International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246/WCAI25-115
Eureka Vision Publication | WCAI25-Conference Proceeding

Original Article

AI Copilots for Clinical Documentation to Reduce Physician Burnout

Sai Nitesh Palamakula
Software Engineer, Microsoft Corporation, Charlotte, NC, USA.

Abstract - Clinical documentation has become a principal contributor to physician burnout, with research showing that doctors devote more than a third, and sometimes over half, of their professional time to electronic health record (EHR) tasks and associated administrative responsibilities. This paper delves into the deployment of AI Copilot systems based on large language models (LLMs), capable of transforming voice or shorthand physician notes into structured, standards-compliant documentation. The analysis emphasizes edge/cloud hybrid deployment architectures, security and regulatory compliance, integration strategies, and system evaluation metrics. By grounding its exploration in real-world outcomes and consensus guidelines, this paper proposes a comprehensive framework for designing, deploying, and evaluating AI Copilots that reduce documentation burden and physician burnout while preserving clinical safety, accuracy, and trustworthiness.

Keywords - Physician burnout, clinical documentation, large language models (LLMs), artificial intelligence (AI) scribes, speech-to-text, edge-cloud hybrid deployment, EHR integration, evaluation metrics, security, compliance, medical informatics.

1. Introduction

The burden of clinical documentation is a leading driver of physician dissatisfaction and burnout worldwide, particularly in technologically advanced health systems where EHR adoption is nearly ubiquitous. Studies have consistently quantified the excessive time allocation to documentation, often exceeding the time doctors spend in direct patient care [1][2]. In the United States, this burden is particularly acute, with outpatient notes averaging fourfold the length of those in peer nations and documentation for billing and regulatory compliance consuming disproportionate work hours [3][4]. The administrative demands, fragmented workflows, and poorly optimized EHR interfaces are not merely sources of inefficiency; they directly undermine physician well-being, hinder patient engagement, and increase the risk of medical error [5][6].

Recent advancements in artificial intelligence especially the maturation of large language models have fueled the development of AI Copilots designed to automate and augment the clinical documentation process. These systems promise to ambiently transcribe and structure live conversations or shorthand notes, converting them into interoperable, standards-based clinical records for seamless EHR integration [7][8]. However, realization of their full potential requires careful system architecture, robust evaluation, adherence to privacy and ethical frameworks, and alignment with the end-user clinical workflow [9][10]. This paper examines the current landscape of LLM-based AI Copilots, focusing on secure edge/cloud hybrid deployments. It assesses the evidence for their efficacy in reducing burnout, explores the technical and organizational challenges of real-world implementation, and details comprehensive evaluation strategies for ensuring clinical safety, accuracy, and trust.

2. Purpose and scope

2.1. Purpose

This paper aims to address a critical issue in modern healthcare: the excessive burden of clinical documentation that significantly contributes to physician burnout. By investigating the role of AI Copilots powered by large language models (LLMs), the paper explores how real-time conversion of voice and shorthand notes into structured clinical documentation can alleviate administrative pressures. The proposed systems leverage ambient AI capabilities, integrated with secure edge/cloud hybrid platforms, to ensure low latency, high accuracy, and compliance with healthcare privacy standards.

2.2. Scope

This paper presents a multidisciplinary exploration of LLM-based AI Copilot technologies designed to streamline clinical documentation and mitigate physician burnout. It begins by examining the causes and consequences of documentation overload in healthcare environments, then surveys current AI-powered documentation systems with a focus on ambient scribing and real-world deployments. The technical core includes a detailed breakdown of edge/cloud hybrid architectures, illustrating performance and compliance benefits critical to clinical applications. This is followed by a comprehensive system design, encompassing audio capture, speech recognition, LLM-driven summarization, and generation of interoperable clinical notes compatible with EHR systems. Evaluation strategies are proposed using NLP-based benchmarks, clinical note validation tools, and physician satisfaction metrics to assess system impact and trustworthiness. Finally, the paper considers deployment scalability, specialty customization, and regulatory challenges, offering insights into the ethical, legal, and organizational

barriers to adoption. As a design-oriented analysis, the scope is deliberately framed to support technical architects, clinical informaticians, and healthcare IT leaders in building safe, efficient, and clinically integrated AI documentation systems.

3. Related work

3.1. Physician Documentation Workload and Burnout

Empirical assessments of clinical documentation burden consistently demonstrate that documentation occupies 34–55% of physicians' workdays [11][12]. U.S. physicians not only face lengthier documentation requirements but also report higher dissatisfaction and after-hours workload, with 77% of surveyed clinicians attributing late work hours to documentation, and 82% disagreeing that their documentation workload is appropriate [3][13]. The most time-consuming components relate to compliance, billing, and regulatory mandates, which undermine the perceived value of EHR systems [2][14]. Burnout prevalence among physicians has risen to epidemic proportions, with rates exceeding 50% for both trainees and practicing doctors. Burnout metrics are most measured using standardized instruments such as the Maslach Burnout Inventory, Oldenburg Burnout Inventory, and the Stanford Professional Fulfillment Index, each revealing strong associations between high documentation workload and adverse outcomes including medical errors, depersonalization, and intent to leave the profession [13][15][16]. Conceptual frameworks for physician burnout identify excessive time pressure, lack of autonomy, inefficient workflows, and the cognitive load imposed by EHR interactions as primary antecedents [14][17]

3.2. Evolution of AI Copilots and AI-Driven Documentation

Efforts to mitigate documentation burden have progressed from medical scribes and rudimentary dictation tools to advanced ambient AI scribe systems based on LLMs [18][19][20]. Traditional scribes, while effective, are costly, raise patient privacy concerns, and are unsustainable at scale [21]. AI-based solutions now leverage real-time speech-to-text, NLP for context extraction, and LLM-based summarization to generate structured clinical notes [22][23]. Notably, leading deployments such as Penda Health's AI Consult have demonstrated clinically meaningful reductions in diagnostic and treatment errors, with a 16% relative reduction in diagnostic errors and a 13% reduction in treatment errors, validating the potential of LLM copilots to improve clinical quality as well as workflow efficiency [7][16]. Microsoft's Dragon Copilot and similar platforms now integrate seamlessly with major EHR systems, supporting multi-participant, multilingual documentation and offering up to 70% improvement in clinician work-life balance, with real-world deployments indicating 13–26 additional appointment slots per provider per month [24][25][26]. Studies on ambient scribe platforms report significant time savings, improved note completeness, and higher user satisfaction, though all emphasize the critical role of physician oversight to guarantee accuracy and clinical appropriateness [27][28].

4. System architecture

The overall framework is visualized in Fig. 1, illustrating LLM based natural language understanding for clinical documentation.

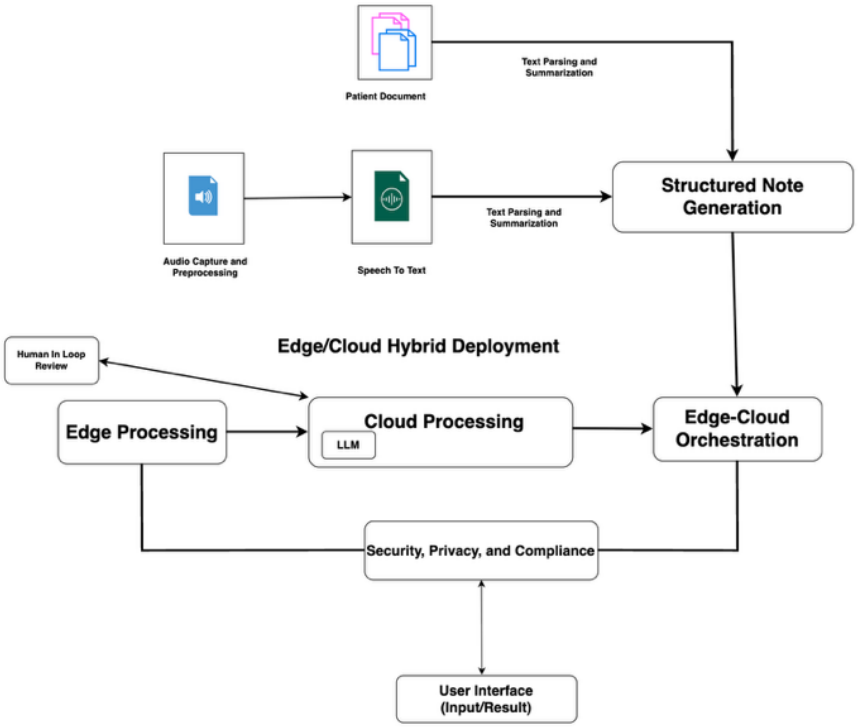

Figure 1. System Architecture of LLM-Based AI Copilot for Clinical Documentation on Edge/Cloud Hybrid Platforms.

4.1. LLM Based AI Copilot Overview

AI Copilot architectures for clinical documentation harness a combination of automatic speech recognition (ASR), LLM-based natural language understanding, and structured output generation. The overall pipeline can be logically decomposed into several modules:

- **Audio Capture and Preprocessing:** Captures real-time talks or short dictations, with clear sound and speaker identification [19][37].
- **System Speech-to-Text (STT) Engine:** Converts audio streams to textual transcripts. Advanced models such as Whisper v3 or Deep gram's Nova Medical use domain-specific fine-tuning to improve medical terminology recognition and reduce word error rates [22][38].
- **Text Parsing and Summarization:** Utilizes LLMs to extract salient clinical concepts, organize the conversation into structured note formats.
- **Structured Note Generation:** Outputs EHR-ready JSON, CDA, or FHIR-compliant representations for direct ingestion into clinical systems. Integration with EHR APIs, such as those specified by HL7-FHIR DocumentReference, enables seamless storage, retrieval, and audit [29][25].
- **Human-in-the-Loop Review:** Provides interfaces for physician review, editing, and sign-off. Importantly, clinical users retain final accountability for note approval, with AI copilots serving as decision support rather than autonomous agents [40][41].

4.2. Edge/Cloud Hybrid Deployment

Clinical environments demand both high performance and robust compliance. Edge/cloud hybrid deployment models are increasingly favored for AI Copilot systems, offering the following advantages:

- **Edge Processing**: Handles latency-sensitive operations such as real-time audio capture and transcription locally, ensuring minimal delays and high availability even during network disruptions. Sensitive data can remain within the institution, aligning with privacy requirements [42][43].
- **Cloud Processing**: Executes resource-intensive LLM operations (e.g., summarization, context resolution) in the cloud, leveraging scalable GPU resources and up-to-date models for improved accuracy and handling complex NLP tasks [42][44].
- **Edge-Cloud Orchestration:** Dynamic task allocation facilitates workload optimizationpushing low-latency tasks to the edge and leveraging cloud capabilities for large-scale analytics, benchmarking, and rapid iteration of AI models [43][45].
- **Security, Privacy, and Compliance**: End-to-end encryption, fine-grained access controls, and auditable logging are enforced across the hybrid infrastructure. Systems adhere to regulatory requirements (HIPAA, GDPR), with options for local PHI processing and secure, policy-governed cloud connectivity [46][47][48].

The overall model inference workflow on edge/cloud nodes is visualized in Fig. 2.

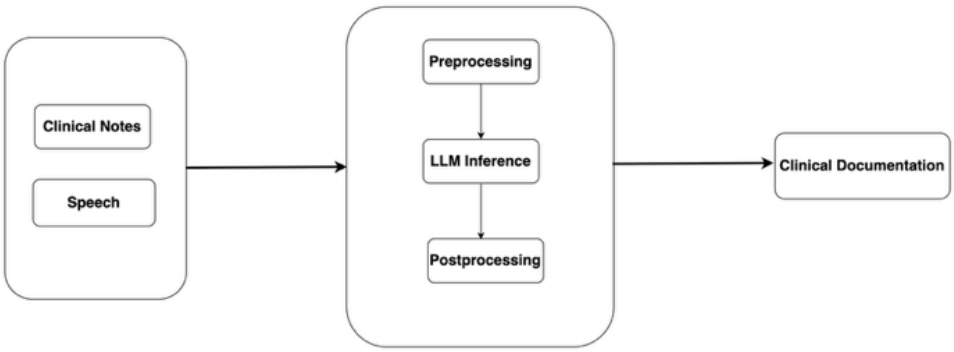

Figure 2. Model Inference Workflow on Edge/Cloud Nodes

4.3. Technical Infrastructure

- **Infrastructure Flexibility:** Support for hospital-owned on-premises edge nodes as well as multi-tenant, secure cloud resources, with detailed configuration of data residency [43][44][31].
- **Failover and Redundancy:** High availability through local processing fallback, automated reconnection, and disaster recovery planning [42][43].

5. Evaluation strategy

A robust evaluation framework is crucial for validating AI Copilot performance, usability, and safety.

5.1. Evaluation Frameworks

Recent consensus recommendations argue for a blended evaluation strategy [37][39][49]:

- Automated metrics should be used alongside human review to capture both quantitative performance and clinical relevance.
- Benchmarks must include both transcription (audio-to-text) and summarization (text-to-note) stages, with error analysis at each phase.
- Progressive and post-market monitoring, including bias evaluation and subgroup analysis, are required for safe deployment [53][54].
- Continuous feedback loops allow rapid identification and correction of systematic failure modes during deployment.

5.2. Performance Metrics

Table I provides an overview of the key evaluation metrics

Table 1. Evaluation Metrics

Metric	Description
Word Error Rate	Time elapsed from telemetry signal capture to fault classification
Note Consistency	Time required to validate and execute infrastructure fixes after fault detection
PDQI-9 Score	Clinician-based rating for note completeness and usability
Time Saved	Number of telemetry events processed per second per agent under high load
Error Rate	Tracks hallucinations, omissions, and incorrect clinical details.
Adoption Rate	Reflects real-world usage of the AI Copilot in clinical settings.

6. Technical considerations

The proposed modular fault remediation framework introduces transformative capabilities for distributed systems, yet several practical and theoretical limitations remain. These constraints warrant further investigation to support broader adoption and long-term sustainability.

6.1. Latency and Real-Time Performance

Temporal responsiveness is paramount AI Copilots must maintain sub-second (sub-500ms) latency between audio input and structured note output to avoid workflow disruptions [40]. GPU-accelerated ASR and LLM inference pipelines, coupled with streaming architectures, ensure real-time operation, especially critical in high-volume or urgent care settings [40][41].

- **Infrastructure Flexibility:** Support for hospital-owned on-premises edge nodes as well as multi-tenant, secure cloud resources, with detailed configuration of data residency [43][44][31].
- **Failover and Redundancy:** High availability through local processing fallback, automated reconnection, and disaster recovery planning [42][43]

6.2. Accuracy, Adaptation, and Specialty Context

- **Medical Terminology:** GPU-accelerated ASR and LLM inference pipelines, coupled with streaming architectures, ensure real-time operation, especially critical in high-volume or urgent care settings [40][41].
- **Contextual Reasoning:** Prompt engineering and section-specific modeling (e.g., K-SOAP, specialized LLM prompts) enhance consistency and factual accuracy [19][20][39].
- **Specialty Differentiation**: Adaptive model configurations or specialty-specific deployments are necessary to minimize errors in highly specialized fields (e.g., cardiology, oncology) [41][42].

6.3. Reliability and Resilience

- **Error Handling:** Robust fallback mechanisms reroute tasks to cloud or human reviewers on failure, with clear flagging of low-confidence outputs [41][43].
- **Customization and Learning:** Systematic mechanisms for user customization and model updating (e.g., template adjustments, feedback-driven model retraining) support sustained accuracy and physician trust [26][34].

6.4. Security, Privacy, and Compliance

- **PHI Safeguard:** Edge processing for de-identification, encrypted cloud transmission, and access controls are standard [42][48].
- **Auditability and Traceability:** Audit logs for all PHI-accessing operations, with version control and rollback [31][34].

7. Challenges and limitations

7.1. Workflow and Adoption Challenges

- Misalignment with clinical workflows slows implementation
- Clinician skepticism toward AI accuracy and reliability

- Risk of overreliance on automation may hinder critical judgment

7.2. Efficiency vs Incentives

- Efficiency gains may unintentionally increase patient volume
- Misaligned incentives can reduce time spent per patient
- Physician well-being may suffer without supportive redesign

7.3. Technical Barriers

- Speech recognition drops in noisy clinical settings
- Limited performance across medical specialties
- Possibility of hallucinations introduces clinical risk

7.4. Ethical & Equaity Concerns

- Transparency and patient consent must be continuously maintained
- Accountability for errors remains a legal and moral challenge
- Bias mitigation and fairness across populations is still unresolved

8. Conclusion

LLM-based AI Copilots represent a transformative intervention for reducing the documentation workload and associated burnout among physicians. Advances in speech recognition, NLP, and secure hybrid edge/cloud architectures have enabled real-world deployment of ambient AI scribes that demonstrably decrease errors, improve physician satisfaction, and open pathways to higher-value patient care. However, actualizing these benefits depends on careful system engineering, user-centered implementation, rigorous evaluation, and sustained alignment with privacy, ethical, and regulatory frameworks. A successful AI Copilot system is characterized by: sub-second, high-accuracy transcription; clinically-aware note generation; seamless and secure EHR integration; customizable, specialty-sensitive interfaces; robust human oversight; and transparent governance. The deployment of these systems requires addressing organizational, technical, and regulatory challenges, from workflow integration and cultural change to the standardization of evaluation metrics and continuous risk management. Future research is called to establish open, multi-institutional datasets supporting representative benchmarking across specialties and settings; to refine evaluation frameworks that blend automated and clinical user-centric metrics; and to develop policy and governance models that ensure both innovation and safety. Ultimately, AI copilots will advance best when they operate not as opaque black boxes, but as transparent, trustworthy partners in the collaborative enterprise of patient care.

References

[1] Gaffney A. et al., “Medical Documentation Burden Among US Office-Based Physicians in 2019,” JAMA Intern. Med., vol. 182, no. 5, 2022.
[2] Moy A. J., Schwartz J. M., Chen R.J. et al., "Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review", JAMIA, vol. 28, no. 5, pp. 998–1008, 2021.
[3] Rossetti S., et al., "AMIA Pulse Survey on Excessive Documentation Burden," Medical Economics, Jun 2024.
[4] Gallani S., Moura L., Sonnefeldt K., “Can AI Save Physicians from Burnout?” Harvard Business School Working Knowledge, Aug 13, 2024.
[5] Perkins S.W., et al., "Improving Clinical Documentation with Artificial Intelligence: A Systematic Review,” AHIMA, Vol. 21, Issue 2, Oct 2024.
[6] Govender T., "The Role of AI in Solving Physician Documentation Challenges," ClinIntell Blog, Feb 13, 2025.
[7] OpenAI, “Pioneering an AI clinical copilot with Penda Health," July 22, 2025.
[8] Microsoft, "Meet Microsoft Dragon Copilot: Your new AI assistant for clinical workflow," Microsoft Industry Blogs, Mar 3, 2025.
[9] Lekadir K., et al., “FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare,” BMJ 2025; 388:e081554.
[10] FSMB, “Navigating the Responsible and Ethical Incorporation of Artificial Intelligence into Clinical Practice,” FSMB House of Delegates, Apr 2024.
[11] Sinsky C., Colligan L., Li L., et al., “Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in Four Specialties," Ann Intern Med, vol. 165, no. 11, pp. 753-760, 2016.
[12] Arndt B.G., Beasley J.W., Watkinson M.D., et al., "Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations,” Ann Fam Med, vol. 15, no. 5, pp. 419-426, 2017.
[13] National Academy of Medicine, "Valid and Reliable Survey Instruments to Measure Burnout, Well-Being, and Other Work-Related Dimensions," NAM.edu..
[14] Moreno-Jiménez B., et al., “The Physician Burnout Questionnaire: A New Definition and Measure,” TPMap, 2014.
[15] Maslach C., Jackson S.E., “Maslach Burnout Inventory. Manual,” Consulting Psychologists Press, 1981.
[16] OpenAI, “Penda Health Case Study,” July 22, 2025.

[17] Demerouti E., Bakker A. B., Nachreiner F., Schaufeli W. B., "The job demands-resources model of burnout," J Appl Psychol, 2001.
[18] Perkins S.W., Muste J.C., Alam T., Singh R.P., “Improving Clinical Documentation with Artificial Intelligence: A Systematic Review,” AHIMA, 2024.
[19] Li Y., Wu S., Smith C., Lo T., Liu B., “Improving Clinical Note Generation from Complex Doctor-Patient Conversation,” arXiv:2408.14568, Jun 2025.
[20] Brake N., Schaaf T., "Comparing Two Model Designs for Clinical Note Generation; Is an LLM a Useful Evaluator of Consistency," NAACL 2024, pp. 352–363.
[21] Yu J.J., Xie R., Toma A., "Automated Clinical Note Generation from Doctor-Patient Conversations using Large Language Models," MEDIQA-Chat 2023 Results.
[22] Deepgram, “How Speech-to-Text Transformed Healthcare and Medical Transcription,” Jan 2025.
[23] Adedeji A., Joshi S., Doohan B., “The Sound of Healthcare: Improving Medical Transcription ASR Accuracy with Large Language Models,” arXiv:2402.07658, Feb 2024.
[24] HealthTech Magazine, “Helpful Tips for Hospitals When Implementing Microsoft Dragon Copilot,” Jul 17, 2025.
[25] HL7 FHIR, "US Core Implementation Guide v8.0.0: Clinical Notes."
[26] Microsoft Learn, “Use unstructured clinical notes enrichment (preview) in healthcare data solutions,” Nov 2024.
[27] BMC Med Inform Decis Mak, "Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review," Jul 2025.
[28] Simbo AI, "Addressing Concerns: Evaluating the Limitations and Challenges Associated with AI-Facilitated Clinical Documentation," Feb 2025.
[29] HL7 Confluence, "Clinical Notes – Structured Documents," 2024.
[30] Iodine Software, "Clinical Documentation Improvement Challenges: Understanding the Obstacles," Jun 5, 2025.
[31] Microsoft, "Copilot for Security Now Covered by HIPAA BAA," Aug 15, 2024.
[32] Nightfall AI, “Is Microsoft Copilot HIPAA Compliant?” 2025.
[33] NHS England, “Guidance on the use of AI-enabled ambient scribing products in health and care settings,” Apr 27, 2025.
[34] Lekadir, K. et al., “FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare,” BMJ 2025.
[35] Helmholtz Munich, “International Experts Establish FUTURE-AI Guidelines for Trustworthy Healthcare AI,” 2025.
[36] Future-AI, “Best practices for trustworthy AI in medicine,” 2025.
[37] Gebauer S., et al., “Benchmarking And Datasets For Ambient Clinical Documentation: A Scoping Review of Existing Frameworks And Metrics For AI-Assisted Medical Note Generation,” medRxiv, Jan 2025.
[38] BMC Med Inform Decis Mak, "Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review," Jul 2025.
[39] Yim W., Fu Y., Abacha A.B., Snider N., Lin T., Yetisgen M., "ACI-BENCH: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation," Microsoft Research, Sep 2023; GitHub Project.
[40] Simplismart, “Agentic AI Medical Scribe Stack for Sub-Second Latency," Jun 16, 2025.
[41] Simbo AI, "A Comprehensive Look at Technical Requirements for Deploying AI Medical Scribes in Modern Healthcare Environments," 2025.
[42] RapidAI, "Edge Cloud," 2025.
[43] HealthTech Magazine, "Maximizing AI Deployment Value in Healthcare Requires a Hybrid Edge-to-Cloud Strategy," Jun 6, 2024.
[44] Forbes, “Powering Possibilities In Healthcare With AI And Edge Computing,” Jul 23, 2025.
[45] Yim W., et al., "Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation," Nature Scientific Data, 2023; GitHub Repository.
[46] AHIMA, "Ethical Standards for Clinical Documentation Integrity (CDI) Professionals," 2020.
[47] FSMB, “Navigating the Responsible and Ethical Incorporation of Artificial Intelligence into Clinical Practice,” FSMB House of Delegates, Apr 2024.
[48] JAMA, “Ethical Obligations to Inform Patients About Use of AI Tools,” 2025.
[49] Gebauer S., et al., “Benchmarking And Datasets For Ambient Clinical Documentation: A Scoping Review,” medRxiv, Jan 2025.
[50] Yim W., et al., "ACI-BENCH," GitHub Repository, 2023.
[51] Microsoft Research, "ACI-BENCH: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation," Sep 2023.
[52] Future-AI, “Home best practices,” 2025.
[53] Lekadir, K. et al., “FUTURE-AI: international consensus guideline,” BMJ 2025.
[54] Helmholtz Munich, “International Experts Establish FUTURE-AI Guidelines,” 2025.
[55] IMO Health, "The future of clinical documentation is ambient, automated, and AI-powered," May 20, 2025.
[56] Tebra, “The future of clinical documentation: How AI-generated medical notes are transforming patient care,” Jun 1, 2025.
[57] Juno Health, “AI and the Future of EHRs: Beyond Documentation to Clinical Decision Support,” May 6, 2025.