AWS Healthscribe

AWS AI Service Cards

AWS AI Service Cards: AWS Healthscribe

Table of Contents

AWS Healthscribe

An AWS AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the responsible design and use of the service. A Service Card will evolve as AWS receives customer feedback, and as the service progresses through its lifecycle. AWS recommends that customers assess the performance of any AI service on their own content for each use case they need to solve. For more information, please see AWS Responsible Use of AI Guide and the references at the end. Please also be sure to review the AWS Responsible AI Policy, AWS Acceptable Use Policy, and AWS Service Terms for the services you plan to use.

This AI Service Card applies to the version of AWS Healthscribe that is current as of November 28, 2023.

Overview

AWS Healthscribe, a new HIPAA-eligible machine learning (ML) capability, empowers healthcare software vendors to build clinical applications that automatically generate preliminary clinical notes by analyzing patient-clinician conversations. AWS Healthscribe combines speech recognition and generative artificial intelligence (AI) to reduce the burden of clinical documentation by transcribing patient-clinician conversations and generating easy-to-review draft clinical notes. With AWS Healthscribe, healthcare software providers can use a single API to automatically create robust transcripts, extract key details (e.g., medical terms and medications), identify speaker roles, classify dialogues, and create summaries from patient-clinician discussions that can then be entered into an electronic health record (EHR) system. AWS Healthscribe enables responsible deployment of AI systems by citing the source of every line of generated text from within the original conversation transcript, making it easier for clinicians to review clinical notes before entering them into the EHR. Built with security and privacy in mind, AWS Healthscribe gives customers control over where their data is stored, encrypts data in transit and at rest, and does not use inputs or outputs generated through the service to train its models. This AI Service Card describes two of AWS Healthscribe's key features, transcription and clinical note generation, implemented by the Transcribe::StartMedicalScribeJob API.

We assess the transcription quality of AWS Healthscribe by measuring how accurately the words in the transcript match those spoken in the audio recording, as determined by a human listener. When a speaker says "The patient has high blood pressure," we expect the transcript to reflect the

actual words spoken, not contain errors like "The patent has a hyper tension." Three types of errors can occur: substitutions (like "hyper" for "high"), insertions of extra words (like "a"), and deletions of existing words (like "blood"). Correctly transcribed words are considered hits. Quality metrics like precision, recall, F1, and word error rate (WER) depend on the number of hits and errors.

The ASR system in AWS Healthscribe is trained to distinguish between variations that are inherent to the speaker and the content of the speech (intrinsic variations) and variations that are unrelated to the speech itself and should be ignored (confounding variations). Some examples of intrinsic variations include (1) differences in grammar and sentence structure; (2) the use of specific medical terminology and acronyms; and (3) tone of speech. Some examples of confounding variations include (1) background noise, echo, or room reverberation that can distort the recorded speech; (2) the quality and positioning of recording devices; (3) different regional dialects and accents; and (4) overlapping speech or interruptions. The system is specifically optimized and tested on data collected with these variations to improve robustness.

The quality of clinical note summarization by AWS Healthscribe is evaluated on factual completeness, factual correctness, and usability. Factual completeness measures the percentage of clinical facts covered in the AI-generated summary compared to a clinical note written by the clinician. Factual correctness gauges the accuracy of the facts in the AI-generated summary, comparing them against the original conversation. The measurement involves identifying any discrepancies in the AI-generated summary not present in the original transcript. Usability assesses the effectiveness of the AI-generated summary in meeting the needs of clinicians. It’s measured based on clinicians’ survey feedback on various dimensions such as coherence, medical terminology, professionalism, and clinical utility.

The performance of AWS Healthscribe’s clinical note summarization capability is influenced by both intrinsic and confounding factors. Intrinsic factors are directly related to the content and include: (1) the complexity and depth of the original conversation; (2) the use of specific medical terminology and acronyms; and (3) the coherence and consistency of the facts presented. Confounding factors, while unrelated to the content, can impact the summarization process. These include: (1) the presence of verbose, lengthy, or redundant content that does not contribute meaningful information to the summary; (2) potential contradictions within the conversation; and (3) the accuracy of the initial transcript, which serves as the basis for the summarization. The system is specifically optimized and tested on data collected with these variations to improve robustness.

Intended use cases and limitations

AWS Healthscribe is designed to help healthcare software providers build clinical applications that can automatically generate preliminary clinical notes and consultation transcripts from audio recordings of patient-clinician conversations. Medical professionals can use these auto-generated notes to quickly review, edit, and finalize their clinical documentation, thereby improving their productivity and consultation experience. AWS Healthscribe also provides evidence mapping which links every AI-generated sentence with relevant parts of the transcript, making it easier to review and finalize the documentation.

Currently, the service supports medical conversations in US-English for general medicine and orthopedics practices in batch mode. AWS Healthscribe generates summaries for specific sections of a clinical note, including chief complaint, history of present illness, assessment, plan, among others (for a complete list, see AWS Healthscribe Clinical Documentation file in the *Amazon Transcribe Developer Guide*). The service can detect and assign participant roles, as either a patient or a clinician, for up to four speakers in the conversation transcript.

However, customers should be aware of the system's limitations. The AI-generated clinical notes are intended to provide a preliminary summary and require human review and editing before final documentation. Given its probabilistic nature, the AI system might hallucinate or misinterpret information from the transcript. The system is reliant on verbal information; therefore, any observations not verbally expressed during the visit won't be captured, potentially affecting the completeness of the note. Transcript accuracy, and hence the clinical note, can be compromised by background noise, room echo, or unclear speech due to speech impediments or situational factors like patient discomfort. Furthermore, the coherence and accuracy of facts in the AI-generated clinical note can be affected if the consultation discussions are overly complex, lengthy, interspersed with overlapping speech or interruptions, or contain contradictory information. Additionally, while evidence mapping offers valuable assistance in contextual understanding and review of the AI-generated content, it is important to remember that these are probabilistic insights generated by AI, so there could be inaccuracies.

Design of AWS Healthscribe

Machine learning

AWS Healthscribe is built using custom-trained speech recognition and generative AI technologies. It works as follows: (1) Identify and separate non-speech and speech cues from the audio input. (2) Generate a transcription associated with the audio input. (3) Extract

features such as speaker roles, transcript segments, and clinical entities. (4) Generate structured clinical summaries along with evidence mapping, which links every sentence in the summary to the relevant conversation dialogue in the transcript. For more information, see the AWS Healthscribe in the *Amazon Transcribe Developer Guide*.

Performance expectations

Intrinsic and confounding variation will differ between customer applications. This means that performance will also differ between applications, even if they support the same use case. Consider two healthcare applications A and B. Application A builds clinical documentation assistance feature for in-clinic consultation setting. It accommodates multiple users in the audio recording, who may be located at varying distances from the microphone. On the other hand, Application B enables clinical documentation insights for a tele-health service. It captures users speaking close to their microphone, with one voice per recording channel, and minimal background noise interference. Because application A and B have differing kinds of inputs, they will likely have differing accuracy rates, even assuming that each application is deployed perfectly using AWS Healthscribe.

Test-driven methodology

We use multiple medical audio datasets to evaluate performance. No single evaluation dataset can represent all possible customer use cases. That is because evaluation datasets vary based on their demographic makeup (the number and type of defined groups), the amount of confounding variation (quality of content, fit for purpose), the types and quality of labels available, and other factors. Groups in a dataset can be defined by acoustic features (such as pitch, tone, intonation), demographic attributes (such as dialect, gender, age and ancestry), confounding variables (channel-specific, e.g., recording equipment varieties, and lexical complexity-related, e.g., consultation length, the number of specialized terms used, etc.), or a mix of all three. Different evaluation datasets vary across these and other factors. Because of this, all metrics – both overall and for groups – vary from dataset to dataset. Taking this variation into account, our development process examines AWS Healthscribe performance using multiple evaluation datasets, takes steps to increase accuracy for groups on which the AI performed least well, works to improve the suite of evaluation datasets, and then iterates.

Fairness and bias

Our goal is for AWS Healthscribe to (1) accurately transcribe medical consultation audio, and (2) accurately transcribe and summarize clinical notes for the full diversity of patient/clinician interactions. To achieve this, we use the iterative development process described above. As part of this process, we build datasets to capture a diverse range of human voices

and acoustic features under a wide range of confounding factors. We routinely test on datasets for which we have reliable and self-reported demographic labels. We define speaker groups by demographic attributes such as ancestry, age, and gender. As an example, on one dataset of scripted clinician/patient conversations, which reduces the variation in conversational difficulty that is introduced by unscripted speech, we find that across 28 demographic groups (such as Female+Asian ancestry, Male+European ancestry), the F1 word recognition (all words, including medical terms) accuracy was 84% or higher for every group of speakers. We use the same dataset to measure how complete the generated summaries are compared to a corresponding set of facts extracted by human clinicians, and how faithful the generated summaries are compared to the consultation transcripts. In this test we find the summarization F1 score of factual completeness and factual correctness to be 72%. Because results depend on AWS Healthscribe, the customer workflow and the test dataset, customers should test AWS Healthscribe on their own content, and follow the recommendations in the workflow design section below.

Explainability

AWS Healthscribe returns timestamps and confidence scores for each word transcribed in the audio. Customers can use the timestamps to listen to the segments of the input audio and verify the transcript content. For the note generation capability, AWS Healthscribe provides the summary along with evidence mapping, which links every sentence in the summary to the relevant dialogue in the transcript. Customers can use the evidence mapping to verify and understand the context behind the insight.

Robustness

To test the robustness of AWS Healthscribe's transcription and summarization capability, we evaluate performance on a wide variety of datasets across different medical consultation settings. The AI system is trained to be resilient under various acoustic environments, such as recording quality, background noise and room reverberation. Furthermore, it is optimized to handle consultations of varying lengths, across numerous medical conditions, and adapt to the unique speaking styles of various clinicians and patients.

Privacy and security

AWS Healthscribe processes only audio input data. Audio inputs are never included in the output returned by the service. Inputs and outputs are never shared between customers. AWS Healthscribe does not train on customer content. For more information, see Section 50.3 of the AWS Service Terms and the AWS Data Privacy FAQs. For service-specific privacy and security information, see the AWS Healthscribe FAQs.

Transparency

Where appropriate for their use case, customers who incorporate AWS Healthscribe in their workflow are encouraged to disclose their use of ML and ASR technology to end users and other individuals impacted by the application, and give their end users the ability to provide feedback to improve workflows. In their documentation, customers can also reference this AI Service Card.

Governance

We have rigorous methodologies to build our AWS AI services responsibly, including a working backwards product development process that incorporates Responsible AI at the design phase, design consultations and implementation assessments by dedicated Responsible AI science and data experts, routine testing, reviews with customers, and best practice development, dissemination, and training.

Deployment and performance optimization best practices

We encourage customers to build and operate their applications responsibly, as described in the AWS Responsible Use of AI Guide. This includes implementing Responsible AI practices to address key dimensions including fairness and bias, robustness, explainability, privacy and security, transparency, and governance.

Workflow design: The performance of any application using AWS Healthscribe depends on the design of the customer workflow. Conditions like background noise, recording device, and others are discussed in the Intended Use Cases section. Depending on the application, these conditions may be optimized by AWS Healthscribe customers, who define the workflow where audio is captured from end users. Evidence mapping, human oversight, workflow consistency and periodic testing for performance drift are also critical considerations that are under the control of customers, and that contribute to accurate, fair outcomes.

1. **Recording conditions:** Ideal audio inputs have moderate to minimal background noise. Workflows should include steps to address variation in use case specific recording conditions.
2. **File type / Sample rate:** For best results, use a lossless audio format, such as FLAC or WAV, with PCM 16-bit encoding. AWS Healthscribe supports sample rates of 16,000 Hz or higher.
3. **Custom vocabularies:** AWS Healthscribe recognizes vocabulary used in a variety of speaker communities (dialect regions). In cases where customers want to provide additional support for words specific to their domain or situation such as brand names or proper nouns and acronyms,

customers can deploy custom vocabularies to improve transcription accuracy for such words. For more information, see Custom vocabularies in the *Amazon Transcribe Developer Guide*.

4. **Human oversight:** The results produced by AWS Healthscribe are probabilistic and accuracy may be impacted by the various confounding factors described above. AWS Healthscribe should not be used to fully automate clinical documentation workflows, but rather to provide assistance to clinicians or medical scribes in their documentation process. We recommend providing evidence mapping capability in the workflow to help users easily understand the source of insight for context and validation. Use of AWS Healthscribe is subject to the AWS Responsible AI Policy. Review this policy prior to using AWS Healthscribe output, including as part of implementing appropriate human oversight, testing, and other use case-specific safeguards.
5. **Consistency:** Customers should set and enforce policies for the kinds of audio inputs permitted, and for how humans use their own judgment to assess AWS Healthscribe output. These policies should be consistent across all demographic groups. Inconsistently modifying audio inputs could result in unfair outcomes for different demographic groups.
6. **Performance drift**: A change in the kinds of audio that a customer submits to AWS Healthscribe and updates to the models powering the features may lead to different outputs over time. To address these changes, customers should consider periodically retesting the performance of AWS Healthscribe, and adjusting their workflow if necessary.

Further information

- For service documentation, see AWS Healthscribe in the *Amazon Transcribe Developer Guide*.
- For details on privacy and other legal considerations, see the following AWS policies: Acceptable Use, Responsible AI, Legal, Compliance, and Privacy.
- For help optimizing workflows, see Generative AI Innovation Center, AWS Customer Support, AWS Professional Services, Ground Truth Plus, and Amazon Augmented AI.
- If you have any questions or feedback about AWS AI service cards, please complete this form.

Glossary

Controllability: Steering and monitoring AI system behavior.

Privacy & Security: Appropriately obtaining, using and protecting data and models.

Safety: Preventing harmful system output and misuse.

Fairness: Considering impacts on different groups of stakeholders.

Explainability: Understanding and evaluating system outputs.

Veracity & Robustness: Achieving correct system outputs, even with unexpected or adversarial inputs.

Transparency: Enabling stakeholders to make informed choices about their engagement with an AI system.

Governance: Incorporating best practices into the AI supply chain, including providers and deployers.