AI Medical Scribes: A Revolution in Healthcare Documentation

Maxime C. Cohen and Elie M. Toubiana
July 31, 2024

Executive Summary. Over the last few decades, electronic health records (EHRs) have significantly altered physicians' workflows. Although promising improved patient care, EHRs have imposed a substantial documentation burden, with studies showing that physicians spend nearly half their time on administrative tasks, contributing to stress and burnout. AI solutions, particularly AI medical scribes, offer a way to alleviate this burden by reducing the time spent on documentation and allowing physicians to focus more on patient care. This article discusses the current state of AI medical scribes, including a pilot study conducted in a large hospital to evaluate their impact on efficiency and quality of care.

1. Introduction

The healthcare industry has long grappled with the administrative burden placed on physicians, particularly when it comes to documenting patient encounters and maintaining accurate medical records (Momenipour and Pennathur, 2019). In an effort to reduce the administrative burden on clinicians, the concept of medical scribes was introduced. Traditionally, (human) scribes—often assistants or medical students with sufficient medical knowledge—have been employed to document patient encounters. However, this solution has its limitations, as it requires that scribes have a deep understanding of medical terminology and nuances. An important part of healthcare documentation is that clinicians write a summary consultation note into the EHR after each patient visit. However, remembering the details from each patient encounter, when physicians often see more than 25 patients a day, is both stressful and challenging for clinicians.[1]

Different approaches exist: while some physicians hire a human scribe whose presence is required during the consultation, others prefer to write the note themselves (e.g., by typing it into the EHR or by writing it by hand). Another possibility is to use speech recognition (SR) software that connects with the EHR interface. Studies have shown significant improvements in report turnaround time (RTT) with the use of SR systems. For example, in radiology, using SR reduced RTT by 81% (Johnson et al., 2014). Similarly, the average RTT in surgical pathology was reduced from four days to three days, with the proportion of reports completed within one day increasing from 22% to 36% (Johnson et al., 2014). Some physicians like to write each note at the end of each consultation, whereas others prefer to batch several notes (e.g., spending time in

[1] https://x.com/EasyOrtho1/status/1397234841645686795; https://forums.studentdoctor.net/threads/note-anxiety-writing-notes-in-busy-clinic-is-driving-me-crazy-any-recommendations.1325096/

the evening to write the notes for all the patients seen on that day). In each case, physicians must carefully remember important medical details, such as symptoms, medical history, and allergies.

Regardless of the method used to write the consultation note, writing it takes several precious minutes. Many studies have shown that physicians spend as much time on documentation as they do on patient interaction (see, e.g., Ammenwerth and Spötl, 2009; Momenipour and Pennathur, 2019), and anecdotal evidence suggests that physicians spend between three and six minutes writing a consultation note. Physicians are not enthusiastic about this part of their job, since it is perceived as a low-added-value task that can induce fatigue and burnout. In addition, when writing notes, physicians are not giving full attention to the patient.

As a further complication, many countries are dealing with a severe shortage of physicians. A U.S. study forecasts a national deficit of 139,160 physician jobs by 2030, with significant regional disparities exacerbating this issue (Zhang et al., 2020). This shortage amplifies the need for efficient solutions to alleviate the administrative burden on existing healthcare professionals and allow physicians to be more efficient by spending less time on documentation.

Recent advancements in generative AI and large language models (LLMs) have introduced a new solution called AI medical scribes, which promises to shorten the time spent on documentation. These AI-powered tools can record, transcribe, and summarize conversations between physicians and patients in nearly real time. The task of summarization is a great example where AI tools perform well (Liu et al., 2023; Zhang et al., 2024). AI medical scribes can automatically generate a consultation summary note. This AI-generated note is then presented to the physician, who can make the final touches by editing specific parts of the note. By automating this task, AI medical scribes can dramatically reduce the time that clinicians spend on documentation, thereby allowing them to reallocate their time to what they do best—namely, seeing and diagnosing patients. AI medical scribes have the potential to save precious minutes for each patient encounter, thereby potentially enhancing access to the healthcare system. They can also reduce clinicians' fatigue and provide a better experience for patients. Adopting AI medical scribes is a great example of how AI can serve as a copilot to complement humans and gain efficiency. At the same time, several challenges and pitfalls are also in order.

2. What is an AI Medical Scribe?

As discussed, the traditional scribe model involves trained professionals who accompany physicians during patient consultations, either in person or virtually, to document patient encounters. Scribes meticulously record patient histories, physical examinations, and treatment plans, allowing physicians to focus on patient care rather than administrative tasks. This model, while effective, is labor intensive, costly, and subject to human error and variability. According to a recent study, traditional scribes significantly improve clinical workflow and patient

satisfaction, demonstrating their safety and efficacy in practice (Gidwani et al., 2017). However, at the same time, the scribe's presence can also be intrusive during sensitive patient interactions. This is where AI medical scribes can be game changers.

AI medical scribes are built in a three-step process. First, the conversation between the healthcare provider and the patient is recorded using an ambient microphone, which can be the microphone on the physician's phone, tablet, or computer, or a special microphone worn by the physician. Recording the conversation is obviously a sensitive topic and requires formal patient consent in most jurisdictions. Most AI medical scribes have the ability to delete the conversation on the fly, once the summary is generated. Second, this audio recording is sent to a speech-to-text (STT) software, such as Microsoft Azure, Google AI, and Amazon Transcribe, which converts the spoken words into written text. Third, this text is processed by an LLM (e.g., GPT-4, Llama 3, Claude), which summarizes and organizes the data into a specific format suitable for medical documentation, while aiming to ensure that the information is accurately and coherently captured for the healthcare provider's use. An illustration of the three steps is provided in Figure 1.

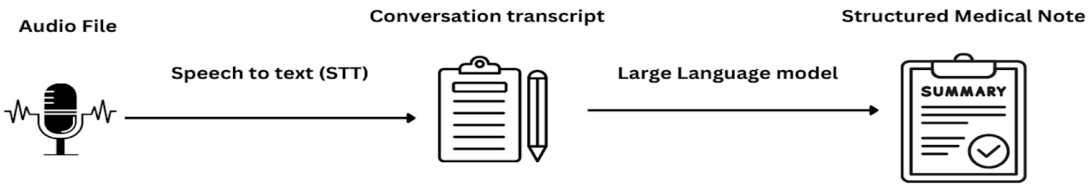

Figure 1. Illustration of the three steps involved in an AI medical scribe.

In summary, AI scribes rely on two technologies to generate a note: an STT tool to convert the audio of the conversation to text and an LLM to convert that text into a structured medical note. LLMs, such as OpenAI's GPT-4 and Meta's Llama 3, are trained on vast datasets, enabling them to understand and generate human-like text. These models can often process and interpret the nuances of medical terminology, context, and patient interactions to generate accurate and coherent clinical notes. LLMs can be further fine-tuned for specialized tasks, such as specifically summarizing medical content. This process involves training the model on examples of conversations and their corresponding medical notes, thereby enhancing the model's proficiency specifically for this task. This process of fine tuning is proprietary to each AI medical scribe provider and can substantially affect the quality and accuracy of the output.

STT technology, by contrast, converts spoken language into written text. Companies like Microsoft Azure, Google Cloud, and OpenAI Whisper,[2] have developed sophisticated STT systems that can handle various accents, dialects, and even medical jargon with high accuracy. When integrated, STT systems transcribe the spoken words of physicians and patients in real time, while LLMs structure these transcriptions into formatted clinical notes, trying to ensure

[2] https://cdn.openai.com/papers/whisper.pdf

they are accurate, comprehensive, and compliant with medical standards. A common metric used to evaluate the accuracy of STT systems is the word error rate (WER), which measures the percentage of errors in the transcribed text compared to the original spoken content. The WER is calculated by summing up the total number of substitutions (incorrect words), deletions (missing words), and insertions (extra words), and then dividing it by the total number of words in the reference text. The formula can be written as follows:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Words in Reference}}$$

For example, if the reference text has 100 words and there are five substitutions, three deletions, and two insertions in the transcription, then the WER would equal 10%. A lower WER indicates a higher accuracy, with 0% corresponding to a perfect transcription. The WER helps compare the performance of different STT systems and identify areas for improvement in the transcription process. The current state of the WER is around 7%, as tested on large samples of open data,[3] but this may vary depending on the spoken accent and the specific setting (e.g., the background noise and the microphone quality may affect the WER). While most AI scribes can operate in multiple languages, the WER can vary quite dramatically from one language to another.[4]

3. Models, Existing Providers, and Evaluation Metrics

The business models for AI scribe services typically involve a subscription-based approach, in which healthcare providers need to pay a monthly fee to access the technology. This fee often varies based on the size of the practice, the volume of documentation generated, and the level of customization required.

While many providers are sharing the AI medical scribe market, this technology may soon become a commodity, especially given the fragmented nature of the U.S. market with over 500 EHR systems. To avoid commoditization, providers must focus on unique features, superior user experience, customization to specific needs, and robust integration capabilities. The complexity of integrating with numerous EHRs requires investment in interoperability standards like HL7 FHIR, strategic partnerships with EHR vendors, and robust API development. Customization and scalability are crucial, as are offering and constantly developing value-added services such as advanced analytics and compliance reporting. Addressing these integration challenges through comprehensive support, user training, and a direct collaboration with healthcare providers can enhance the appeal of AI scribes, ensuring they meet diverse needs and improve overall efficiency and care quality.

[3] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard
[4] https://cdn.openai.com/papers/whisper.pdf

While providing a comprehensive list of all the providers in this space is beyond the scope of this article, competing service providers that offer AI medical scribe services include the following: Ambience, Augmedix Healthcare, AutoScribe, Brios.ai, Chartx.ai, ChatLabs AI, DeepCura, DeepScribe, Freed, Heidi, InteliDoc-AI, Lindy, MedScribe, Mikata, Nabla, NoteMD, Nuance (DAX), Saykara, Scribeberry, ScribeEMR, ScribeMD.ai,[5] Suki, Tali AI, and Zirr AI.

These providers typically offer two types of pricing models: a per-provider monthly fee (often between $79 and $700) and a per-visit fee (between $0.20 and $8). Most providers offer a variety of services based on price, integration capabilities, and deployment environments. For example, HCA Healthcare, Inc., in their partnership with Augmedix, included a data-sharing program that allowed them to run analytics on the data captured by Augmedix.[6]

To evaluate and assess the quality and impact of using an AI medical scribe, several performance metrics should be considered. We next discuss three types of metrics: physician documentation quality instrument-9 (PDQI-9), patient satisfaction, and clinician satisfaction.

PDQI-9 is a standardized tool used to assess the quality of clinical documentation in healthcare settings (Stetson et al., 2012). Specifically, it is a nine-item questionnaire designed to evaluate various aspects of physician documentation to ensure that it meets high quality standards. Each item is rated on a five-point Likert scale, ranging from "strongly disagree" to "strongly agree." The PDQI-9 assesses the following key quality dimensions to verify that the note is:

1. **Up-to-date:** Evaluates whether the note includes the most recent test results and recommendations. An up-to-date note ensures that all information reflects the current status of the patient, facilitating appropriate clinical decisions and care.
2. **Accurate:** Assesses the factual correctness of the note. An accurate note contains error-free information, providing a reliable account of the patient's condition, history, and treatment.
3. **Thorough:** Measures the completeness of the note. A thorough note documents all significant issues related to the patient, covering every aspect necessary for comprehensive patient care.
4. **Useful:** Evaluates the relevance and value of the information provided. A useful note offers critical information and analysis that are essential for patient care and decision making.
5. **Organized:** Assesses the structure and formation of the note. An organized note is well formed and structured in a way that aids the reader in understanding the patient's clinical course and necessary actions.

[5] The second author of this article is the founder and CEO of ScribeMD.ai, and the first author is an advisor.

[6] https://www.googlecloudcommunity.com/gc/Community-Blogs/The-future-of-clinical-data-building-an-intelligence-engine-with/ba-p/753668

6. **Comprehensible:** Evaluates the clarity of the note. A comprehensible note is clear, without ambiguity, and easy to understand, ensuring that all sections are straightforward and logical.
7. **Succinct:** Measures the brevity and conciseness of the note. A succinct note is brief, to the point, and free from redundancy, providing necessary information without unnecessary details.
8. **Synthesized:** Evaluates the integration and combination of information from various sources. A synthesized note pulls together data from different healthcare providers, test results, and previous records to present a comprehensive view of the patient's health, aiding in accurate diagnosis and effective treatment planning.
9. **Internally Consistent:** Assesses the internal coherence of the note. An internally consistent note does not contain contradictions and ensures that all parts of the note are in agreement, presenting a cohesive narrative of the patient's condition and care.

PDQI-9 is commonly used in research and clinical practice to improve the quality of physician documentation, enhance communication among healthcare providers, and ultimately improve patient care outcomes. One can, of course, also consider additional metrics, such as the lag time (i.e., the time it takes between the end of the patient encounter and generation of the note) and the editing time (i.e., the time it takes to edit and finalize the AI-generated note).

The second type of metric aims to capture patient satisfaction. While this is naturally subjective and highly dependent on the context, a common way to evaluate patient satisfaction is by conducting surveys or interviews with the patients after the consultation. Here are examples of questions that can be included in such surveys:

- Did you feel that the doctor spent enough time with you?
- Did the doctor maintain eye contact and actively listen to your concerns?
- Were you able to discuss all of your health concerns without feeling rushed?
- On a 1-to-5 scale, how satisfied are you with your visit?

Finally, the third type of metric involves clinician satisfaction. Ultimately, the adoption of new technologies depends on the satisfaction of its end-users, which in the case of AI medical scribes are the clinicians. This aspect can be measured both before and after using the AI medical scribe to determine whether a significant improvement can be measured in day-to-day life. This evaluation can include both quantitative measurements, such as the average time spent per patient visit, the average time spent writing a summary consultation note, and the quality of the note (e.g., the WER). It can also include qualitative survey questions, such as the following:

- How often do you experience burnout or fatigue related to documentation tasks?
- How much time do you typically spend on documentation per patient encounter?
- Do you feel that the AI scribe has allowed you to spend more time focusing on your patients?

- Do you feel that using the AI scribe helped you save time on documentation? If yes, how many minutes?
- Was the quality of the AI-generated note lower, comparable, or higher than your standards?
- Do you have any concerns about integrating an AI scribe into your practice?

Evaluating the performance of an AI medical scribe by considering PDQI-9, patient satisfaction, and clinician satisfaction would allow a holistic evaluation before considering a large-scale deployment. A more rigorous assessment would be for each patient encounter to include the use of both the AI scribe and the regular business-as-usual scribe concurrently. Then, one can compare the quality and accuracy of both notes by computing the WER and checking whether the AI-generated notes include omissions or errors.

4. Integration with Healthcare Systems

A successful implementation of AI medical scribes hinges on their seamless integration with existing healthcare systems, particularly with EHRs. Ensuring compatibility, developing effective workflow integration strategies, and addressing data security, privacy, and compliance considerations are critical components of this process.

Compatibility with EHRs

The EHR U.S. market is highly fragmented and is characterized by a diverse array of vendors and systems that often struggle to communicate effectively. This fragmentation arises from several factors, including the rapid adoption of EHRs following federal incentive programs, the varying needs of different healthcare providers and specialties, and the lack of standardized interoperability protocols. Large healthcare systems tend to use EHRs from market-leading developers like Epic and Cerner, whereas smaller practices and independent physicians often opt for different smaller vendors or custom solutions. This disparity has led to significant challenges in data exchange and integration, particularly for small practices that may lack the resources to implement more advanced interoperability features. The fragmented nature of the market has also contributed to a divide in healthcare quality and efficiency, as patients treated by large, integrated practices benefit from more seamless information sharing, while those seen by smaller practices may experience gaps in care coordination.

AI medical scribes must be compatible with a wide range of EHR systems to be adopted and effective. This compatibility would ensure that AI-generated clinical notes can be effortlessly integrated into existing digital records, thereby preserving the continuity and accessibility of patient information. Many AI scribing solutions are designed with interoperability in mind, allowing them to interface with popular EHRs, such as Epic, Cerner, and Allscripts (in the U.S.

market, for example). This interoperability is crucial for maintaining the efficiency of clinical workflows and minimizing disruptions during the transition to AI-assisted documentation.

Workflow integration strategies

Integrating AI medical scribes into clinical workflows requires careful planning and execution. Successful integration strategies involve the following key steps:

1. **Assessment and customization:** Each healthcare facility must assess its needs and workflow patterns and customize the AI scribe solution accordingly. This ensures that the AI system is complementing the current workflow rather than imposing an additional burden.
2. **Training and support:** Physicians and staff need adequate training to effectively use AI scribes. Ongoing technical support is also essential to address any issues that arise and to continually optimize the system's performance.
3. **Phased implementation:** A phased approach to implementation allows for gradual adaptation and improvement. Starting with a small pilot program can help identify potential challenges and areas for improvement before a full-scale rollout.
4. **Continuous feedback and improvement:** Regular user feedback can help refine the AI system, making it more intuitive and responsive to the needs of healthcare providers.

Data security, privacy, and compliance

The integration of AI medical scribes into healthcare systems raises critical considerations regarding data security, privacy, and compliance. Ensuring that these AI systems comply with regulatory standards such as the Health Insurance Portability and Accountability Act (HIPAA) is paramount. At a high level, the following are four important considerations:

1. **Data encryption and access control:** AI scribes must employ robust encryption methods to protect patient data both in transit and at rest. Strict access controls are necessary to ensure that only authorized personnel can access sensitive information.
2. **Regular security audits:** Regular security audits can help identify and mitigate potential vulnerabilities in the AI system. These audits should be part of a comprehensive cybersecurity strategy.
3. **Compliance with regulatory standards:** AI scribes must adhere to all relevant regulatory standards to ensure that patient data is handled in a way that is both secure and compliant. This includes HIPAA, as well as any local or international regulations that may apply in each jurisdiction.
4. **Patient consent and transparency:** Clear communication with patients about the use of AI in their care journey, along with receiving informed consent, is crucial for maintaining trust and compliance with legal requirements.

In summary, the integration of AI medical scribes with healthcare systems involves ensuring EHR compatibility, developing effective workflow integration strategies, and addressing data security, privacy, and compliance considerations. By carefully managing these aspects, healthcare providers can leverage AI scribes to enhance clinical documentation, improve workflow efficiency, and maintain the highest standards of patient care and data protection.

5. Benefits of AI Medical Scribes

Enhanced efficiency and productivity

The integration of AI medical scribes can significantly enhance efficiency and productivity within clinical settings. A recent study conducted by the Permanente Medical Group (Tierney et al., 2024) provided an AI scribe to 3,442 physicians who used it in 303,266 patient encounters during a 10-week period. This study demonstrated the capacity of AI scribes to handle high volumes of clinical interactions effectively and a notable reduction in the time physicians spent on clinical documentation. By transcribing real-time conversations between clinicians and patients, the technology minimizes the need for manual labor. This streamlined process allows physicians to focus more on direct patient care, thereby increasing their overall productivity.

Reduced clinician burnout and improved satisfaction

AI medical scribes can play a crucial role in alleviating clinician burnout and enhancing job satisfaction. The burden of extensive clinical documentation is a well-documented contributor to physician burnout, often leading to dissatisfaction and decreased quality of life. The aforementioned study highlighted that physicians using ambient AI scribes experienced a significant reduction in after-hours documentation, commonly referred to as “pajama time” (Saag et al., 2019). This reduction in clerical workload allows physicians to maintain a better work–life balance and can reduce the stress associated with excessive documentation. Feedback from physicians indicated a favorable response to AI scribe technology, with many clinicians noting improved patient interactions and a more enjoyable clinical experience due to the reduced need for manual note-taking (Tierney et al., 2024).

Financial impact and cost savings

The financial benefits of using AI medical scribes extend beyond direct cost savings to include long-term economic advantages for healthcare organizations. By enhancing physician efficiency and reducing the time spent on documentation, healthcare providers can potentially increase patient throughput and optimize resource utilization. Adopting AI scribes can potentially reduce the time spent working with EHRs, both during and outside patient encounters. This reduction translates into more efficient use of clinician time, allowing for more patient appointments and

potentially higher revenue. Additionally, the reduction in clinician burnout can lead to lower turnover rates and decreased recruitment and training costs for healthcare organizations.

Improved documentation accuracy and patient care quality

If fine-tuned properly, AI medical scribes can lead to improved documentation accuracy and, consequently, an enhanced quality of patient care. The technology's ability to transcribe conversations and summarize key clinical content ensures that important details are captured accurately and efficiently. Tierney et al. (2024) found that AI-generated clinical documentation maintained high quality, with an average score of 48 out of 50 in key domains such as synthesis, internal consistency, and succinctness. Moreover, early assessments indicated that patients experienced better interactions with their physicians when AI scribes were used, as they reported increased time spent communicating directly with their doctors and less time with physicians looking at computer screens. This improvement in the physician–patient relationship is crucial for effective and empathetic patient care.

In summary, the integration of AI medical scribes offers numerous benefits, including enhanced efficiency and productivity, reduced clinician burnout, positive financial impact, improved documentation accuracy, and patient care quality. These advantages underscore the potential of AI technology to transform clinical workflows and improve healthcare delivery.

6. Challenges and Pitfalls

AI medical scribes offer promising advancements in healthcare documentation and allow healthcare professionals to focus more on patient care. However, several pitfalls must be addressed to ensure their effective and safe implementation.

Privacy concerns

One of the primary concerns with AI medical scribes is the privacy and security of patient data. Medical records contain highly sensitive information, and any breach or misuse can have severe consequences for patients and healthcare providers alike. Ensuring that AI systems follow stringent data protection regulations, such as HIPAA in the U.S. or PIPEDA in Canada, is crucial. As mentioned, advanced encryption methods and secure data handling protocols must be in place to safeguard patient information.

Conversations between doctors and patients also often include highly sensitive information. Most AI medical scribe systems have to self-host an LLM or have a direct relationship with the LLM host to ensure that the data is not used for training purposes. Some language models are designed to be small enough to run on devices, which can reduce privacy risks by keeping the data processing step local and avoiding external exposure.

Furthermore, AI medical scribes typically operate on private servers, with many providers offering a "bring your own cloud" model. This approach allows healthcare organizations to host an instance of the AI medical scribe on their own servers, thereby ensuring that patient data remains within their controlled environment. This setup aligns with hospitals' practices for in-house EHR systems, emphasizing data security and compliance with regulations.

Training and support for healthcare professionals

Effective training and ongoing support are crucial for a successful implementation of AI medical scribes. One can consider a multifaceted training approach, including live and recorded webinars, to educate physicians on the use of AI scribes. The training should emphasize best practices for safe and effective use, as well as the importance of receiving patient consent.

In addition to initial training, continuous support is necessary to address technical issues and challenges that arise during regular use. Providing at-the-elbow support from technology experts and maintaining internal support sites can help ensure that healthcare professionals feel confident in using AI scribes. Including physicians' feedback is also important for constant improvement and fine-tuning of the tool (e.g., physicians can request specific formatting preferences for the consultation notes and changes to the user interface).

In conclusion, while AI medical scribes offer numerous benefits, addressing technical challenges, ethical considerations, barriers to adoption, and the need for training and support are critical to realizing their full potential. By proactively addressing these challenges, healthcare organizations can enhance the effectiveness of AI scribes and ensure their successful integration into clinical practice.

Hallucinations

Another critical issue is the potential for AI hallucinations, where the AI may generate incorrect or fabricated information that still seems very credible (Xu et al., 2024). These inaccuracies can arise from various factors, including biases in training data, limitations of natural language processing algorithms, or misunderstandings of medical terminology. Hallucinations in medical documentation can lead to serious consequences, such as misdiagnoses or inappropriate treatment plans, ultimately compromising patient safety. Continuous monitoring and validation of AI outputs by trained professionals are necessary to mitigate these risks. Since the AI-generated summary notes are presented to physicians only as a recommendation, physicians need to carefully check them (and edit them as they see fit) before approving them.

Liability issues

Liability in the context of AI medical scribes presents a complex challenge. Determining who is responsible when the AI system makes an error is not straightforward. Is it the healthcare

provider who relied on the AI-generated notes, the software developer who created the AI system, or the institution that implemented it? While the answer to this question is beyond the scope of this article, clear guidelines and frameworks need to be established to address such liability issues. Moreover, healthcare professionals should be adequately trained to understand the limitations of AI systems.

Technical challenges and areas for improvement

Despite the promising potential of AI medical scribes, several technical challenges persist that need to be carefully addressed. As discussed, a major technical hurdle is the integration of AI scribes directly into EHRs. Tierney et al., 2024 noted that the lack of direct EHR integration resulted in workflow inefficiencies, as physicians had to manually input AI-generated summaries into patient records. Such inefficiencies amplify physicians' burdens and reduce the likelihood that they will adopt these technologies. Future enhancements should focus on seamless EHR integration to further streamline documentation processes.

A clear area for improvement is the AI's capability to handle diverse clinical scenarios accurately (e.g., short vs. long consultation, new vs. recurring patients, chronic vs. acute disease, inpatient vs. outpatient settings). Instances of hallucination, where the AI scribe generates incorrect or misleading information, are likely to be observed when adopting a new generative AI technology. Thus, continuous refinements and fine-tuning of AI algorithms and natural language processing techniques are essential to minimize such errors and improve the overall reliability of AI-generated documentation.

Environmental impact

The training and inference of LLMs used in AI medical scribes may contribute to greenhouse gas emissions. The computational power required for these processes consumes a large amount of energy, raising concerns regarding their environmental sustainability. As the healthcare industry adopts AI technologies, it is essential to consider and mitigate their environmental impact through the use of renewable energy sources and more efficient algorithms. Recently, several smaller custom AI models have been developed to mitigate this issue.

Equity across languages and accents

AI medical scribes must also address issues of equity in their deployment. LLMs often perform better with certain languages and accents, reflecting biases present in the training data. This can lead to disparities in the quality of service provided to patients from diverse linguistic backgrounds. Ensuring that AI systems are trained on diverse datasets and can accurately interpret and transcribe a wide range of languages and accents is critical to providing equitable

healthcare solutions. Failure to address these biases could exacerbate existing health disparities and undermine the inclusivity of AI medical technologies.

7. Pilot Study

In this section, we present a pilot study conducted by ScribeMD.ai. Our goal is to evaluate the performance and potential impact of integrating an AI medical scribe into the existing workflow of clinicians.

Context and pilot design

The study was conducted across multiple sites in emergency departments of a large network hospital in South Africa. The pilot was deployed for three weeks, from July 4 to July 24, 2024. Prior to starting the pilot study, we conducted interviews with several physicians, who clearly reported high levels of stress (and even feelings of burnout) with regard to documentation, and especially regarding the requirement of submitting summary consultation notes in a timely manner. The pilot study aimed to determine whether AI medical scribes could alleviate the documentation burden on healthcare providers, reduce fatigue, and ultimately improve the efficiency of medical documentation without sacrificing quality and accuracy. To assess the impact, we utilized both quantitative metrics and qualitative assessments.

The design of the pilot study was as follows: We recruited 41 emergency room (ER) physicians working in eight emergency departments from a large hospital network in South Africa. Surveys were administered before the pilot to collect data on their current documentation practices and challenges, as well as their perceptions of AI scribes. Examples of questions from this survey included: *"How much time do you typically spend on documentation per patient encounter?"* and "*How often do you experience burnout or fatigue related to documentation tasks?"*

The pilot study included 2,150 patient visits for which the AI scribe was used. Out of these 2,150 visits, we randomly selected 50 for conducting further extensive analyses.[7] The average consultation length was 4.3 minutes for both the full sample and the random subsample (this number is typical for ER patients and can be much higher in different healthcare settings). The physicians were asked to use an AI scribe for all their patient visits. Each patient visit was one observation that includes an audio file of the consultation recording.

All 50 consultation recordings were manually scrubbed of any personal identifiable information (PII) by removing parts where sensitive personal information could have been linked to a specific patient, such as name, home address, and date of birth. Prior to the study, formal consent

[7] The reason that we only considered a small random subset of 50 visits is because the subsequent analyses are both costly and time consuming. Repeating the analyses using a larger sample of visits is left as future research.

was collected from the participating physicians. At the end of the study, we compared the notes generated by the AI scribe relative to the ones generated by three human scribes.[8]

The audio consultation was recorded using an iPad 9th generation, and in some cases using an external Bluetooth microphone device to capture the audio with a higher quality. The physicians were provided training sessions on how to use the AI scribing application and on where to place the iPad to best capture the audio data. The LLM used to generate the AI medical note was Anthropic's Claude 3.5 Sonnet hosted on Amazon Web Services. To obtain the best results and meet physicians' expectations, the AI model used custom prompting strategies to generate each section of the medical note. For example, the AI-generated note is structured into several sections, including chief complaint, past medical history, family history, and allergies (see an illustration in Figures 2 and 3).

We recruited three experienced human medical scribes to generate medical notes for the 50 randomly selected patient visits. Subsequently, two additional experienced human scribes were enlisted to evaluate the quality of the notes produced by the three recruited human scribes as well as the AI-generated notes (we thus hired a total of five different human medical scribes). The scribes were provided with audio recordings of the visits, which had been sanitized of PII. These exact same recordings were used as input to the AI system to ensure fair comparisons. The human scribes had no control over the placement of the microphone during the recordings, ensuring that all notes were produced asynchronously under consistent audio conditions. Unlike in a real-world synchronous situation, where a scribe can move around to ensure they hear the conversation well or ask the patient or physician to repeat a specific piece of information that was misheard, our experimental setting did not allow for such adjustments.

The two additional experienced scribes were provided with blinded data, which included the audio recordings, the notes produced by the three human scribes, and the AI-generated notes. These scribes were instructed to evaluate and grade each note based on the PDQI-9 criteria, assigning a score between 1 and 5 for each dimension by comparing the audio recording to the actual note. We then computed the average of both evaluators. The PDQI-9 criteria, which were thoroughly explained to the evaluators, encompass various aspects of note quality, hence ensuring a comprehensive assessment. We next discuss the results across three assessments: quality of the note (PDQI-9), time to generate the note, and qualitative clinical surveys.

Quality of the note

Interestingly, the AI scribe had a comparable performance to that of the three human scribes on most PDQI-9 criteria (and even outperformed it in some dimensions). A summary of the score comparisons can be found in Figure 2, with further details below:

[8] An additional interesting (and practically relevant) comparison we did not perform involves considering the notes generated by the AI scribe with human edits.

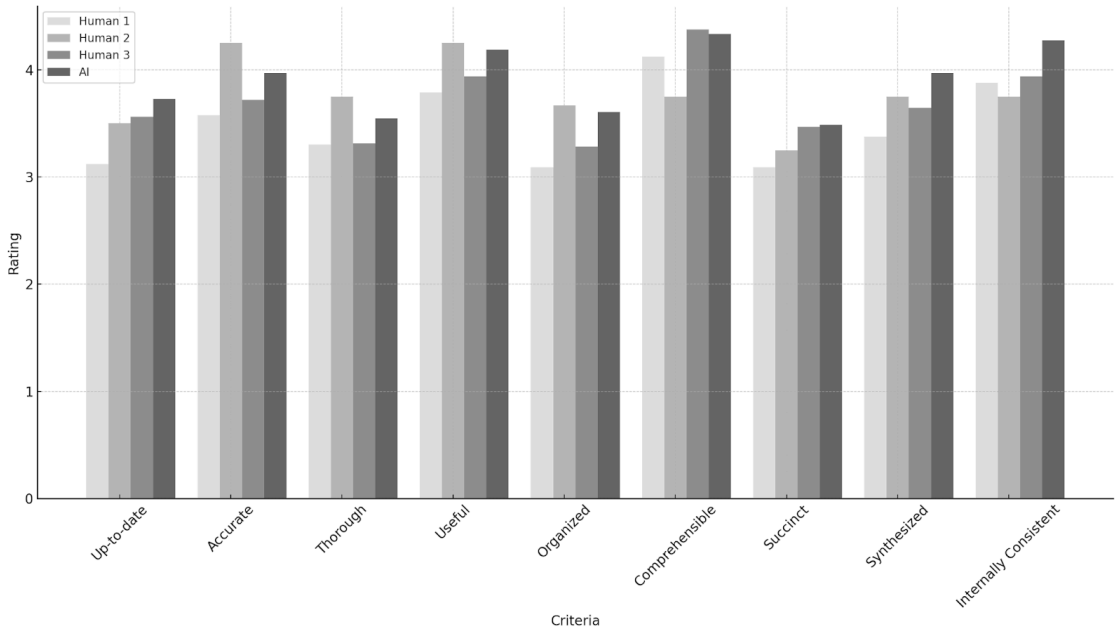

Figure 2. Comparison of note quality between human scribes and an AI scribe.

- **Up-to-date:** The AI received a score of 3.73, which is higher than all three human medical scribes (3.12, 3.50, and 3.56).
- **Accurate:** The AI scored 3.97, outperforming two human scribes (3.58 and 3.72) but lower than Human 2 (4.25).
- **Thorough:** The AI achieved a score of 3.55, surpassing Human 1 (3.30) and Human 3 (3.31), while being lower than Human 2 (3.75).
- **Useful:** The AI's score of 4.19 was higher than all Humans 1 and 3 (3.79 and 3.94) and a bit lower than Human 2 (4.25).
- **Organized:** The AI's score of 3.61 was higher than Human 1 (3.09) and HUMAN 3 (3.28) but a bit lower than Human 2 (3.67).
- **Comprehensible:** The AI achieved a score of 4.33, which is very close to the highest human score of 4.38 from Human 3, and higher than the two other human scores (4.12 and 3.75).
- **Succinct:** The AI's score of 3.48 surpassed all three human scores (3.09, 3.25, and 3.47), though it is very close to that of Human 3.
- **Synthesized:** The AI scored 3.97, higher than all three humans (3.38, 3.75, and 3.65).
- **Internally Consistent:** The AI received a score of 4.27, which is higher than all three human scores (3.88, 3.75, and 3.94).

Interestingly, the AI attained among the highest scores and consistently outperformed (i.e., across all nine dimensions) one of the humans. While none of the four options was the highest across all dimensions, the average score of the AI happens to be the highest (3.9 versus 3.48, 3.77, and 3.69), with the caveats that some dimensions may carry more weight than others, so that the

average is an imperfect measure. This comparison suggests that AI medical scribes have the potential to enhance the quality of clinical documentation in ER settings, potentially improving patient care and reducing the documentation burden on healthcare providers.

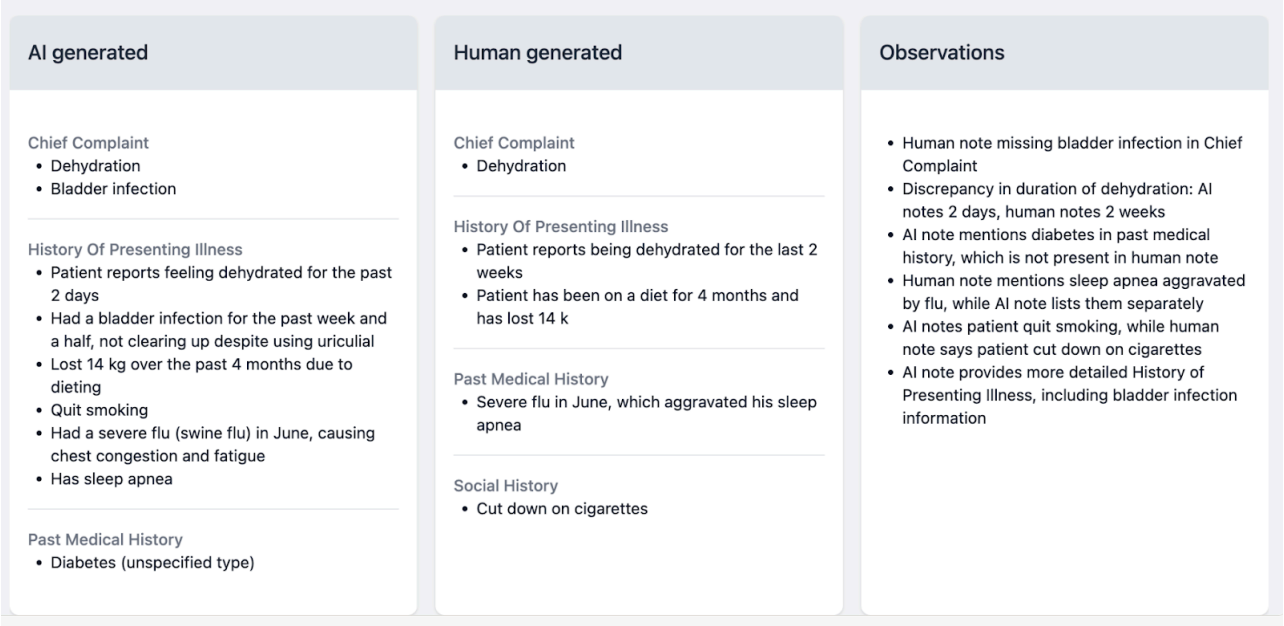

Figure 3. Example of an AI-generated note (left) and a human-generated note (right) for the same patient encounter.

Figure 3 presents an example of an AI-generated note and a human-generated note for the same patient encounter. While the AI-generated note includes more details (e.g., the human-generated note omitted the bladder infection), it is worth noting that this specific example includes a hallucination as the patient mentioned cutting down smoking, whereas the AI-generated note mentions instead that the patient quit smoking. This reinforces the importance of treating the AI-generated note as a first draft only and the need for physicians to double check the information and make the required edits to it before generating a final AI-assisted note. In addition, feedback can be collected about the edits performed by the clinicians, ultimately allowing the AI model to improve and become more accurate.

Another example displayed in Figure 4 illustrates that the human scribe overlooked crucial details, including information about the patient's high cholesterol. This omission underscores the potential for human error in manual scribing and highlights the importance of accurate and comprehensive documentation in patient care. With careful development and implementation, AI scribes have the potential to assist humans and help mitigate this issue.

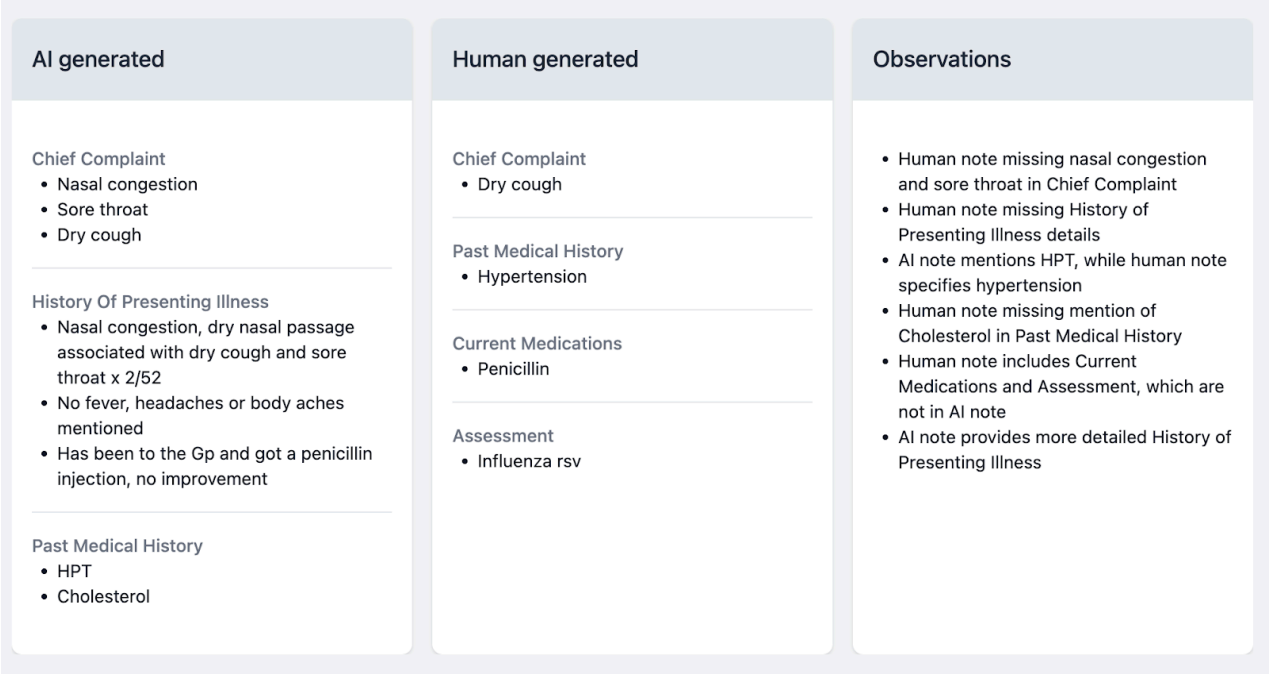

Figure 4. Second example of an AI-generated note (left) and a human-generated note (right) for the same patient encounter.

Time to generate the note

An additional important evaluation metric is the time it takes to generate the note and submit it to the EHR. In our study, the comparative analysis between AI-generated and human-generated medical notes revealed a significant disparity in efficiency, with AI consistently outperforming humans in terms of speed. That being said, in a typical process, the AI-generated note will need to be screened (and potentially edited) by clinicians, so that the total time is likely to be longer. The visualization in Figure 5 highlights the difference between the time to generate the AI and human notes, showing that both the average and the median times to generate a note are considerably lower for AI than for humans. The average time for AI to generate a note in this case was 18.69 seconds, with a standard deviation of 13.70 seconds (and a 95-th percentile of 31.5 seconds), indicating a consistently fast performance. By contrast, the average time for the three humans was 437.14 seconds, with a standard deviation of 85.60 seconds (and a 95th percentile of 564.0 seconds), reflecting greater variability and often a substantially longer duration.

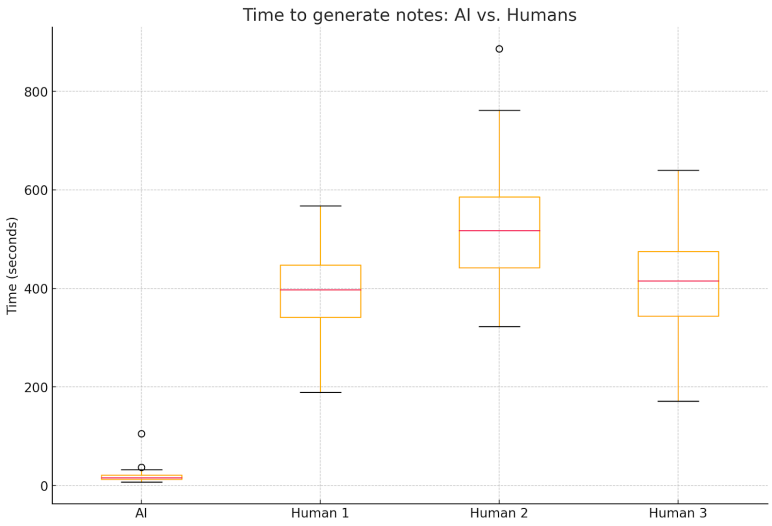

Figure 5. Time to generate a medical note (in seconds) from AI and from humans in our pilot study.

As mentioned, we could not measure the time it took for physicians to edit the note because, unfortunately, in all 50 visits, the note was not edited within the medical scribe application but was likely edited in the EHR. Given the lack of data access for the total time to generate the final AI-assisted note, we instead measured the perceived time spent on documentation via a qualitative survey (see Figure 6). The survey answers indicated a significant reduction in the perceived time, with the percentage of respondents who reported spending more than five minutes per encounter dropping from 70.73% to 40%.

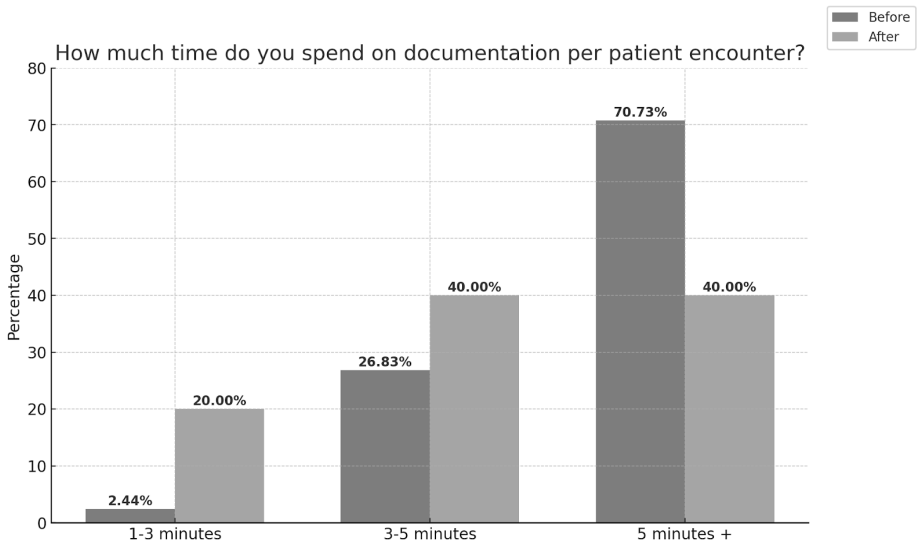

Figure 6. Survey responses about the typical time spent on documentation per patient encounter.

Qualitative clinical survey results

We next present the results for the impact on physician fatigue and on patient care. The results of the pilot study highlight the significant impact that AI scribes can have on improving documentation practices in emergency departments, thus potentially enhancing clinicians' quality of life.

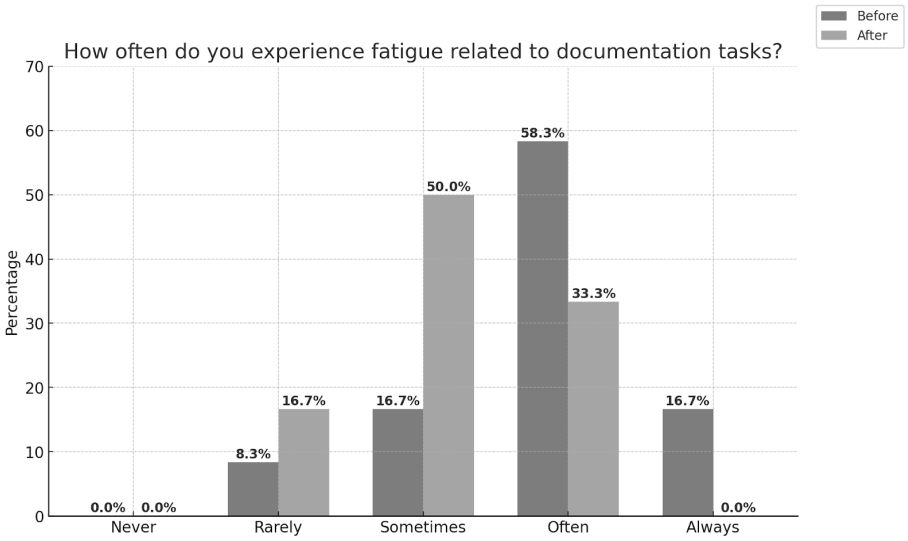

Figure 7. Survey responses about fatigue before pilot (left) and after pilot (right).

As shown in Figure 7, the percentage of clinicians reporting that they "rarely" experienced fatigue related to documentation increased from 8.3% before the pilot to 16.7% after the pilot. In addition, after the pilot, we observed that the majority of physicians still mentioned experiencing documentation fatigue "sometimes" or "often." Although these results have some limitations, they still hold promising value.

This reduction in fatigue (and hence burnout) is crucial, as it can lead to a better overall job satisfaction and an improved quality of patient care. Despite initial concerns regarding the accuracy and reliability of AI scribes, user feedback indicated satisfactory performance in these areas, which is encouraging for the future integration of this technology in clinical settings.

In terms of patient care, all but one physician answered "positively" to the question: "*Do you feel that the AI scribe has positively or negatively impacted your patient interactions?"*[9] This is an additional promising result. While we could not directly survey patients, follow-up studies could potentially ask patients about their experience and their perception of the care.

[9] The physician who noted that they were negatively impacted provided the following feedback: "*You need to put the tablet between yourself and the patient, which creates a barrier. In emergency situations, it is nice to start examining while trying to get history and the tablet is in the way. I find myself having to run back to the room to fetch the tablet."*

Finally, we highlight that 61% of physicians reported issues with the AI not transcribing the audio data properly, with one physician noting *"Sometimes it won't scribe anything. Noticed when a patient speaks too softly."* Three physicians also mentioned technical issues due to lack of a good Wi-Fi connection. All the reported issues were technical in nature and could easily be solved (e.g., by using a high-quality microphone and a better Wi-Fi connection).

Takeaways

The feasibility of integrating AI medical scribes into emergency department workflows is promising according to the above pilot study. At the same time, additional studies should be conducted with a larger number of patients to assess the robustness of our findings and confirm the positive assessment. In the pilot study, healthcare providers generally accepted the AI technology positively, with some expressing concerns that can be addressed through training and quick iterative improvements. The potential of AI scribes to enhance efficiency by reducing documentation time and allowing providers to concentrate more on patient care is significant.

Moving forward, continued monitoring and iterative improvements will be essential to address any concerns and enhance the accuracy and reliability of AI scribes. Expanding the pilot study to include more sites, a larger participant pool, and alternative clinical settings will provide more comprehensive data on the impact and effectiveness of AI scribes. This expansion will help solidify the benefits of AI scribes, pave the way for broader implementation in various clinical settings, and benefit a larger number of patients and healthcare professionals.

8. The Future of AI Medical Scribes

The field of AI medical scribing is rapidly evolving, primarily driven by the advancements in LLMs. These models, including OpenAI's GPT-4 and Meta's Llama 3, have shown remarkable improvements in understanding and generating human-like text, which is pivotal for medical scribing applications. The pace at which LLM technology is advancing is unprecedented, with new iterations and enhancements emerging frequently. This rapid development suggests a future in which the capabilities of AI medical scribes will continue to expand and improve rapidly.

Local processing and enhanced privacy

One of the most promising aspects of the ongoing advancements in LLMs is the potential for local processing. Currently, many AI scribing solutions rely on cloud-based servers to handle the computationally-intensive tasks of processing and generating text. However, as personal computers become more powerful, the expectation is that the entire scribing process could soon be managed locally. This shift would eliminate the need to send sensitive patient data to external servers, thereby significantly enhancing privacy and security. The ability to run these models locally on personal devices would also reduce latency and improve the overall user experience, making AI medical scribes more reliable and accessible.

Potential extensions and applications

The future of AI medical scribes is not limited to transcription and documentation. Several potential extensions can significantly augment their usefulness in clinical settings. These include the following:

1. **Clinical decision support:** AI medical scribes can be integrated with clinical decision support systems to provide real-time assistance to healthcare providers. By analyzing patient data and cross-referencing it with medical databases, AI scribes can offer diagnostic recommendations, flag potential drug interactions, and recommend treatment plans, thereby enhancing clinical decision-making.
2. **Billing and invoicing:** Another area ripe for integration is medical billing and invoicing. AI scribes can automate the extraction of billing codes from clinical note information to ensure accuracy and compliance with healthcare regulations. This automation can streamline the billing process and reduce administrative burden, leading to faster reimbursements and improved financial management for healthcare providers.
3. **Recommending actions:** AI scribes can incorporate a function to recommend future necessary actions, such as referrals to healthcare specialists, ordering specific lab tests, and prescribing medication. These added functionalities can make the medical process more holistic and reduce the burden placed on patients who must navigate the complex healthcare ecosystem.
4. **Direct patient communication:** AI medical scribes can also extend their functionalities to facilitate direct communication with patients. This includes generating patient summaries, follow-up email instructions, and educational materials based on the clinical encounter. By providing patients with clear and concise information, AI scribes can enhance patient understanding and engagement, thereby ultimately contributing to better health outcomes.

9. Summary and Recommendations

AI medical scribes offer a promising solution for reducing the administrative burden experienced by physicians. By leveraging AI and LLM technologies, these tools can efficiently handle documentation tasks, freeing up valuable time for clinicians. However, an important point to acknowledge is that AI technologies are not without risks. LLMs are still prone to hallucinations and mistakes, which can lead to inaccuracies in medical documentation and ultimately have negative consequences.

To maximize the benefits of AI medical scribes while mitigating the risks, the following recommendations can be considered:

1. **Ethical and clinical oversight:** Establish a robust framework for ethical and clinical oversight to address any concerns related to patient privacy and data security.

2. **Continuous monitoring and evaluation:** Regularly assess the performance of AI medical scribes to ensure accuracy and reliability. For example, one can implement systematic mechanisms that will promptly identify and correct errors. For instance, AI models can rely on human expert feedback to ensure that the models are creating notes accurately and are better aligned to what a physician would expect.
3. **Integration with clinical workflows:** Seamlessly integrate AI scribes into existing clinical workflows to enhance efficiency without disrupting patient care. Otherwise, physicians are unlikely to adopt such tools.
4. **Training and support for clinicians:** Provide adequate training and support to clinicians to effectively utilize AI medical scribes by ensuring that they understand how to review and validate AI-generated documentation.

By addressing these key areas, AI medical scribes can significantly alleviate the documentation burden experienced by physicians and ultimately mitigate the impact of the physician shortage and improve the overall quality of care.

References

Ammenwerth, E., Spötl, H. P. (2009). The time needed for clinical documentation versus direct patient care. Methods of information in medicine, 48(01), 84-91.

Gidwani, R., Nguyen, C., Kofoed, A., Carragee, C., Rydel, T., Nelligan, I., ... Lin, S. (2017). Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial. The annals of family medicine, 15(5), 427-433.

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., Dawson, L. (2014). A systematic review of speech recognition technology in health care. BMC medical informatics and decision making, 14, 1-14.

Liu, Y., Fabbri, A. R., Liu, P., Radev, D., Cohan, A. On learning to summarize with large language models as references. arXiv 2023. arXiv preprint arXiv:2305.14239.

Mishra, P., Kiang, J. C., Grant, R. W. (2018). Association of medical scribes in primary care with physician workflow and patient experience. JAMA internal medicine, 178(11), 1467-1472.

Momenipour, A., Pennathur, P. R. (2019). Balancing documentation and direct patient care activities: A study of a mature electronic health record system. International journal of industrial ergonomics, 72, 338-346.

Saag, H. S., Shah, K., Jones, S. A., Testa, P. A., Horwitz, L. I. (2019). Pajama time: working after work in the electronic health record. Journal of general internal medicine, 34, 1695-1696.

Stetson, P. D., Bakken, S., Wrenn, J. O., Siegler, E. L. (2012). Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). Applied clinical informatics, 3(02), 164-174.

Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., ... Lee, K. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM catalyst innovations in care delivery, 5(3), CAT-23.

Xu, Z., Jain, S., Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817.

Zhang, X., Lin, D., Pforsich, H., Lin, V. W. (2020). Physician workforce in the United States of America: forecasting nationwide shortages. Human resources for health, 18, 1-9.

Zhang, H., Yu, P. S., Zhang, J. (2024). A systematic survey of text summarization: From statistical methods to large language models. arXiv preprint arXiv:2406.11289.