Epidemiology & Health Data Insights
(eISSN 3080-8111)

*Review Article*

# Autonomous Medical Documentation Pipelines: Integrating Large Language Models and Cloud Speech Services to Reduce Clinician Administrative Burden and EHR Workflow Bottlenecks

**Heidi Heather Henry Heimbruch[1], Williams Temidayo Solomon[2], Salamah Abimbola Junaid[3], Daniel Obinna Eke[1], Leo Tata[4], Olaitan Ebenezer Oluwadare[5], Habib Shehu[6]**

[1]*Department of Nursing, Myrtle E. and Earl E. Walker College of Health Professions, Maryville University of St. Louis, St. Louis, MO, USA*
[2]*Department of Medicine, Howard University College of Medicine, Washington, DC, USA*
[3]*Emergency Medicine Department, Northern General Hospital, Sheffield NHS Foundation Trust, Sheffield, England*
[4]*College of Health and Human Sciences, Iowa State University, Ames, IA, USA*
[5]*Applied Mathematics Department, School of Physics, Engineering, Mathematics and Computer Science, Delaware State University, Dower, DE, USA*
[6]*Department of Information Technology, Kampala International University, Kampala, Uganda*

Received: May 29, 2026
Accepted: Jun 19, 2026

Corresponding author's email:
habib.shehu@studmc.kiu.ac.ug

**Citation:** Heimbruch HHH, Solomon WT, Junaid SA, Eke DO, Tata L, Oluwadare OE, et al. Autonomous Medical Documentation Pipelines: Integrating Large Language Models and Cloud Speech Services to Reduce Clinician Administrative Burden and EHR Workflow Bottlenecks. Epidemiology and Health Data Insights. 2026;2(4):ehdi045.
https://doi.org/10.63946/ehdi/18797



## ABSTRACT

**Background:** Autonomous medical documentation has become a central challenge in modern healthcare, as clinicians increasingly spend substantial portions of their workday on electronic health record (EHR) tasks rather than direct patient care, contributing to dissatisfaction, burnout, and workflow inefficiencies. Large language models (LLMs) and cloud-based speech services offer a potential solution by automating aspects of note generation, structured documentation, and coding support while preserving clinician oversight.

**Objectives:** To map and characterize the emerging evidence on autonomous or semi-autonomous documentation pipelines that integrate LLMs, cloud speech services, and EHR workflows, focusing on technical architecture, workflow integration, safety, governance, and clinician experience.

**Methodology:** A systematic search of PubMed, ACM Digital Library, and Dimensions AI was conducted for studies published from 2019 to March 2026, supplemented by reports and policy documents. Eligible studies included empirical research, reviews, and implementation reports addressing documentation efficiency, accuracy, and clinician outcomes.

**Findings:** Evidence indicates that layered pipelines combining speech processing, LLM-driven note generation, retrieval of structured EHR data, and FHIR-based integration can reduce documentation time, decrease after-hours charting, improve note quality, and enhance clinician satisfaction. Human oversight remains essential to mitigate risks from hallucination, transcription errors, and workflow misalignment. Governance, consent, and data security are critical for safe adoption.

**Conclusion:** Overall, autonomous documentation pipelines are most effective when implemented as assisted automation tools embedded within context-sensitive clinical workflows, with iterative evaluation and robust governance. These insights provide a foundation for future research, clinical validation, and scalable deployment strategies.

**Keywords**: Clinical Documentation; Large Language Models; Cloud Speech Services; Electronic Health Records; AI Scribes; Workflow Optimization; Clinician Burnout

## Introduction

Clinical documentation is central to high-quality patient care, continuity, communication, quality measurement, and legal compliance. The shift from paper to electronic health records (EHRs) was originally intended to improve accessibility and data exchange, but in practice it has also introduced a substantial administrative burden for clinicians [1]. Clinicians across specialties now routinely report spending a disproportionate share of their workday interacting with EHR systems rather than with patients, contributing directly to dissatisfaction, burnout, and reduced capacity for direct clinical care [2].

Studies indicate that up to half of a clinician's time may be consumed by charting, order entry, inbox management, and billing documentation [3]. These demands are compounded by complex regulatory requirements, structured reporting obligations, and reimbursement processes, all of which reduce workflow flexibility and create inefficiencies [4]. Consequently, technological solutions that can reduce clerical workload while preserving accuracy and safety are increasingly sought.

Artificial intelligence (AI) and large language models (LLMs) have emerged as promising tools to automate aspects of clinical documentation [5]. LLMs generate human-like text, summarize information, extract key data, and structure content for clinical use [6]. Their integration into healthcare has moved rapidly from theoretical frameworks to pilot implementations across diverse clinical environments [7]. Parallel advancements in cloud speech and automatic speech recognition (ASR) technologies have improved transcription accuracy and domain-specific term recognition, enabling near-real-time capture of clinician–patient interactions [8].

Ambient documentation systems, which passively record encounters and generate draft notes for review, illustrate how these technologies can operate synergistically. Evidence suggests AI scribes and related systems reduce documentation time, decrease after-hours charting, and improve same-day note completion, while preserving clinician oversight [9]. However, limitations in methodology, short-term evaluation, and variable note quality underscore the need for rigorous investigation [4]. Persistent challenges include LLM hallucinations, accuracy variation across settings, privacy concerns, and integration complexities with heterogeneous EHR systems [5, 3, 6]. FHIR-based interoperability offers promise, but real-world implementation remains limited [10].

Accordingly, this scoping review aims to map and characterize the emerging evidence on autonomous medical documentation pipelines that integrate LLMs, cloud speech services, and EHR workflows. The review focuses on system conceptualization, design, implementation, and evaluation, with particular attention to documentation burden, workflow efficiency, interoperability, coding support, governance, and safety. Specifically, it addresses the following questions: what types of autonomous or semi-autonomous documentation pipelines have been reported; how have LLMs, cloud speech services, and EHR integration been combined; what outcomes have been evaluated; and what evidence gaps remain for safe, scalable adoption.

## Methodology

### Study Design

This study was conducted as a scoping review to map and characterize the emerging evidence on autonomous medical documentation pipelines integrating large language models (LLMs), cloud speech services, and electronic health record (EHR) workflows. A scoping review design was chosen because the literature is heterogeneous, encompassing technical, clinical, and implementation studies, and because the aim is to identify the breadth, range, and key concepts of the evidence. The review sought to clarify how autonomous or semi-autonomous documentation systems are conceptualized, implemented, and evaluated, and to highlight gaps in clinical, technical, and governance knowledge.

### Information Sources

Relevant literature was identified through searches of PubMed, ACM Digital Library, and Dimensions AI, selected for their coverage of healthcare, engineering, and technology research. In addition to the reports retrieved from these primary sources, reports and policy documents from reputable organizations were searched. This inclusion ensures that the review incorporates both academic studies and practical reports that can provide insights into real-world applications and challenges.

### Search Strategy

The search focused on publications from 2019 to the search date (March 29, 2026), reflecting recent advances in LLMs, cloud speech technologies. For PubMed, search strategies combined MeSH terms with title and abstract keywords, while ACM Digital Library and Dimensions AI relied on keywords in titles and abstracts. Boolean operators "AND" and "OR" were applied to combine concepts of clinical documentation, AI-enabled transcription, workflow, and clinician

burden. Table 1 detailed the search strings adopted for each database.

**Inclusion and Exclusion Criteria**

Studies were included if they examined AI-enabled clinical documentation, cloud speech recognition, EHR integration, or related workflow outcomes. Eligible sources comprised empirical studies, reviews, implementation reports, and technical papers that provided evidence on efficiency, documentation quality, or clinician experience. Studies were excluded if they focused on general AI applications unrelated to clinical documentation, non-clinical speech recognition, opinion pieces without evidence, or publications outside the 2019–2026 range.

**Study Selection**

Search results were screened for relevance based on titles and abstracts, followed by full-text review. This screening was handled by two independent reviewers and discrepancies were resolved by discussion. Sources were selected if they contributed evidence to the review objectives. The study selection process was documented in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram (Figure 1).

**Data Collection and Presentation of Findings**

Data were systematically extracted from each included study using a structured charting framework capturing study design, setting, technologies used, pipeline components, and key outcomes relating to efficiency, safety, and workflow integration. The data were iteratively refined through cross-checking for consistency. Findings were then grouped thematically to describe technical architecture, workflow integration, safety and accuracy, governance, clinician acceptance, and research gaps, providing a comprehensive overview of the current landscape of autonomous documentation systems.

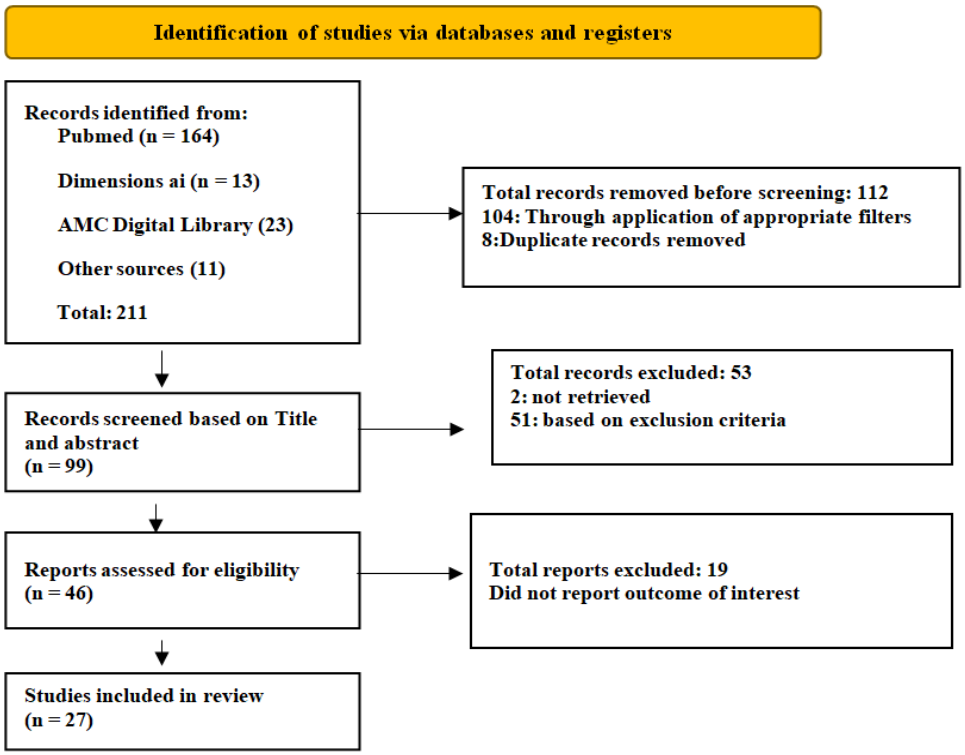


**Figure 1:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram.

## Results

**Overview of the Included Studies**

The systematic search identified a total of 211 records. Following removal of 104 records through appropriate database filters and 8 duplicates removed in reference manager (Zotero), 99 records were screened at title and abstract stage, of which 46 were assessed at full-text. 27 publications ultimately met all eligibility criteria and were included in this review. Table 1 summarized the characteristics of the included studies.

The included studies reflect a rapidly expanding and methodologically diverse body of evidence on autonomous clinical documentation pipelines. Across the twenty-seven studies, there is a clear predominance of evidence syntheses, including scoping and systematic reviews, alongside an emerging but substantial proportion of empirical and implementation-based investigations. Most studies concentrate on large language models, speech-to-text systems, and retrieval-augmented generation, either

individually or as integrated components within broader documentation ecosystems. A significant cluster of the literature also draws attention to interoperability frameworks such as FHIR, underscoring the centrality of structured data exchange in enabling real-world deployment of these systems within electronic health record environments.

In terms of application domains, ambulatory and emergency care settings feature most prominently, reflecting where documentation burden is most acutely experienced and where AI-enabled solutions are being most actively tested. Several studies report measurable reductions in documentation time, clinician workload, and after-hours charting, while others focus more critically on safety, hallucination risk, and governance considerations. Notably, patient-facing and ethical dimensions such as consent and trust appear consistently across the literature, indicating a maturing awareness of implementation realities beyond technical performance. Collectively, the evidence portrays a field in transition, moving from isolated technological innovation toward integrated, workflow-sensitive documentation ecosystems shaped by both efficiency gains and persistent safety and governance challenges.

**Table 1: Characteristics of the included studies**

| Study ID | Study Type | Focus Area | Technology Component | Clinical/Workflow Domain | Key Outcomes Reported | Key Findings | Relevance to Autonomous Documentation Pipeline |
|---|---|---|---|---|---|---|---|
| Woo et al. (2026) [6] | Scoping review | LLMs in documentation | LLM systems | Clinical documentation | Efficiency, safety, readability | LLMs improve note quality but have safety risks | Establishes overall evidence base |
| Saadat et al. (2025) [8] | Technical review | AI documentation enhancement | AI-assisted documentation tools | Clinical note-taking | Error reduction, interoperability | Improves documentation quality and workflow | Supports AI-driven documentation systems |
| Artsi et al. (2025) [10] | Systematic review | Real-world LLM workflows | LLM integration | Clinical workflows | Implementation outcomes | LLMs feasible but variable across settings | Shows real-world deployment challenges |
| Klusty et al. (2025) [11] | Technical study | Automated transcription | Speech-to-text systems | Clinical transcription | Accuracy, feasibility | Enables automated clinical transcription | Core speech layer of pipeline |
| Gargari & Habibi (2025) [12] | Narrative review | RAG in healthcare | Retrieval-Augmented Generation | Knowledge grounding | Model accuracy, hallucination reduction | RAG improves contextual reliability | Retrieval layer support |
| HL7 (FHIR v5.0) [13] | Technical standard | Health interoperability | FHIR API | EHR integration | Data exchange capability | Enables standardized EHR connectivity | Core integration infrastructure |
| Ng et al. (2026) [14] | Systematic review | Clinical speech AI | Speech recognition systems | Documentation workflows | Accuracy, usability | Variable accuracy across settings | Speech-to-text limitations |
| Brown et al. (2020) [15] | Foundational ML paper | LLM architecture | Transformer LLMs | NLP generation | Language generation capability | Basis for modern LLM systems | Underpins LLM layer |
| Neupane et al. (2024) [16] | Experimental study | Clinical summarization | LLM summarization tools | Clinical notes | Summary quality | LLMs generate structured summaries | Note generation layer |
| Bednarczyk et al. (2025) [17] | Scoping review | Text summarization | LLM summarization | Clinical documentation | Quality of summaries | Promising but variable quality | Evidence synthesis support |
| Hou et al. (2025) [18] | Experimental study | Clinical coding | Fine-tuned LLMs | Billing/coding workflows | ICD accuracy | High coding performance after tuning | Coding automation layer |

| Study ID | Study Type | Focus Area | Technology Component | Clinical/Workflow Domain | Key Outcomes Reported | Key Findings | Relevance to Autonomous Documentation Pipeline |
|---|---|---|---|---|---|---|---|
| Li et al. (2024) [19] | Framework study | Interoperability | FHIR-GPT | EHR integration | Data integration efficiency | LLMs can integrate with FHIR APIs | EHR interoperability layer |
| Neha et al. (2025) [20] | Comprehensive review | RAG systems | Retrieval systems | Clinical AI systems | Grounding improvement | RAG improves factual accuracy | Retrieval enhancement |
| ONC (2026) [21] | Policy brief | Health API adoption | FHIR APIs | Health systems | Adoption trends | Increasing API use in hospitals | System-level interoperability |
| Nellutla (2021) [22] | Technical framework | Compliance systems | DevOps pipelines | Governance | HIPAA compliance | Enables continuous compliance | Governance layer |
| Tajirian et al. (2025) [23] | Mixed-methods | EHR burden | EHR systems | Clinical workflow | Time burden, burnout | High documentation burden persists | Baseline workflow inefficiency |
| Ma et al. (2025) [24] | Observational study | Ambient AI scribes | AI scribe systems | Clinical documentation | Time reduction | Reduced documentation time | Core autonomous pipeline evidence |
| Olson et al. (2025) [25] | Multi-site study | Burnout reduction | Ambient AI scribes | Ambulatory care | Burnout, workload | Reduced burnout and workload | Workforce impact evidence |
| Song et al. (2025) [26] | Experimental study | ED documentation | LLM assistant | Emergency care | Time efficiency | Faster discharge documentation | High-intensity workflow use |
| Anderson et al. (2025) [27] | Simulation study | AI scribe safety | Ambient scribe tools | Clinical documentation | Errors, completeness | Risk of omissions and inaccuracies | Safety validation layer |
| Lawrence et al. (2025) [28] | Survey study | Consent ethics | AI documentation systems | Governance | Consent perception | Consent is essential for adoption | Ethical governance layer |
| Leiserowitz et al. (2025) [29] | Survey study | Patient attitudes | Voice AI systems | Patient interaction | Trust, acceptance | Patients generally receptive | User acceptance evidence |
| Ramsay et al. (2025) [30] | Mixed-methods | AI procurement | AI deployment systems | Health systems | Adoption barriers | Governance challenges affect rollout | Implementation constraints |
| Alboksmaty et al. (2025) [31] | Systematic review | Voice-to-text AI | Speech recognition tools | Primary care | Care quality | Improves efficiency but safety unclear | Speech layer evidence |
| Topaz et al. (2025) [32] | Review | AI scribe risks | AI scribes | Clinical documentation | Risk analysis | Identifies hallucination and safety risks | Safety risk framework |
| Palm et al. (2025) [33] | Validation study | Note quality | LLM scribes | Clinical documentation | Note quality metrics | LLM notes comparable but imperfect | Output validation layer |
| Klusty et al. (2025) [34] | Technical study | Clinical transcription | Automated transcription | Speech layer | Accuracy | Clinical transcription architecture | Reinforces speech-to-text layer |

**Technical Architecture of Autonomous Documentation Pipelines**

The reviewed studies consistently describe autonomous documentation pipelines as layered systems, combining cloud speech processing, large language model (LLM) generation, retrieval of structured EHR data, and integration back into EHRs [6, 11–13]. The cloud speech layer captures clinician–patient interactions, performing speech-to-text conversion, speaker diarization, noise handling, and recognition of medical vocabulary [8, 11, 14]. The LLM layer generates structured clinical notes, organizes content into templates such as SOAP, extracts key findings, proposes coding, and provides patient-facing instructions [6, 15–18]. The retrieval layer accesses prior notes, labs, medications, allergies, and problem lists, often using retrieval-augmented generation (RAG) techniques to anchor note content in patient-specific context [12, 19, 20]. Integration into EHRs relies heavily on FHIR APIs, enabling secure data exchange, context retrieval, and note write-back [13, 21].

Figure 2 illustrates this pipeline, showing the flow from audio capture to final EHR update with clinician oversight. Evidence indicates that pipeline design is critical: well-coordinated systems reduce manual effort while preserving safety and clinical relevance [6, 11, 12]. Deployment models varied across studies, with cloud-hosted, hybrid, and private solutions reported depending on institutional infrastructure and data governance requirements.

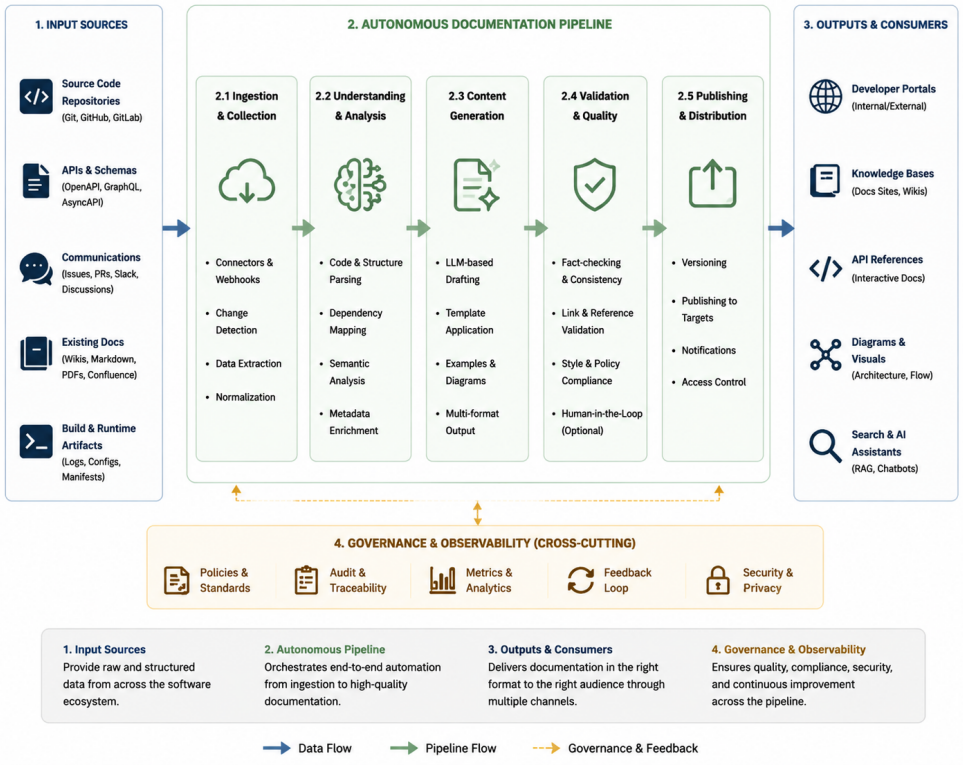


**Figure 2:** Technical Architecture of Autonomous Documentation Pipelines

Table 2 presents the technological features, integration models, and study settings for each pipeline reported in the included literature. This mapping demonstrates the diversity of system architectures and highlights common functional layers.

**Table 2: Evidence Matrix of LLM and Cloud Speech Documentation Studies**

| Author (Year) | Country / Health System | Clinical Setting | Technology Used | Documentation Task Automated | Type of Integration | Outcome Measured | Reported Benefits | Reported Risks / Limitations | Relevance to Autonomous Documentation Pipelines |
|---|---|---|---|---|---|---|---|---|---|
| Woo et al., 2025 [6] | Multiple countries (literature-based) | Mixed medical and nursing settings | LLM-based documentation systems | Note generation, discharge summaries, encounter | Varied | Efficiency, readability, note quality, safety | Up to 40% time savings, improved readability | Hallucination, privacy, bias, weaker performance in complex cases, limited | Establishes broader evidence base for LLM-driven |

| Author (Year) | Country / Health System | Clinical Setting | Technology Used | Documentation Task Automated | Type of Integration | Outcome Measured | Reported Benefits | Reported Risks / Limitations | Relevance to Autonomous Documentation Pipelines |
|---|---|---|---|---|---|---|---|---|---|
| | | | | documentation | | | | standardization | documentation systems |
| Song et al., 2025 [26] | South Korea, tertiary academic hospital | Emergency department | On-site LLM assistant | Discharge note drafting | Direct use during documentation task | Writing time, note quality | Median note-writing time reduced from 69.5s to 32s, no decline in assessed quality | Single clinical context, task-specific design | Demonstrates LLM utility in high-volume ED documentation |
| Ma et al., 2025 [27] | USA, large academic medical center | Ambulatory specialties | Ambient AI scribe powered by LLM | Visit note generation | Routine patient encounters | Time per note, daily documentation, after-hours documentation, total EHR time | Reduced time per note by 0.57 min, daily documentation by 6.89 min, after-hours by 5.17 min, total EHR by 19.95 min/day | Short evaluation period, heterogeneous clinician use | Shows AI integration can reduce documentation burden in real-world workflows |
| Olson et al., 2025 [25] | USA, six health systems | Ambulatory care | Ambient AI scribe | Clinical note drafting | Real-world deployment | Burnout, cognitive load, after-hours work, patient attention | Burnout declined 51.9% → 38.8%, improved patient focus, reduced after-hours work | Self-reported outcomes, short follow-up | Connects AI documentation to clinician well-being and workflow relief |
| Albo ksmaty et al., 2025 [32] | Multiple countries | Primary care / outpatient | AI-powered voice-to-text systems | Clinical consultation documentation | Varied | Efficiency, care quality, patient-centeredness, safety | Improved documentation speed, administrative burden, patient interaction | Safety evidence inconclusive, heterogeneous study designs | Demonstrates role of cloud speech as front-end for documentation pipelines |
| Ng et al., 2025 [14] | Multiple countries | Mixed clinical settings | AI-based speech recognition | Speech transcription for documentation | Varied | Accuracy, adaptability, workflow integration | Reduced manual entry, faster documentation | Errors with specialized terms, accents, inconsistent workflow integration, | Highlights importance of high-quality speech capture for downstream |

| Author (Year) | Country / Health System | Clinical Setting | Technology Used | Documentation Task Automated | Type of Integration | Outcome Measured | Reported Benefits | Reported Risks / Limitations | Relevance to Autonomous Documentation Pipelines |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | need for human review | LLM processing |
| Anderson et al., 2025 [28] | USA | Simulated ambulatory encounters | Ambient digital scribe platforms | Transcript and note generation | Platform-based simulation | Note completeness, omission/commission errors, safety | Generated notes clinically useful | Varying quality across platforms, omissions, inaccuracies | Reinforces need for clinician oversight and safety validation |
| Hou et al., 2025 [18] | USA, research environment | Clinical documentation datasets | Fine-tuned LLM for coding | ICD-10 code generation | Coding-focused model | Exact match, category match | 69.2% exact match, 87.2% category match after domain-specific fine tuning | Tested as a model task, not mature live workflow | Extends pipeline concept from note generation toward coding assistance |

**Workflow Integration and Implementation Models**

Autonomous documentation systems have been implemented in ambulatory care, emergency departments, inpatient units, and specialty clinics [11, 23, 27, 25]. Across these settings, studies report reductions in documentation time, after-hours charting, total EHR interaction, and clicks per encounter. Song et al. (2025) [26] found that LLM-assisted discharge notes in an emergency department reduced median note-writing time from 69.5 seconds to 32.0 seconds. Ma et al. (2025) [27] reported that ambient AI scribes reduced daily documentation by nearly 7 minutes and after-hours work by 5 minutes, improving clinician focus on patient care. Olson et al. (2025) [25] observed a decrease in clinician burnout from 51.9% to 38.8% after 30 days of AI scribe use.

Integration challenges were noted when pipelines interfaced with heterogeneous EHR systems, requiring adaptations to local workflows and attention to interoperability [11, 14, 21]. These studies suggest that workflow improvements rely on embedding automation into existing processes rather than merely providing transcription tools, emphasizing the alignment of AI systems with clinician task flow [6, 27].

**Safety and Accuracy of AI Documentation**

The literature reports variability in note quality and accuracy. LLM-generated notes improve efficiency but can contain hallucinations, omissions, or misclassified findings [6, 28, 33, 34]. Speech recognition systems may introduce errors due to accents, background noise, or specialized terminology, which propagates downstream to the LLM layer [11, 14].

Coding and billing assistance was tested in some pipelines. Hou et al. (2025) [18] reported 69.2% exact match and 87.2% category match for ICD-10 coding after fine-tuning LLMs. Anderson et al. (2025) [28] found that ambient digital scribes required clinician review to ensure completeness and safety. Overall, studies emphasize that human oversight remains essential, and AI outputs must be treated as draft content until clinician verification. Table 2 captures reported outcomes, including risks and limitations related to accuracy and workflow reliability.

**Governance, Privacy, and Ethical Considerations**

Studies highlight the importance of consent, privacy, and accountability in pipeline deployment [6, 10, 29, 30]. Lawrence et al. (2025) [29] and Leiserowitz et al. (2025) [30] note that clear patient consent for ambient recording is critical. Privacy risks exist at multiple points, including audio capture, cloud storage, transcription, model processing, and EHR write-back [6, 10]. Vendor accountability, encryption, audit trails, and role-based access were consistently recommended for mitigating risk [12, 31].

Table 3 presents a governance and implementation readiness checklist summarizing risks, safeguards, responsible actors, and monitoring indicators. It provides a structured overview of

operational requirements for safe and compliant deployment of autonomous documentation systems.

**Table 3: Evidence Matrix of LLM and Cloud Speech Documentation Studies**

| Governance Domain | Main Risk | Pipeline Stage Affected | Required Safeguard | Responsible Actor | Monitoring Indicator | Evidence Needed Before Scale-Up |
|---|---|---|---|---|---|---|
| Consent and Recording | Weak or poorly understood consent | Audio capture | Clear patient education, explicit consent, opt-out route | Clinician, Health System | Consent completion rate, patient concerns logged | Demonstrated patient understanding and ability to decline |
| Data Privacy | Unauthorized access, unclear retention, secondary data use | Audio, transcript, storage, model processing | Encryption, access logs, retention rules, vendor data agreements | Privacy Officer, IT, Vendor | Access anomalies, privacy incidents | Approved privacy review and data flow map |
| Speech Recognition Accuracy | Misheard terminology, accent-related errors | Speech-to-text conversion | Local validation, specialty vocabulary testing, error review | Informatics Team, Clinical Leads | Correction rate, critical term error rate | Evidence of acceptable accuracy across real care settings |
| LLM Hallucination | Fabricated findings, plans, or consent language | Note generation | Source grounding, constrained templates, mandatory clinician review | AI Governance Group, Clinicians | Hallucination audit results | Demonstrated safety and note quality in representative cases |
| Clinical Coding Error | Incorrect or inflated coding suggestions | Coding and billing support | Human validation, billing audit | Coding & Compliance Team | Coding mismatch rate | Concordance with accepted coding standards |
| EHR Integration Failure | Wrong patient write-back, delayed posting, note misplacement | EHR integration | Identity checks, sandbox testing, rollback procedures | EHR Team, IT | Failed update rate, correction tickets | Stable integration performance in pilot testing |
| Bias Across Accents or Languages | Unequal documentation quality | Speech and generation layers | Stratified testing across user and patient groups | Equity Lead, Data Science Team | Subgroup accuracy differences | Evidence of acceptable performance across diverse populations |
| Clinician Over-Reliance | Automation bias, reduced critical review | Review and approval stage | Training, attestation prompts, visible edit workflow | Clinical Leadership | Review completion rate, edit behavior | Competency evidence for users before routine deployment |
| Vendor Accountability | Opaque updates, unstable service, unclear liability | Entire pipeline | Contracted update disclosure, audit rights, service reliability terms | Procurement, Legal, Info Governance | Service outages, unreported model changes | Supplier documentation and governance acceptance |
| Audit and Feedback | Repeated errors remain unnoticed | Post-submission, monitoring | Audit trail, incident reporting, periodic quality review | Quality & Safety Committee | Incident closure time, audit frequency | Regular reports demonstrating corrective action and learning |

**Clinician Acceptance and User Experience**

Clinician perception varied by setting and system maturity. Studies report increased satisfaction, reduced cognitive load, and improved focus on patients when AI scribes or LLM pipelines were implemented [25, 27, 32]. Trust and acceptance depended on perceived reliability, accuracy, and the ability to review and correct notes [6, 10, 29, 30]. User experience was sensitive to workflow alignment and EHR integration; poor integration diminished efficiency gains and clinician trust [11, 14].

**Evidence Gaps and Research Needs**

Despite promising results, the evidence base remains limited by small sample sizes, short follow-up

periods, focus on high-resource settings, and concentration in specific specialties [6, 25, 32]. There is a scarcity of data from rural or low-resource environments and limited long-term safety or real-world deployment studies. Few studies evaluated patient-centered outcomes, cost-effectiveness, or workflow redesign beyond transcription automation. These gaps highlight areas for future research.

## Discussion

The findings of this scoping review elucidate the evolving but still emergent nature of autonomous documentation pipelines that combine large language models (LLMs), cloud speech services, and EHR workflows. Overall, the evidence aligns with earlier reviews indicating that AI-powered documentation systems can streamline routine text generation and reduce clinician time on electronic tasks [35], while also illustrating persistent challenges in accuracy, integration, and real-world deployment. Efficiency gains, as observed in reductions in documentation time and after-hours work, appear tied to systems that align closely with clinician workflows and enable iterative human review rather than standalone automation. This suggests that the value of autonomy in this context lies more in assisted automation than in unsupervised note creation.

Technically, layered architectures that incorporate cloud speech processing, LLM drafting, and retrieval of structured EHR data appear essential to contextualized documentation, but their benefits are contingent on robust interoperability and workflow alignment. Systematic reviews of AI speech recognition emphasize that performance variability across clinical environments and terminology can undermine utility unless models are calibrated to domain-specific vocabularies [22]. Similarly, a scoping by Gebauer (2025) [36] on evaluation frameworks highlights the lack of standardized metrics for assessing completeness, factuality, and clinical relevance—metrics crucial for comparing systems and fostering iterative improvements.

Safety and accuracy remain central concerns. Ambient AI scribes may produce structured notes comparable to clinician-authored documents on quality instruments, but they also exhibit tendencies toward hallucination and verbosity, which necessitate rigorous review workflows [34]. These observations align with emergent research noting potential patient safety risks flagged by end-user feedback, particularly when transcription errors involve critical clinical details [37]. Implementation studies also reveal that the impact of automation on after-hours EHR use is complex, with some clinicians experiencing an increase in documentation time when review processes are not sufficiently streamlined, underscoring the importance of workflow design and human-in-the-loop structures [22, 36].

Ethical and governance considerations stand out as prerequisites for safe adoption. Frameworks for evaluating ambient digital scribing tools explicitly incorporate human evaluation, automated metrics, and iterative testing to ensure clinical and factual fidelity [38]. Effective governance structures must address consent, privacy, and model monitoring while ensuring accountability for final documentation, as pipelines that process identifiable health information raise significant confidentiality concerns. These findings echo broader work on AI in clinical workflows that emphasize regulatory compliance and post-deployment monitoring as critical to maintaining safety and trust [10].

Clinician acceptance is influenced by both perceived reliability and integration friction. Studies outside of core documentation pipeline research report that workflows with poor integration into existing EHR interfaces may negate potential efficiency gains, as clinicians spend time reconciling draft text or correcting structure [35]. Qualitative synthesis suggests clinicians appreciate draft generation but are cautious about over-reliance on outputs that may propagate errors or undermine clinical reasoning.

Despite demonstrable promise, the evidence base remains constrained by methodological heterogeneity, small sample sizes, and limited specialty-specific evaluation, particularly in real-world, low-resource, or rural contexts. Longitudinal research is needed to assess sustained impacts on burnout, workflow, patient experiences, and system-level outcomes. Moreover, standardized evaluation frameworks that combine clinical quality metrics with usage analytics will be essential to advancing the field beyond pilot implementations toward safe, scalable deployment.

## Conclusion

This review found that autonomous medical documentation pipelines integrating large language models and cloud speech services hold substantial promise for reducing clinician administrative burden and improving EHR workflow efficiency. Evidence indicates that these systems can streamline note generation, decrease after-hours charting, and enhance clinician focus on patient care, particularly when

designed with layered architectures that combine high-accuracy speech transcription, structured data retrieval, and contextualized LLM drafting. However, it was also found that variability in transcription accuracy, hallucinations, and integration challenges remain significant barriers to fully realizing these benefits, highlighting the necessity of clinician oversight and human-in-the-loop review. Ethical, privacy, and governance considerations were consistently emphasized, demonstrating that safe deployment requires robust consent procedures, auditability, and clear accountability for final documentation. Practically, institutions should prioritize pilot implementations that integrate automated workflows into existing clinical processes, provide training for end-users, and establish monitoring frameworks to assess performance, safety, and equity. Future research is needed to evaluate long-term impacts on clinician burnout, patient outcomes, workflow optimization, and cost-effectiveness, as well as to expand evidence in low-resource and rural settings. Collectively, these findings suggest that autonomous documentation technologies are most effective when positioned as assisted automation tools embedded within carefully governed and context-sensitive clinical workflows.

## Acknowledgements

**Author Contributions:** H.H.H.H. conceptualized the study and led manuscript development. W.T.S., S.A.J., D.O.E., and L.T. contributed to the literature search, data extraction, and synthesis of evidence. O.E.O. provided expertise on health information technology, interoperability, and technical review of the manuscript. H.S. supervised the study, contributed to study design, critically revised the manuscript for intellectual content, and served as corresponding author. All authors participated in drafting, reviewing, and approving the final version of the manuscript.

**Funding:** This research received no external funding.

**Disclosures:** The authors declare no conflicts of interest related to this work. The authors alone are responsible for the content and writing of this manuscript.

**Disclosure of AI Use:** Artificial intelligence (AI) tools, including large language model–based writing assistance, were used solely to support language editing, grammar refinement, formatting, and organizational improvements during manuscript preparation. All scientific content, interpretation of evidence, critical analysis, and final editorial decisions were performed and verified by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the manuscript content.

**Acknowledgments:** The authors thank the researchers, clinicians, and organizations whose published work contributed to the evidence synthesized in this review. No additional acknowledgments are declared.

## References

1. Sasseville M, Yousefi F, Ouellet S, et al. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. Healthcare (Basel). 2025;13(12):1447. doi:10.3390/healthcare13121447
2. Olakotan O, Samuriwo R, Ismaila H, Atiku S. Usability Challenges in Electronic Health Records: Impact on Documentation Burden and Clinical Workflow: A Scoping Review. Journal of Evaluation in Clinical Practice. 2025;31(4):e70189. doi:10.1111/jep.70189
3. Sarraf B, Ghasempour A. Impact of artificial intelligence on electronic health record-related burnouts among healthcare professionals: systematic review. Front Public Health. 2025;13. doi:10.3389/fpubh.2025.1628831
4. Zhao J, Liu H, Chen Y, Song F. Application of artificial intelligence tools and clinical documentation burden: a systematic review and meta-analysis. BMC Med Inform Decis Mak. 2025;26(1):29. doi:10.1186/s12911-025-03324-w
5. Al-Garadi M, Mungle T, Ahmed A, Sarker A, Miao Z, Matheny ME. Large Language Models in Healthcare. arXiv.org. February 6, 2025. doi:10.48550/arXiv.2503.04748
6. Woo BFY, Cato K, Cho H, You SB, Song J. The use of large language models in clinical documentation: A scoping review. International Journal of Nursing Studies. 2026;176:105322. doi:10.1016/j.ijnurstu.2025.105322
7. Winkler C. Leveraging Large Language Models in Healthcare: From Speech Documentation to Conversational Agents. In: Scholz S, Wüchner-Fuchs M, Höller K, eds. Advancements in Digital Health and Care: Empowering Healthcare Through Innovation, Strategies and

Ethical Considerations. Springer Nature Switzerland; 2026:187-206. doi:10.1007/978-3-032-16837-5_16

8. Saadat S, Khalilizad Darounkolaei M, Qorbani M, Hemmat A, Hariri S. Enhancing Clinical Documentation with AI: Reducing Errors, Improving Interoperability, and Supporting Real-Time Note-Taking. InfoScience Trends. 2025;2(3):1-13. doi:10.61186/ist.202502.01.01
9. Razaghi M, Hafez A, Farina JM, et al. Transforming clinical documentation with ambient artificial intelligence (AI) scribes: a narrative review of technology, impact, and implementation. Cardiovascular Diagnosis and Therapy. 2026;16(1):11-11. doi:10.21037/cdt-2025-454
10. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, Klang E. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front Digit Health. 2025;7. doi:10.3389/fdgth.2025.1659134
11. Klusty MA, Logan WV, Armstrong SE, et al. Toward Automated Clinical Transcriptions. AMIA Jt Summits Transl Sci Proc. 2025;2025:235-241. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12150720/
12. Gargari OK, Habibi G. Enhancing medical AI with retrieval-augmented generation: A mini narrative review. Digit Health. 2025;11:20552076251337177. doi:10.1177/20552076251337177
13. Index - FHIR v5.0.0. Accessed March 29, 2026. https://fhir.hl7.org/fhir/index.html
14. Ng SI, Xu L, Siegert I, et al. An End-to-End Overview of Clinical Speech AI. IEEE Trans Audio Speech Lang Process (2025). 2026;34:1016-1048. doi:10.1109/taslpro.2026.3660470
15. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. arXiv. Preprint posted online July 22, 2020:arXiv:2005.14165. doi:10.48550/arXiv.2005.14165
16. Neupane S, Tripathi H, Mitra S, et al. ClinicSum: Utilizing Language Models for Generating Clinical Summaries from Patient-Doctor Conversations. Proc IEEE Int Conf Big Data. 2024;2024:5050-5059. doi:10.1109/bigdata62323.2024.10825266
17. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, et al. Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review. Journal of Medical Internet Research. 2025;27(1):e68998. doi:10.2196/68998
18. Hou Z, Liu H, Bian J, He X, Zhuang Y. Enhancing medical coding efficiency through domain-specific fine-tuned large language models. npj Health Syst. 2025;2(1):14. doi:10.1038/s44401-025-00018-3
19. Li Y, Wang H, Yerebakan HZ, Shinagawa Y, Luo Y. FHIR-GPT Enhances Health Interoperability with Large Language Models. NEJM AI. 2024;1(8):AIcs2300301. doi:10.1056/AIcs2300301
20. Neha F, Bhati D, Shukla DK. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. AI. 2025;6(9). doi:10.3390/ai6090226
21. Hospital Use of APIs to Enable Data Sharing between EHRs and Third-Party Technology. ONC Health IT Research & Analysis. Accessed March 29, 2026. https://healthit.gov/data/data-briefs/hospital-use-of-apis-to-enable-data-sharing-between-ehrs-and-third-party-technology/
22. Nellutla N. Continuous Compliance Pipelines for HIPAA-Aligned Healthcare DevOps Systems. International Journal of Science and Engineering Applications. 2021;10(12). doi:10.7753/IJSEA1012.1006
23. Tajirian T, Lo B, Strudwick G, et al. Assessing the Impact on Electronic Health Record Burden After Five Years of Physician Engagement in a Canadian Mental Health Organization: Mixed-Methods Study. JMIR Human Factors. 2025;12(1):e65656. doi:10.2196/65656
24. 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

25. 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. doi:10.1001/jamanetworkopen.2025.34976
26. Song JW, Park J, Kim JH, You SC. Large Language Model Assistant for Emergency Department Discharge Documentation. JAMA Netw Open. 2025;8(10):e2538427. doi:10.1001/jamanetworkopen.2025.38427
27. 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
28. Anderson TN, Mohan V, Dorr DA, Ratwani RM, Biro JM, Gold JA. Evaluating the Quality and Safety of Ambient Digital Scribe Platforms Using Simulated Ambulatory Encounters. Mayo Clinic Proceedings: Digital Health. 2025;3(4):100292. doi:10.1016/j.mcpdig.2025.100292
29. Lawrence K, Kuram VS, Levine DL, et al. Informed Consent for Ambient Documentation Using Generative AI in Ambulatory Care. JAMA Netw Open. 2025;8(7):e2522400. doi:10.1001/jamanetworkopen.2025.22400
30. Leiserowitz G, Mansfield J, MacDonald S, Jost M. Patient Attitudes Toward Ambient Voice Technology: Preimplementation Patient Survey in an Academic Medical Center. JMIR Medical Informatics. 2025;13(1):e77901. doi:10.2196/77901
31. Ramsay AIG, Crellin N, Lawrence R, et al. Procurement and early deployment of artificial intelligence tools for chest diagnostics in NHS services in England: a rapid, mixed method evaluation. eClinicalMedicine. 2025;89:103481. doi:10.1016/j.eclinm.2025.103481
32. Alboksmaty A, Aldakhil R, Hayhoe BWJ, Ashrafian H, Darzi A, Neves AL. The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: a systematic review. eBioMedicine. 2025;118:105861. doi:10.1016/j.ebiom.2025.105861
33. Topaz M, Peltonen LM, Zhang Z. Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. NPJ Digit Med. 2025;8(1):569. doi:10.1038/s41746-025-01895-6
34. Palm E, Manikantan A, Mahal H, Belwadi SS, Pepin ME. Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. Front Artif Intell. 2025;8:1691499. doi:10.3389/frai.2025.1691499
35. 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
36. Gebauer S. Benchmarking And Datasets For Ambient Clinical Documentation: A Scoping Review Of Existing Frameworks And Metrics For AI-Assisted Medical Note Generation. medRxiv. Preprint posted online January 29, 2025:2025.01.29.25320859. doi:10.1101/2025.01.29.25320859
37. Dai J, Huang A, Nasrallah C, et al. Patient Safety Risks from AI Scribes: Signals from End-User Feedback. arXiv.org. December 1, 2025. Accessed May 17, 2026. https://arxiv.org/abs/2512.04118v1
38. Wang H, Yang R, Alwakeel M, et al. An evaluation framework for ambient digital scribing tools in clinical applications. npj Digit Med. 2025;8(1):358. doi:10.1038/s41746-025-01622-1