clinicians accessed the tools through the Haiku mobile application or through a browser-based interface connected to Epic. In all workflows, AI-generated text was reviewed, edited as needed, and signed within the EHR, and clinicians retained full responsibility for the final documentation.[14,15] Clinicians received onboarding and workflow guidance as part of the pilot rollout, with ongoing technical support available during the study period. Consistent with institutional procedures for the pilot, clinicians informed patients at the visit that audio-assisted note drafting might be used, that the resulting documentation would be clinician-reviewed before filing, and that identifiable audio was not retained. Liability and regulatory considerations were managed through institutional compliance and IT review, and the EHR audit trail preserved both the initial AI-generated draft and the final signed version. We report overall pooled summaries and vendor-stratified analyses where applicable and do not disclose the mapping between the vendor labels.

We extracted all ambient AI note sections from Epic between September 2024 and August 2025. In routine clinical use, ambient AI generates draft text at the note-section level, and documentation is completed and signed at the encounter note level. Clinicians can selectively choose one or more AI-drafted note sections to insert into an EHR note. Four types of note sections can be generated by the ambient AI tools, including History of Present Illness (HPI), Assessment & Plan (A&P), Physical Exam, and Results. We analyzed workflow outcomes that are naturally defined per encounter (e.g., turnaround time and encounter-level unedited use) at

the note level by aggregating section-level data within each unique note ID, and then conducted section-level analyses to characterize where edits concentrate across note-section types.

Prevalence of Unedited Note Sections by Clinical Context. To quantify how often ambient AI drafts were accepted as-is, we first computed a note-level no-edit indicator (1 = all AI-generated sections within a note were signed without edits; 0 = otherwise) and summarized the prevalence of unedited use. We then conducted a section-level drill-down by defining a binary section-level indicator (1 = no-edit, 0 = edited) and stratifying unedited prevalence by note section, care setting (primary vs specialty care), and encounter type. For each dimension, we used χ^2 tests of independence to compare the distribution of unedited vs edited sections across categories, reporting χ^2 statistics, degrees of freedom, and p-values.[84] We also conducted vendor-stratified χ^2 tests as sensitivity analyses.

Unedited Use as Personal Style vs Specialty Norms. To distinguish individual documentation style from broader specialty norms, we fit mixed-effects logistic regression models separately within each vendor, restricted to clinicians who have used ambient AI to draft over 100 note sections.[85] The binary outcome indicated whether a section was submitted without edits. Fixed effects included note section, encounter type, and clinician group (specialty vs primary care). We specified random intercepts for individual clinician and specific medical specialty (e.g., cardiology, neurology, etc.). [86] Including clinician group as a fixed effect captures broad differences between primary and specialty care, while the specialty random intercept captures residual variation across specific specialties beyond that grouping. We summarized variance attributable to each level by computing intraclass correlation coefficients (ICCs).

Linguistic Metrics Before and After Editing. For note sections that were edited, we computed eight metrics for both AI draft and final version: character count, word count, sentence count,

unique word count, type–token ratio (unique words divided by total words), average word length, Flesch Reading Ease, and Flesch–Kincaid Grade Level. [87] Within each note section we calculated a before–after difference for each metric by subtracting the AI draft from the clinician-final text (Δ = Final − AI). To quantify overall editing, we used paired t-tests to test whether the mean Δ differed from zero. [88] Within each vendor, we used paired t-tests to assess whether mean Δ differed from zero. To compare editing magnitude between vendors, we compared mean Δ between Vendor A and Vendor B using Welch's two-sample t-test. To examine variation by note section, we compared Δ across HPI, A&P, Physical Exam, and Results using one-way ANOVA with Tukey's HSD post-hoc tests for pairwise contrasts. [89]

Quantifying Edit Intensity with Myers Diff. For all edited sections, we aligned the AI draft and final note using a token-level Myers diff algorithm. [90] This approach measures edits at the token level, where each token corresponds to a word or punctuation unit. We defined total edit intensity as the total number of tokens changed, calculated as the sum of inserted, deleted, and replaced tokens.

Factors Associated With Longer Turnaround Time. To identify which notes were more likely to have turnaround over 24 hours, we fit a logistic regression model using the binary turnaround indicator as the outcome, and encounter type, clinician specialty, and the transformed number of AI-generated note sections included in each note as predictors. Vendor type was included as a covariate, and we tested whether the association between section count and long turnaround differed by vendor using the vendor-by-log(1+sections) interaction by comparing nested models using a likelihood-ratio test. Encounter types and specialties with fewer than 20 notes, or with 0% or 100% turnaround prevalence over 24 hours, were collapsed into pooled "Other" categories. [91]

Turnaround Time and Editing Intensity. A binary turnaround indicator was coded as 1 when time from AI draft generation to note signing exceeded 24 hours, and 0 otherwise. Because signing occurs at the note level and the draft and signing dates were identical across sections within a note, we conducted turnaround analyses at the note level by aggregating section-level edits within each note. To test whether later submissions received more edits, we first compared total note-level edit intensity (sum of inserted, deleted, and replaced tokens across all sections in the note) between notes signed within 24 hours and those signed after 24 hours using descriptive statistics and a Mann–Whitney U test. [92,93] We repeated these comparisons stratified by vendors. For clinicians who have used ambient AI to draft over 100 note sections, we then fit a mixed-effects model with log-transformed total edit intensity per note as the outcome and turnaround time indicator as the primary predictor, adjusting for encounter type, clinician group (primary vs specialty care), and the log-transformed number of AI-generated note sections included in each note as fixed effects, and including an individual clinician-level random intercept. [94] To avoid unstable estimates due to sparse categories or perfect prediction, encounter types, specialties, and section categories with small counts were collapsed into an "Other" category prior to modeling.

Factors Associated With Heavier Editing Intensity. To examine which note sections, encounter types, and specialties were associated with heavier editing, we fit a regression model with log-transformed edit intensity as the outcome, and note section, encounter type, and clinician specialty as categorical predictors, including vendor type as a covariate, and using clinician-clustered robust standard errors to account for within-clinician correlation (i.e., clustering on clinician to allow correlation among sections authored by the same clinician). [95] Encounter types and specialties with fewer than 50 sections were pooled into "Other" categories.

Ethical Oversight and Data Security. The work was approved by the University of California, Irvine Institutional Review Board (IRB #7123). All data processing occurred within HIPAA-compliant secure environments, including UCI Health's Protected Virtual Computing Environment (PVCE) and Amazon Web Services (AWS) HIPAA-aligned infrastructure.

Results

Dataset Characteristics

The dataset included 23,760 unique notes containing 48,155 AI-drafted note sections, authored by 225 clinicians for 16,520 patients (Table 5). We summarized characteristics at the note-section level because clinicians can selectively insert and edit individual AI-generated smart elements, and a single note may contain multiple AI-drafted sections. When stratified by vendor, Vendor A contributed 12,439 notes and Vendor B contributed 11,321 notes. Full vendor-stratified cohort characteristics are provided in Table 5.

Table 5. Study Cohort and Dataset Characteristics

Category	Subcategory	Vendor A	Vendor B
General	Study period	2024-10-28 – 2025-07-29	2024-09-26 – 2025-07-25
	Total notes (N)[a]	12439	11321
	Total note sections	25101	23054
	Unique clinicians (N)	119	144
	Unique patients (N)	8845	8392
Section name (per note section) [b]	HPI	12,174 (48.5%)	11,793 (51.2%)
	A&P	9,423 (37.5%)	6,968 (30.2%)
	Results	2,200 (8.8%)	2,292 (9.9%)

	Physical Exam	1,304 (5.2%)	2,001 (8.7%)
Clinician license (per note section) [b]	MD, DO	23,713 (94.5%)	20,225 (87.7%)
	NP, PA	1,209 (4.8%)	2,446 (10.6%)
	OD, MD-PhD, LVN	179 (0.7%)	383 (1.7%)
Note type (per note section) [b]	Progress Note	24,661 (98.2%)	22,345 (96.9%)
	Consults / History and Physical (H&P) / Procedure Note	440 (1.8%)	709 (3.1%)
Encounter type (per note section) [b]	Office Visit	21,881 (87.2%)	20,802 (90.2%)
	Video Visit / Telemedicine	2,530 (10.1%)	1,998 (8.7%)
	Other encounter types [c]	690 (2.7%)	254 (1.1%)
Clinician specialty (per note section) [b]	Family Practice	8,162 (32.5%)	3,857 (16.7%)
	Internal Medicine	5,409 (21.5%)	972 (4.2%)
	Geriatric Medicine	30 (0.1%)	3,987 (17.3%)
	Nurse Practitioner	1,013 (4.0%)	375 (1.6%)
	Physician Assistant	109 (0.4%)	1,247 (5.4%)
	Rheumatology	5,356 (21.3%)	546 (2.4%)
	Orthopaedic Surgery	286 (1.1%)	4,175 (18.1%)
	Hand Surgery	395 (1.6%)	3,591 (15.6%)
	Neurology	15 (<0.1%)	1,545 (6.7%)
	Ob/Gyn	109 (0.4%)	851 (3.7%)
	Gastroenterology	857 (3.4%)	0
	Physical Medicine and Rehab	563 (2.2%)	10 (<0.1%)
	Neurosurgery	520 (2.1%)	15 (<0.1%)
	Cardiology	478 (1.9%)	0
	Other specialty [d]	1,030 (4.1%)	259 (1.1%)

[a] *To reduce the risk of sensitive content, we excluded VIP patients, confidential/sensitive encounters or notes, sensitive departments (e.g., psychiatry/behavioral therapy/substance use/addiction), psychotherapy notes, and any notes explicitly marked sensitive.*

[b] *Unless otherwise noted, percentages (%) are n of note sections within each vendor (Vendor A: N = 25,101; Vendor B: N = 23,054).*

[c] *Other encounter types = Integrative Medicine, Procedure Visit, OB Visit, Erroneous Encounter, Tumor Board.*

[d] *Vendor A other specialty (count=11): Colon and Rectal Surgery, Dermatology, Endocrinology, Hospitalist, Infectious Diseases, Integrative Medicine, Optometry, Otolaryngology, Surgical Oncology, Urology, Vascular Surgery. Vendor B other specialty (count=11): Colon and Rectal Surgery, Emergency Medicine, General Surgery, Head and Neck Surgery, Hematology/Oncology, Ophthalmology, Orthopedics, Otolaryngology, Psychiatry, Urology, Vascular Neurology*

How Often Are AI Drafts Signed Without Edits?

Overall, 15.6% of the notes were signed without any edits (3,698/23,760), with slightly higher no-edit rates for Vendor A than Vendor B (17.4%, 2,164/12,439 vs 13.6%, 1,534/11,321). Across all note sections, 14,568 of 48,155 sections (30.3%) were signed without any provider-edited text; section-level no-edit rates were likewise higher for Vendor A than Vendor B (32.0%, 8,020/25,101 vs 28.4%, 6,548/23,054). Within each vendor, no-edit rate differed substantially by note section (Vendor A: $\chi^2(3)=4,686.2$, $p<0.001$; Vendor B: $\chi^2(3)=1,794.8$, $p<0.001$). For Vendor A, no-edit rates were 8.7% for A&P, 40.2% for HPI, 54.3% for Physical Exam, and 72.5% for Results; for Vendor B, rates were 19.2% for A&P, 25.1% HPI, 41.6% for Physical Exam, and 62.0% for Results (Table 6). No-edit rates also varied by encounter type within both vendors (Vendor A: $\chi^2(7)=152.9$, $p<0.001$; Vendor B: $\chi^2(5)=55.2$, $p<0.001$). Office visits had a 33.0% no-edit rate for Vendor A and 28.4% for Vendor B; video visits were lower (26.3% and 24.2%), while telemedicine encounters were 32.2% and 36.0%, respectively. Differences by clinician group were significant for Vendor A (primary care 36.2% vs specialty care 26.1%; $\chi^2(1)=275.5$, $p<0.001$) but not for Vendor B (29.6% vs 29.1%; $\chi^2(1)=0.7$, $p=0.418$). We conducted a supplementary specialty-level analysis and found substantial heterogeneity in section-level no-edit rates across clinician specialties overall ($\chi^2(33)=3528.4$, $p<0.001$) and within each vendor

when analyzed separately (Appendix 1. Supplementary Table S1). Vendor-stratified supplementary analyses use the analytic subset with Vendor A/B mapping and complete covariates; denominators therefore differ from full cohort totals in Table 5.

Table 6. Section-Level No-Edit Rates by Clinician Group, Encounter Type, and Note Section

		Vendor A		Vendor B	
Dimension	Category	Total sections, n	No-edit sections, n (% within Vendor)	Total sections, n	No-edit sections, n (% within Vendor)
Clinician group	Primary Care	14,723	5,323 (36.2%)	10,438	3,093 (29.6%)
	Specialty Care	10,046	2,624 (26.1%)	11,239	3,273 (29.1%)
Encounter type	Office Visit	21,881	7,221 (33.0%)	20,802	5,917 (28.4%)
	Video Visit	1,824	479 (26.3%)	1,118	270 (24.2%)
	Telemedicine	706	227 (32.2%)	880	317 (36.0%)
	Integrative Medicine	614	77 (12.5%)	0	0
	Procedure Visit	58	12 (20.7%)	155	27 (17.4%)
	OB Visit	9	2 (22.2%)	94	14 (14.9%)
	Erroneous Encounter	3	2 (66.7%)	5	3 (60.0%)
	Tumor Board	6	0 (0.0%)	0	0
Note section	HPI	12,174	4,894 (40.2%)	11,793	2,960 (25.1%)
	A&P	9,423	822 (8.7%)	6,968	1,336 (19.2%)
	Results	2,200	1,596 (72.5%)	2,292	1,420 (62.0%)
	Physical Exam	1,304	708 (54.3%)	2,001	832 (41.6%)

** Categories with 0 total sections in a vendor stratum were excluded from χ^2 tests*

** Percentages are calculated within each vendor and category: no-edit sections / total sections for that vendor.*

Is "no-edit" behavior driven by individual writing styles or specialty norms?

Unedited submissions were driven more by individual clinician style than by specialty norms in both vendor analyses. In mixed-effects logistic regression models fit separately for Vendor A and Vendor B (adjusting for note section and encounter type), specialty-care clinicians had slightly lower odds of signing AI note drafts without edits than primary-care clinicians (Vendor A: β_specialty = −0.13; Vendor B: β_specialty = −0.14). Random-effects variance was concentrated at the clinician level in both vendors (Vendor A: Var_clinician = 3.84, ICC = 0.54; Vendor B: Var_clinician = 2.47, ICC = 0.43), with negligible variance attributable to specialty (Vendor A: Var_specialty = 0.0023, ICC ≈ 0.00; Vendor B: Var_specialty = 0.0018, ICC ≈ 0.00). These findings indicate that "no-edit" behavior varied primarily across individual clinicians rather than across specialties after accounting for section type and encounter context (Appendix 1. Supplementary Table S2).

Linguistic Differences Between AI Drafts and Final Note Sections

Paired comparisons showed small but statistically significant differences between AI drafts and final note sections across most linguistic metrics (Table 7). In the vendor-stratified analysis, Table 7 summarizes 33,502 paired AI–final section comparisons in total (Vendor A: 16,996; Vendor B: 16,506). Across both vendors, final notes generally had fewer sentences and changes in readability and lexical diversity, although the magnitude and direction of changes varied by vendor. We additionally tested whether the before–after change (Final–AI) differed between vendors using a Welch two-sample t-test (i.e., a difference-in-differences comparison of mean

deltas): the vendor gap in mean deltas was significant for most metrics ($p<0.001$), but not for sentence count ($p=0.455$) and only marginal for unique word count ($p=0.071$).

Table 7. Linguistic differences comparing AI drafts and clinician-final note sections

	Vendor A				Vendor B				P for between-vendor difference in mean Δ (Welch t-test)
Metric	AI mean	Final mean	Mean diff (Final-AI)	P (Final vs AI)	AI mean	Final mean	Mean diff (Final-AI)	P (Final vs AI)	
Average word length (characters)	5.57	5.442	-0.129	<0.001	5.229	5.174	-0.055	<0.001	<0.001
Character count	1275.403	1237.21	-38.193	<0.001	1282.75	1290.729	7.979	0.069	<0.001
Flesch–Kincaid grade level	11.395	11.218	-0.176	<0.001	10.353	10.643	0.291	<0.001	<0.001
Flesch Reading Ease	37.204	39.494	2.29	<0.001	45.155	45.159	0.004	0.958	<0.001
Sentence count	15.089	13.907	-1.182	<0.001	16.204	14.97	-1.234	<0.001	0.455
Type–token ratio	0.643	0.67	0.027	<0.001	0.649	0.654	0.005	<0.001	<0.001
Unique word count	114.572	116.473	1.901	<0.001	118.492	121.248	2.756	<0.001	0.071
Word count	187.864	185.162	-2.702	<0.001	201.696	203.326	1.63	0.017	<0.001

Across note sections, patterns of linguistic change differed by vendor (Table 8). For Vendor A, edits were most pronounced in A&P, with large decreases in character count ($\Delta = -164.2$), word count ($\Delta = -17.9$), and sentence count ($\Delta = -3.17$), whereas changes in other sections were smaller and often positive (e.g., Results showed increases in character count and word count).

For Vendor B, changes varied across sections: HPI showed decreases in character count, word count, and sentence count, whereas Results showed increases in all three measures. Overall, section-level differences in changes were statistically significant within both vendors (all $p < 0.001$).

Table 8. Section-Level Changes in Linguistic Metrics Between AI Drafts and Clinician-Final Note Sections

	Vendor A Δ (Final–AI)				Vendor B Δ (Final–AI)			
Metric	A&P	HPI	Results	Phys Exam	A&P	HPI	Results	Phys Exam
Average word length (characters)	-0.198	-0.019	-0.357	-0.251	-0.023	-0.064	-0.016	-0.171
Character count	-164.158	83.693	194.841	73.422	111.632	-85.34	134.826	119.107
Flesch–Kincaid grade level	-0.086	-0.201	-2.138	0.77	0.318	-0.039	1.146	2.011
Flesch Reading Ease	3.28	0.748	10.805	-1.59	-0.382	1.277	-4.771	-4.192
Sentence count	-3.171	0.867	1.168	0.327	-0.998	-1.816	1.009	0.354
Type–token ratio	0.049	0.012	-0.072	-0.021	0.007	0.015	-0.042	-0.043
Unique word count	-4.028	7.294	16.862	7.531	9.455	-3.72	12.1	12.45
Word count	-17.87	10.803	39.361	11.625	16.572	-12.153	21.776	18.757

**Omnibus tests comparing Δ (Final − AI) across note sections were significant within each vendor (all $p < 0.001$).*

Which notes are more likely to have turnaround over 24 hours?

Across notes that were edited before being signed (n = 20,062), 3,194 (15.9%) had turnaround over 24 hours. In vendor-adjusted models, the association between the number of AI-generated

note sections within a note and long turnaround differed by vendor (Vendor × log[1+sections] likelihood-ratio test $p < 0.001$). For Vendor A, a higher section count was associated with greater odds of turnaround over 24 hours (log[1+sections]: $\beta = 0.51$, $p = 0.010$), and this association was significantly stronger for Vendor B (interaction: $\beta = 1.06$, $p < 0.001$). Encounter type also showed systematic differences: relative to office visits (reference), telemedicine ($\beta = -0.52$, $p < 0.001$) and video visits ($\beta = -0.34$, $p = 0.001$) had lower odds of turnaround over 24 hours; integrative medicine visits ($\beta = 0.72$, $p = 0.002$) and OB visits ($\beta = 1.66$, $p < 0.001$) had higher odds, while procedure visits were not significantly different ($\beta = 0.12$, $p = 0.821$). At the specialty level, family practice clinicians were less likely to have a turnaround over 24 hours ($\beta = -3.03$, $p < 0.001$). Several specialties had longer turnaround time, including dermatology ($\beta = 1.46$, $p < 0.001$) and infectious diseases ($\beta = 2.50$, $p < 0.001$) (Appendix 1. Supplementary Table S3).

Do notes signed after 24 hours have more or fewer edits?

To examine whether notes signed later had more or fewer edits, we compared total note-level edit intensity between notes signed within 24 hours versus after 24 hours. Among edited notes with valid turnaround times (n = 20,062), 16,868 notes (84.1%) were signed within 24 hours of draft generation and 3,194 (15.9%) were signed after 24 hours. Long-turnaround (>24h) was associated with lower edit loads after adjustment for encounter type, clinician group, vendor, and the log-transformed number of AI-generated note sections per note (log[1+sections]). Compared with ≤24h, >24h submissions had a reduction in log edit intensity ($\beta = -0.269$, $p < 0.001$). The turnaround-by-vendor interaction was not statistically significant ($\beta = 0.073$, $p = 0.196$). The clinician random intercept variance was estimated near zero, indicating minimal residual

clustering by individual clinician behavior after these adjustments (Appendix 1. Supplementary Table S4).

Which sections, encounter types, and specialties show more edits?

This analysis included 33,496 edited note sections from 215 clinicians. In the adjusted model, edit intensity differed strongly by note section: compared with A&P, edit intensity was lower for HPI ($\beta = -0.51$, $p = 0.002$), physical exam ($\beta = -1.36$, $p < 0.001$), and results ($\beta = -1.22$, $p < 0.001$). Specialty remained associated with edit intensity relative to cardiology, with higher edit intensity in epilepsy ($\beta = 1.60$, $p < 0.001$), neurology ($\beta = 1.09$, $p = 0.001$), and physical medicine and rehabilitation ($\beta = 0.56$, $p = 0.035$), and lower edit intensity in family practice ($\beta = -0.80$, $p = 0.003$), geriatrics ($\beta = -0.87$, $p = 0.019$), hand surgery ($\beta = -0.84$, $p = 0.004$), hospitalist medicine ($\beta = -1.60$, $p < 0.001$), integrative medicine ($\beta = -0.99$, $p = 0.002$), internal medicine ($\beta = -0.53$, $p = 0.027$), obstetrics/gynecology ($\beta = -0.82$, $p = 0.004$), and ophthalmology ($\beta = -1.18$, $p = 0.003$). Vendor was not a significant predictor of edit intensity after adjustment ($\beta = 0.14$, $p = 0.591$), and encounter type was not significant in the adjusted model (all $p \geq 0.21$) (Appendix 1. Supplementary Table S5).

Discussion

This study provides empirical evidence on ambient AI documentation usage and edit patterns among clinicians in routine ambulatory care. By linking AI drafts to clinician-final text and quantifying edits, we describe where clinicians accept AI drafts as-is, where they invest effort to revise, and how these patterns relate to turnaround time and care settings, offering a practical evaluation framework that can be reused for operational monitoring and model improvement. [96]

We found that most notes containing AI drafts (84.4%) were edited by clinicians before signing. Even though no-edit rates differed across note sections and specialties, our findings suggested substantial clinician-level heterogeneity with minimal residual clustering by specialty after accounting for section and encounter types. This suggests that the tendency to sign AI-drafted sections without edits is primarily a clinician-level phenomenon. Health systems could therefore target implementation support and workflow coaching to clinicians with unusually high or low rates of unedited submission, rather than applying training or configuration changes uniformly by specialty. [97,98] These findings also reinforce that ambient AI remains a human-in-the-loop technology, where clinician review and editing are an essential part of safe use and accountability.

When edits occurred, changes were generally modest and consistent, with the greatest editing concentrated in A&P, followed by HPI. This gradient aligns with A&P as the most information-rich note section and the primary site of clinical reasoning, risk management, and decision-making. [99] In practice, this suggests prioritizing A&P and HPI for model refinement, such as clearer problem organization, and more specific and actionable plans with fewer generic statements, while concentrating implementation monitoring and quality checks on note sections where clinicians invest the most editing effort. [48]

Additionally, notes signed more than 24 hours after AI draft generation had lower edit intensity, suggesting that prolonged turnaround is not solely explained by extensive text revision. Longer turnaround may reflect other workflow factors (e.g., clinic scheduling, team documentation practices). Accordingly, efforts to reduce prolonged turnaround may benefit from workflow supports such as end-of-visit signing prompts and tighter integration between the ambient AI tool and EHR workflows. [100,101] In addition, because clinical notes support billing and

compliance, longer AI-generated drafts should not be assumed to meet specialty-specific documentation requirements; implementation should include billing/compliance guardrails and specialty-specific guidance alongside quality monitoring. Differences in editing patterns across vendor deployments likely reflect a combination of model behavior, configuration choices, and workflow integration. Given that health systems typically implement a single ambient AI solution, our findings suggest the need for specialty- and section-specific customization and monitoring (e.g., A&P vs HPI).

This study has several limitations. Including only clinical notes from ambulatory settings within a single academic health system may limit generalizability to other institutions and clinical settings. Observed editing patterns may be shaped by local implementation factors, including system configuration, workflow integration, and clinician training and support. Because these factors vary across health systems, they may influence both the amount and type of clinician editing. Our analyses quantified the frequency and intensity of edits rather than the clinical correctness of the final notes or downstream care outcomes; therefore, we could not determine whether high no-edit use reflected high-quality drafts or missed errors, or whether higher edit levels reflected necessary clinical corrections or individual documentation style. [102]

Future work should link editing patterns to independent assessments of note quality and safety, and distinguish stylistic edits of individual clinicians from clinically impactful changes to support targeted model refinement and note quality monitoring. Furthermore, prospective studies that evaluate vendor-specific performance and systematically document local implementation details could better uncover how vendor and deployment factors contribute to observed editing patterns. Overall, our proposed methods suggest an approach for health systems deploying clinical AI applications beyond ambient documentation, such as AI-drafted patient message

replies: routinely track no-edit rates, edit intensity, and turnaround time to identify clinician- and content-specific friction points, and tailor implementation support to individual workflow variation rather than assuming uniform specialty-level patterns. [67,103,104]

Conclusion

In this large-scale evaluation of two EHR-integrated ambient AI documentation systems in ambulatory settings, we found that clinicians edited the majority of AI-drafted notes before signing, and that signing drafts without edits appeared to reflect individual clinician practice style more than specialty-level norms. Documentation turnaround time was not explained by higher editing burden, as notes signed more than 24 hours after AI draft generation tended to have lower edit intensity. Section-level drill-down analysis further showed that when edits occurred, revisions were generally modest and concentrated in information-dense note sections, particularly the A&P, with specialty care clinicians editing more than primary care. Vendor-level differences suggest that operational monitoring and implementation support may require vendor-specific tuning. Together, these findings provide a comprehensive understanding of clinician editing behavior for ambient AI drafts to inform targeted model refinement, clinician-facing implementation support, and ongoing governance of ambient AI documentation in real-world practice.

CHAPTER 4: What Do Clinicians Edit in Ambient AI-Drafted Clinical Documentation? A Qualitative Content Analysis

Abstract

Objective

Ambient artificial intelligence (AI) documentation is increasingly used to draft clinical notes from patient-provider conversations, but how clinicians revise and finalize these drafts is not well understood. This qualitative content analysis study characterizes real-world edits to AI-generated drafts and identifies opportunities for improvement of AI design and the implementation process.

Materials and Methods

Eight coders analyzed clinical documentation generated by ambient AI from 200 clinical encounters. We developed an inductive coding framework with 11 codes across three categories: clinical content, terminology, and language style. Interrater reliability was assessed using Cohen's kappa. We then applied thematic analysis to synthesize patterns across the coded edits.

Results

The most frequently edited content pertained to clinical facts including orders (e.g., procedures, lab tests) (40.0%), symptoms (30.3%), medication prescriptions (27.3%), and diagnosis descriptions (25.9%). In comparison, edits related to terminology use (11.6%) and language style (7.2%) were less frequent. The results of our thematic analysis show that most edits can be categorized into one of the following five types: to correct factual errors, to address needs of medical specialty, to express diagnostic certainties, to convert patient expressions into objective assessments recorded in medical terms, and to reorganize or condense content.

Conclusion and Discussion

Clinicians routinely revise ambient AI drafts to improve accuracy and clinical specificity. Future work on AI development and clinical implementation should emphasize specialty customization and support personalized documentation practices, alongside clinician education that promotes robust and consistent review routines to ensure documentation quality.

Introduction

Ambient artificial intelligence (AI) systems are increasingly integrated into clinical workflows to support documentation by generating note drafts. [25,27] Early studies suggest they may reduce documentation burden, improve efficiency, and enhance note quality. [12,41] However, despite growing adoption, the real-world reliability and clinical appropriateness of AI-drafted documentation remain underexplored.

Prior evaluations have largely emphasized clinician and patient perspectives through surveys and interviews, focusing on usability, trust, and time savings rather than the content of the generated text. [17,19,24,42,105] A few studies compared AI-generated notes with reference text using automated performance and text-similarity metrics including the F1 score, overlap-based measures such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and embedding-based scores. However, these approaches typically produce aggregate, note-level outcomes without explicating what content was changed. [43,45,106] Expert-based note quality assessments, such as structured reviewer ratings and instruments like the Physician Documentation Quality Instrument, provide

an overall quality assessment but are often conducted on simulated or curated data, offering limited insight into the specific edits clinicians make or the reasons behind them. [48]

Limited work has compared real-world ambient AI drafts and final notes committed to electronic health record (EHR) systems to directly characterize editing behavior. Understanding these edits is essential because they show where AI drafts fall short and can reveal systematic gaps with implications for clinical safety, billing, and communication. [107] They can also inform model improvement and implementation support to better align ambient AI with clinical documentation practice.

To address this gap, we conducted a content analysis of 314 ambient AI-generated note sections from 200 paired ambient AI drafts and clinician-finalized versions, including 1,804 word-level edit operations (insertions, deletions, and replacements) across 33 specialties and 73 clinicians in an ambient AI pilot program. Using qualitative coding and thematic analysis, we examined the types of textual changes clinicians made, the content areas most frequently modified, and the broader themes that describe how clinicians refine AI-generated drafts into finalized documentation. This study offers a fine-grained empirical characterization of real-world clinician editing in ambient AI documentation, with the goal of informing more reliable and more workflow-aligned documentation support systems.

Methods

This study employed qualitative coding and thematic analysis to characterize clinician edits to AI-generated clinical note drafts at the content level. The research was conducted as part of a quality improvement initiative at the University of California, Irvine Medical Center (UCI

Health), involving clinicians across ambulatory primary and specialty care clinics using two commercially available ambient AI systems that were integrated into the Epic electronic health record (EHR) via mobile and web interfaces. Vendors are referred to as Vendor A and Vendor B to preserve contractual and institutional confidentiality. The study protocol was approved by the University of California, Irvine Institutional Review Board (IRB #7123). All data processing occurred within HIPAA-compliant secure computing environments, including UCI Health's Protected Virtual Computing Environment (PVCE) and Amazon Web Services (AWS) infrastructure.

Data Source and Sampling Strategy. Ambient AI drafts clinical content by sections, covering history of present illness (HPI), assessment and plan (A&P), physical exam, and results, which clinicians can selectively insert into EHR notes. The overall corpus of this study consisted of all AI-generated draft note sections and their corresponding clinician-finalized versions recorded between September 26, 2024, and August 5, 2025, from the University of California, Irvine Medical Center. To construct an annotation dataset for content analysis that reflects the full corpus, we drew a stratified sample of 200 notes, each containing one or more AI-drafted note sections with clinician edits, representing 33 clinical specialties. Sampling quotas were aligned with the overall corpus distribution, ensuring the inclusion of at least two notes per specialty or one in cases where only a single note was available. Quotas were also designed to achieve a balanced representation from both vendor systems in the annotated sample.

Edit Detection Using the Myers Diff Algorithm. We detected editing activity in the sample by quantifying insertions, deletions, and replacements between each ambient AI draft and the clinician-finalized note section. For each draft–final pair, we applied a word-level Myers diff algorithm to align tokens and identify edit operations, recording the span and sequence of each

change. [90] We also generated side-by-side HTML views with highlighted edits, allowing for direct comparison of the AI output against clinician revisions. Figure 2 provides a synthetic example for illustration only and does not contain patient data.

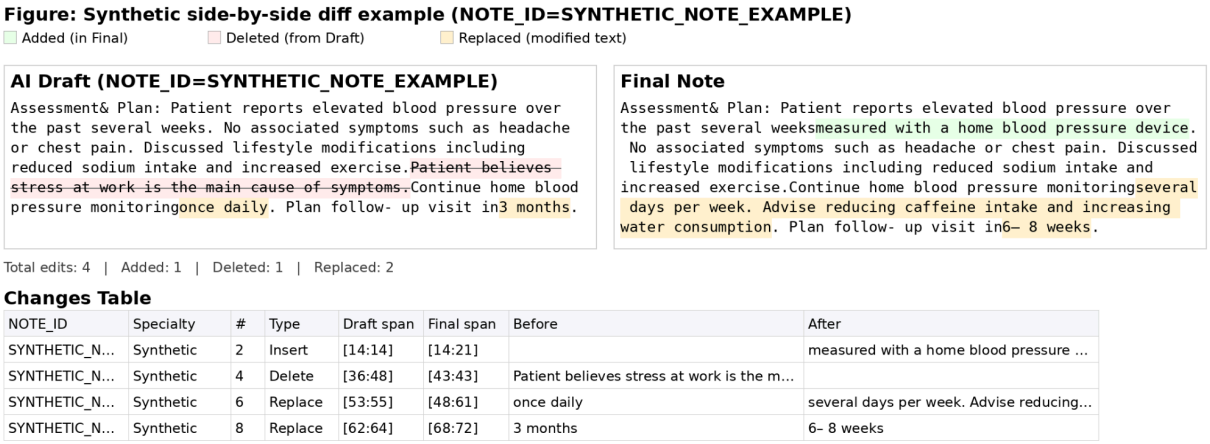

Figure: Synthetic side-by-side diff example (NOTE_ID=SYNTHETIC_NOTE_EXAMPLE)

Added (in Final) Deleted (from Draft) Replaced (modified text)

AI Draft (NOTE_ID=SYNTHETIC_NOTE_EXAMPLE)

Assessment& Plan: Patient reports elevated blood pressure over the past several weeks. No associated symptoms such as headache or chest pain. Discussed lifestyle modifications including reduced sodium intake and increased exercise.~~Patient believes stress at work is the main cause of symptoms.~~Continue home blood pressure monitoringonce daily. Plan follow- up visit in3 months.

Final Note

Assessment& Plan: Patient reports elevated blood pressure over the past several weeksmeasured with a home blood pressure device. No associated symptoms such as headache or chest pain. Discussed lifestyle modifications including reduced sodium intake and increased exercise.Continue home blood pressure monitoringseveral days per week. Advise reducing caffeine intake and increasing water consumption. Plan follow- up visit in6– 8 weeks.

Total edits: 4 | Added: 1 | Deleted: 1 | Replaced: 2

Changes Table

NOTE_ID	Specialty	#	Type	Draft span	Final span	Before	After
SYNTHETIC_N...	Synthetic	2	Insert	[14:14]	[14:21]		measured with a home blood pressure ...
SYNTHETIC_N...	Synthetic	4	Delete	[36:48]	[43:43]	Patient believes stress at work is the m...	
SYNTHETIC_N...	Synthetic	6	Replace	[53:55]	[48:61]	once daily	several days per week. Advise reducing...
SYNTHETIC_N...	Synthetic	8	Replace	[62:64]	[68:72]	3 months	6– 8 weeks

Figure 2. Synthetic Example of Side-by-Side Visualization of Clinician Edits to an AI-Generated Clinical Note

Sentence-Level Content Analysis. We defined an edit unit as the smallest contiguous set of sentences that fully captured each diff-detected edit span, including adjacent or overlapping edits, to preserve clinical context and meaning. To align draft and final text at the sentence level, we used a dynamic approach that allowed flexible mappings, including one-to-one, one-to-many (split), many-to-one (merge), and many-to-many alignments. Each edit unit was labeled by its alignment pattern and quantified by the number of insertions, deletions, and replacements it contained, capturing both within-sentence phrasing changes and larger structural revisions.

Coding Framework and Procedure. An inductive coding framework was developed to characterize the types and functions of clinician edits. [108] Two researchers (YG, DH) independently reviewed a subset of edit units to develop an initial codebook. To support comprehensive inductive discovery across heterogeneous documentation patterns, we developed the codebook using the pooled set of note section pairs from 200 notes spanning both vendor

systems and a broad range of clinical specialties. The codebook was then iteratively refined through consensus meetings and multiple rounds of coding with eight coders, all with experience in health informatics research and qualitative analysis. The full coding team then conducted an initial calibration round in which all eight coders applied the draft codebook to the same set of edit units; discrepancies were resolved through adjudication meetings and email discussion, and the codebook was iteratively refined. Interrater reliability was assessed using Cohen's κ across major code categories. [109] We finalized the codebook after achieving acceptable interrater agreement across coders. The final refined framework consisted of three top-level code categories capturing distinct categories of editing behavior (entity-level clinical content changes, terminology and wording adjustments, or language and style modifications), along with 11 subcodes representing more granular phenomena. [110,111]

Using the finalized codebook, coders applied labels to each sentence-level edit unit. We treated each code as a binary label, present or absent, for each sentence-level edit unit, allowing for multiple codes per unit. Across eight coders, we computed Cohen's kappa for each code for each annotator pair and macro-averaged kappas across codes with equal weight; overall reliability was the mean of the 28 pairwise macro-averaged kappa values. Because edit units often reflected specialty-specific documentation practices and patient-specific clinical context, we anticipated partial overlap at code boundaries and therefore used iterative calibration and consensus adjudication to ensure a stable codebook prior to thematic synthesis. All coding was conducted in structured spreadsheets using the finalized codebook. To maintain independent coding, each coder worked in a separate spreadsheet. For reference when needed, coders could consult color-coded HTML diff visualizations of the paired draft and final text to view each edit unit alongside the exact textual changes within the full note section.

Table 9. Qualitative coding schema

Primary Code Category	Code	Example(s)	Description
Entity-level: Clinical Content	Medication (E–Med)	Before: "Keppra one pump BID"; After: "Keppra 500 mg BID"	Edits that add, remove, or modify medication information, including medication name, dose, frequency, route, formulation, timing, or adherence details.
	Symptom (E–Sym)	Before: "Cramping"; After: "Cramping at night"	Edits that add, remove, or modify symptom mentions, including presence/absence, severity, location, timing, duration, triggers, or associated features.
	Diagnosis (E–Dx)	Before: "epilepsy"; After: "refractory focal epilepsy"	Edits that add, remove, or modify diagnoses/conditions or diagnostic labels, including specificity (type/subtype), qualifiers, or diagnostic framing.
	Procedure / test / lab order (E–Test)	Before: "EEG not done"; After: *(deleted)*	Edits to tests or procedures and related orders, including labs and imaging, whether performed in-clinic or externally, such as changes to test type, status, referrals, timing/location, or results.
	Demographics / ID (E–ID)	Before: "[WRONG-SPELLED NAME]"; After: "[CORRECTED NAME]"	Edits that add, remove, or modify patient identifiers or demographic details (e.g., age, name, pronouns, PHI placeholders)
	Social context (E–Soc)	Before: *(none)*; After: "lives with spouse"	Edits that add, remove, or modify social history/context (e.g., tobacco/alcohol/substance use, living situation, education, occupation, support system, marital status, habits/behaviors) [112]
	Past medical history (E-Hx)	Before: (none); After: "Previously saw nephrologist at [LOCATION]"	Edits that add, remove, or modify past medical history, including comorbidities, prior events, prior specialty care, surgeries, and relevant historical conditions.
Terminology and Wording	Lay–jargon substitution	Before: "pain"; After: "myalgia"	Edits that substitute clinical jargon with lay language (or vice versa) while

	(T–Lay)		preserving the underlying meaning.
	Abbreviation Expansion/Contraction (T–Abbr)	Before: “right arm”; After: “R arm”	Edits that expand, introduce, or standardize abbreviations and acronyms.
Language and Style	Empathy / rapport / politeness (S–Emp)	Before: *(none)*; After: “It was a pleasure to see Ms. [NAME]”	Edits that change interpersonal tone (rapport, empathy, politeness, patient-centered phrasing) without changing core clinical content.
	Certainty / hedging (S–Cert)	Before: “due to”; After: “likely due to”	Edits that increase or decrease expressed certainty, probability, attribution, or diagnostic commitment (e.g., hedging, qualifiers, speculative language).

Thematic Analysis. After all edit units were coded using the finalized codebook, we conducted a thematic synthesis to identify higher-level patterns in clinician revision behavior. Vendor identities were masked during codebook development and edit-unit coding to reduce bias and support consistent application of the codebook. After coding was complete, vendors were specified for a vendor-stratified descriptive summary of code distributions and to assess whether themes were represented in both vendor groups. Two lead researchers (YG, DH) independently reviewed code distributions, representative edit units, and analytic memos to examine how frequently co-occurring codes reflected shared underlying revision behavioral patterns. Codes were iteratively grouped based on conceptual similarity and the clinical intent they represented (e.g., factual correction, clinical precision, documentation standardization). Through this process, we aggregated code-level edits into five higher-level themes that capture recurring clinician revision patterns in AI-drafted notes. Themes were refined through iterative team discussion and disagreements in theme boundaries were resolved through consensus. The resulting themes

represent interpretive syntheses of multiple code types and are intended to characterize how clinicians systematically revise AI-drafted documentation.

Results

Dataset Characteristics

A total of 314 note section pairs from 200 notes containing ambient AI-drafted sections and finalized versions were included in the annotated sample, representing 73 clinicians, 200 patients, and 33 clinical specialties. Clinicians could generate more than one ambient AI section draft within the same note (for example, only A&P, or both HPI and A&P). We coded section-level draft–final pairs, grouped by note ID for tracking. The complete sampled notes for qualitative annotation included a history of present illness section (n = 149) and an assessment and plan section (n = 135), while fewer included physical examination (n = 18) or results (n = 12). Across all sampled draft–final section pairs, clinicians made 1,804 token-level modifications, including 556 insertions, 393 deletions, and 855 replacements. The sample covered 33 specialties; family practice contributed the most notes (n = 54), followed by rheumatology (n = 22) and internal medicine (n = 20), with the remaining notes distributed across a broad set of lower-volume specialties (Table 10).

Table 10. Characteristics for paired ambient AI drafts and clinician-finalized notes

Category	**Subcategory**	**Overall**	**Vendor A**	**Vendor B**
Overall (sample-level)	Study period	2024-09-26 to 2025-08-05		
	# of notes	200	109	91
	# of clinicians	73	35	38
	# of patients	200	109	91
AI-drafted Note Section Pairs	# of History of Present Illness (HPI) pairs	149	73	76
	# of Assessment & Plan (A&P) pairs	135	87	48
	# of Physical Exam pairs	18	11	7
	# of Results pairs	12	5	7
Token-level editing activity	# of Insertions	556	311	245
	# of Deletions	393	198	195
	# of Replacements	855	466	389
Specialty Distribution (n = # of note)	Family Medicine	54	39	15
	Rheumatology	22	20	2
	Internal Medicine	20	17	3
	Orthopaedic Surgery	15	1	14
	Geriatric Medicine	15	0	15
	Others	74	32	42

** Other specialties (Overall n=74; Vendor A n=32; Vendor B n=42) included: Hand Surgery (A=1, B=9); Neurology (A=0, B=7); Nurse Practitioner (A=4, B=2); Gastroenterology (A=3, B=0); Ob/Gyn (A=0, B=3); Physician Assistant (A=0, B=3); Cardiology (A=2, B=0); Colon and Rectal Surgery (A=0, B=2); Dermatology (A=2, B=0); Emergency Medicine (A=0, B=2); Endocrinology (A=2, B=0); Epilepsy (A=0, B=2); General Surgery (A=0, B=2); Head and Neck Surgery (A=0, B=2); Hematology/Oncology (A=0, B=2); Hospitalist (A=2, B=0);*

Infectious Diseases (A=2, B=0); Integrative Medicine (A=2, B=0); Neurosurgery (A=2, B=0); Ophthalmology (A=2, B=0); Optometry (A=2, B=0); Otolaryngology (A=2, B=0); Physical Medicine and Rehab (A=2, B=0); Psychiatry (A=0, B=2); Urology (A=1, B=1); Vascular Neurology (A=0, B=2); Orthopedics (A=0, B=1); Vascular Surgery (A=1, B=0).

Distribution of Qualitative Edit Codes

We coded 713 edit units extracted from 314 AI draft–final note section pairs. Vendor A contributed 373 edit units and Vendor B contributed 340 edit units. After an initial training phase and two rounds of independent coding with consensus discussions, interrater reliability across eight coders reached a moderate level of agreement (overall Cohen's $\kappa = 0.568$), which is consistent with multi-label coding of heterogeneous clinical edits; disagreements were resolved through adjudication and consensus discussions. Because a single edit unit could receive multiple codes, results are reported as the frequency of edit units containing each code; percentages therefore do not sum to 100 percent. Table 11 summarizes the distribution of edit codes overall and by vendor.

The most common edit types involved clinical content. Procedure, test, and lab orders edits (E–Test) appeared in 285 edit units (40.0%). A common pattern was clarifying order-specific metadata, including the test type, the intended location or service, and the timing or status of the order when that detail affected interpretability. This was followed by symptom changes (E–Sym) in 216 (30.3%). For symptom edits, clinicians commonly refined the symptom description to improve clinical specificity. This included adjusting the symptom label, and specifying severity or pattern when the draft was vague. Clinicians also removed symptom descriptions that were not part of the patient's current presentation or that appeared to be carried over from prior context without supporting evidence from the encounter. Medication modifications (E–Med) appeared in 195 (27.3%), and edits most often made the regimen actionable by adding or correcting dose, frequency, route, or administration details. Clinicians also standardized

medication nomenclature, such as correcting the drug name, aligning the term with common clinical usage, or clarifying which formulation was intended. A smaller but recurring pattern was medication reconciliation, where clinicians removed a medication that was not relevant to the current plan or clarified whether the patient was currently taking it. Diagnosis refinements (E–Dx) appeared in 185 (25.9%). Clinicians frequently revised the diagnostic framing by narrowing a broad label to a more specific condition, aligning the diagnosis term with the symptom description, or clarifying whether the diagnosis was established versus under evaluation. Edits involving past medical history (18.9%) and social context information (17.8%) were also frequent, whereas terminology and wording edits such as jargon versus lay (7.2%) and abbreviation changes (4.5%) and language and style edits such as certainty or hedging (4.1%) and empathy and rapport (3.1%) were less common.

Table 11. Distribution of qualitative edit codes overall and by vendor

Code	**Vendor A, n (%)**	**Vendor B, n (%)**	**Total, n (%)**
E-Test	160 (42.9%)	125 (36.8%)	285 (40.0%)
E-Sym	121 (32.4%)	95 (27.9%)	216 (30.3%)
E-Med	120 (32.2%)	75 (22.1%)	195 (27.3%)
E-Dx	111 (29.8%)	74 (21.8%)	185 (25.9%)
E-Hx	73 (19.6%)	62 (18.2%)	135 (18.9%)
E-Soc	60 (16.1%)	67 (19.7%)	127 (17.8%)
E-ID	37 (9.9%)	26 (7.6%)	63 (8.8%)
T-Lay	25 (6.7%)	26 (7.6%)	51 (7.2%)

T-Abbr	22 (5.9%)	10 (2.9%)	32 (4.5%)
S-Cert	17 (4.6%)	12 (3.5%)	29 (4.1%)
S-Emp	11 (2.9%)	11 (3.2%)	22 (3.1%)

** Percentages are calculated using the total number of coded edit units within each subgroup as the denominator (Vendor A: N=373; Vendor B: N=340; Overall: N=713). Because an edit unit can receive multiple codes, counts may sum to >100%.*

Thematic Synthesis of Clinician Edits

Editing activity varied by note section, with the highest concentration in narrative sections such as HPI and Assessment and Plan and fewer edits in Physical Exam and Results. Vendor-stratified summaries showed broadly similar code distributions, and all five themes were observed in notes from both vendors. Given the qualitative aim and descriptive nature of vendor stratification, we present themes pooled across vendors for interpretability. Representative before/after examples are shown in Appendix 2.

Theme 1- Clinicians correct factual errors in AI drafts.

These edits corrected misspelled names and pronouns introduced during transcription and fixed inaccurate time-stamp information. They also refined event descriptions by updating prior medical history details, such as changing a hospitalization to a more accurate emergency department visit, and by correcting anatomic and severity information (e.g., laterality, joint location, stenosis severity). Clinicians further corrected medication and test/procedure details, including drug names and dosing frequency, sedation type, and imaging modality. These edits improved the accuracy of the clinical note across aspects including medical history, demographics, and treatments.

Theme 2- Clinicians refine generic drafts into specialty-appropriate documentation.

Clinicians added clinical details commonly required for specialty documentation, such as symptom onset and trajectory, medication regimen specifics (dose, schedule, and whether use was as needed), and test or treatment context that shaped the plan. In neurology/epilepsy, edits emphasized temporal progression and neuroanatomical localization, reframing brief summaries into problem-oriented assessments that track symptom course, differentiate neurologic problems, and document medication intervals or titration. For example, a neurology progress note was expanded by adding detailed trauma history and symptom characterization across neurologic problems, and rewriting the assessment into a specialized, multi-layered plan. In procedure-oriented specialties such as hand surgery and orthopaedics, edits replaced general descriptions with structured, objective findings and procedure- or imaging-specific terminology, making decisions explicit through digit-level exam metrics or dated radiology findings. A hand surgery note was refined by replacing general symptom and exam language with digit-specific, quantified findings (e.g., finger-level localization, monofilament scores) and by stating clear procedural intent, such as proceeding with right index and ring trigger finger release. In cardiology, clinicians expanded symptom documentation beyond brief general phrasing by specifying exertional context and exercise tolerance, linking dyspnea, chest symptoms, or palpitations to what the patient can do and under what conditions, so the note reflected cardiovascular-style functional assessment rather than a generic symptom checklist.

Theme 3- Clinicians revise overly certain language to match the available evidence with clinical wording.

Across notes, clinicians often replaced definitive causal or diagnostic wording with more appropriately qualified language when the documentation did not support a firm conclusion, such

as changing “due to” to “likely due to” and revising “no retinal or optic nerve disease” to “no obvious retinal or optic nerve disease.” They also clarified when a condition was being evaluated rather than established, for example, by reframing plans aiming to “definitively rule out or confirm AFib” as monitoring in the setting of “no confirmed diagnosis of AFib.” Similarly, diagnostic labels were adjusted to match test evidence, such as shifting a stated dermatomyositis diagnosis to a suspected drug eruption after a negative myositis panel.

Theme 4- Clinicians normalize conversational narratives into standardized chart language.

Clinicians frequently edited transcript-like language to align the note with clinician-facing documentation standards by removing or reframing patient-attributed causal explanations and speculative risk phrasing, and by replacing them with objective, clinically interpretable statements. For example, subjective clauses such as “which she believes are related to her thyroid condition” or “she recalls an instance when…” were removed or rewritten as verifiable clinical information, such as the current levothyroxine dose or the patient’s expressed concern about thyroid level fluctuations. More broadly, conversational and vague phrasing was rewritten into conventional chart language and medical terminology, such as changing “she’s doing okay” to “patient reports stable condition” and replacing lay descriptions like “a disintegrating disc” with “a bulging disc”.

Edits also standardized abbreviations and shorthand, for example, using “MCV” and inserting routine clinical abbreviations such as “EGD”, “abnl”, and “CIN II”. Clinicians increased technical specificity by naming the exact device, procedure, or structure, such as revising “partial amputation” to “ray amputation,” and specifying the ECU tendon. Medication references were similarly clarified by replacing generic wording with the specific drug name, such as “used the

medication" to "used Qsymia" and "Lexapro" to "escitalopram," with accompanying refinements to dosing or decision language.

Theme 5- Clinicians reorganize, relocate, and condense AI drafts.

Clinicians frequently restructure AI-generated drafts to make the note easier to navigate and use, reorganizing content to improve scanability and clarity of clinical plans while trimming material that did not match the visit. Dense narrative was often converted into structured formats, for example, with long paragraphs rewritten as bullet lists, problem narratives turned into numbered plans, and free text reorganized under problem-oriented headings. Clinicians also relocated information to more appropriate sections, such as moving social or surgical history into labeled past history fields and separating imaging descriptions into a dedicated radiology-style section with dated, objective findings. Edits also reflected the process of relevance triage where duplicative or boilerplate text was consolidated, and content that appeared misaligned with the clinical context was removed, including passages on specialty-irrelevant medication management.

Discussion

Clinicians substantially modify AI-generated drafts, and those edits follow repeated patterns that align with the five themes in our qualitative analysis, which point to specific opportunities for improvement in ambient AI documentation tools and their implementation.

Clinicians often correct inaccuracies or hallucinations, especially in discrete details that can change clinical meaning, such as names, medication dosages, and test or procedure attributes. Even when the overall narrative in AI drafts was broadly accurate, these errors are difficult for

downstream readers to detect and can propagate if copied forward. This finding suggests that improving factual reliability of ambient AI tools should be a top priority, paired with workflow features that help clinicians quickly verify the fields that most often need to be double-checked. [113] At present, most ambient AI tools do not directly access the patient's chart, so certain errors, such as names or other identifiers, cannot be automatically cross-checked even when the correct information already exists in the EHR. If chart integration becomes feasible, many of these errors could be corrected upstream. In the meantime, for institutions integrating ambient AI into routine practice, training for clinicians should go beyond a generic requirement to "review the draft" and instead teach a consistent verification routine that prioritizes the most error-prone details. [98,114]

Clinicians' tailoring AI drafts to be more specialty-specific suggests that a single, generic note structure and style is often insufficient for specialty documentation needs. An actionable response is to support specialty-tuned drafting behavior through templates and prompts that reliably elicit the details that matter for a given discipline, rather than broadly relying on the subjective–objective–assessment–plan (SOAP) structured format. [115–117] For institutions, this implies that ambient AI adoption should be paired with specialty-level configuration and evaluation aligned with each specialty's documentation norms and clinical priorities. [118]

Overly certain language in ambient AI drafts can outpace the available clinical evidence, creating extra work for clinicians and increasing the risk of carrying forward unverified conclusions. Beyond clinical risk, this pattern may raise medicolegal concerns if AI-generated phrasing systematically nudges notes toward stronger diagnostic commitments than the encounter supports. [119] In our data, clinicians spent substantial effort dialing back overly certain AI language. Actionably, AI developers can in response design models that default to appropriately qualified language and conventional diagnostic framing, and that avoid stating conclusions more

strongly than the note supports. [120] At the practice level, this theme points to a clear focus for governance and education, reinforcing a consistent review habit around diagnostic uncertainty and ensuring the system's outputs do not nudge documentation toward over-commitment when diagnoses are not yet established.

Clinicians often rewrote transcript-like text into a usable clinical record, shifting conversational phrasing into objective, clinically interpretable documentation. This mismatch may be amplified by how the system structures conversational content into the Subjective section (SOAP). [115] In our data, clinicians frequently condensed and reorganized that section, suggesting potential misalignment between how the draft is assembled and how clinicians typically write and review notes. Ambient AI developers can strengthen attribution-aware writing that keeps a clear boundary between patient reports and clinician assessment without adding interpretation that the note does not support. Terminology normalization can also help increase precision without causing note bloat. [37,61,121,122] For clinicians, this theme is a reminder that review is not only error correction; it is also ensuring the note reads as a professional record that others can quickly interpret and act on.

Recurring patterns of note restructuring and condensing suggest that draft organization is a major determinant of whether ambient notes are usable at the point of care. For example, ambient capture can include social history and social determinants of health (SDOH) information that is clinically relevant because it can shape access to care, home and community support, and adherence. However, in the context of certain encounters, some of these elements may be secondary to the immediate diagnostic and clinical decision making, particularly when captured with high granularity (e.g., transportation needs narratives). [123] Clinicians often shortened or removed these details when they did not inform the assessment or plan. These edits reflect a

familiar tradeoff in real-world documentation. As notes get longer, they become harder to review efficiently, and important information can be harder to find. [82,124] An actionable implication is to prioritize tool behaviors that produce more structured drafts, route content into predictable sections, and give clinicians simple ways to control verbosity without losing clinically important context. At the organizational level, this supports setting expectations for note structure and routinely monitoring note length and usability as part of ambient AI governance.

This study has several limitations. Our qualitative analysis drew on the AI draft–final note section pairs from 200 notes, and some specialties were represented by only a small number of cases, so we may not fully capture specialty-specific nuances. The dataset also comes from a single health system and a single deployment context, which may limit how well these patterns generalize to other institutions, care settings, or ambient AI products. Vendor-stratified summaries are descriptive, and the study was not designed or powered for formal between-vendor comparisons. The modest qualitative sample and broad thematic categories might limit sensitivity to subtle vendor-specific differences. Interrater agreement was moderate (overall Cohen's $\kappa = 0.568$). This is consistent with a multi-coder, multi-label codebook applied to heterogeneous notes spanning diverse specialties and patient contexts, where code boundaries can overlap for clinically adjacent edits. We mitigated this through iterative calibration, adjudication, and consensus discussions, and we generated themes based on convergent patterns across codes and examples rather than any single coder's label. [125] We only analyzed text changes without the source audio, user interface cues, or clinicians' stated intent. Therefore, we cannot determine why a given edit was made or whether it reflects individual style, local documentation norms, or risk tolerance. Finally, we did not assess the clinical correctness of the

finalized notes or downstream impacts such as care outcomes, time burden, or corresponding change in billing codes.

Taken together, our findings translate routine clinician edits into concrete actionable insights for both tool improvement and real-world implementation. The patterns we observed suggest that ambient AI should be evaluated and tuned for specialty needs and clinician-to-clinician variation. Future work should incorporate targeted semi-structured interviews to clarify the rationales and workflow factors behind common edits and to validate our interpretation of these patterns. Future work should also investigate whether specific edit patterns are associated with continuity of care. For example, edits that clarify the problem list or tighten the assessment and plan may make notes more usable for downstream readers. Studies should also examine operational performance, including turnaround time, downstream review effort or follow-up work, and whether term-level changes affect billing or coding outcomes. Finally, future work should evaluate safety-critical signals, especially the correction of factual inaccuracies and the moderation of overly certain language when diagnoses remain uncertain.

Conclusion

Our qualitative content analysis of clinician edits to AI drafts provides a comprehensive view of what clinicians changed and where improvement is needed. Clinicians corrected mistakes, replaced overly generic statements with specialty-appropriate detail, and tempered overly certain language to better match the available evidence using standard clinical wording. They also rewrote conversational, patient-attributed narrative into professional documentation and reorganized drafts into clearer, more structured notes. Ambient AI tools should be designed around how clinicians document in different specialties and should avoid wording that goes

beyond the evidence in the encounter. Health systems can use recurring edit patterns to identify target implementation support and training that improve day-to-day documentation efficiency.

CHAPTER 5: Understanding Clinician Edits to Ambient AI Draft Notes: A Feasibility Analysis Using Large Language Models

Abstract

Ambient AI documentation tools generate draft notes that clinicians can review and edit before signing off in electronic health records. Scalable computational approaches to characterize how clinicians modify drafts remain limited, yet are essential for evaluating and improving AI effectiveness. We examined the feasibility of a few-shot prompted large language model (LLM) for categorizing sentence-level edits between AI drafts and final documentation. We developed five label-specific binary models targeting medication, symptom, diagnosis, orders/tests/procedures, and social history edits, and refined prompts using adversarial negatives and verification gates. Evaluation was performed against a human-annotated corpus. Medication and symptom models achieved promising performance (F1=0.787 and 0.780), whereas remaining models were precision-limited. Errors clustered in long, complex edits and category-boundary ambiguity. Therefore, prompt engineering is reliable for categorizing edits with explicit clues, while for complex context-dependent categories, it is better suited for triage by flagging edits for human review.

Introduction

Ambient artificial intelligence (AI) documentation systems can draft clinical notes by transcribing and summarizing clinical visit audio, shifting documentation into a workflow in which clinicians review and revise AI-generated content before committing them to electronic health record (EHR) systems [27,41,83,118]. Prior studies have emphasized benefits such as reduced documentation burden, improved efficiency, and favorable user experience [17,22,42,126]. However,

these evaluation metrics offer limited visibility into where AI drafts fall short. The edits clinicians make to AI drafts can provide a direct lens on system reliability and usefulness by revealing what the draft omitted, or framed in ways that clinicians judged insufficient for clinical communication and note quality. In Chapter 4, we compared ambient AI-drafted notes with clinician-finalized versions, derived sentence-level edit units that retain local context to interpret each modification, and manually annotated each sentence unit of analysis with an editing category taxonomy [127,128]. This framework supports fine-grained characterization of what clinicians change, but extending it to routine deployment remains challenging.

Manual annotation is resource-intensive, making it difficult to quantify how edit patterns vary across specialties and note sections [129]. Traditional natural language processing (NLP) classifiers also face practical and methodological barriers in this setting since edited content is often heterogeneous, with multiple changes spanning different clinical concepts. For example, diagnosis-related edits may hinge on subtle framing or implied reasoning that is difficult to capture with keyword-based signals. In addition, supervised classifiers typically require substantial task-specific training data and ongoing maintenance as documentation templates, clinical language, and ambient AI outputs evolve, limiting their usefulness for rapid, iterative measurement in real-world deployments [130]. Large language models (LLMs) prompting offer a pragmatic path toward scaling edit categorization because these labels depend on clinical meaning rather than surface similarity alone (e.g., lexical overlap or edit distance) [131–133]. In practice, feasibility of prompt engineering depends on how explicitly the target information is expressed in the local text unit and the amount of clinical context required to interpret it. For instance, medication edits often contain explicit cues such as drug names and regimen attributes,

whereas diagnosis-related edits may be subtle, context-dependent, and difficult to separate from symptom descriptions or narrative reframing [134].

This paper evaluates the feasibility of using few-shot prompting LLMs to categorize clinicians' edits to AI drafts using sentence-level edit units derived from paired draft–final notes. Building on our prior edit unit cohort and label reference annotations in Chapter 4, we reframe the task as five label-specific binary models (medications, symptoms, diagnoses, orders/procedures, and social context), each required to output a structured decision with evidence content from the *before* and *after* note text [127]. We designed a prompt refinement strategy for performance control and auditability, including adversarial negatives and an evidence verification gate. We operationalize feasibility in terms of (i) categorization performance (precision, recall, and F1) and (ii) operational output completion rate under runtime constraints. Importantly, we evaluate feasibility under realistic deployment constraints for protected health information (PHI) including limited GPU resources under HIPAA-compliant environments, and restricted use of external APIs or additional training levers. Our goal is to identify which edit categories support high-precision automated monitoring versus those better suited for high-recall candidate retrieval with targeted human review.

Methodology and Materials

Data and Task Formulation

We used data from University of California, Irvine Health, where two commercial ambient AI systems were piloted in outpatient settings from late 2023 through mid 2025. The source data include ambient AI-generated draft note sections and the corresponding clinician-finalized note sections. Clinicians can selectively choose among four note sections to insert into the notes and

review, including History of Present Illness, Physical Exam, Assessment & Plan, and Results. This feasibility study built on the paired draft–final section dataset and sentence-level edit units developed in Chapter 4 [127]. In Chapter 4, we sampled notes to represent various specialties and note sections, and aligned AI-drafted and clinician-finalized note sections to derive sentence-level edit units that retain sufficient surrounding context to interpret each change. The sentence-level edit units served as the unit of analysis for this study and were constructed to be non-overlapping within a section by merging adjacent/overlapping text spans into a single contiguous unit. Each unit of analysis consists of a *Before* span from the AI draft and an *After* span from the clinician-final text. The resulting gold-standard corpus comprised 314 draft–final note section pairs from 200 clinical encounters, and 713 sentence-level edit units.

In this feasibility study, we operationalized edit classification as a set of label-specific binary categorization tasks at the edit-unit level. Each model evaluated whether a target edit category was present in the *Before* and *After* text for a given edit unit. We focused on five clinically meaningful categories identified that are routinely documented across outpatient encounters and are common targets of clinician revision: medication-related edits (E-Med), symptom-related edits (E-Sym), diagnosis-related edits (E-Dx), test/order/procedure-related edits (E-Test), and social-context edits (E-Soc). We separated the annotated edit units into three non-overlapping subsets: (1) a training set (n=313) used only to select few-shot examples and draft initial prompts, (2) a development set (n = 200) used for iterative prompt refinement and error analysis, and (3) a held-out test set (n = 200) that was frozen and used only for final evaluation and reporting performance. We performed this random split once on the full annotated edit-unit dataset and reused the same partitions for all five models. Edit category definitions and

illustrative de-identified examples are summarized in Table 12. This study is approved by University of California, Irvine Institutional Review Board (UCI IRB #7123).

Table 12. Edit category definitions and de-identified *Before/After* examples for binary classification

Edit category	**Definition**	**Example Sentence (AI-drafted)**	**Example Sentence (clinician-finalized)**
Medication-related edits (E-Med)	Any edit that adds, removes, or changes medication-related information, including details such as regimen attributes (e.g., dose, frequency, route, formulation, timing), medication changes (e.g., start, stop, switch, titrate), adherence status, or medication list content.	*"Current medications include amlodipine and gabapentin."*	*"Current medications include amlodipine."*
Symptom-related edits (E-Sym)	Any edit that adds, removes, or modifies patient-reported or clinician-documented symptoms, including presence or absence, severity, duration, timing, location, quality, triggers or relievers, associated symptoms, or symptom course.	*"Reports intermittent chest tightness when climbing stairs; denies shortness of breath."*	*"Reports intermittent chest tightness with exertion; denies dyspnea at rest."*
Diagnosis-related edits (E-Dx)	Any edit that adds, removes, or changes diagnostic or condition labeling or diagnostic framing, such as refining diagnostic specificity (e.g., subtype, acuity, stage), changing problem naming, or revising assessment language in a way that alters the condition being described.	*"Assessment: possible viral illness."*	*"Assessment: acute upper respiratory infection."*
Test/order/procedure-related edits (E-Test)	Any edit involving tests, labs, or procedures, including orders, referrals, when content is added, removed, changed, including	*"Plan: order laboratory testing."*	*"Plan: order CBC and CMP."*

	planned and completed.		
Social-context edits (E-Soc)	Any edit that adds, removes, or revises social history information, such as living situation, employment or finances, insurance, transportation needs, caregiving support, education or literacy, food or housing insecurity, or substance use.	*(null)*	*"Denies tobacco use; drinks alcohol socially."*

** null indicates the span was absent on that side of the aligned draft–final pair.*

LLM inference and prompt design

We conducted few-shot, prompt-based inference using the open-weight instruction-tuned model meta-llama/Llama-3.2-3B-Instruct implemented with Hugging Face Transformers [135,136]. All experiments were run on a single NVIDIA T4G GPU (16 GB VRAM) with NVIDIA driver version 580.95.05 and CUDA 13.0. Inference was performed with PyTorch 2.9.1 and Hugging Face Transformers 4.57.6 (Accelerate 1.12.0) using the standard *model.generate()* API. We used deterministic decoding (do_sample=False) with max_new_tokens=100 and a per-request time limit of 30 seconds. Prompts were processed one edit unit per request (batch size = 1). All data processing and model inference were conducted within a HIPAA-compliant, institution-approved AWS environment with appropriate access controls. A system message specified the target edit category and required a structured JSON output, and a user message provided metadata (note ID, section name, edit-unit order) along with the edit-unit text [137]. Edit-unit order is a within-section sequence number used to uniquely index multiple edit units within the same note section.

Prompts were refined iteratively on the development set through error analysis, including adding targeted adversarial negative examples reflecting common confusions for each category and introducing a brief structured self-check that required the model to apply inclusion and

exclusion criteria before deciding [138,139]. In addition, we implemented a prompt-level verification gate by requiring that positive predictions be supported by qualifying, verbatim evidence excerpts from the edit unit. Specifically, the prompt instructed the model to output *present=true* only when it could quote category-appropriate "anchor" evidence (e.g., medication anchors for medication-related edits, diagnostic labels for diagnosis-related edits) and to output *present=false* with evidence marked as N/A otherwise. During the development stage, we also experimented with prompts that requested stepwise rationale; the final prompts retained only the structured self-check and the evidence-based output constraints, without collecting reasoning details [140]. For each model, we used a set of few-shot examples consisting of approximately 4–6 positives and 18–20 targeted negatives; the composition of examples was adjusted during development to address observed error patterns, and then frozen for the held-out test evaluation. All models shared the same prompt template and JSON output schema, with category-specific components limited to the operational definition, boundary/anchor rules, and examples. We summarize prompt structure in Figure 3. Full prompts are available from the first author upon reasonable request.

System prompt Structure:

- **Task:** Decide whether the edit unit contains a medication-related change
- **Operational definition:** What counts as a medication-related edit
- **Boundary rules**: What does not count (common confusions and hard negatives)
- **Mandatory anchor requirement**: Evidence must include a qualifying "medication anchor" that is part of the change
- **Decision rule:** When to output *present=true* vs *present=false* (default to *false* if unsure)
- **Evidence rule:** Verbatim snippets; require *Before* and *After* when available; "N/A" if *present=false*
- **Output format:** JSON only (fixed keys: code, present, evidence)
- **Few-shot examples:** Small set of positives + adversarial negatives

User message (example)

- Note ID: <id>
- Section: <section>
- Edit unit order: <order>
- Before: "Continue metformin 500 mg BID."
- After: "Increase metformin to 1000 mg BID."

Figure 3. Prompt structure for edit classification (example: medication-related edits).

Structured outputs, post-processing, and evaluation

Models were instructed to return a single JSON object containing a binary decision and a short evidence field with verbatim excerpts from the edit unit when *present=true*. Because model generations might produce malformed JSON, we implemented a tolerant parser that extracts the first JSON-like object from the response and supports lightweight normalization (e.g., trimming non-JSON prefixes/suffixes and correcting minor quoting/punctuation issues) before parsing. In the final held-out evaluation, all outputs were valid JSON and no further edits were required. Instances that produced no output due to runtime limits or other execution errors were logged and excluded from metric computation. Prompt refinement was completed before running the held-out test set. Final performance was assessed on the held-out set of 200 edit units. We report label-specific precision, recall, and F1 on instances with successfully parsed outputs. We manually reviewed false positives (FPs) and false negatives (FNs) for each best-performing model and summarized recurring failure modes. Figure 4 summarizes the end-to-end workflow.

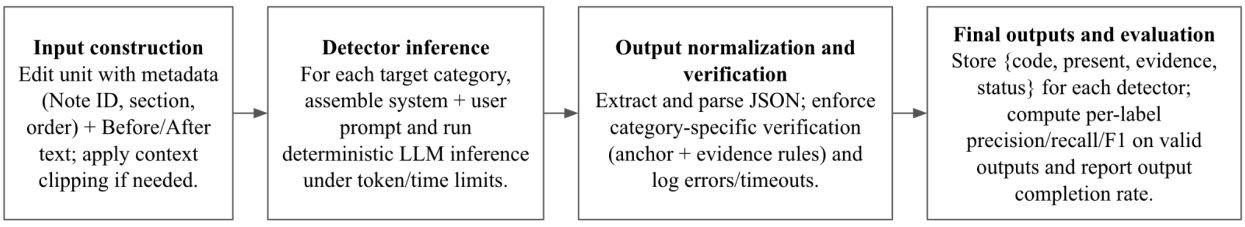

Figure 4. End-to-end workflow for LLM-based edit-category classification and verification.

Results

We evaluated few-shot prompting for five edit types using an iterative development protocol on a development set, followed by a locked evaluation on a held-out test set. We present results in three parts. First, we use E-Med as an illustrative case study to show how targeted prompt

refinements improved precision and recall over successive iterations. Second, we report held-out test performance for the best-performing model for each edit type. Third, we characterize dominant failure modes through qualitative error analysis, highlighting where the prompt-only model falls short.

E-Med prompt refinement case study

We curated few-shot prompt examples from the training set and iteratively evaluated prompt variants on a held-out development set (n=200) that was not used to construct the examples. After each run, we revised the prompt to address the dominant error pattern observed in the prior iteration. Starting from a definition-only zero-shot baseline, the model showed recall that was markedly low. Adding a small set of labeled examples shifted the model toward the intended decision boundary, substantially improving recall. However, expanding example coverage further increased sensitivity at the expense of over-triggering, causing a precision drop driven by medication-related language that did not meet the definition. To restore specificity, we introduced an explicit "exclusion" section plus adversarial negatives targeting the most frequent false-positive patterns. Finally, we added a mandatory pre-output verification gate requiring (i) an explicit medication-related anchor and (ii) evidence that is consistent with the edited span before allowing *present=true*. As summarized in Table 13, these successive refinements improved overall development-set performance, increasing F1 from 0.400 in the zero-shot baseline to 0.800 in the final prompting.

Table 13. E-Med prompt refinement progression (detailed key changes)

Prompt version	**Key changes to E-Med prompt**	**Precision**	**Recall**	**F1**
v0 (zero-shot)	Definition-only baseline prompt:	0.667	0.286	0.4

	● Task framing and E-Med definition. ● Inclusion/exclusion guidance without any labeled examples; no explicit checklist.			
v1	Introduce labeled examples to calibrate decision boundaries: ● Added 1 positive and 4 negative examples from the training pool to reduce under-detection by grounding the abstract definition in concrete patterns.	0.5	0.643	0.563
v2	Increase coverage of medication-change patterns: ● Expanded to 4 positive and 8 negative examples to cover more surface forms (e.g., start/stop, dose change, frequency change, adherence/titration language, med list edits).	0.519	1	0.683
v3	Target dominant false positives with explicit "what does NOT count.": ● Added a structured exclusion section derived from observed FPs (e.g., allergy-only mentions, problem/diagnosis changes without med action, labs/vitals without med change, generic "meds reviewed" without specific change). ● Added 10 more adversarial negative examples that look medication-adjacent but should be *present=false.*	0.632	0.857	0.727
v4	Add mandatory pre-output verification gate (rule-based checklist). *Before* allowing *present=true*, the model must confirm: ● Explicit medication anchor exists (e.g., medication name, dose, frequency, route, formulation, titration, or adherence). ● The anchor is part of the edit change (added/removed/modified between "*Before*" and "*After*"). ● Evidence snippets contain only medication-related text.	0.769	0.833	0.8
Final (locked test)	Frozen v4 prompt ran on the held-out set.	0.774	0.8	0.787

Best-performing prompt-based model performance on the held-out test set

Table 14 summarizes performance for the best-performing model for each edit type on the held-out test set (n=200). Overall, the strongest performance was observed for E-Med (Precision=0.774, Recall=0.800, F1=0.787) and E-Sym (Precision=0.657, Recall=0.959, F1=0.780). E-Dx achieved moderate performance (Precision=0.560, Recall=0.836, F1=0.671). In contrast, E-Test (Precision=0.523, Recall=0.831, F1=0.642) and E-Soc (Precision=0.483, Recall=0.933, F1=0.636) showed lower precision despite high recall, indicating that these categories remain more challenging for prompt-only classification and are more prone to FPs in fully automated use.

Table 14. Best-performing model performance by edit type on the held-out test set

Edit Category	N_TP	N_FP	N_FN	N_TN	Prevalence (gold positives)	N failed (out-of-memory)	Precision	Recall	F1
E-Med	72	21	18	89	45.00%	0	0.774	0.8	0.787
E-Sym	71	37	3	89	37.00%	0	0.657	0.959	0.780
E-Dx	56	44	11	88	33.50%	1	0.560	0.836	0.671
E-Test	69	63	14	52	41.50%	2	0.523	0.831	0.642
E-Soc	56	60	4	80	30.00%	0	0.483	0.933	0.636

** TP = true positives; FP = false positives; TN = true negatives; FN = false negatives. Each edit type was evaluated on up to 200 test edit units. TP+FP+TN+FN may be <200 when some units encountered run-time errors and were excluded from scoring.*
** Prevalence is reported relative to the full test set. Other metrics are computed over the successfully processed units.*

Error patterns and boundary failures in prompt-based edit classification

To investigate where prompt-based models succeed and fail, we reviewed FPs and FNs for the best-performing model in each edit category (Table 15). Across models, the dominant FP pattern reflects boundary ambiguity in clinical language. This was most prominent among categories including symptoms, diagnoses, and orders/tests/treatment plans. Edits in assessment and problem-list text or care-plan statements were frequently "pulled" into the wrong category when explicit anchors were weak. For example, the symptom classification model tended to over-trigger on problem-list or plan phrasing that implies a symptom context but does not directly state a symptom change, while diagnosis and test models were prone to over-interpreting management-oriented wording (medications, procedures, referrals) as diagnostic or test changes. E-Soc model results showed a similar pattern when general clinical narrative or planning language was interpreted as social determinants of health (SDOH) in the absence of a clear social history anchor.

Long, complex edit units and template-like multi-item lists within a unit are associated with a consistent failure mode across categories: true anchors were present but buried among other problems, plans, or list items, and the model either missed them (FN) or extracted an

overly broad evidence span that included nearby text that was not part of the target change (FP). This pattern was most visible in care-plan style units that bundle multiple actions, counseling statements, and follow-up instructions. In these blocks, the model may detect that a label-specific anchor is *present*, but it often cannot localize the specific edited subspan, leading to overly broad or misaligned evidence. Across diagnosis- and test-related edits, errors clustered when temporality or status determined the label. In many cases, the edit unit did not explicitly indicate whether something was planned, performed, resulted, or part of prior history. When temporality cues were implicit, models tended to either miss the change (FN) or misattribute it to a related category (FP), especially when the text is written in a compressed "plan" style that blends diagnostic framing, workup, and management. Another cross-category error pattern is when the edit is delete-only or otherwise short and fractured. For example, in delete-only units where the *After* text is empty, the model may be left with only a single short line (e.g., "Continue vitamin D supplementation."). In these cases, the model may infer the label from a minimal span without surrounding narrative, making it harder to distinguish content type.

Table 15. Summary of model performance and dominant error modes by edit category

Edit Category	**Dominant error direction**	**Top FP modes (n)***	**Top FN modes (n)***	**Key drivers**
E-Med	Balanced	Procedure/injection or follow-up language mis-triggering (5); diagnosis/differential wording mis-triggering (4); in long mixed edit units, evidence pulls non-target content (4).	Delete-only medication statements missed (8); medication embedded in long mixed content missed(6); implicit/uncertain medication decision language missed (e.g., "agents discussed…may reconsider") (4).	FNs occur in delete-only or long mixed procedural blocks, and with implicit/generic medication decisions. Residual FPs reflect boundary bleed into procedure/diagnosis wording and occasional evidence-span drift.

E-Sym	Precision-limited	Diagnosis/assessment/problem-list edits mis-triggering (10); treatment/medication/procedure/care-plan language mis-triggering (10); non-symptom clinical history (e.g., "utilizing an assistive device…") mis-triggering (6).	Symptom content embedded in surgical/history narrative missed (2).	Boundary overlap (symptoms vs diagnoses/treatments) drives mislabeling when the edited text does not explicitly mention a patient symptom or symptom denial, especially in template-like lists or delete-only units.
E-Dx	Precision-limited	Treatment/medication/procedure/surgery language mis-triggering (12); symptom-only or laterality/narrative updates mis-triggering (8); test/imaging/workup statements mis-triggering (6).	Temporality updates to prior diagnoses/illness history missed (4); delete-only diagnosis statements removed missed (4); diagnosis cues embedded in long exam/history expansions missed (2).	Diagnosis is tightly coupled with symptoms, tests, and management; single edit units often lack context to separate "diagnostic framing" from workup/plan language. Performance degrades on delete-only and long mixed-content edits.
E-Test	Precision-limited	Medication edits mis-triggering (15); Dx/assessment rewording mis-triggering (9); non-order narrative edits (10).	Referrals/therapy/follow-up actions missed (6); procedure/surgery/pre-op items missed (3); surveillance/awaiting results missed (2).	Boundary overlap (tests vs medication/diagnosis) and limited temporality cues (planned vs performed vs results). Delete-only edits reduce available evidence, amplifying misses and spurious triggers.
E-Soc	Precision-limited	Clinical diagnosis/assessment/problem-list edits mis-triggering (e.g., "Her anemia has worsened slightly"; "Venous insufficiency, possible calciphylaxis") (16); care-plan / workup language (referral/PT/surgery planning/tests) mis-triggering (8); minor clinical narrative/date/laterality edits mis-triggering (4).	Obvious "social history" additions missed in long *before* and *after* context. (3).	Over-broad trigger space and weak SDOH anchor enforcement. Occasional evidence-text misalignment (pipeline/merge) can further inflate false positives.

** Top error modes shown; remaining errors were heterogeneous.*

Discussion

Scaling edit classification between ambient AI drafts and clinician-final notes is essential for understanding how clinicians refine AI-generated documentation and where these AI systems fall short [107]. It can support continuous monitoring, guide targeted model improvements, and ultimately inform safer and more efficient workflows. In this feasibility study, we evaluated

few-shot prompting models using Llama-3.2-3B-Instruct, a 3B-parameter instruction-tuned large language model, deployed in a HIPAA-compliant environment with limited computing power and without additional labeled training data or EHR context [141]. We found that prompt design choices, especially iteratively adding adversarial negative examples and applying an explicit verification gate before output, helped manage precision–recall tradeoffs and improved performance. The models performed best for edit types with clear, text-anchored signals, especially medication- and symptom-related edits. Performance was substantially weaker for more context-dependent categories such as social history, diagnosis, and test/order/procedure-related edits, indicating that prompt-only methods are not equally feasible across edit types and that some categories will require richer context or human review to support reliable downstream use.

These differences have direct implications for what can be used safely in practice. For categories with stronger performance, the most defensible near-term use is high-precision labeling to support scalable measurement and targeted quality monitoring [142]. To avoid over-reliance on noisy outputs in future deployments, such use should be paired with safeguards such as confidence-thresholded acceptance, explicit reporting of coverage, and abstention for ambiguous edit units such as delete-only changes or long mixed-content spans [143]. Feasibility should therefore be judged by whether the model can achieve sufficiently high precision at the coverage needed for the intended workflow, not by overall F1 alone. In contrast, the context-dependent categories that remained precision-limited in our evaluation should not be used for fully automated evaluation because false positives can systematically distort downstream estimates [144]. Despite the labor required, manual review remains the most reliable way to characterize how ambient AI drafts perform and how clinicians revise them, and it is

essential for establishing a trustworthy reference for evaluation. Until additional controls are introduced, these models are better positioned for triage workflows that surface candidate edits for targeted human review [145]. From an operational perspective, the value of these labels also depends on the balance between time saved and time spent verifying outputs; triage deployments should therefore report or plan for the associated workload [146].

Failure cases in this study are concentrated in edit units where the available evidence is inherently weak, especially delete-only changes where the final text is missing and long mixed-content spans where multiple clinical concepts are interrelated. These boundary conditions suggest clear design implications for prompt-only classification at scale. One path is to develop prompts for known difficult structures, such as list-like templates, and for ambiguous contents of the note, such as distinguishing planned tests from performed procedures and requiring explicit anchor spans before accepting a positive prediction [147]. A complementary path is to move beyond prompt-only methods via parameter-efficient fine-tuning on labeled edit units, particularly for cases that rely on implicit clinical reasoning [148]. These adjustments can target the specific situations where edit unit context is least informative and where precision failures are most costly for downstream use [144].

We used a fixed prompt design to reflect realistic governance and deployment constraints; however, retrieval-based prompting may improve borderline, context-sensitive categories with high anchor variability, without requiring additional labeled training data [149]. This direction is most relevant for complex edits, where dynamically retrieving label-specific examples could provide more appropriate reference cases than a fixed example set. At the same time, we used a 3B-parameter instruction-tuned large language model in a HIPAA-compliant environment with limited GPU memory, which contributed to out-of-memory failures and

constrained prompt length and runtime. In the final prompt iterations, inference approached roughly 30 seconds per edit unit as prompts grew longer. This underscores a practical tradeoff in prompt engineering where adding detailed instructions, verification gates, and an expanded list of examples can improve performance but reduces throughput and scalability. Future work should therefore evaluate larger LLMs with stronger reasoning capability, redesign the task with hierarchical labels for vague boundaries such as separating test orders from procedures, and consider domain knowledge adaptation such as SDOH ontology where richer context may be necessary [150,151]. In addition, lightweight rule-based components such as dictionary matching may further improve robustness when edit units mix multiple clinical concepts [152]. Finally, scaling to tens of thousands of note sections will likely require more efficient inference strategies such as batching and optimized serving. Overall, our findings suggest that prompt-only classification is most feasible for edit types with explicit textual anchors, while context-dependent categories require further controls or added context before they can support reliable automated measurement.

Conclusion

This feasibility study shows that few-shot, prompt-based LLM classifiers have the potential to scale edit-type categorization of clinician modifications to ambient AI draft notes, but performance varies substantially by content category. Under privacy and deployment constraints, models on medication- and symptom-related edits achieved strong held-out performance after iterative prompt refinement with hard negatives and an evidence-based anchor verification gate. In contrast, diagnosis, order/test/procedure, and social history classifiers remained precision-limited because sentence-level edits often blend clinically related content and key cues,

and rationales are frequently implicit. These findings support a practical workflow: edits with explicit textual cues can be monitored automatically, whereas categories that typically require broader clinical context are better suited for labeling likely edit types for targeted human review. This work demonstrates a deployable pathway for scalable, auditable edit detection and classification that can help health systems and ambient AI vendors monitor common revision patterns in AI drafts, prioritize quality improvement targets, and evaluate changes over time. Future work should test larger LLMs with parameter-efficient fine-tuning, add lightweight rule-based signals, and redesign labels to support hierarchical drill-down categorization for mixed content edits.

CHAPTER 6: Clinicians' Rationale for Editing Ambient AI-drafted Clinical Notes: Persistent Challenges and Implications for Improvement

Abstract

Objective

The use of ambient AI documentation tools is rapidly growing in US hospitals and clinics. Such tools generate the first draft of clinical notes from scribed patient-provider conversations, which clinicians can then review and edit before signing them into electronic health records (EHR). Understanding how and why clinicians make modifications to AI-generated drafts is critical to improving AI design and clinical efficiency, yet it has been under-studied. This study aims to address this gap.

Materials and Methods

We conducted semistructured interviews with 30 clinicians from the University of California, Irvine Health who used a commercial ambient AI tool in routine outpatient care. We invited them to describe how and why they edited AI drafts based on both their personal experience and review of some real-world examples identified from our previous studies.

Results

Modifications to AI drafts were primarily made to improve clinical accuracy and specialty-specific precision, reduce medico-legal and liability risk, and meet billing, coding, and documentation standards. Such editing was necessary due to reasons such as transcription errors,

speaker attribution mistakes, overconfident statements without evidence, missing key clinical details, and AI's lack of information about the patient context.

Conclusion and Discussion

Improving ambient AI documentation will require coordinated effort from vendors, institutions, and clinicians. Key targets include core model reliability (e.g., transcription accuracy), specialty- and encounter-level customization, clinician-level personalization, more effective EHR integration, and institutional support (e.g., training, governance, and standardized review guidance), complemented by clinicians' adaptive communication strategies that strengthen human–AI collaboration.

Background and Significance

Ambient artificial intelligence (AI) documentation tools are rapidly becoming one of the most widely adopted generative AI tools in clinical care, which use visit audio to draft clinical notes and aim to reduce documentation burden and clinician burnout. [12,17,25,41] Early evaluations suggest adoption may improve documentation efficiency and productivity, supporting health enterprise growth. [24,42,153,154] Many clinicians report positive experiences, such as feeling more present with patients and reducing after-hours documentation. [22,42] At the same time, emerging concerns highlight the need for additional investigation of ambient AI performance and how it impacts care in routine clinical settings. Clinicians report persistent challenges, including lengthy or underspecified note sections, problems with accuracy, and barriers in non-English encounters. [22,155,156] Importantly, the central role of clinical documentation in medico-legal accountability has

raised concerns about downstream effects on billing, including potential upcoding, higher spending, and productivity pressures that may undermine clinician well-being. [157] Continued ethical and legal questions remain, including consent and transparency for patients and staff, and uncertainty about liability and retention when AI is used in clinical care. [158]

Prior work analyzing ambient AI use demonstrates that clinicians frequently edit AI drafts and that editing varies by note section, specialty, and individual documentation style. [128] Content analyses have also described recurring edit patterns, such as correcting factual errors, adding specialty-specific details, and reorganizing content for clarity. [127] However, these studies mainly characterize what changed in the note and do not fully capture why clinicians conduct specific edits or what persistent challenges are shaping those decisions. Because editing burden directly mediates the efficiency gains promised by ambient AI, understanding the drivers of edits is essential to realizing both clinician-experience and financial return-on-investment benefits.

We conducted semi-structured interviews with clinicians who used ambient AI tools in routine care and applied thematic analysis to examine (1) clinicians' intent behind common edits (2) and persistent challenges that shape editing decisions and trust. By linking clinicians' explanations to recurring edit patterns observed in our prior studies [127,128], our findings identify practical targets for improving ambient AI note quality, workflow integration, and governance.

Methods

Participant recruitment and interview procedures

We recruited clinicians from the University of California, Irvine Health (UCI Health), where a commercial ambient AI documentation tool was implemented in outpatient settings across multiple provider roles at the end of 2023. Using an internal e-mail list of active users, we identified 87 clinicians with over 50 ambient AI–assisted encounters and 14 informatics clinician champions who used ambient AI and were extensively engaged in medical informatics research. We invited eligible clinicians via e-mail and interviewed 30 participants between January and February 2026; recruitment stopped when we judged thematic saturation had been achieved. Interviews were conducted remotely via Zoom and each lasted approximately 45 minutes. Participants provided verbal consent for audio recording. Recordings were stored in the Zoom cloud and transcripts were generated via Zoom automated transcription with only obvious transcription errors corrected. Transcripts were de-identified prior to coding, and access to the dataset was restricted to study authors. Participants received a $50 Starbucks gift card after completing the interview.

We used a semi-structured interview method to ensure consistent coverage of core topics while allowing participants to expand on experiences they considered relevant. In developing the interview guide, we drew on our prior content analysis of ambient AI draft–final note pairs, which identified five common edit patterns, including factual corrections, specialty-specific refinement, diagnostic certainty calibration, professionalization of transcript-like text, and restructuring and condensing. These patterns informed an a priori framework that guided both interview domains and initial coding. We presented standardized, de-identified editing examples

which aligned with these edit patterns and asked clinicians to interpret the edits and explain what drove their decisions. The interview protocol and example edits are provided in Appendix 3. This study was reviewed and approved by the University of California, Irvine Institutional Review Board (UCI STUDY00000824-IRB).

Qualitative coding and data analysis

We conducted thematic analysis to characterize clinicians' perspectives, starting from the a priori framework described above and iteratively refining codes as new concepts emerged. Analyses were conducted in Atlas.ti (version 9.24.3; ATLAS.ti Scientific Software Development GmbH, Berlin, Germany). Three authors (YG, DH, ZY) jointly developed and refined the codebook through iterative rounds of double-coding (two interviews per round) and consensus discussions to resolve discrepancies and clarify definitions. One author (YG) then applied the refined codebook across the full dataset and led synthesis, supported by analytic memos and an audit trail. Within the a priori domains, we identified inductive subthemes by examining code prevalence, co-occurrence, and variation across interviews. To strengthen interpretive validity, we conducted member checking with two interview participants (EC, ST) who reviewed summary interpretations of key findings.

Results

Study Overview

A total of 30 clinicians participated in the interview study, yielding an overall response rate of 29.7%. Participants included 15 primary care clinicians and 15 subspecialists. Most were attending physicians (n=27); the remainder were nurse practitioners (n=2) and a physician

assistant (n=1). Subspecialties represented included rheumatology (n=3), neurology (n=2), obstetrics and gynecology (n=2), neurosurgery, oncology, colon and rectal surgery, gastroenterology, integrative medicine and psychiatry, ophthalmology, radiation oncology, and hematology/oncology (n=1 each).

We analyzed interview transcripts, guided by findings from our prior note-pair content analysis, with a focus on emerging themes related to (1) clinicians' revision intents, (2) the precipitating challenges that limited AI effectiveness, and (3) recommendations for improvement. In the following sections, we summarize the key intents driving edits and the recurring challenges clinicians described as barriers to achieving those intents. We then present clinicians' recommendations for tool and implementation improvements.

Clinician Editing Intents for Ambient AI Documentation

Correct factual, transcription, and attribution errors to protect record accuracy

Clinicians edit ambient AI drafts to prevent inaccuracies. Several clinicians emphasized that transcription mistakes were clinically consequential, especially when small lexical differences changed meanings, such as misspellings of "less common medications" including infusion agents used for chemotherapy (P17, P19, P20, P26). They linked terminology use errors to real-world speech conditions, including accent or unclear audio (P1, P11, P13, P17, P20–22, P27). One clinician summarized this limitation as it "can't get it exactly" (P21). Additionally, identity or speaker attribution issues occurred, such as misspelled names, incorrect pronouns, or content assigned to the wrong speaker. When family members, partners, staff, or interpreters were present, clinicians reported the draft could carry over non-patient discussion, include symptoms described by someone other than the patient, or mismatch speakers with content (P1, P10, P12,

P14, P21, P27). In some cases, misattribution led to clinically incorrect history, such as assigning a spouse's transplant status to the patient, or pulling in a chemotherapy-order discussion that occurred when a nurse entered the room to update messages for another patient (P5, P10, P15, P18, P29), so clinicians needed to correct the inaccuracies manually.

Achieving specialty-specific precision and completeness in drafted notes

Clinicians described editing ambient AI drafts to make them more specialty-specific and complete, particularly when drafts "sounded" generic and did not reflect specialty priorities or clinicians' differential reasoning (P14–16, P17–19, P22–23, P25, P27–28, P30).

Subspecialists emphasized that drafts often lacked differential framing and appropriate prioritization, often requiring rewriting note content to match how they practice (P16–17, P19, P23, P25, P27–28, P30). One participant summarized the mismatch as the AI drafts "sounded like an internal medicine doctor… I'm a subspecialist" (P16). A psychiatry participant described the A&P drafts as "severely lacking" because their work involves "much more nuance" including "reading between the lines and inferring what patients are saying versus what they may mean" (P17). Rheumatology clinicians noted that important specialty constructs were often under-captured, such as documenting "low, moderate, or high disease activity". They also reported that counseling was often summarized only as a vague mention of "risks," rather than documenting specific risk-benefit discussions, side effects, or consent-related details (P15, P19, P24).

Beyond specialty-appropriate framing, clinicians also edited drafts to address omissions of clinically salient information. For completeness, participants noted that edits were often triggered by omissions of clinical details that the tool failed to capture, including missed

review-of-systems questions, inconsistent capture of advance care planning and Medicare Annual Wellness functional/social assessments, and omission of negative screening histories unless a positive finding was present (P7–9, P13–14, P19, P21, P26, P30). Several described adding pertinent negatives (e.g., denied symptoms or cancer history) that they reported eliciting during the visit but that were not captured in the draft, to preserve clinical reasoning; as one clinician put it, "If I'm editing the history, I'm usually adding pertinent negatives that I feel the AI did not catch" (P11).

Participants emphasized that clinical specificity is not simply adding more detail, but preserving distinctions that affect interpretation and downstream management. They described these distinctions as involving symptom characterization, chronology, and diagnostic labeling, which were sometimes omitted or blurred in AI drafts (P22, P27–28). For example, one clinician noted the draft might say "fractured wrist" without "specifying where exactly, or the exact type of surgery," reducing usefulness for follow-up care (P22). Clinicians noted that the draft could compress step-by-step symptom narratives into vague summaries and might retain an earlier version of the patient story unless later clarifications were explicitly restated for the record (P27, P28).

Clinicians also described restructuring content that the draft placed in the wrong note section or attributed to the wrong diagnosis. Several noted blurred boundaries between subjective report and clinician interpretation, with labs, imaging, or assessment statements appearing in the HPI (P4, P5, P13, P17, P20). One clinician explained that the AI draft mis-presented the clinician diagnosis "as if the patient said it" (P13). Others reported related sectioning issues, including placing patient-reported results or diagnoses into Results or A&P without preserving them as history (P7), or failing to integrate later-elicited history back into the HPI (P11). At the

assessment level, clinicians said the draft could "latch[ed] on" to secondary issues and turn keywords into standalone diagnoses (P6, P9, P29), sometimes splitting a single condition into multiple diagnoses (P9) or adding historical diagnoses mentioned during the visit to A&P (P29).

Coding requirements, medico-legal safeguards, and evidence-aligned certainty

Clinicians edited ambient AI drafts to ensure billing, compliance, and medico-legal requirements were met, particularly in the A&P where coding depends on diagnostic specificity, documented decision-making, and severity-of-illness detail (P1, P5, P13, P25, P28). Several noted that the tool might suggest a diagnosis but still "doesn't fill out the billing form" for level-of-service requirements, prompting them to add severity considerations and ensure documentation supported billing and risk-adjustment expectations (P1, P13, P25). Some explicitly linked these edits to documenting reasoning, describing it as important "from billing medico-legal standpoints" (P2, P28). When key elements were missing, participants described targeted additions, including details reflecting time spent, counseling, and options discussed, and relevant screening or history elements discussed during the encounter (P5, P7–9, P18–19, P28).

For example, oncology clinicians highlighted that consent documentation often requires substantially more specificity for billing than what is naturally said during patient-facing counseling: "for billing purposes, they want us to put in all the detail… list everything, just to be able to bill" (P5). Several clinicians further emphasized that time-based billing hinges on capturing not only actions taken but also counseling and options considered: "talking about things that are not done is still time we spent and options we offered… it justifies billing," and "the more detail we have in the note, the more coding and billing we can do to justify the level of the visit" (P18). Additionally, a neurosurgeon noted that documentation sometimes must "satisfy

several criteria before the insurance company would approve an operation," and described ensuring the history clearly captured "a certain number of weeks of physical therapy or conservative measures… [and] prior trial of conservative measures" when the draft was not clear (P3).

Clinicians also described routine edits to align certainty with evidence and to reduce liability and patient confusion (P1–2, P4, P6, P11, P14–16, P17–19, P22, P24, P26–27, P29-30). They emphasized that documentation should reflect what is known versus still being evaluated, particularly early in a workup (P1, P7–9, P14–16, P18–19, P22, P24, P26–27). Several felt the draft could flatten probabilistic reasoning into definitive statements, for example, "I say 'possibly this'… and the AI writes it as if it's definitive" (P19), or "if I don't explicitly say 'these tests are to confirm,' then it writes the diagnosis as confirmed" (P18). To restore evidentiary alignment, clinicians inserted qualifiers and conditional plans and avoided prematurely closing the diagnostic narrative, including when unconfirmed problems appeared in the A&P or when an intended hedge such as "likely a disc herniation" was missing (P1, P8, P14–16, P15, P27). They also described hedging as a communication and risk-management practice, such as using "low suspicion" as "medico-legal protection" (P2, P24).

Edit for professional, standardized notes and efficient clinical review

Across interviews, clinicians described editing ambient AI drafts to meet professional documentation standards so notes read like clinician-authored records rather than verbatim narratives (P2, P8, P9, P11–15, P17, P21, P22, P24, P25, P27–29). For example, participants wanted notes that "look professional when somebody reads it" (P14) and reflected how they were taught to write, including that it "shouldn't read like a transcript" (P21). Edits often focused

on avoiding informal wording and colloquialisms, and standardizing terminology, including converting spoken exam language into "medical terminology" (P4, P9, P11, P13, P17, P21, P23, P24, P27, P29). Some preferred more understandable and actionable language for the after-visit summary when the note was facing patients who may have limited health literacy (P29).

In addition, AI drafts were often "too wordy and paragraphy" (P2) and included redundancy, low-yield detail, or repeated plans, so clinicians condensed content, trimmed small expansions that accumulated, and shortened content they felt were "too long" for other providers or patients to read (P5, P7, P9, P13–15, P21, P23, P24).

Workflow and EHR integration constraints shaped how much restructuring was needed to form a standardized final note (P1, P3, P4, P6, P8, P9, P12, P16, P18–20, P22, P23, P25, P28, P29). Clinicians described barriers when AI drafts did not fit EHRs' problem-oriented structure that "saves under that particular diagnosis" (P4), requiring "copy and paste" steps (P4, P19). Additionally, clinicians deleted duplication when information was already documented elsewhere outside the AI smart elements (P8, P29). Others pointed to integration boundaries and chart access, including uncertainty about what the tool pulled beyond the conversation (P3, P24) and requests for the "ability to read the EHR" to support reusable history, continuity across visits, and real-time review (P1, P16, P18, P23).

Clinician Expectations and Recommendations to Improve Ambient AI Draft Usability

We group recommendations by the scope of change and who would need to act on them, ranging from individual clinician behaviors and targeted system or configuration adjustments to broader workflow redesign and EHR integration. This framing highlights actionable informatics targets

for individual clinicians, vendors, and health systems. Figure 5 provides an overview of the recurring challenges clinicians described and the improvement targets they recommended.

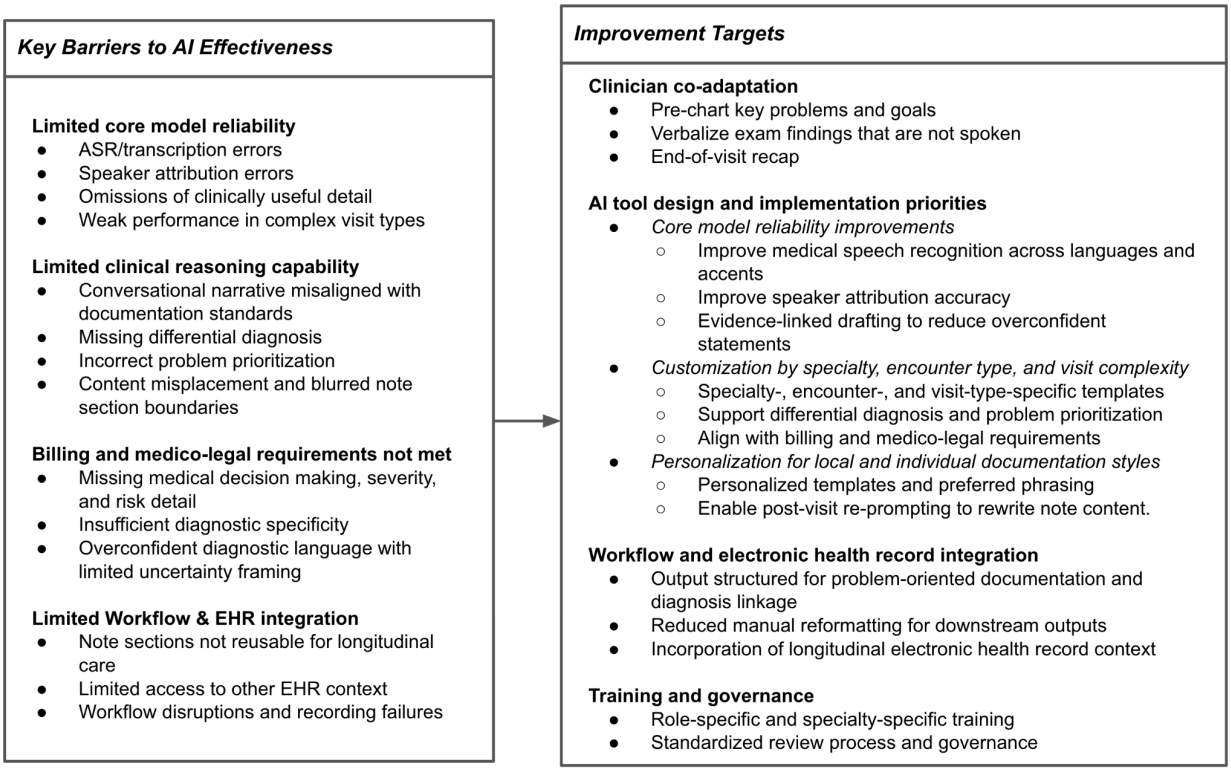

Figure 5. Challenges shaping ambient AI note editing and improvement targets

Precharting, in-visit verbalization, and post-visit strategies

To reduce editing load during note review, clinicians described adapting how they communicate so the ambient system can capture clinically relevant details more reliably (P3, P4, P6–14, P15–20, P26). A common practice was to verbalize information that is otherwise nonverbal, explicitly stating laterality and location when patients gesture (P3, P10, P11, P12, P13, P14, P17, P18, P19, P20). For example, "If a patient points, I'll say out loud, 'It looks like there's a rash on your right arm,' so AI can capture it" (P10). Others reported narrating exam maneuvers and findings in real time ("Okay, I'm doing a rectal exam… now I'm doing anoscopy…") to populate the physical exam more completely (P6), or deliberately speaking standardized exam descriptors

aloud ("regular rhythm," "lungs are clear") (P15). Outside the encounter, participants described pre-charting (P26, P9) and end-of-visit "wrap-up" strategies to reduce downstream edits, such as "summarizing the encounter at the end" (P6) or recapping problem-by-problem plans ("Okay, for the UTIs we're going to do this…") (P16).

Core AI model reliability, customization, and personalization for clinician-ready drafts

Clinicians' vendor-facing recommendations clustered into three layers: improving core default behaviors, enabling specialty and encounter-type customization, and providing personalization controls. First, they described foundational model behaviors that should improve by default to reduce downstream edits. They emphasized fidelity to the encounter, preferring the system to "record what the clinician says" rather than produce "its own version" (P14). Several recommended accurate speaker attributions so the draft reflects who performed the assessment (P2, P5, P10, P15, P29). They also recommended conservative certainty as a default because "we tend to hedge more in medicine" (P2) and anticipated "fewer corrections…if hedging terminology is included by default" (P12), with certainty wording scaled to evidentiary strength (P26). In addition, participants highlighted gaps in medication and supplement handling (P17) and inconsistent recognition of the same drug across mentions (P18), motivating iterative updates of medication library (P21). Second, participants recommended specialty- and encounter-type customization, particularly for contexts where a generic draft does not fit, such as learner settings where "it has to be shorter," with "a template for the AI to fit into an attestation" (P1), and specialties with "pretty standardized…how we were trained" documentation that may require different prompting to support the expected structure (P17). Third, clinicians wanted personalization functions, including the ability to "tweak it…more verbose, less verbose, what to

include" and "give it more instruction" (P6), as well as re-writing controls such as "re-prompting ability like in ChatGPT" for targeted revisions (P1, P21, P30).

EHR integration to streamline workflow and reuse prior chart context

Clinicians recommended tighter EHR integration to reduce workflow friction and better support longitudinal care (P1, P4, P6, P8, P16–20, P22, P24–26, P28, P30). They suggested minimizing extra clicks and copy/paste for tasks such as after-visit summaries ("copy and paste…one paragraph") (P8, P19). Participants also wanted AI outputs to sync with Epic structures (e.g., problem list and diagnosis-linked A&P) and reuse prior chart context—similar to dot phrases and structured history—to improve chronic disease management and longitudinal review ("a whole different experience") (P16, P20, P22, P28). Additional integration expectations included drafting orders mentioned in conversation ("put that prescription in"), summarizing relevant chart review for complex histories, and improving specialty alignment so irrelevant cross-specialty content is omitted by default (P20, P24, P1, P19, P30). Finally, participants noted that more integrated capture workflows could reduce dependence on personal mobile devices and related interruptions (e.g., battery drain) ("phone died…charge during the visit") (P1, P2, P18, P29).

Institutional Training and Governance to Reduce Uncertainty and Misuse

Clinicians emphasized that onboarding for ambient AI was often generic and recommended role-, setting-, and specialty-specific training, alongside clearer institutional guidance on when and how to obtain and document patient consent for ambient AI recording and documentation (P1, P2, P6, P24). They wanted practical instruction on "how to make it work," noting that limited guidance can lead clinicians to rely on trial-and-error and may contribute to frustration or

abandonment when workflows vary across settings and are not equally supported (P1). Participants also highlighted broader education gaps in AI-assisted documentation that shape how clinicians evaluate and edit AI drafts (P2). From a governance perspective, they recommended standardizing consent processes, so the burden does not fall on individual clinicians each visit, and ensuring training covers routine workflows such as pre-charting and after-visit tasks (P6, P24).

Discussion

Prior studies describe what clinicians edit in ambient AI notes. Our study is among the first to systematically characterize the underlying clinical, legal, and workflow rationales that drive those edits, translating edit behaviors into actionable design and governance. Clinicians whose documentation required more complex differential reasoning and were subject to greater billing and medicolegal scrutiny reported spending more effort correcting errors, restoring missing or misplaced clinical details, and rewriting content to align with specialty norms and evidentiary support. As a result, clinicians often either edit extensively or limit use to certain note sections they consider more reliable, which helps explain why adoption and perceived value vary by specialty and encounter type. [128]

Clinicians' selective use across note sections often reflects systematic gaps between conversational summarization and the requirements of clinical documentation. [159] Accurate conversation capture does not guarantee clinically appropriate notes because documentation requires additional judgments about what is clinically relevant, where it belongs in the note, who the source is, and how strongly a statement should be made given available evidence and uncertainty. [48,63,160] These demands are more pronounced in subspecialty care, where notes

require more longitudinal decision-making and specialty-specific differential reasoning while also meeting billing, legal, and institutional expectations. [76,157,158] In this setting, recurring limitations such as transcription errors and omissions of key clinical details can cascade into substantial downstream editing. The problem is further intensified when drafts are overly long, fail to distinguish clinically relevant content, or do not align with EHR workflows. [161]

Improving ambient AI tools will require coordinated, multi-level efforts. At the foundational model level, systems should prioritize robust transcription and accurate speaker attribution. [27] When they fail, downstream content becomes harder to trust and more time-consuming to review. Drafts should also default to safer clinical language by tying statements to evidence and using conservative certainty when the basis is unclear. [162–164] Beyond baseline reliability, ambient drafting should better reflect specialty and encounter type needs through configurable templates and encounter presets, such as new visits versus follow-up visits, chronic disease management versus procedures, and visits involving learner supervision versus independent practice. [165,166] However, we do not know how comprehensive or deliberate clinicians are in verbalizing the information they intend to include in the note; ambient drafts can only draw from what is stated in the encounter transcript. When key procedural rationale or context is not explicitly articulated, drafts may default to hedging or inference rather than documenting clear justification. Future work should evaluate "speak-to-document" practices and test whether specialty- and visit-type templates can cue more structured, note-ready clinical language at the point of care. Furthermore, ambient AI should be implemented as an EHR-integrated workflow rather than a standalone note generator. Access to the patient history, medication list, and recent notes is essential for longitudinal care and can prevent redundant information that clinicians routinely delete in drafts. [167,168] Finally, because many edits reflect personal preferences, tools should support

clinician-level personalization through reusable templates, and controlled learning from prior edits, paired with pre-prompting or re-prompting functions to reduce repetitive rewriting and corrections. [169]

In addition, institutions should provide thorough implementation support through specialty-, role-, and workflow-specific onboarding [170,171], alongside clear governance guidance and a standardized review checklist specifying how and what clinicians should verify. [172] Clinicians may also benefit from actively adopting "AI-aware" communication and review behaviors to improve draft quality and support human–AI collaboration.

Our study has several limitations. These insights reflect outpatient workflows and a self-selected group of participating clinicians and may underrepresent non-adopters and settings such as inpatient or emergency care where documentation practices and operational constraints differ. Findings are also limited to a single institution and a single ambient AI vendor, which may affect generalizability. Despite these limitations, the results suggest that sustained reductions in editing burden will require closer alignment between ambient AI outputs and real-world documentation workflows. Such alignment may help shift ambient AI from producing generic summaries that require substantial rewriting to generating drafts that are faster to review, and more consistently usable across specialties.

Conclusion

When editing the ambient AI documentation drafts, clinicians aim to safeguard factual accuracy, preserve specialty-appropriate reasoning and terminology, and ensure documentation meets billing, compliance, and medico-legal expectations. Based on these findings, we recommend improving core draft reliability by increasing transcription and speaker attribution accuracy and

using default language that appropriately reflects clinical uncertainty. We also recommend adaptable specialty- and encounter-specific templates, stronger EHR integration to incorporate longitudinal context, and institutional training and governance to support consistent, safe AI-assisted documentation practices.

CHAPTER 7: Summary and Conclusions

Ambient AI documentation is evolving rapidly, and major health systems are scaling these tools to alleviate documentation burden and improve efficiency [27,107]. Its real-world effectiveness is reflected not only in adoption, but also in how clinicians revise and verify AI-generated content under clinical, legal, and workflow constraints, which ultimately shapes the quality of clinical documentation. It is essential to understand how documentation practice changes and to assess note quality through real-world clinician editing of AI drafts in order to determine whether these systems deliver both clinician-experience benefits and financial return on investment [153]. This dissertation contributes an end-to-end, multi-level evaluation that integrates clinician perceptions, EHR-based time-efficiency metrics, and large-scale comparisons between ambient AI drafts and clinician-finalized notes. Using a mixed-methods approach that combines qualitative analysis with computational measurements, it characterizes how ambient AI fits into documentation workflows, what clinicians edit in drafts during practice, and why they make those revisions. Across studies, the central contribution of this dissertation is translating clinician perspectives, observed real-world edit patterns, and the clinical, legal, and workflow rationales underlying those edits into actionable design, implementation, and governance targets for safer and more effective deployment at scale.

RQ1. How do clinicians perceive the impact of ambient AI on documentation efficiency, workload, and patient care?

Analyses of EHR-based efficiency metrics in Chapter 2 showed significant reductions in note-writing time after adopting ambient AI, even as note length increased. Pre–post survey results also indicated improvements across multiple domains, including lower cognitive demand

and reduced documentation burden, along with perceived gains in clinical efficiency, patient-centered care, and EHR usability.

RQ2. To what extent do clinicians revise ambient AI drafts, and how do editing frequency and intensity vary across individual clinicians, specialties, and note sections?

Chapter 3 found that clinicians routinely revised ambient AI drafts before signing, with edits present in 84.4% of notes. Editing frequency varied across note sections and care settings, but the dominant source of variation was individual clinician documentation style rather than specialty-level norms. Draft-to-final comparisons also showed modest linguistic differences, including slightly higher lexical diversity and modest changes in readability. Editing intensity was greatest in the Assessment and Plan section, consistent with clinicians concentrating effort where clinical reasoning and decisions are documented.

RQ3. What changes do clinicians make to AI-drafted notes, and do these changes follow systematic patterns?

Chapter 4 characterized how clinicians revise ambient AI drafts, producing a human-annotated corpus and identifying recurring edit patterns. Clinicians most often modified core clinical facts such as medication- and diagnosis-related content, and edited to refine writing style and note structure. Edits consistently reflected practical aims, including correcting inaccuracies, aligning wording with specialty documentation norms, calibrating diagnostic certainty to match the available evidence, translating conversational language into professional clinical documentation, and reorganizing or condensing text to improve clarity.

RQ3a. How feasible is it to use computational methods (e.g., LLM-based detectors) to measure edit behaviors at scale for ongoing monitoring and improvement?

Building on the human-annotated edit corpus developed in Chapter 4, Chapter 5 examined whether edit patterns can be measured at scale using few-shot, prompt-based LLM classifiers. The study developed and iteratively refined models for five clinically meaningful edit types, using adversarial negative examples and an explicit evidence verification gate to improve robustness and auditability. Performance was strongest for edit types with clear textual anchors, such as medication- and symptom-related changes, while categories that rely more heavily on complex clinical context, including diagnoses, orders/tests/procedures, and social history, remained more challenging to classify. Chapter 5 shows that prompt-based LLMs can support an automated monitoring pipeline by reliably measuring edit types with clear textual cues, while serving primarily as a screening step for more context-dependent categories that require human confirmation.

RQ4. What are clinicians' rationales for revising ambient AI drafts, and what do they believe is needed for effective ambient AI use?

Chapter 6 addressed RQ4 through interviews with clinicians who use ambient AI as part of routine practice. Clinicians described editing as a necessary step to ensure clinical accuracy and specialty-appropriate precision, manage medico-legal risk, and meet billing, coding, and professional documentation requirements. They attributed much of the editing burden to recurring tool limitations, including transcription and speaker attribution errors, overly certain statements, omissions of important clinical details, and limited access to patient historical context at the time of drafting. Looking forward, clinicians emphasized the need for stronger EHR

integration, more robust specialty customization and personalization controls, and expanded functionality beyond documentation to better support clinical workflows.

Taken together, the findings suggest several priorities for making ambient documentation both more effective and more governable in routine practice. More fundamentally, these findings raise the question of what a clinical note is intended to accomplish. A clinical note is not simply a transcript of an encounter or an end product to be optimized for completeness. Its core clinical function is to preserve and communicate the information, reasoning, decisions, and plan needed to support ongoing care, while also serving patient-facing, billing, medico-legal, and institutional functions. Accordingly, note quality does not have a one-size-fits-all definition. What counts as a "good" note depends on its intended readers and downstream uses, as well as the encounter type, specialty, and local documentation requirements. More information is not necessarily better if clinically important reasoning becomes harder to identify, just as a shorter note is not better if it omits information needed for follow-up, accountability, or patient understanding. The clinician edits observed across this dissertation reflect this balancing process: clinicians did not only correct errors, but also selected, qualified, synthesized, and reorganized information to make the final record clinically usable. In this context, editing behavior becomes a particularly informative outcome because it provides direct evidence of how well AI-generated content meets these purposes and how much additional work is required before a note is safe and acceptable to sign.

Ambient AI also creates an opportunity to reconsider whether a single conventional note should remain the universal target for documentation. Because AI systems can generate different representations of the same encounter, future systems could potentially support purpose-specific

views for clinician follow-up, specialty care, patients and caregivers, or administrative requirements, rather than requiring one document to satisfy all of these needs simultaneously. This possibility suggests that ambient AI evaluation should move beyond asking whether a draft reproduces a conventional note and instead assess whether the resulting documentation communicates the right information, at the appropriate level of detail and certainty, for its intended user and task. Ultimately, future work should also examine whether improvements in documentation quality translate into safer or more effective patient care, a downstream relationship that the present dissertation was not designed to establish.

Furthermore, the dissertation shows that improving ambient AI effectiveness requires joint effort from all stakeholders. From an evaluation standpoint, health systems need repeatable, scalable measurement that can be run over time to track AI performance, usage drift, and downstream editing and coding workload. Ambient AI vendors should prioritize stronger customization for different specialties and visit contexts, and offer selective personalization functions with clear safeguards so flexibility does not come at the cost of safety or consistency. Beyond AI model performance, implementation support, especially training and governance, is an essential component of real-world effectiveness. Clinicians' input suggests that guidance on how to review and verify AI-generated content is often lacking or inconsistently applied, which may lead to added review burden, documentation compliance risks, and downstream effects on documentation and coding behavior. Moving forward, health systems will need clearer expectations for review and accountability, paired with training that reflects specialty workflows, encounter types, and, where appropriate, individual practice needs.

In closing, this dissertation demonstrates that ambient AI can improve clinical documentation efficiency and support better care delivery, but its safe and effective integration into clinical

practice will require ongoing improvement and careful governance. Meaningful progress will depend on coordinated attention to clinicians, patients, healthcare institutions, AI and EHR infrastructure, and the regulatory and economic conditions that shape documentation. Grounded in real-world evidence and clinical practice, this dissertation offers a path forward for developing ambient AI solutions that are more effective, trustworthy, and responsive to the needs of clinical practice.

REFERENCES

1. Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J. 2024;11(3):100157. doi:10.1016/j.fhj.2024.100157

2. DeepScribe, Inc. The Digital Scribe Revolution: How AI is Changing Clinical Documentation. Published online 2020. https://www.deepscribe.tech/resources/digital-scribe-whitepaper

3. Rights (OCR) O for C. HITECH Act Enforcement Interim Final Rule. October 28, 2009. Accessed November 5, 2024. https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.html

4. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. N Engl J Med. 2010;363(6):501-504. doi:10.1056/NEJMp1006114

5. Li C, Parpia C, Sriharan A, Keefe DT. Electronic medical record-related burnout in healthcare providers: a scoping review of outcomes and interventions. BMJ Open. 2022;12(8):e060865. doi:10.1136/bmjopen-2022-060865

6. AMIA Survey Underscores Impact of Excessive Documentation Burden | AMIA - American Medical Informatics Association. Accessed May 12, 2025. https://amia.org/news-publications/amia-survey-underscores-impact-excessive-documentation-burden

7. Chishtie J, Sapiro N, Wiebe N, et al. Use of Epic Electronic Health Record System for Health Care Research: Scoping Review. J Med Internet Res. 2023;25:e51003. doi:10.2196/51003

8. Shanafelt TD, West CP, Dyrbye LN, et al. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians During the First 2 Years of the COVID-19 Pandemic. Mayo Clin Proc. 2022;97(12):2248-2258. doi:10.1016/j.mayocp.2022.09.002

9. Murad MH, Vaa Stelling BE, West CP, et al. Measuring documentation burden in healthcare. J Gen Intern Med. 2024;39(14):2837-2848.

10. Fu T, Berlin S, Gupta A, Sommer J. Automated Incidental Findings Notification Through the Electronic Health Record Utilizing Dictation Macros. J Imaging Inform Med. Published online January 13, 2025. doi:10.1007/s10278-024-01357-7

11. Onitilo AA, Shour AR, Puthoff DS, Tanimu Y, Joseph A, Sheehan MT. Evaluating the adoption of voice recognition technology for real-time dictation in a rural healthcare system: A retrospective analysis of dragon medical one. Chen ACH, ed. PLOS ONE. 2023;18(3):e0272545. doi:10.1371/journal.pone.0272545

12. Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal. 2024;5(3). doi:10.1056/CAT.23.0404

13. Galloway JL, Munroe D, Vohra-Khullar PD, et al. Impact of an Artificial Intelligence-Based Solution on Clinicians' Clinical Documentation Experience: Initial Findings Using Ambient Listening Technology. J Gen Intern Med. 2024;39(13):2625-2627. doi:10.1007/s11606-024-08924-2

14. Generative AI for Clinical Conversations | Abridge. Accessed May 5, 2025. https://www.abridge.com/

15. Microsoft Dragon Copilot | Microsoft Cloud for Healthcare. Accessed May 5, 2025. https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot

16. Ma SP, Liang AS, Shah SJ, et al. Ambient artificial intelligence scribes: utilization and impact on documentation time. J Am Med Inform Assoc. 2025;32(2):381-385. doi:10.1093/jamia/ocae304

17. Albrecht M, Shanks D, Shah T, et al. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open. 2024;8(1):ooaf013. doi:10.1093/jamiaopen/ooaf013

18. Cao DY, Silkey JR, Decker MC, Wanat KA. Artificial intelligence-driven digital scribes in clinical documentation: Pilot study assessing the impact on dermatologist workflow and patient encounters. JAAD Int. 2024;15:149-151. doi:10.1016/j.jdin.2024.02.009

19. Haberle T, Cleveland C, Snow GL, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc JAMIA. 2024;31(4):975-979. doi:10.1093/jamia/ocae022

20. Shah SJ, Devon-Sand A, Ma SP, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. J Am Med Inform Assoc. 2025;32(2):375-380. doi:10.1093/jamia/ocae295

21. Stults CD, Deng S, Martinez MC, et al. Evaluation of an Ambient Artificial Intelligence Documentation Platform for Clinicians. JAMA Netw Open. 2025;8(5):e258614. doi:10.1001/jamanetworkopen.2025.8614

22. Shah SJ, Crowell T, Jeong Y, et al. Physician Perspectives on Ambient AI Scribes. JAMA Netw Open. 2025;8(3):e251904. doi:10.1001/jamanetworkopen.2025.1904

23. Owens LM, Wilda JJ, Grifka R, Westendorp J, Fletcher JJ. Effect of Ambient Voice Technology, Natural Language Processing, and Artificial Intelligence on the Patient–Physician Relationship. Appl Clin Inform. 2024;15(04):660-667. doi:10.1055/a-2337-4739

24. Guo Y, Hu D, Wang J, et al. Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload. arXiv. Preprint posted online April 2, 2025:arXiv:2504.13879. doi:10.48550/arXiv.2504.13879

25. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-World Evidence Synthesis of Digital Scribes Using Ambient Listening and Generative Artificial Intelligence for Clinician Documentation Workflows: Rapid Review. JMIR AI. 2025;4:e76743.

26. Gebauer S. Benchmarking And Datasets For Ambient Clinical Documentation: A Scoping Review Of Existing Frameworks And Metrics For AI-Assisted Medical Note Generation. medRxiv. Published online 2025:2025-01.

27. Hundal J, Jain M, McCollom J. Ambient Artificial Intelligence in Health Care Documentation: A Review of Tools, Integration, and Clinical Implications. AI Precis Oncol. Published online 2025.

28. Bank AJ, Gage R. Annual impact of scribes on physician productivity and revenue in a cardiology clinic. Clin Outcomes Res. Published online September 2015:489. doi:10.2147/CEOR.S89329

29. Arya R, Salovich DM, Ohman‐Strickland P, Merlin MA. Impact of Scribes on Performance Indicators in the Emergency Department. Acad Emerg Med. 2010;17(5):490-494. doi:10.1111/j.1553-2712.2010.00718.x

30. Wolff B. Artificial intelligence and natural language processing in modern clinical neuropsychology: A narrative review. Clin Neuropsychol. Published online August 18, 2025:1-25. doi:10.1080/13854046.2025.2547934

31. Dragon Medical One | Microsoft for Healthcare. Accessed October 23, 2025. https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-medical-one

32. Saxena K, Diamond R, Conant RF, Mitchell TH, Gallopyn IG, Yakimow KE. Provider Adoption of Speech Recognition and its Impact on Satisfaction, Documentation Quality, Efficiency, and Cost in an Inpatient EHR. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:186-195.

33. Devine EG, Gaehde SA, Curtis AC. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. J Am Med Inform Assoc JAMIA. 2000;7(5):462-468. doi:10.1136/jamia.2000.0070462

34. Kumah-Crystal YA, Pirtle CJ, Whyte HM, Goode ES, Anders SH, Lehmann CU. Electronic Health Record Interactions through Voice: A Review. Appl Clin Inform. 2018;9(3):541-552. doi:10.1055/s-0038-1666844

35. Thaker VV, Lee F, Bottino CJ, et al. Impact of an Electronic Template on Documentation of Obesity in a Primary Care Clinic. Clin Pediatr (Phila). 2016;55(12):1152-1159. doi:10.1177/0009922815621331

36. Fichadia PA, Virmani M, Shah P, et al. Utilization and efficacy of DotPhrases in the electronic medical record for improving physician documentation. Bayl Univ Med Cent Proc. 2024;37(4):692-696. doi:10.1080/08998280.2024.2352993

37. Vawdrey DK, Cauthorn C, Francis D, Hackenberg K, Maloney G, Hohmuth BA. A Practical Approach for Monitoring the Use of Copy-Paste in Clinical Notes. AMIA Annu Symp Proc AMIA Symp. 2021;2021:1178-1185.

38. Can Computer-Assisted Physician Documentation Alleviate Administrative Burden? Accessed October 23, 2025. https://psdconnect.org/journal/can-computer-assisted-physician-documentation-alleviate-administrative-burden

39. Computer Assisted Coding Software: What is CAC in Healthcare. ForeSee Medical. Accessed October 23, 2025. https://www.foreseemed.com/computer-assisted-coding

40. Kakaday R, Herrera EZ, Coskey O, Hertel AW, Kaiser P. The STREAMLINE Pilot – Study on Time Reduction and Efficiency in AI-Mediated Logging for Improved Note-Taking Experience. Appl Clin Inform. Published online March 17, 2025:a-2559-5791. doi:10.1055/a-2559-5791

41. Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. NEJM Catal. 2025;6(5). doi:10.1056/CAT.25.0040

42. Guo Y, Wang J, Hu D, et al. Evaluating ambient artificial intelligence documentation: effects on work efficiency, documentation burden, and patient-centered care. J Am Med Inform Assoc. Published online 2025:ocaf180.

43. Yim W wai, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Sci Data. 2023;10(1):586.

44. Abacha AB, Yim W wai, Fan Y, Lin T. An empirical study of clinical note generation from doctor-patient encounters. In: 2023:2291-2302.

45. Sezgin E, Sirrianni JW, Kranz K. Evaluation of a digital scribe: conversation summarization for emergency department consultation calls. Appl Clin Inform. 2024;15(03):600-611.

46. Arko Iv L, Hudelson C, Kumar J, et al. Documenting Care with AI: A Comparative Analysis of Commercial Scribe Tools. In: Househ MS, Tariq ZUA, Al-Zubaidi M, Shah U, Huesing E, eds. Studies in Health Technology and Informatics. IOS Press; 2025. doi:10.3233/SHTI250857

47. Croxford E, Gao Y, Pellegrino N, et al. Development and Validation of the Provider Documentation Summarization Quality Instrument for Large Language Models. arXiv. Preprint posted online January 17, 2025:arXiv:2501.08977. doi:10.48550/arXiv.2501.08977

48. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform. 2012;3(2):164-174. doi:10.4338/aci-2011-11-ra-0070

49. Crossley GM, Howe A, Newble D, Jolly B, Davies HA. Sheffield Assessment Instrument for Letters (SAIL): performance assessment using outpatient letters. Med Educ. 2001;35(12):1115-1124. doi:10.1046/j.1365-2923.2001.01065.x

50. Sellam T, Das D, Parikh AP. BLEURT: Learning Robust Metrics for Text Generation. arXiv. Preprint posted online May 21, 2020:arXiv:2004.04696. doi:10.48550/arXiv.2004.04696

51. Ganesan K. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. arXiv. Preprint posted online 2018. doi:10.48550/ARXIV.1803.01937

52. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. arXiv. Preprint posted online February 24, 2020:arXiv:1904.09675. doi:10.48550/arXiv.1904.09675

53. Yujian L, Bo L. A Normalized Levenshtein Distance Metric. IEEE Trans Pattern Anal Mach Intell. 2007;29(6):1091-1095. doi:10.1109/TPAMI.2007.1078

54. Nalawati RE, Dini Yuntari A. Ratcliff/Obershelp Algorithm as An Automatic Assessment on E-Learning. In: 2021 4th International Conference of Computer and Informatics Engineering (IC2IE). IEEE; 2021:244-248. doi:10.1109/IC2IE53219.2021.9649217

55. Clinchant S, Perronnin F. Textual Similarity with a Bag-of-Embedded-Words Model. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval. ACM; 2013:117-120. doi:10.1145/2499178.2499180

56. Wu C, Wang B. Extracting Topics Based on Word2Vec and Improved Jaccard Similarity Coefficient. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). IEEE; 2017:389-397. doi:10.1109/DSC.2017.70

57. Specia L, Farzindar A. Estimating machine translation post-editing effort with HTER. Published online 2010.

58. Koponen M. Comparing human perceptions of post-editing effort with post-editing operations.

59. Fu S, Wen A, Schaeferle GM, et al. Assessment of Data Quality Variability across Two EHR Systems through a Case Study of Post-Surgical Complications. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2022;2022:196-205.

60. Huang AE, Hribar MR, Goldstein IH, Henriksen B, Lin WC, Chiang MF. Clinical Documentation in Electronic Health Record Systems: Analysis of Similarity in Progress Notes from Consecutive Outpatient Ophthalmology Encounters. AMIA Annu Symp Proc AMIA Symp. 2018;2018:1310-1318.

61. Lowry LZ, Ramaiah M, Prettyman SS, et al. Examining the Copy and Paste Function in the Use of Electronic Health Records. National Institute of Standards and Technology; 2017:NIST IR 8166. doi:10.6028/NIST.IR.8166

62. Zheng K, Mei Q, Yang L, Manion FJ, Balis UJ, Hanauer DA. Voice-dictated versus typed-in clinician notes: linguistic properties and the potential implications on natural language processing. AMIA Annu Symp Proc AMIA Symp. 2011;2011:1630-1638.

63. Palm E, Manikantan A, Pepin ME, Mahal H, Belwadi SS. Assessing the Quality of AI-Generated Clinical Notes: A Validated Evaluation of a Large Language Model Scribe. arXiv. Preprint posted online May 15, 2025:arXiv:2505.17047. doi:10.48550/arXiv.2505.17047

64. Bracken A, Reilly C, Feeley A, Sheehan E, Merghani K, Feeley I. Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review. J Med Syst. 2025;49(1):28. doi:10.1007/s10916-025-02157-4

65. Wang Z, West CP, Vaa Stelling BE, et al. Measuring Documentation Burden in Healthcare. Agency for Healthcare Research and Quality (US); 2024. Accessed May 5, 2025. http://www.ncbi.nlm.nih.gov/books/NBK608551/

66. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med. 2021;4(1):57. doi:10.1038/s41746-021-00432-5

67. Hassan H, Zipursky AR, Rabbani N, et al. Special Topic on Burnout: Clinical Implementation of Artificial Intelligence Scribes in Healthcare: A Systematic Review. Appl Clin Inform. Published online April 30, 2025:a-2597-2017. doi:10.1055/a-2597-2017

68. Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). Biometrika. 1965;52(3/4):591. doi:10.2307/2333709

69. KLAS Arch Collaborative. KLAS Research. Accessed May 12, 2025. https://engage.klasresearch.com/klas-arch-collaborative/

70. Woolson RF. Wilcoxon Signed‐Rank Test. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. 1st ed. Wiley; 2005. doi:10.1002/0470011815.b2a15177

71. Wing C, Simon K, Bello-Gomez RA. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. Annu Rev Public Health. 2018;39(1):453-469. doi:10.1146/annurev-publhealth-040617-013507

72. Clarke V, Braun V. Thematic analysis. J Posit Psychol. 2017;12(3):297-298. doi:10.1080/17439760.2016.1262613

73. Maleki Varnosfaderani S, Forouzanfar M. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. Bioeng Basel Switz. 2024;11(4):337. doi:10.3390/bioengineering11040337

74. Akbar F, Mark G, Warton EM, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. J Am Med Inform Assoc. 2021;28(5):923-930. doi:10.1093/jamia/ocaa229

75. Garcia P, Ma SP, Shah S, et al. Artificial Intelligence-Generated Draft Replies to Patient Inbox Messages. JAMA Netw Open. 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201

76. Cohen IG, Ritzman J, Cahill RF. Ambient Listening—Legal and Ethical Issues. JAMA Netw Open. 2025;8(2):e2460642. doi:10.1001/jamanetworkopen.2024.60642

77. De Cremer D, Kasparov G. The ethical AI—paradox: why better technology needs more and not less human responsibility. AI Ethics. 2022;2(1):1-4. doi:10.1007/s43681-021-00075-y

78. Harishbhai Tilala M, Kumar Chenchala P, Choppadandi A, et al. Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review. Cureus. 2024;16(6):e62443. doi:10.7759/cureus.62443

79. Kaplan B, Litewka S. Ethical Challenges of Telemedicine and Telehealth. Camb Q Healthc Ethics. 2008;17(4):401-416. doi:10.1017/S0963180108080535

80. Martinez-Martin N, Luo Z, Kaushal A, et al. Ethical issues in using ambient intelligence in health-care settings. Lancet Digit Health. 2021;3(2):e115-e123. doi:10.1016/S2589-7500(20)30275-2

81. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon. 2024;10(4):e26297. doi:10.1016/j.heliyon.2024.e26297

82. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: A systematic review. Health Policy. 2018;122(8):827-836. doi:10.1016/j.healthpol.2018.05.014

83. Guo Y, Hu D, Wang J, et al. Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload. In: Househ MS, Tariq ZUA, Al-Zubaidi M, Shah U, Huesing E, eds. Studies in Health Technology and Informatics. IOS Press; 2025. doi:10.3233/SHTI250921

84. Cochran WG. Some Methods for Strengthening the Common χ 2 Tests. Biometrics. 1954;10(4):417. doi:10.2307/3001616

85. Mixed Effects Logistic Regression | Stata Data Analysis Examples. Accessed December 13, 2025. https://stats.oarc.ucla.edu/stata/dae/mixed-effects-logistic-regression/

86. Josef Perktold, Skipper Seabold, Kevin Sheppard, et al. statsmodels/statsmodels: Release 0.14.2. Published online April 17, 2024. doi:10.5281/ZENODO.593847

87. Kincaid JP, Fishburne Jr, Robert P. R, Richard L. C, Brad S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel: Defense Technical Information Center; 1975. doi:10.21236/ADA006655

88. Xu M, Fralick D, Zheng JZ, Wang B, Tu XM, Feng C. The Differences and Similarities Between Two-Sample T-Test and Paired T-Test. Shanghai Arch Psychiatry. 2017;29(3):184-188. doi:10.11919/j.issn.1002-0829.217070

89. Nanda A, Mohapatra DrBB, Mahapatra APK, Mahapatra APK, Mahapatra APK. Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. Int J Stat Appl Math. 2021;6(1):59-65. doi:10.22271/maths.2021.v6.i1a.636

90. The Myers Difference Algorithm. Accessed December 13, 2025. https://www.nathaniel.ai/myers-diff/

91. Blizzard L, Hosmer W. Parameter Estimation and Goodness-of-Fit in Log Binomial Regression. Biom J. 2006;48(1):5-22. doi:10.1002/bimj.200410165

92. Ross A, Willson VL. Independent Samples T-Test. In: Basic and Advanced Statistical Tests. SensePublishers; 2017:13-16. doi:10.1007/978-94-6351-086-8_3

93. McKnight PE, Najab J. Mann‐Whitney U Test. In: Weiner IB, Craighead WE, eds. The Corsini Encyclopedia of Psychology. 1st ed. Wiley; 2010:1-1. doi:10.1002/9780470479216.corpsy0524

94. Oberg AL, Mahoney DW. Linear Mixed Effects Models. In: Ambrosius WT, ed. Topics in Biostatistics. Vol 404. Methods in Molecular BiologyTM. Humana Press; 2007:213-234. doi:10.1007/978-1-59745-530-5_11

95. Colin Cameron A, Miller DL. A Practitioner's Guide to Cluster-Robust Inference. J Hum Resour. 2015;50(2):317-372. doi:10.3368/jhr.50.2.317

96. Sloss EA, Abdul S, Aboagyewah MA, et al. Toward alleviating clinician documentation burden: a scoping review of burden reduction efforts. Appl Clin Inform. 2024;15(03):446-455.

97. Mullankandy S, Mukherjee S, Ingole BS. Applications of AI in Electronic Health Records, Challenges, and Mitigation Strategies. In: 2024 International Conference on Computer and Applications (ICCA). IEEE; 2024:1-7. doi:10.1109/ICCA62237.2024.10927863

98. Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. Visc Med. 2020;36(6):443-449. doi:10.1159/000511930

99. Stupp D, Barequet R, Lee IC, et al. Structured Understanding of Assessment and Plans in Clinical Documentation. Health Informatics. Preprint posted online April 17, 2022. doi:10.1101/2022.04.13.22273438

100. Nguyen OT, Kunta AR, Katoju S, et al. Electronic Health Record Nudges and Health Care Quality and Outcomes in Primary Care: A Systematic Review. JAMA Netw Open. 2024;7(9):e2432760. doi:10.1001/jamanetworkopen.2024.32760

101. Lindsay MR, Lytle K. Implementing Best Practices to Redesign Workflow and Optimize Nursing Documentation in the Electronic Health Record. Appl Clin Inform. 2022;13(3):711-719. doi:10.1055/a-1868-6431

102. Luo YF, Sun W, Rumshisky A. A Hybrid Normalization Method for Medical Concepts in Clinical Narrative using Semantic Matching. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2019;2019:732-740.

103. Hu D, Guo Y, Cho HN, et al. When AI Writes Back: Ethical Considerations by Physicians on AI-Drafted Patient Message Replies. arXiv. Preprint posted online 2025. doi:10.48550/ARXIV.2508.13217

104. Hu D, Guo Y, Zhou Y, Flores L, Zheng K. A systematic review of early evidence on generative AI for drafting responses to patient messages. Npj Health Syst. 2025;2(1):27. doi:10.1038/s44401-025-00032-5

105. Nguyen OT, Turner K, Charles D, et al. Implementing Digital Scribes to Reduce Electronic Health Record Documentation Burden Among Cancer Care Clinicians: A Mixed-Methods Pilot Study. JCO Clin Cancer Inform. 2023;(7):e2200166. doi:10.1200/CCI.22.00166

106. Jorg T, Kämpgen B, Feiler D, et al. Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing. Insights Imaging. 2023;14(1):47. doi:10.1186/s13244-023-01392-y

107. Suhail D, Wong ZY, Ubhi J, Kungwengwe G, Faderani R, Mosahebi A. Current evidence and future directions of metrics used to evaluate ambient clinical documentation: A scoping review. Int J Med Inf. 2026;205:106113. doi:10.1016/j.ijmedinf.2025.106113

108. Fereday J, Muir-Cochrane E. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. Int J Qual Methods. 2006;5(1):80-92. doi:10.1177/160940690600500107

109. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012;22(3):276-282.

110. Colicchio TK, Dissanayake PI, Cimino JJ. The anatomy of clinical documentation: an assessment and classification of narrative note sections format and content. AMIA Annu Symp Proc AMIA Symp. 2020;2020:319-328.

111. Campbell JR, Carpenter P, Sneiderman C, et al. Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity. J Am Med Inform Assoc. 1997;4(3):238-251. doi:10.1136/jamia.1997.0040238

112. Navarro V. What We Mean by Social Determinants of Health. Int J Health Serv. 2009;39(3):423-441. doi:10.2190/HS.39.3.a

113. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications: A Position Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. Yearb Med Inform. 2019;28(01):128-134. doi:10.1055/s-0039-1677903

114. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, Van Der Schaar M. AI education for clinicians. eClinicalMedicine. 2025;79:102968. doi:10.1016/j.eclinm.2024.102968

115. Cameron S, Turtle‐Song I. Learning to Write Case Notes Using the SOAP Format. J Couns Dev. 2002;80(3):286-292. doi:10.1002/j.1556-6678.2002.tb00193.x

116. Pearce PF, Ferguson LA, George GS, Langford CA. The essential SOAP note in an EHR age. Nurse Pract. 2016;41(2):29-36. doi:10.1097/01.NPR.0000476377.35114.d7

117. Seo JH, Kong HH, Im SJ, et al. A pilot study on the evaluation of medical student documentation: assessment of SOAP notes. Korean J Med Educ. 2016;28(2):237-241. doi:10.3946/kjme.2016.26

118. Acampora G, Cook DJ, Rashidi P, Vasilakos AV. A Survey on Ambient Intelligence in Healthcare. Proc IEEE. 2013;101(12):2470-2494. doi:10.1109/JPROC.2013.2262913

119. McCradden MD, Stedman I. Explaining decisions without explainability? Artificial intelligence and medicolegal accountability. Future Healthc J. 2024;11(3):100171. doi:10.1016/j.fhj.2024.100171

120. Hatherley JJ. Limits of trust in medical AI. J Med Ethics. 2020;46(7):478-481. doi:10.1136/medethics-2019-105935

121. Kahn D, Stewart E, Duncan M, et al. A Prescription for Note Bloat: An Effective Progress Note Template. J Hosp Med. 2018;13(6):378-382. doi:10.12788/jhm.2898

122. Liu J, Capurro D, Nguyen A, Verspoor K. “Note Bloat” impacts deep learning-based NLP models for clinical prediction tasks. J Biomed Inform. 2022;133:104149. doi:10.1016/j.jbi.2022.104149

123. Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: measures and associated insights. J Am Med Inform Assoc. 2021;28(12):2608-2616. doi:10.1093/jamia/ocab194

124. Hultman GM, Marquard JL, Lindemann E, Arsoniadis E, Pakhomov S, Melton GB. Challenges and Opportunities to Improve the Clinician Experience Reviewing Electronic Progress Notes. Appl Clin Inform. 2019;10(03):446-453. doi:10.1055/s-0039-1692164

125. Cole R. Inter-Rater Reliability Methods in Qualitative Case Study Research. Sociol Methods Res. 2024;53(4):1944-1975. doi:10.1177/00491241231156971

126. Olson KD, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open. 2025;8(10):e2534976-e2534976.

127. Guo Y, Hu D, Yang Z, et al. What Do Clinicians Edit in Ambient AI-Drafted Clinical Documentation? A Qualitative Content Analysis. Health Informatics. Preprint posted online January 6, 2026. doi:10.64898/2026.01.05.26343473

128. Guo Y, Hu D, Zhou Y, et al. From Conversation to Chart: An Analysis of Clinician Edits to Ambient AI Draft Notes. Health Informatics. Preprint posted online January 6, 2026. doi:10.64898/2026.01.05.26343471

129. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc. 2010;17(6):646-651. doi:10.1136/jamia.2009.001024

130. Kostina A, Dikaiakos MD, Stefanidis D, Pallis G. Large Language Models For Text Classification: Case Study And Comprehensive Review. arXiv. Preprint posted online January 14, 2025:arXiv:2501.08457. doi:10.48550/arXiv.2501.08457

131. Zhang Y, Wang M, Li Q, Tiwari P, Qin J. Pushing The Limit of LLM Capacity for Text Classification. In: Companion Proceedings of the ACM on Web Conference 2025. ACM; 2025:1524-1528. doi:10.1145/3701716.3715528

132. Wang Z, Pang Y, Lin Y, Zhu X. Adaptable and Reliable Text Classification using Large Language Models. In: 2024 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE; 2024:67-74. doi:10.1109/ICDMW65004.2024.00015

133. Abburi H, Suesserman M, Pudota N, Veeramani B, Bowen E, Bhattacharya S. Generative AI Text Classification using Ensemble LLM Approaches. arXiv. Preprint posted online 2023. doi:10.48550/ARXIV.2309.07755

134. Dong H, Falis M, Whiteley W, et al. Automated clinical coding: what, why, and where we are? Npj Digit Med. 2022;5(1):159. doi:10.1038/s41746-022-00705-7

135. meta-llama/Llama-3.2-3B-Instruct · Hugging Face. December 6, 2024. Accessed February 11, 2026. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

136. Chen WY, Liu YC, Kira Z, Wang YCF, Huang JB. A Closer Look at Few-shot Classification. arXiv. Preprint posted online 2019. doi:10.48550/ARXIV.1904.04232

137. Shorten C, Pierse C, Smith TB, et al. StructuredRAG: JSON Response Formatting with Large Language Models. arXiv. Preprint posted online 2024. doi:10.48550/ARXIV.2408.11061

138. Xu J, Le H. Generating Representative Samples for Few-Shot Classification. arXiv. Preprint posted online 2022. doi:10.48550/ARXIV.2205.02918

139. Yang Y, Huang P, Cao J, Li J, Lin Y, Ma F. A prompt-based approach to adversarial example generation and robustness enhancement. Front Comput Sci. 2024;18(4):184318. doi:10.1007/s11704-023-2639-2

140. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. Advances in Neural Information Processing Systems. Vol 35. Curran Associates, Inc.; 2022:24824-24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

141. IT Sr Technical Specialist, Labcorp, USA, Surisetty LS. GenAI & Healthcare APIs: Securing LLM Access to Sensitive Medical Records. Int J Innov Res Sci Eng Technol. 2026;15(1). doi:10.15680/IJIRSET.2026.1501004

142. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017;24(2):423-431. doi:10.1093/jamia/ocw105

143. Shekhar S, Ghavamzadeh M, Javidi T. Binary Classification with Bounded Abstention Rate. arXiv. Preprint posted online 2019. doi:10.48550/ARXIV.1905.09561

144. Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak. 2017;17(1):155. doi:10.1186/s12911-017-0556-8

145. Kumar S, Datta S, Singh V, Datta D, Kumar Singh S, Sharma R. Applications, Challenges, and Future Directions of Human-in-the-Loop Learning. IEEE Access. 2024;12:75735-75760. doi:10.1109/ACCESS.2024.3401547

146. Jeong J, Kim S, Pan L, et al. Reducing the workload of medical diagnosis through artificial intelligence: A narrative review. Medicine (Baltimore). 2025;104(6):e41470. doi:10.1097/MD.0000000000041470

147. Olex AL, McInnes BT. Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be. J Biomed Inform. 2021;118:103784. doi:10.1016/j.jbi.2021.103784

148. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv. Preprint posted online October 16, 2021:arXiv:2106.09685. doi:10.48550/arXiv.2106.09685

149. Rubin O, Herzig J, Berant J. Learning To Retrieve Prompts for In-Context Learning. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2022:2655-2671. doi:10.18653/v1/2022.naacl-main.191

150. Kwon W, Li Z, Zhuang S, et al. Efficient Memory Management for Large Language Model Serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles. ACM; 2023:611-626. doi:10.1145/3600006.3613165

151. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc. 2021;28(12):2716-2727. doi:10.1093/jamia/ocab170

152. Xie T, Gao Z, Ren Q, et al. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. arXiv. Preprint posted online 2025. doi:10.48550/ARXIV.2502.14768

153. Holmgren AJ, Fenton CL, Thombley R, et al. Ambient Artificial Intelligence Scribes and Physician Financial Productivity. JAMA Netw Open. 2026;9(1):e2553233. doi:10.1001/jamanetworkopen.2025.53233

154. Shah SJ, Garcia P. Ambient AI Scribes—What Is the Return on Investment? JAMA Netw Open. 2026;9(1):e2553238. doi:10.1001/jamanetworkopen.2025.53238

155. Van Tiem J, Cramer E, Iverson C, et al. Listening to the note: clinician perspectives on ambient artificial intelligence scribes in medical documentation. J Am Med Inform Assoc. 2026;33(2):255-262. doi:10.1093/jamia/ocaf214

156. Bakken S. Clinician, patient, and organizational perspectives on ambient AI scribes. J Am Med Inform Assoc. 2026;33(2):253-255. doi:10.1093/jamia/ocaf231

157. Nong P, Neprash HT. Unintended Consequences of Using Ambient Artificial Intelligence Scribes for Billing. JAMA Health Forum. 2026;7(1):e255771. doi:10.1001/jamahealthforum.2025.5771

158. Anderson TN, Mohan V, Gold JA. Ethical considerations for clinical adoption of ambient digital scribe technology. J Am Med Inform Assoc. Published online December 26, 2025:ocaf227. doi:10.1093/jamia/ocaf227

159. Choudhury A. Factors influencing clinicians' willingness to use an AI-based clinical decision support system. Front Digit Health. 2022;4:920662. doi:10.3389/fdgth.2022.920662

160. Burke HB, Sessums LL, Hoang A, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc. 2015;22(1):199-205. doi:10.1136/amiajnl-2014-002726

161. Leung TI, Coristine AJ, Benis A. AI Scribes in Health Care: Balancing Transformative Potential With Responsible Integration. JMIR Med Inform. 2025;13(1):e80898.

162. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181-186. doi:10.1136/jamia.2010.007237

163. Colicchio TK, Cimino JJ. Clinicians' reasoning as reflected in electronic clinical note-entry and reading/retrieval: a systematic review and qualitative synthesis. J Am Med Inform Assoc. 2019;26(2):172-184. doi:10.1093/jamia/ocy155

164. Mehta R, Radhakrishnan N, Warring C, et al. The Use of Evidence-Based, Problem-Oriented Templates as a Clinical Decision Support in an Inpatient Electronic Health Record System. Appl Clin Inform. 2016;07(03):790-802. doi:10.4338/ACI-2015-11-RA-0164

165. Iannello J. Template Design and Analysis: Integrating Informatics Solutions to Improve Clinical Documentation. Fed Pract. 2020;(Vol 37 No 11). doi:10.12788/fp.0072

166. Epstein JA, Cofrancesco J, Beach MC, et al. Effect of Outpatient Note Templates on Note Quality: NOTE (Notation Optimization through Template Engineering) Randomized Clinical Trial. J Gen Intern Med. 2021;36(3):580-584. doi:10.1007/s11606-020-06188-0

167. Morin O, Vallières M, Braunstein S, et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nat Cancer. 2021;2(7):709-722. doi:10.1038/s43018-021-00236-2

168. Ye J, Woods D, Jordan N, Starren J. The role of artificial intelligence for the application of integrating electronic health records and patient-generated data in clinical decision support. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2024;2024:459-467.

169. Mishra V, Liebovitz D, Quinn M, Kang L, Yackel T, Hoyt R. Factors That Influence Clinician Experience with Electronic Health Records. Perspect Health Inf Manag. 2022;19(1):1f.

170. Patel MR, Balu S, Pencina MJ. Translating AI for the Clinician. JAMA. 2024;332(20):1701. doi:10.1001/jama.2024.21772

171. Misra R, Keane PA, Hogg HDJ. How should we train clinicians for artificial intelligence in healthcare? Future Healthc J. 2024;11(3):100162. doi:10.1016/j.fhj.2024.100162

172. Brunner J, Cannedy S, McCoy M, Hamilton AB, Shelton J. Software is Policy: Electronic Health Record Governance and the Implications of Clinical Standardization. J Gen Intern Med. 2023;38(S4):949-955. doi:10.1007/s11606-023-08280-7 1. Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J. 2024;11(3):100157. doi:10.1016/j.fhj.2024.100157

APPENDIX 1: Supplementary Tables for Chapter 3

General notes for Supplementary Tables S1–S5.

(1) Vendor-stratified analyses use the analytic subset with Vendor A/B mapping and required covariates; therefore denominators may differ from the full cohort totals reported in the main manuscript (Chapter 3, Table 5).

(2) "No-edit" at the section-level indicates missing/empty provider-edited final text for that section under our operational definition (whitespace-only and "nan/none/null" strings treated as missing).

(3) Rare categories were collapsed into "Other_" to address sparse cells and quasi/perfect separation; coefficients for "Other_" categories are not interpreted.

(4) Very small p-values are reported as $p < 0.001$ (rather than 0.0) to avoid numeric underflow artifacts.

Supplementary Table S1. Counts of Overall AI-drafted note sections finalized without edits ("no-edit") versus edited, by clinician specialty.

Clinician Specialty	Edited	No-edit	Total	No-edit %
Family Practice	8827	3192	12019	26.6
Internal Medicine	3229	3152	6381	49.4
Rheumatology	4397	1505	5902	25.5
Orthopaedic	3095	1366	4461	30.6

Surgery				
Geriatric Medicine	2615	1402	4017	34.9
Hand Surgery	2488	1498	3986	37.6
Unknown/Not recorded	1454	255	1709	14.9
Neurology	1525	35	1560	2.2
Nurse Practitioner	1173	215	1388	15.5
Physician Assistant	901	455	1356	33.6
Ob/Gyn	785	175	960	18.2
Gastroenterology	655	202	857	23.6
Physical Medicine and Rehab	540	33	573	5.8
Neurosurgery	292	243	535	45.4
Cardiology	465	13	478	2.7
Ophthalmology	164	281	445	63.1

Epilepsy	240	7	247	2.8
Otolaryngology	23	202	225	89.8
Endocrinology	160	6	166	3.6
Hospitalist	67	97	164	59.1
Dermatology	141	16	157	10.2
Colon and Rectal Surgery	92	43	135	31.9
Integrative Medicine	96	37	133	27.8
Emergency Medicine	37	52	89	58.4
Infectious Diseases	55	15	70	21.4
Optometry	7	38	45	84.4
General Surgery	22	12	34	35.3
Psychiatry	14	4	18	22.2
Urology	2	12	14	85.7
Vascular Neurology	11	0	11	0
Hematology/On	5	1	6	16.7

cology				
Orthopedics	2	2	4	50
Vascular Surgery	3	1	4	25
Head and Neck Surgery	2	1	3	33.3
Surgical Oncology	3	0	3	0

* “No-edit %” = No-edit / (Edited + No-edit) note sections within specialty.

* “Unknown/Not recorded” (shown as <NA> in source data) indicates missing specialty labels.

* χ^2 tests assess association between clinician specialty and no-edit status: Overall $\chi^2=3762.493$, df=34, $p<0.001$.

*Specialty totals differ from vendor-stratified versions of this table if those versions exclude records with missing specialty or vendor labels or use analytic subsets with additional covariate requirements.

Supplementary Table S1-2. Counts of Vendor A AI-drafted note sections finalized without edits (“no-edit”) versus edited, by clinician specialty

Clinician Specialty	Edited	No-edit	Total	No-edit %

Cardiology	465	13	478	2.7
Colon and Rectal Surgery	42	24	66	36.4
Dermatology	141	16	157	10.2
Endocrinology	160	6	166	3.6
Family Practice	5937	2225	8162	27.3
Gastroenterology	655	202	857	23.6
Geriatric Medicine	17	13	30	43.3
Hand Surgery	248	147	395	37.2
Hospitalist	67	97	164	59.1
Infectious Diseases	55	15	70	21.4
Integrative Medicine	96	37	133	27.8
Internal Medicine	2439	2970	5409	54.9
Neurology	15	0	15	0

Neurosurgery	280	240	520	46.2
Nurse Practitioner	906	107	1013	10.6
Ob/Gyn	88	21	109	19.3
Ophthalmology	161	276	437	63.2
Optometry	7	38	45	84.4
Orthopaedic Surgery	233	53	286	18.5
Otolaryngology	18	200	218	91.7
Physical Medicine and Rehab	530	33	563	5.9
Physician Assistant	101	8	109	7.3
Rheumatology	4154	1202	5356	22.4
Surgical Oncology	3	0	3	0
Urology	1	3	4	75

Vascular Surgery	3	1	4	25

Supplementary Table S1-3. Counts of Vendor B AI-drafted note sections finalized without edits ("no-edit") versus edited, by clinician specialty

Clinician Specialty	Edited	No-edit	Total	No-edit %
Colon and Rectal Surgery	50	19	69	27.5
Emergency Medicine	37	52	89	58.4
Epilepsy	240	7	247	2.8
Family Practice	2890	967	3857	25.1
General Surgery	22	12	34	35.3
Geriatric Medicine	2598	1389	3987	34.8
Hand Surgery	2240	1351	3591	37.6
Head and Neck Surgery	2	1	3	33.3

Hematology/Oncology	5	1	6	16.7
Internal Medicine	790	182	972	18.7
Neurology	1510	35	1545	2.3
Neurosurgery	12	3	15	20
Nurse Practitioner	267	108	375	28.8
Ob/Gyn	697	154	851	18.1
Ophthalmology	3	5	8	62.5
Orthopaedic Surgery	2862	1313	4175	31.5
Orthopedics	2	2	4	50
Otolaryngology	5	2	7	28.6
Physical Medicine and Rehab	10	0	10	0
Physician	800	447	1247	35.8

Assistant				
Psychiatry	14	4	18	22.2
Rheumatology	243	303	546	55.5
Urology	1	9	10	90
Vascular Neurology	11	0	11	0

Supplementary Table S2. Vendor-stratified binomial mixed-effects GLM for "no-edit" (section finalized without edits).

Parameter (reference levels)	Vendor A: Post. Mean (Post. SD)	Vendor B: Post. Mean (Post. SD)
FIXED EFFECTS (mean structure)		
Intercept (baseline: A&P; baseline encounter type; Primary Care)	-3.0965 (0.0178)	-0.9371 (0.0171)
Section: HPI (vs A&P)	2.3398 (0.0229)	0.6786 (0.0245)
Section: Physical Exam (vs A&P)	2.9516 (0.0611)	1.2532 (0.0489)
Section: Results (vs A&P)	3.2473 (0.0534)	1.9572 (0.0478)
Encounter: Integrative Medicine (vs baseline encounter)	-0.4698 (0.1342)	Not included
Encounter: OB Visit (vs baseline encounter)	-0.0008 (0.8565)	-0.9010 (0.2973)
Encounter: Office Visit (vs baseline encounter)	-0.0842 (0.0189)	-0.9148 (0.0179)

Encounter: Procedure Visit (vs baseline encounter)	-0.4224 (0.3608)	-1.2830 (0.2177)
Encounter: Telemedicine (vs baseline encounter)	0.1214 (0.1214)	-0.4639 (0.0851)
Encounter: Tumor Board (vs baseline encounter)	-2.8837 (1.2099)	Not included
Encounter: Video Visit (vs baseline encounter)	-0.1326 (0.0708)	-0.9609 (0.0864)
Clinician group: Specialty Care (vs Primary Care)	-0.1767 (0.0293)	-0.1396 (0.0237)
RANDOM EFFECTS / VARIANCE STRUCTURE		
Random-effect SD (clinician)	1.9594	1.5731
Random-effect SD (specialty)	0.0479	0.0423
ICC (clinician)	0.538	0.429
ICC (specialty)	0	0
MODEL SAMPLE SIZES		
n_sections	24,769	21,677
n_clinicians	95	77

n_specialties	26	24

*Outcome: no-edit (binary). Reference levels correspond to omitted categories in model coding. For Section_Name2, the reference is A&P. For Clinician_Group, the reference is Primary Care (omitted); Specialty Care is shown as the indicator term. Encounter_Type reference is the omitted category in the analytic dataset and can be verified by listing the first category level used in modeling.

*Fixed effects are posterior means with posterior SDs from vendor-stratified binomial mixed GLMs (logit link). ICCs are computed using residual variance $\pi^2/3$.

*Random effects include clinician-level and specialty-level random intercepts; reported SDs are derived from log-SD variance parameters.

*ICC_specialty is approximately zero in both vendor strata due to very small specialty-level variance.

*Some encounter-type levels appear in only one vendor stratum due to zero counts/lack of representation; those terms are shown as "Not included."

Supplementary Table S3. Association of vendor, visit type, note section, and clinician specialty with long turnaround (>24 hours) among edited notes

Model component	Term (reference	Estimate (beta)	Std. Err.	OR	95% CI (OR)	p-value	Notes

	)						
MAIN EFFECTS (m1)	Intercept	-1.8778	0.258	0.153	0.092 to 0.253	<0.001	Baseline: Vendor A, Office Visit, baseline specialty, baseline section, log1p_n_sections=0
	Vendor B (vs Vendor A)	1.0436	0.248	2.84	1.747 to 4.614	<0.001	Main vendor effect at log1p_n_sections=0
	log1p(n_sections)	0.5109	0.199	1.667	1.129 to 2.461	0.01	Effect of note length/section count in Vendor A
INTERACTION (m1)	Vendor B × log1p(n_sections)	1.0612	0.208	2.89	1.921 to 4.347	<0.001	Additional effect of log1p(n_sections) in Vendor B
VISIT TYPE (vs Office Visit)	Integrative Medicine	0.7214	0.236	2.057	1.295 to 3.269	0.002	

	OB Visit	1.6576	0.445	5.247	2.193 to 12.552	<0.001	
	Procedure Visit	0.1236	0.547	1.132	0.387 to 3.306	0.821	Not statistically significant
	Telemedicine	-0.5184	0.148	0.595	0.446 to 0.795	<0.001	
	Video Visit	-0.3443	0.105	0.709	0.577 to 0.871	0.001	
	Other_Visit_Type (collapsed)	--	--	--	--	--	Collapsed due to sparsity/perfect separation; coefficient not interpreted
NOTE SECTION (vs baseline section)	HPI	0.2346	0.077	1.264	1.087 to 1.471	0.002	
	Other_Section_Name2	-1.3381	1.093	0.262	0.031 to 2.233	0.221	Collapsed due to sparsity; not interpreted

	(collapsed)						
CLINICIAN SPECIALTY (selected; vs reference specialty)	Family Practice	-3.0289	0.18	0.048	0.034 to 0.069	<0.001	
	Hand Surgery	-3.1399	0.197	0.043	0.029 to 0.064	<0.001	
	Orthopaedic Surgery	-2.7103	0.187	0.067	0.046 to 0.096	<0.001	
	Ob/Gyn	-4.6735	0.37	0.009	0.005 to 0.019	<0.001	
	Neurology	-2.4485	0.198	0.086	0.059 to 0.127	<0.001	
	Physician	-2.3183	0.211	0.098	0.065 to	<0.001	

	Assistant				0.149		
	Nurse Practitioner	-2.7847	0.261	0.062	0.037 to 0.103	<0.001	
	Unknown	-1.8259	0.193	0.161	0.110 to 0.235	<0.001	
	Rheumatology	-1.4503	0.174	0.234	0.167 to 0.330	<0.001	
	Internal Medicine	-0.7506	0.167	0.472	0.340 to 0.655	<0.001	
	Gastroenterology	-1.2965	0.258	0.273	0.165 to 0.454	<0.001	
	Geriatric Medicine	-0.3283	0.181	0.72	0.505 to 1.027	0.07	Not statistically significant
	Physical Medicine and Rehab	0.5295	0.252	1.698	1.036 to 2.782	0.036	
	Dermatology	1.4564	0.307	4.29	2.351 to 7.829	<0.001	

	Infectious Diseases	2.5019	0.456	12.206	4.991 to 29.851	<0.001	
	Integrative Medicine (specialty)	-3.1283	0.764	0.044	0.010 to 0.196	<0.001	
	Other_Clinician_Specialty (collapsed)	-6.2304	1.018	0.002	0.0003 to 0.0145	<0.001	Collapsed due to sparsity/perfect separation; interpret cautiously
MODEL COMPARISON	LRT: Vendor × log1p(n_sections)	--	--	--	--	<0.001	LR=25.57, df=1
SAMPLE SIZE	Notes (edited-only with valid turnaroun	--	--	--	--	--	N=20,062; Vendor A=12,439; Vendor B=11,321; >24h proportion=0.159

	d)						

*Model: Binomial GLM (logit). Outcome = long_turnaround_24h (1 if >24 hours). Analytic sample includes edited notes with valid turnaround time (N=20,062); proportion >24h = 0.159.

*Categorical predictors with rare/perfect-separation levels were collapsed into "Other_" to improve model stability; coefficients for "Other_" levels are not interpreted.

*Reference levels: vendor = Vendor A; visit type = Office Visit; section = A&P; clinician specialty = Cardiology.

*The primary model includes the Vendor × log1p(n_sections) interaction (m1). Odds ratios (ORs) and 95% CIs are derived from m1 coefficients.

*Likelihood-ratio test (LRT) compares m1 vs m0 for the Vendor × log1p(n_sections) interaction.

Supplementary Table S4. Note-level edit load by completion turnaround (≤24h vs >24h), overall and by vendor, with mixed-effects model adjusting for vendor and covariates.

Supplementary Table S4-1. Descriptive statistics (total edits per note) by turnaround bucket

Stratum	Turnaround bucket	N notes	Mean edits	SD	Min	P25	Median (P50)	P75	Max
Overall	≤24h	16868	163.8466	211.1563	0	30	85	211	2224

Overall	>24h	3194	140.4383	184.7541	0	28	80.5	178	2148
Vendor A	≤24h	9525	167.7757	203.3184	0	34	94	227	1644
Vendor A	>24h	750	150.2733	179.9645	1	42	94	190	2148
Vendor B	≤24h	7343	158.7498	220.8193	0	27	76	191	2224
Vendor B	>24h	2444	137.4202	186.1311	0	25	76	174	1888

* Descriptive summaries use total edits per note (counts). The adjusted mixed-effects model uses log_edit_total_tokens_note as the dependent variable (log scale). Clinician group effects may be unstable/non-identifiable in this model and are not interpreted when standard errors indicate collinearity.

Supplementary Table S4-2. Nonparametric comparison (Mann–Whitney U) of edit load by turnaround

Comparison	Mann–Whitney U	p-value

Overall (≤24h vs >24h)	28243559	<0.001
Vendor A (≤24h vs >24h)	3579165	0.9257
Vendor B (≤24h vs >24h)	9272159.5	0.01346

Supplementary Table S4-3. Mixed-effects model (note-level; adjusted)

Fixed effect (reference)	Coef.	Std. Err.	z	p-value	95% CI (Coef.)	Interpretation note
Turnaround >24h (vs ≤24h)	-0.269	0.049	-5.528	<0.001	-0.364 to -0.174	Lower edit load on log scale for >24h notes (adjusted)
Vendor B (vs Vendor A)	-0.072	0.049	-1.478	0.139	-0.167 to 0.023	Not statistically significant
Turnaround >24h × Vendor B	0.073	0.056	1.293	0.196	-0.038 to 0.183	No evidence of differential turnaround effect by vendor
log1p(n_sections)	1.685	0.045	37.535	<0.001	1.597 to 1.773	More sections strongly associated with higher edit load

Encounter: Office Visit	-0.112	0.127	-0.883	0.377	-0.362 to 0.137	Not statistically significant
Encounter: Other_encounter_type	-0.194	0.289	-0.674	0.5	-0.760 to 0.371	Not statistically significant
Encounter: Procedure Visit	-0.339	0.176	-1.933	0.053	-0.683 to 0.005	Borderline
Encounter: Telemedicine	-0.281	0.134	-2.09	0.037	-0.545 to -0.017	Statistically significant
Encounter: Video Visit	-0.172	0.129	-1.339	0.18	-0.425 to 0.080	Not statistically significant
Clinician group: Specialty Care (vs Primary Care)	Not interpretable	--	--	--	--	Estimate unstable (extreme SE); omitted from interpretation

*Reference levels: turnaround_24h = ≤24h; vendor = Vendor A; encounter type = Integrative Medicine; clinician group = Primary Care. Coefficients are interpreted relative to these references.

*Model: mixed linear model (REML), dependent variable = log_edit_total_tokens_note

Random intercept: clinician (clinician_id).

*Analytic set: clinicians with ≥100 edited notes (45 clinicians; N=18,027 notes).

Supplementary Table S4-4. Random effects (note-level)

Random effect / variance component	Value
Number of clinicians (random intercept groups)	45
Clinician-level variance (random intercept)	0.0000
Residual variance (scale)	0.9683
ICC_clinician	0.0000

*Model: Linear mixed-effects model estimated using REML; outcome = log(edit_total_tokens_note); random intercept for clinician (clinician_id). Analytic set: clinicians with ≥100 edited notes (45 clinicians; N=18,027 notes).

*Note: The estimated clinician-level random-intercept variance was at the boundary (≈0), yielding ICC_clinician ≈ 0 (i.e., minimal residual clustering by clinician after accounting for fixed effects).

Supplementary Table S5. Predictors of higher editing intensity among edited AI-drafted note sections (OLS; clinician-clustered SE), adjusted for vendor

Outcome: log_edit_total_tokens (edited sections only; >0 edits)

Analytic sample: 33,496 sections; 215 clinicians

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
MODEL / SAMPLE	Outcome: log_edit_total_tokens	--	--	--	--	--	OLS with clinician-clustered SE
MODEL / SAMPLE	Edited sections with >0 edits	--	--	--	--	--	N=33,496; clinicians=215
MODEL / SAMPLE	Model fit	--	--	--	--	--	R^2=0.268; Adj R^2=0.267; F=157.5; Prob(F) <0.001
VENDOR	Vendor B (vs Vendor A)	0.1441	0.268	0.537	0.591	-0.382 to 0.670	Not statistically significant
NOTE SECTION (vs reference section)	HPI	-0.5058	0.161	-3.136	0.002	-0.822 to -0.190	Lower editing intensity vs reference section
NOTE SECTION (vs reference	Phys Exam	-1.3638	0.311	-4.379	<0.001	-1.974 to -0.753	Lower editing intensity vs

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
section)							reference section
NOTE SECTION (vs reference section)	Results	-1.2208	0.219	-5.584	<0.001	-1.649 to -0.792	Lower editing intensity vs reference section
VISIT TYPE (collapsed; vs reference visit type)	OB Visit	0.4354	0.351	1.24	0.215	-0.253 to 1.124	Not statistically significant
VISIT TYPE (collapsed; vs reference visit type)	Office Visit	-0.0712	0.328	-0.217	0.828	-0.714 to 0.572	Not statistically significant
VISIT TYPE (collapsed; vs reference visit type)	Other_Visit_Type	-0.9046	0.482	-1.876	0.061	-1.850 to 0.040	Borderline; very small n (9)

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
VISIT TYPE (collapsed; vs reference visit type)	Procedure Visit	-0.3062	0.348	-0.881	0.378	-0.988 to 0.375	Not statistically significant
VISIT TYPE (collapsed; vs reference visit type)	Telemedicine	-0.0923	0.348	-0.265	0.791	-0.773 to 0.589	Not statistically significant
VISIT TYPE (collapsed; vs reference visit type)	Video Visit	0.0426	0.316	0.135	0.893	-0.577 to 0.662	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Colon and Rectal Surgery	-0.6454	0.497	-1.298	0.194	-1.620 to 0.329	Not statistically significant
CLINICIAN SPECIALTY	Dermatology	0.2788	0.17	1.638	0.101	-0.055 to 0.612	Not statistically significant

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
(collapsed; vs reference specialty)							
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Emergency Medicine	-1.776	0.303	-5.864	<0.001	-2.370 to -1.182	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Endocrinology	0.5103	0.33	1.546	0.122	-0.137 to 1.157	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Epilepsy	1.6049	0.318	5.04	<0.001	0.981 to 2.229	Higher editing intensity vs reference specialty

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Family Practice	-0.8017	0.267	-3.007	0.003	-1.324 to -0.279	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Gastroenterology	-0.6801	0.356	-1.911	0.056	-1.378 to 0.017	Borderline
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	General Surgery	-0.3281	0.303	-1.083	0.279	-0.922 to 0.266	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs reference	Geriatric Medicine	-0.8677	0.369	-2.35	0.019	-1.591 to -0.144	Lower editing intensity vs reference specialty

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
specialty)							
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Hand Surgery	-0.8354	0.289	-2.894	0.004	-1.401 to -0.270	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Hospitalist	-1.603	0.165	-9.692	<0.001	-1.927 to -1.279	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Infectious Diseases	0.1681	0.189	0.889	0.374	-0.202 to 0.538	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs	Integrative Medicine	-0.9944	0.316	-3.148	0.002	-1.613 to -0.375	Lower editing intensity vs reference

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
reference specialty)							specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Internal Medicine	-0.5298	0.239	-2.214	0.027	-0.999 to -0.061	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Neurology	1.0933	0.319	3.428	0.001	0.468 to 1.718	Higher editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Neurosurgery	-0.9023	0.524	-1.722	0.085	-1.929 to 0.125	Not statistically significant
CLINICIAN SPECIALTY	Nurse Practitioner	-0.5537	0.459	-1.207	0.227	-1.453 to 0.345	Not statistically significant

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
(collapsed; vs reference specialty)							
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Ob/Gyn	-0.8186	0.281	-2.914	0.004	-1.369 to -0.268	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Ophthalmology	-1.1835	0.403	-2.937	0.003	-1.973 to -0.394	Lower editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Orthopaedic Surgery	-0.6115	0.322	-1.9	0.057	-1.242 to 0.019	Borderline

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Other_Clinician_Specialty	-0.62	0.421	-1.474	0.14	-1.444 to 0.204	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Otolaryngology	0.4671	0.567	0.824	0.41	-0.644 to 1.579	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Physical Medicine and Rehab	0.5566	0.263	2.113	0.035	0.040 to 1.073	Higher editing intensity vs reference specialty
CLINICIAN SPECIALTY (collapsed; vs reference	Physician Assistant	-0.8368	0.59	-1.418	0.156	-1.994 to 0.320	Not statistically significant

Component	Predictor (reference)	Coef. (β)	Std. Err.	z	p-value	95% CI (β)	Interpretation note
specialty)							
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Rheumatology	0.4819	0.477	1.009	0.313	-0.454 to 1.418	Not statistically significant
CLINICIAN SPECIALTY (collapsed; vs reference specialty)	Unknown	-0.0548	0.333	-0.164	0.869	-0.708 to 0.598	Not statistically significant

* Reference categories (omitted): vendor = Vendor A; section = A&P; visit type = Integrative Medicine; clinician specialty = Cardiology.

* Model: OLS regression with clinician-clustered (cluster-robust) standard errors (215 clinicians). Outcome = log_edit_total_tokens among edited sections with >0 edits (N=33,496).

APPENDIX 2: Representative examples of clinician edits to ambient AI-drafted notes by five qualitative themes.

Theme 1- Clinicians correct factual errors in AI drafts		
Edit type	**AI draft text** [a]	**Clinician-final text** [a]
Demographic correction	"Seen by [PROVIDER_NAME_A]."	"Seen by [PROVIDER_NAME_B]."
Demographic correction	"The patient is [AGE_A] years old."	"The patient is [AGE_B] years old."
Temporal correction	"...in March"	"...in next March"
Tense correction	"The patient has been taking..."	"The patient had been taking..."
Pronoun correction	"...he..."	"...she..."
Clinical event type correction	"a recent hospitalization"	"a recent ER visit on 12/15/24"
Medication detail correction	"has been using fluconazole 200 mg weekly"	"has been using fluconazole 200 mg a few times per week"
Medication name correction	"Zepal"	"albuterol"
Laterality correction	"Cerumen impaction… more on the left"	"Cerumen impaction… more on the right"
Anatomic site correction	"MCP joint"	"DIP joint"
Procedure/test label correction	"colonoscopy"	"sigmoidoscopy"
Test detail correction	"MRI of brain"	"MRA of head and neck"
Device name correction	"a cardio device"	"a Kardia device"

Followup detail correction	“return in 6 week”	“return in 2–3 Month”
Theme 2- Clinicians refine generic drafts into specialty-appropriate documentation		
Edit type	**AI draft text**	**Clinician-final text**
Diagnosis specificity	“headache”	“possibly migraine with aura”
Diagnosis specificity	“hypoglycemia and hyperglycemia”	“hyperglycemia and rare hypoglycemia”
Diagnosis specificity	“Psoriatic Arthritis…”	“Rheumatoid arthritis, double seropositive… high titer anti-CCP”
Symptom specificity	“noise sensitivity”	“light and sound sensitivity”
Symptom specificity	“pain in back and neck”	“neck pain without acute changes”
Symptom specificity	“Achilles”	“Calf pain”
Temporal course specification	“headaches”	“persistent daily headaches since November 2024”
Medication regimen specification	“Vitamin B2”	“Vitamin B2 400 mg per day”
Medication regimen specification	“Prescribe meloxicam once daily…”	“Restart meloxicam 15 mg daily with dinner.”
Theme 3- Clinicians revise overly certain language to match the available evidence with clinical wording		
Edit type	**AI draft text**	**Clinician-final text**
Certainty calibrated to evidence	“due to”	“likely due to”
Certainty calibrated to evidence	“No retinal or optic nerve disease.”	“No obvious retinal or optic nerve disease.”

Certainty calibrated to evidence	“to definitively rule out or confirm AFib”	“as there is currently no confirmed diagnosis of AFib.”
Theme 4- Clinicians convert transcript-like narrative into professional documentation		
Edit type	**AI draft text**	**Clinician-final text**
Clinical register normalization	“She’s doing okay”	“Patient reports stable condition”
Medical terminology standardization	“a disintegrating disc”	“a bulging disc”
Clinician framing / documentation precision	“He is on HIV medications.”	“He is compliant with HIV medications.”
Abbreviation standardization	“mean corpuscular volume”	“MCV”
Technical specificity increase	“potential removal of the stones”	“potential removal of the dropped gall stones in hepatorenal fossa”
Procedure terminology standardization	“partial amputation”	“ray amputation”
Generic to specific entity	“used the medication”	“used Qsymia”
Medication name normalization	“Lexapro”	“escitalopram”
Patient narrative reframed into clinician documentation	Subjective self-attributions (“which she believes are related…”, “she recalls…”)	Reframed into objective clinician documentation
Theme 5- Clinicians reorganize or condense content to improve clarity [b]		
Edit type	**AI draft structure**	**Clinician-final structure**
Format conversion	Long paragraph	Bullet list (e.g., “• Continue healthy diet…”)

Format conversion	Problem paragraph (prose)	Numbered list (1–7)
Format conversion	Prose narrative	Structured "#condition:" headings
Section relocation	Social + surgical history embedded mid-paragraph	Moved under labeled "past medical history"
Relevance trimming	Large chunks of past history	Deleted and/or reorganized; "psychiatric history" section added/structured

[a] *Examples are de-identified; bracketed tokens denote masked identifiers, and excerpts are paraphrased where needed to protect privacy.*

[b] *For Theme 5, edits span longer passages and section organization; we therefore report structure-level transformations rather than verbatim text.*

APPENDIX 3: Semi-Structured Interview Guide and De-identified Note-Editing Examples

Introduction:

Hi Dr. [Participant Name],

Thank you for participating in this study.

My name is [Interviewer Name], and I'm a PhD student in the Informatics Department at UC Irvine. Our team has been following the UCI ambient AI pilot and ongoing deployment, with a focus on evaluation and improvement. The goal of today's interview is to understand how clinicians review and edit ambient AI–generated note drafts, and the reasoning behind them.

From a previous content analysis study, we find several recurring patterns of the changes that have been made to AI draft notes. I'll later show a small set of examples with slides and ask for your perspective on what drives those changes.

The interview will take about 30-45 minutes. I'll start with a few questions about your experience using the tool, then we'll discuss the edit patterns. Your input will help us develop actionable recommendations for both tool improvement and clinician training.

Verbal Consent

Before we start, I want to assure you that we are not collecting identifiable information in this interview. With your permission, we would like to record this interview so we can transcribe it for analysis. The recording will be stored securely and used only for research purposes. Do I have your permission to record this interview?

[Consent]

[Start the recording]

Section A. Background and overall editing intent

1. **Overall editing amount**

   In your day-to-day use, would you describe your edits to the AI draft as usually light, moderate, or heavy?

   - Probe: When do you typically edit the draft? (during the visit / right after / end of day / later)

2. Where do you spend most of your editing effort? Which note sections usually get the most vs least edits (e.g., HPI, A&P, PE, Results, Meds)?

3. **Top intentions**

   When you edit the AI draft, what are the most common changes you make (for example, medication prescription)?

   - Could you list a few types of changes that come to mind?

Section B. We compared AI drafts vs final notes and found recurring edit patterns. I'll show a few examples and would love to understand the intention behind these changes, and what could reduce the need to edit. (start to show examples of each theme)

4. **Edit Pattern 1: factual errors**

   Do you encounter factual errors in the AI draft?

   - Probe: If so, what kinds are most common?

   - Probe: What do you think contributes to these errors?

- e.g., have you reviewed the transcripts?

Edit type	AI draft text	Clinician-final text
Demographic/pronoun correction	"Seen by [PROVIDER_NAME_A]."	"Seen by [PROVIDER_NAME_B].
	"The patient is [AGE_A] years old."	"The patient is [AGE_B] years old."
	"...he..."	"...she..."
Temporal correction	"...in March"	"...in next March"
Tense correction	"The patient has been taking..."	"The patient had been taking..."
Clinical detail correction	" …using fluconazole 200 mg weekly"	" …using fluconazole 200 mg a few times per week"
	"Zepal"	"albuterol"
	"Cerumen impaction… more on the left"	"Cerumen impaction… more on the right"
	"MCP joint"	"DIP joint"
	"colonoscopy"	"sigmoidoscopy"
	"MRI of brain"	"MRA of head and neck"

5. **Edit Pattern 2: editing clinical details**

   In our review, we often saw clinicians revise clinical details in the draft, such as medication specifics, symptom descriptions, and test/treatment details (SHOW examples). Have you made similar types of edits?

   - Probe: What purpose does that serve for you/ what are you aiming to improve?
   - Probe: Is this more about specialty norms, your personal style, or clinic/institution expectations?

- Probe: In the AI draft, do you ever notice irrelevant details included or important details missing? Can you share 1–2 specific examples?

Edit type	**AI draft text**	**Clinician-final text**
Diagnosis specificity	"headache"	"possibly migraine with aura"
	"hypoglycemia and hyperglycemia"	"hyperglycemia and rare hypoglycemia"
	"Psoriatic Arthritis…"	"Rheumatoid arthritis, double seropositive… high titer anti-CCP"
Symptom specificity	"noise sensitivity"	"light and sound sensitivity"
	"pain in back and neck"	"neck pain without acute changes"
	"Achilles"	"Calf pain"
Temporal course specification	"headaches"	"persistent daily headaches since November 2024"
Medication regimen specification	"Vitamin B2"	"Vitamin B2 400 mg per day"
	"Prescribe meloxicam once daily…"	"Restart meloxicam 15 mg daily with dinner."
Medical terminology standardization	"a disintegrating disc"	"a bulging disc"

6. **Edit Pattern 3: certainty calibration**

   We saw edits that soften or qualify language (SHOW examples, 'due to' → 'likely,' 'rule out' framing). Have you made similar types of edits?

   - Probe: What's the intent behind these changes for you?
   - Probe: Should the AI default to more hedged language, or would that create other problems?

Edit type	**AI draft text**	**Clinician-final text**
Certainty calibrated to evidence	"due to"	"likely due to"

	"No retinal or optic nerve disease."	"No obvious retinal or optic nerve disease."
	"to definitively rule out or confirm AFib"	"as there is currently no confirmed diagnosis of AFib."

7. **Edit Pattern 4: transcript-like → professional documentation**

   We saw conversational/transcript-like text rewritten into more professional clinical documentation. Have you made similar types of edits?

   - Probe: What are you trying to achieve with these edits?
   - Probe: Does the current SOAP-style organization help or hinder, especially in the Subjective section?

Edit type	**AI draft text**	**Clinician-final text**
Rewriting Transcript-Like Text	"She's doing okay"	"Patient reports stable condition"
	Subjective self-attributions ("which she believes are related…", "she recalls…")	Reframed into objective clinician documentation
Language Reframing	"He is on HIV medications."	"He is compliant with HIV medications."
Abbreviation standardization	"mean corpuscular volume"	"MCV"
Procedure terminology standardization	"partial amputation"	"ray amputation"
Generic → specific entity	"used the medication"	"used Qsymia"
	"Lexapro"	"escitalopram"

8. **Edit Pattern 5: reorganize/condense + relevance trimming**

We saw clinicians reorganize and condense drafts (bullets, problem-oriented A/P, moving content, trimming details). Have you made similar types of edits?

- Probe: What's the goal behind these changes?
- Probe: How do you decide what's important enough to keep?
- Probe: Do you feel the AI synthesizes and organizes information the way a clinician would? If not, what would you want it to do differently?

Edit type	**AI draft structure**	**Clinician-final structure**
Format conversion	Long paragraph	Bullet list (e.g., "• Continue healthy diet…")
	Prose narrative	Structured "#condition:" headings
	Problem paragraph (prose)	Numbered list (1–7)
Section relocation	Social + surgical history embedded mid-paragraph	Moved under labeled "past medical history"
Relevance trimming	Large chunks of past medical history	Deleted and/or reorganized; "psychiatric history" section added/structured
	Long descriptions of social history	Condensed and kept physical activity descriptions only

Section C. Close: actionable recommendations

9. **Language and multilingual use:** Do you use the tool in languages other than English? If yes, how does that affect the experience (e.g., accuracy, tone, terminology, workflow)? What issues are different compared to English?
10. **Clinician strategies and workflow adjustments:** Do you do anything differently during the visit or after to help the AI produce a better note?

11. **Change over time:** Since you started using the tool, what has improved over time (either the tool itself or your own workflow with it)? What challenges have persisted?
12. **Specialty-specific needs and recommendations:** For your specialty, what additional content or features would make the tool more useful?