Amazon Transcribe

AWS AI Service Cards

AWS AI Service Cards: Amazon Transcribe

Table of Contents

Amazon Transcribe

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 Service Card applies to the releases of Amazon Transcribe that are current as of May 26, 2026.

Overview of Amazon Transcribe

Amazon Transcribe is a cloud-based automatic speech recognition (ASR) service designed for enterprise use cases. Amazon Transcribe converts spoken audio into text, enabling customers to add speech-to-text capabilities to their voice-enabled applications. Customers can use Amazon Transcribe for tasks such as contact center analytics, voicemail transcription, meeting captioning, closed-captioning for media content, transcribing educational content and search and analytics for audio and video archives. This AI Service Card applies to the use of Amazon Transcribe via the StartTranscriptionJob API (batch mode) and the StartStreamTranscription API (streaming mode). Features covered by this Service Card include automatic speech recognition, language identification, speaker diarization, transcribing digits, custom vocabularies, identifying and redacting personally identifiable information (PII), vocabulary filtering and toxicity detection. Amazon Transcribe is a managed AWS AI service; customers can focus on integrating speech-to-text capabilities into their applications without having to provision or manage any infrastructure. Typically, customers use the AWS Console to develop and test applications, and the API for production workloads at scale.

An Amazon Transcribe *<audio input, configuration parameters, transcript output>* triple is said to be "effective" if a skilled human evaluator decides that the resulting transcript: 1/ accurately reflects the words spoken in the audio; 2/ correctly attributes speech to speakers when diarization is enabled; 3/ appropriately handles formatting such as punctuation, capitalization, and digit transcription when enabled; and 4/ is suitable for the customer's intended downstream use. Otherwise, the output may require refinement or may not fully meet all evaluation criteria for the specific use case. A customer's workflow must decide if a transcript is effective using

human judgment, whether human judgment is applied on a case-by-case basis (as happens when reviewing individual transcripts) or is applied via the customer's choice of an acceptable score on an automated test.

The "overall effectiveness" of Amazon Transcribe for a specific use case is based on the percentage of use-case-specific inputs for which the service returns an effective result. Customers should define and measure effectiveness for themselves for the following reasons. First, the customer is best positioned to know which audio inputs will best represent their use case and should therefore be included in an evaluation dataset. Second, different ASR systems may respond differently to the same audio input, requiring tuning of the configuration and/or the evaluation mechanism. We assess the quality of Amazon Transcribe by measuring how well the words from an ASR transcript match the words spoken in the audio, as transcribed by a human listener. When a speaker says *"This system can really recognize speech"*, we expect the transcript to contain the spoken words, not *"This system can wreck a nice beach"*. Three types of errors may appear in a transcription: substitutions (like "recognize" transcribed as "wreck"), insertions (extra words such as "nice"), and deletions (missing words such as "really"). Correctly-transcribed words are called hits. Quality metrics like precision, recall, F1, and word error rate (WER) depend on the number of hits and errors.

As with all ML solutions, Amazon Transcribe must overcome issues of intrinsic and confounding variation. Intrinsic variation refers to features of the audio input that the service needs to pay attention to, such as the specific words spoken by each individual and their unique pronunciation, intonation, speaking rate, and vocabulary. Confounding variation refers to features of the audio input that the service should ignore, for example, variations in background noise, recording equipment, room acoustics, and transmission quality, which should not affect the transcription of the spoken words. The full set of variations encountered in the input audio include: languages, dialects, accents, speaker age, speaker gender, speaking rate, speaking style (read vs. spontaneous), number of speakers, overlapping speech, background noise, recording device quality and placement (near-field vs. far-field), sampling rate, room reverberation, transmission line noise, and domain-specific vocabulary. Since different Amazon Transcribe features use audio input and configuration parameters differently, customers should experiment as necessary to understand how best to adjust their configuration to achieve an effective result.

Intended Use Cases and Limitations

Core Capabilities: Amazon Transcribe serves a wide range of potential application domains and offers the following core capabilities:

- **Batch and Streaming Transcription.** Amazon Transcribe offers two transcription modes. Batch transcription processes pre-recorded audio files stored in Amazon Simple Storage Service and returns a complete transcript after processing; it benefits from full audio context, making it suitable for call recordings, media archives, and voicemails. Streaming transcription processes live audio in near real-time, returning partial and final results as audio is received; it must produce results with limited future context, so partial results may change as more audio arrives. Streaming is suitable for live captioning, real-time call monitoring, and voice chat. Accuracy may differ between modes for the same audio.
- **Language Identification.** Automatically detects the language spoken in an audio input. When language identification is enabled, the service analyzes the acoustic and phonetic characteristics of the speech to determine which of the candidate languages is most likely being spoken. Customers can optionally specify a set of candidate language codes, and the service returns the identified language along with a confidence score. Language identification can be used in both batch and streaming modes. In batch mode, the service identifies the dominant language of the entire audio file. In streaming mode, the service can identify the language at the start of the stream. Language identification is most accurate when the set of candidate languages is small and when the candidate languages are phonetically distinct from one another. Accuracy may be reduced when candidate languages are closely related (for example, distinguishing between regional variants of the same language) or when speakers switch languages mid-utterance.
- **Speaker Diarization.** Differentiates speakers in an audio input based on their voice characteristics, annotating the transcript to indicate which speaker produced each segment of speech. When speaker diarization is enabled, the service analyzes the audio to identify distinct speakers and labels each transcribed segment with a speaker identifier (Speaker 0, Speaker 1, Speaker 2, and so on). Speaker labels are assigned based on voice characteristics within a single audio file or streaming session; labels are randomly assigned and are not consistent across separate audio files or sessions. Customers can optionally specify the expected number of speakers to improve diarization accuracy. Speaker diarization supports up to 30 unique speakers per audio input. The feature performs best when speakers have distinct voice characteristics and when there is minimal overlapping speech. Performance may degrade when speakers have very similar voice characteristics, when there are many speakers in a single recording, or when speakers frequently interrupt one another. Speaker diarization is available in both batch and streaming modes. Speaker diarization does not perform speaker identification or recognition. It cannot be used to identify who a speaker is, to match a speaker against a known voice profile, or to track the same speaker across separate audio files or sessions.
- **Transcribing Digits.** Converts spoken numbers into their written digit form in the transcription output. For example, when a speaker says "my phone number is five five five one two three

four," the service can output "my phone number is 5551234" rather than the spoken-word form. This feature is part of Amazon Transcribe's inverse text normalization (ITN) capability, which also handles formatting of other spoken elements such as dates, times, currencies, and addresses into their conventional written forms. Digit transcription is enabled by default for supported languages and requires no additional configuration. The accuracy of digit transcription depends on factors such as the clarity of the spoken numbers, the presence of background noise, the context in which numbers are spoken (for example, isolated number sequences vs. numbers embedded in conversational speech), and expected adherence to locale variations. Digit transcription is available for a subset of supported languages.

- **Custom Vocabularies.** Allows customers to provide a list of domain-specific terms, brand names, proper nouns, acronyms, or other words that the service should recognize with higher accuracy. When a custom vocabulary list is provided, the service biases its recognition toward the specified terms, making it more likely to correctly transcribe those words when they occur in the audio. This is particularly useful for words that are uncommon in general speech but frequent in the customer's domain; for example, proprietary product names, medical terminology, legal terms, or financial instrument identifiers. Custom vocabulary is available across all languages, for both Batch and Streaming modes.
- **PII Identification and Redaction.** Redacts personally identifiable information (PII) like names, phone numbers, credit card numbers, Social Security numbers and other PII types from the transcript. When PII identification is enabled, the service identifies PII entities in the transcript under a placeholder tag. In batch mode, customers can choose to receive both the original (unredacted) and redacted transcripts, or only the redacted transcript. In streaming mode, redaction is completed only at the end of the audio segment. The types of PII that Amazon Transcribe can identify and redact vary between batch and streaming transcriptions, and vary by language. PII identification and redaction uses pattern matching and contextual analysis to detect PII entities; it is not guaranteed to identify all PII instances in all contexts. Customers should not rely solely on automated PII redaction for compliance with data protection regulations without additional review processes, particularly for regulated workloads. The accuracy of PII identification may vary depending on factors such as audio quality, speaker clarity, and the context in which PII is spoken.

The components differ in the parameters (for example, API arguments) required to invoke them. For more information about these specifications, see the Amazon Transcribe Documentation.

Supported Languages. Amazon Transcribe supports over 100 languages and locales, with varying levels of feature support and accuracy. Transcription accuracy is highest for English (particularly

US English), which benefits from the largest and most diverse training data, the broadest range of evaluation datasets, and the most extensive demographic fairness testing. Accuracy is strong for other high-resource languages but may not match English performance across all conditions. For lower-resource languages, accuracy may be noticeably lower, and performance may be less consistent across acoustic conditions. Feature availability also varies by language: digit transcription, custom vocabularies, and PII redaction are available for a subset of supported languages, with the broadest feature set available for English. Customers should test Amazon Transcribe on their own audio content for each language they intend to use, and should set accuracy expectations based on their own evaluation rather than extrapolating from English performance.

Supported Input Formats. Amazon Transcribe operates on audio input only. Batch transcription is supported in the following supported formats: WAV, MP3, MP4, M4A, FLAC, OGG, AMR, WebM. Streaming transcription is supported in the following supported formats: PCM, OGG, FLAC, G711_ALAW, G711_ULAW, G729. Sampling frequencies from 8kHz to 48kHz are supported for both modes. The maximum supported audio file size is 2GB, and the maximum supported audio length is 8 hours.

Appropriateness for Use. Because its output is probabilistic, Amazon Transcribe may produce inaccurate transcriptions. Customers should evaluate outputs for accuracy and appropriateness for their use case, especially if they will be directly surfaced to end users or used for consequential decisions. Additionally, if Amazon Transcribe is used in customer workflows that produce consequential decisions – such as decisions that may impact an individual's rights, access to services, financial standing, or employment – customers must evaluate the potential risks of their use case and implement appropriate human oversight, testing, and other use case-specific safeguards to mitigate such risks. Customers should treat Amazon Transcribe outputs as an assistive tool and not as a sole basis for consequential determinations. For more information, see the AWS Responsible AI Policy.

Safety Features. Amazon Transcribe converts speech to text verbatim, including any inappropriate content spoken in the audio. The service is not a generative AI system, does not produce content beyond its transcription of the provided audio, does not use text prompts, and is therefore not susceptible to prompt injection attacks. To help customers manage sensitive content in transcriptions, Amazon Transcribe offers vocabulary filtering, which enables customers to mask or remove words that are sensitive or unsuitable for their audience from transcription results based on a customer-defined word list. Amazon Transcribe also enforces a default vocabulary filter for certain sensitive words. Additionally, customers can use the PII redaction feature to

automatically identify and remove personally identifiable information from transcripts and the Amazon Transcribe Toxicity Detection feature for identifying and classifying toxic voice content.

Considerations When Choosing Use Cases

When assessing an ASR service for a particular use case, we encourage customers to specifically define the use case, i.e., by considering at least the following factors: the **business problem** being solved; the **stakeholders** in the business problem and deployment process; the **workflow** that solves the business problem, with the model and oversight as components; key system **inputs and outputs**; the expected intrinsic and confounding **variation**; and the types of **errors** possible and the relative impact of each.

Consider the following use case of utilizing Amazon Transcribe to support a contact center analytics workflow that helps a customer service operations team analyze agent-customer phone calls. The **business goal** is to generate accurate transcripts of recorded customer calls for downstream analysis, enabling the operations team to derive insights about conversation quality, customer sentiment, and agent performance. The **stakeholders** include the contact center operations manager, who wants to improve agent productivity and customer satisfaction through data-driven insights, and the contact center agents, whose interactions are transcribed and analyzed. The **workflow** is: 1/ the contact center records customer calls, with one speaker per audio channel (agent on one channel, customer on the other); 2/ recorded audio files are uploaded to Amazon Simple Storage Service; 3/ the operations team submits the audio to Amazon Transcribe via the StartTranscriptionJob API, specifying the language, enabling speaker diarization, and optionally enabling PII redaction; 4/ Amazon Transcribe processes the audio and returns a transcript with speaker labels and timestamps; 5/ the operations team uses the transcribed text with downstream analytics tools to generate insights, which supervisors review to improve processes and training. **Input audio** contains information regarding customer inquiries, complaints, account details, agent responses, hold times, and conversational dynamics. **Output transcripts** contain text representations of the spoken words, speaker attribution, word-level timestamps, confidence scores, and optionally redacted PII. Input **variations** include: 1/ different accents, dialects, and speaking rates of callers and agents; 2/ background noise from the caller's environment (home, car, public space) and the contact center floor; 3/ telephony audio quality at 8kHz sampling rate with potential line noise; 4/ domain-specific terminology related to the customer's industry (such as financial product names, insurance policy terms, or technical support jargon); 5/ overlapping speech when the agent and caller talk simultaneously; 6/ varying call durations and conversation styles (scripted vs. unscripted). The **error types**, ranked in order of estimated negative impact on stakeholders, include: 1/ substitution errors on critical terms such as product names, account numbers, or action items, which could lead to incorrect downstream

analytics or decisions (medium-high impact); 2/ deletion errors where important spoken content is missing from the transcript, causing incomplete analysis (medium impact); 3/ insertion errors where words not spoken appear in the transcript, potentially introducing noise into analytics (low-medium impact); 4/ incorrect speaker attribution, where speech is assigned to the wrong speaker, affecting agent performance analysis (medium impact); 5/ PII not correctly identified and redacted, creating a compliance risk (medium-high impact); 6/ PII incorrectly identified and redacted from non-PII content, causing information loss (low-medium impact). With this in mind, we would expect the operations team to test a sample of call recordings through the API, compare the resulting transcripts against human-generated reference transcripts, and review the output for their specific use case.

Amazon Transcribe processes audio that may contain personal information in the form of human speech. Customers must ensure compliance with all laws and regulations applicable to their use of Amazon Transcribe. This includes understanding obligations under privacy, data protection, and communication laws – including consent, eavesdropping, wiretapping, and recording laws – that apply in their jurisdiction and to the individuals whose speech is being processed. Customers should collect and process only audio that is within the reasonable expectations of the individuals whose speech is being recorded and transcribed. This includes ensuring that all necessary and appropriate consents are obtained from speakers before their audio is submitted to the service, both for the recording of their speech and for the processing of that recording by a cloud-based service.

Many applications are designed to capture audio from a specific set of speakers (for example, the two parties on a phone call, or participants in a meeting who have consented to recording). However, microphones may also capture speech from individuals who are not the intended users of the application. To avoid unintentionally processing the speech of non-consenting individuals, customers should consider the following:

- **Microphone placement and scope:** Customers cannot always control who may speak near the recording device. Workflows should be designed to minimize the capture of speech from unintended parties, particularly in public or open environments.
- **Reasonable expectations of speakers:** Amazon Transcribe should be used only in scenarios where the collection and processing of speech data is within the reasonable expectations of the speakers.
- **Disclosure to affected individuals:** Where appropriate, customers should inform individuals that their speech is being recorded and processed using automated speech recognition technology.

Organizations operating in regulated industries or sensitive domains should evaluate any specific legal, regulatory, or policy obligations when using Amazon Transcribe, as it may not be appropriate for every industry or scenario without additional safeguards. Customers who are required to conduct data protection impact assessments or AI risk assessments should consider the information in this Service Card, the data processing descriptions in the Privacy section, and may request additional detail through their AWS account team.

Unsupported Use Cases. Amazon Transcribe is designed and optimized for naturally occurring human speech. The following uses are out of scope and may produce unreliable results:

- **Synthetic speech:** Amazon Transcribe is not designed for audio generated by text-to-speech (TTS) systems, voice synthesis engines, or other speech generation technology. While the service may produce output for synthetic speech inputs, accuracy is not guaranteed and has not been evaluated for this use case.
- **Mechanically or digitally transformed speech:** Audio that has been altered through pitch shifting, speed modification, voice disguising, vocoding, or other signal transformations is out of scope. Such transformations change the acoustic characteristics that the model relies on for accurate recognition.
- **Non-speech audio:** Amazon Transcribe is designed to transcribe human speech. It is not designed to interpret non-speech sounds such as music, tones, alarms, or animal sounds. The service may produce insertion errors (spurious words) when processing audio that contains extended non-speech segments.
- **Speaker identification or biometric recognition:** As noted under Speaker Diarization, the service labels speakers generically (Speaker 0, Speaker 1, etc.) and cannot identify who a speaker is, match a speaker against a known voice profile, or track speakers across separate audio files or sessions.
- **Covert surveillance:** Amazon Transcribe may not be used for covert audio surveillance or monitoring of individuals without their knowledge and consent, or in any manner that violates applicable laws and regulations.

Amazon Transcribe is not intended to support any prohibited practices under applicable AI legislation or any other relevant law. Amazon Transcribe can be integrated into an array of systems such as contact center platforms, content management systems, meeting productivity tools, accessibility solutions, compliance monitoring systems, and media analytics workflows. All Amazon Transcribe use cases must comply with the AWS Acceptable Use Policy.

Design of Amazon Transcribe

Machine learning

Amazon Transcribe performs automatic speech recognition using machine learning, specifically, a neural encoder model combined with a language model. At a high level, the core service works as follows: 1/ Amazon Transcribe receives audio input (along with configuration parameters such as language code and feature selections) via the API or Console; 2/ the service extracts acoustic features from the audio input; 3/ the acoustic features are processed by a neural encoder that generates probabilities over a set of text tokens at each time frame; 4/ a language model is applied to score and rank candidate transcriptions based on these probabilities; 5/ a beam-search decoding algorithm selects the highest-ranking transcription; 6/ additional feature-specific processing is applied as requested (described below); 7/ the final output, including the transcript, word-level timestamps, confidence scores, and any requested feature outputs, is returned to the customer.

Language identification uses a downstream model that consumes high-level representations from the acoustic encoder to produce language-level output probabilities. The model is trained on multilingual data with balanced sampling across languages to ensure all languages are represented during training. **Speaker diarization** analyzes the audio to differentiate and cluster speakers based on unique voice characteristics, used solely to annotate the transcript with speaker labels and discarded after processing. Speaker labels are assigned within a single audio file or session and are not linked to any speaker identity. **Custom vocabularies** are implemented through a contextual adapter module that biases the encoder's output probabilities toward recognizing customer-specified terms. When a customer provides a custom vocabulary list, the terms are encoded into embedding vectors, and these vectors are used to boost the probability of the corresponding terms during beam-search decoding. **PII identification and redaction** is applied as a post-processing step on the transcription output. The service uses pattern matching and contextual analysis to identify PII entities in the transcript text and replaces them with placeholder tags when redaction is enabled. **Word-level timing** is generated by an alignment model trained on top of the ASR encoder, which produces frame-level token probabilities that are used to estimate the start and end time of each word in the transcript through forced alignment. **Confidence scores** are generated for each word based on the model's output probabilities and are calibrated during each model update cycle to ensure consistency across model versions.

Amazon Transcribe models are trained and tested on licensed and proprietary datasets, off-the-shelf purchased datasets, and open-source datasets in the public domain. The training data

covers a diverse range of acoustic conditions, speaker demographics, languages, and use cases. Training data from licensed and purchased datasets was collected from speakers who consented for their voice recordings to be used for commercial purposes. Training data from open-source datasets was released under a commercial-usage license. Amazon Transcribe may use customer content to develop and improve the service. Customers can opt out of having their content used for service improvement via AWS Organizations or other opt-out mechanisms we may provide. When customers opt out, their content is not used for service improvement. Regardless of opt-out status, customer content is never shared between customers. Amazon Transcribe is available pursuant to the AWS Customer Agreement or other relevant agreements with AWS.

Controllability

We say that Amazon Transcribe exhibits a particular "behavior" when it generates the same kind of output for the same kinds of audio inputs and configuration parameters (for example, language code, sample rate, speaker diarization settings, custom vocabulary, and PII redaction settings). For a given model architecture, the control levers that we have over the behaviors are primarily a/ the training data corpus (which we curate from a variety of licensed, proprietary, and publicly available sources), b/ the language models and vocabulary that determine how the service ranks candidate transcriptions, c/ the safety filters (such as the default vocabulary filter) applied to post-process outputs, and d/ customer-facing configuration options that allow customers to tailor the service behavior to their use case. Customers can further steer the behavior of Amazon Transcribe to their specific needs through several mechanisms:

- **Custom vocabularies** bias the model toward recognizing specific terms that are important to the customer's domain, improving accuracy for those terms without requiring model retraining.
- **PII redaction** allows customers to control what types of sensitive information are identified and removed from outputs, with configurable entity type selection.
- **Vocabulary filtering** allows customers to control what words are masked or removed from outputs, based on a customer-defined word list.
- **Speaker diarization settings** allow customers to specify the expected number of speakers, improving speaker attribution accuracy when the number of speakers is known in advance.
- **Language identification settings** allow customers to specify a set of candidate languages, narrowing the identification task and improving accuracy.
- **Confidence score thresholds** can be implemented by customers in their downstream workflows to flag or escalate low-confidence segments for human review.

These mechanisms serve as compensating controls that customers can use to address known model limitations for their specific use case. The effectiveness of these controls should be evaluated by the customer on their own content.

Performance expectations

Intrinsic and confounding variation differ between customer applications. This means that performance will also differ between applications, even if they support the same use case. Consider two applications A and B. Application A enables live captioning for a television news broadcast, with professional presenters speaking clearly into high-quality microphones, scripted speech with domain-specific terminology, and minimal background noise. Application B transcribes customer service calls at a contact center, with callers speaking on mobile phones in noisy environments, unscripted conversational speech with filler words and interruptions, and telephony audio at 8kHz sampling rate. Because A and B have differing kinds of inputs, they will likely have differing error rates, even assuming that each application is deployed perfectly using Amazon Transcribe. In addition, streaming transcription provides real-time results but may have different accuracy characteristics compared to batch transcription, which can process the entire audio file before returning results. Furthermore, performance differs across languages: English (particularly US English) generally delivers the highest accuracy due to the largest and most diverse training data, while lower-resource languages may exhibit higher error rates and greater sensitivity to adverse acoustic conditions. As a result, the overall utility of Amazon Transcribe will depend both on the service and on the workflows it enables. Performance results depend on a variety of factors including Amazon Transcribe itself, the customer workflow, and the evaluation dataset; we recommend that customers test Amazon Transcribe using their own content.

Test-driven methodology

We use multiple datasets, automated evaluations and human inspection to evaluate the performance of Amazon Transcribe models. No single evaluation dataset suffices to completely capture performance. This is because evaluation datasets vary based on use case, intrinsic and confounding variation, the quality of ground truth available, and other factors. Our development testing involves automated testing against publicly available and proprietary datasets, benchmarking against proxies for anticipated customer use cases, evaluation across multiple languages and acoustic conditions, and iterative refinement based on results. Our development process examines Amazon Transcribe's performance using these tests, takes steps to improve the model and the suite of evaluation datasets, and then iterates. In this Service Card, we provide examples of test results to illustrate our methodology.

Automated Evaluations: Automated testing provides consistent comparisons between candidate models by applying standardized metrics across evaluation datasets. The primary metric for transcription accuracy is word error rate (WER), which measures the aggregate of insertion, deletion, and substitution errors divided by the total number of words in the reference transcript. We also measure F1 scores for evaluating the accuracy of specific output elements – such as custom vocabulary terms, named entities, digit transcription, punctuation, and capitalization – which balance the percentage of predicted items that are correct (precision) against the percentage of correct items that are included in the prediction (recall). For features such as speaker diarization, custom vocabulary recognition, punctuation, and capitalization, we use feature-specific metrics including diarization accuracy, custom vocabulary F1 scores, and punctuation and casing F1 scores. Models are evaluated on datasets totaling thousands of hours of audio data, covering multiple languages, acoustic conditions (including telephony at 8kHz and broadband at 16kHz), and speaker demographics. Evaluation datasets include internally commissioned datasets representative of customer use cases, purchased off-the-shelf datasets, and publicly available benchmarking datasets used by the scientific community. All evaluation datasets are versioned, archived, and subject to legal review.

Model Quality Monitoring: Amazon Transcribe employs automated monitoring to track service behavior over time. Canary systems periodically evaluate the service by measuring specific metrics (including WER, latency, and processing speed) against pre-determined reference datasets. These systems automatically trigger alerts when service metrics demonstrate any significant drift or anomalous behavior. Mechanisms are in place to roll back the service to a previously known stable state if needed.

Input Validation: Amazon Transcribe validates the format of the input audio (for example, encoding, sample rate, and channel configuration) before processing it. Non-adherence to prescribed format specifications will result in an error response. All data used in model development is validated and checked for quality and consistency before use in training or evaluation.

Safety

Safety is a shared responsibility between AWS and our customers. Our goal for safety is to mitigate key risks of concern to our enterprise customers, and to society more broadly. Amazon Transcribe is an ASR service that converts speech to text verbatim. It is not a generative AI service and does not generate any content that is not spoken in the input audio. This fundamental design characteristic means that the service does not produce hallucinated, fabricated, or novel content. Amazon Transcribe does not use text prompts and is not susceptible to prompt injection attacks.

In a case where a user speaks inappropriate content, the service will transcribe that content verbatim. To help customers manage such scenarios, Amazon Transcribe provides the following safety-relevant features:

Vocabulary Filtering: Customers can define a list of words (e.g., offensive, profane, or otherwise inappropriate words) to mask or remove from transcription results. The default vocabulary filter removes certain sensitive words; customers can supplement this with their own word lists. Note that the default vocabulary filter may not include words from all supported languages; customers should review and supplement the filter for their specific language and use case.

PII Redaction: Customers can automatically identify and redact personally identifiable information from transcripts, including names, addresses, credit card numbers, Social Security numbers, and other PII types. However, PII identification is not guaranteed to detect all PII instances in all contexts, and false positives (non-PII content incorrectly identified as PII) may also occur. Customers operating in regulated environments should not rely solely on automated PII redaction without implementing additional review processes appropriate to their compliance requirements.

Toxicity Detection: Amazon Transcribe Toxicity Detection (available as a separate feature) leverages both audio and text-based cues to identify and classify toxic voice content across seven categories including sexual harassment, hate speech, threats, abuse, profanity, insults, and graphic content. Toxicity Detection is not guaranteed to detect all instances of toxic language in all contexts, and false positives (non-toxic content incorrectly identified as toxic) may also occur. Customers operating in regulated environments should not rely solely on automated toxicity detection without implementing additional review processes appropriate to their compliance requirements. Please see the Service Card for Toxicity Detection for additional useful information.

Customers are responsible for end-to-end testing of their applications on datasets representative of their use cases and any additional safety mitigations, and deciding if test results meet their specific expectations of safety, fairness, and other properties, as well as overall effectiveness.

Fairness

Amazon Transcribe is designed to work well for speakers across the variety of pronunciations, intonations, vocabularies, and grammatical features that speakers of each supported language may use. We consider speaker communities defined by regions – for example, speakers of

different English varieties such as US, British, Australian, Indian, and South African English – and communities defined by multiple dimensions of identity, including accent or ancestry, age, and gender. To measure accuracy across these dimensions, we perform extensive evaluations covering both demographic variation and confounding factors such as acoustic conditions, pronunciation, and speaking rate. For controlled comparison across demographic groups in English, we use evaluation datasets designed to minimize confounding variation, enabling fair comparisons across dimensions such as accent/ancestry, age, and gender.

We find that Amazon Transcribe performs reasonably well across demographic attributes. As an example, on one dataset of read speech with 23 demographic groups, defined by ancestry, gender, and regional dialect (such as Liberian English, New England English), we find the median WER to be 2.4%, and statistically significant differences ($p < 0.001$) in accuracy between native (WER of 0.9%) and non-native English speakers (WER of 4.8%). Along the gender dimension, the median error rate for females is slightly lower (less than 1 error difference) than that of males, and the error distributions indicate these differences are statistically significant ($p < 0.001$).

For non-English languages, transcription accuracy varies across languages and locales. This variation is expected and reflects differences in the volume and diversity of training data available for each language, the linguistic characteristics of each language, and the representativeness of available evaluation datasets. In general, languages with larger training datasets and more diverse acoustic coverage perform better than languages with limited training data. We apply quality thresholds as launch criteria for all supported languages, and we continuously work to improve accuracy across all languages by expanding training data diversity and volume. For most non-English evaluation datasets, fine-grained demographic information (such as speaker gender, age, and accent) is not available, which limits the measurement of within-language demographic fairness for those languages. We are investing in expanding the demographic coverage of datasets for non-English languages. Customers should evaluate Amazon Transcribe on their own content for each language and should not assume that performance observed for one language will generalize to another.

Because results depend on Amazon Transcribe, the customer workflow and the evaluation dataset, we recommend that customers additionally test Amazon Transcribe on their own content.

Robustness

We maximize robustness with a number of techniques, including using large training datasets that capture many kinds of variation across many speakers, and simulating a variety of acoustic

conditions during training (such as different noise types, reverberation levels, and telephony codecs). Ideal audio inputs to Amazon Transcribe contain audio with high recording quality, low background noise, and low room reverberation. However, Amazon Transcribe is trained to be resilient even when inputs vary from ideal conditions and can perform well in noisy and multi-speaker settings. We measure model robustness by evaluating performance across a range of acoustic conditions, including clean speech, moderate and high background noise, far-field recording, overlapping speech, and telephony audio. Amazon Transcribe models degrade gracefully under adverse conditions rather than producing catastrophic errors. Additionally, given the architecture of the model (a neural encoder combined with a language model operating on continuous-valued audio inputs) the risk of adversarial attacks such as training data extraction or membership inference is significantly lower compared to text-based generative models. We are not aware of published work demonstrating such attacks on encoder-only ASR architectures of this type.

Explainability

When Amazon Transcribe transcribes audio, it assigns a confidence score to each word in the transcript, indicating the service's confidence that the word was correctly recognized. Confidence scores are calibrated during each model update cycle to ensure consistency across model versions, which supports the stability of downstream workflows that rely on specific confidence thresholds. Customers can use these confidence scores to identify portions of the transcript that may require additional review, to implement quality thresholds appropriate for their use case, or to trigger human review for low-confidence segments. Additionally, Amazon Transcribe provides word-level timing information, indicating the start and end time of each recognized word in the audio. This "acoustic grounding" can help customers identify challenging portions of the audio where the service may have made errors, and facilitates alignment of transcripts with the source audio for verification purposes. If customers enable alternative transcriptions (available in batch mode), Amazon Transcribe returns alternative versions of the transcript that have lower confidence levels. Customers can explore alternative transcriptions to gain more insight into candidate words and phrases that were generated for each audio input.

Privacy

Amazon Transcribe 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. Amazon Transcribe may use customer content to develop and improve the service. Customers can opt out of having their content used for service improvement via AWS Organizations or other opt-out mechanisms we may provide. When customers opt out, their content is not used for service development or improvement. Regardless of opt-out status, inputs and outputs are

never shared between customers. 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 Amazon Transcribe FAQs and Amazon Transcribe Security.

How Amazon Transcribe processes data. Amazon Transcribe processes audio input directly within the AWS infrastructure in the AWS Region where the customer is using the service. The processing flow differs by mode:

- **Batch transcription:** Customers specify their own Amazon Simple Storage Service (Amazon S3) storage location for both the input audio files and the output transcript files. Amazon Transcribe reads the audio from the customer-specified location, processes it, and writes the transcript output back to the customer-specified location. The customer controls the storage and retention of both input and output data.
- **Streaming transcription:** Audio is streamed to the service in real-time. The audio is processed on AWS server memory, and transcription results are returned to the customer in real-time within the same connection. The service does not separately store audio input or transcription output data at rest beyond the duration of the streaming session.
- **Speaker diarization:** When customers enable speaker diarization, the speech recognition engine analyzes voice characteristics in the audio to differentiate between speakers. These voice characteristic signals are used solely for the purpose of annotating the transcript with speaker labels and are not retained after processing is complete. Speaker diarization does not support speaker identity recognition and cannot track speakers across separate audio files.
- **Language identification:** When language identification is enabled, the service analyzes acoustic characteristics of the audio to determine the spoken language. This analysis is performed as part of the transcription process and no additional data is stored.

Data access. Only authorized personnel have access to content processed by Amazon Transcribe, in accordance with AWS security policies. AWS implements appropriate technical and physical controls, including encryption, designed to prevent unauthorized access to or disclosure of customer content. For more information, see AWS Data Privacy FAQs.

Security

Amazon Transcribe uses AWS owned encryption by default to protect data at rest. Customers can optionally use customer-managed keys via AWS Key Management Service (AWS KMS) for an additional layer of encryption. This allows customers to control the encryption keys used to protect their transcription output. Amazon Transcribe employs Transport Layer Security (TLS) 1.2 to encrypt data in transit, ensuring secure communication between the service and client applications.

Amazon Transcribe supports AWS PrivateLink for both Batch and Streaming APIs. This allows customers to access Amazon Transcribe privately from their Amazon Virtual Private Cloud (VPC) without using public IPs and avoid traffic traversing the public internet, enhancing network security. Amazon Transcribe integrates with AWS Identity and Access Management (IAM) to enable fine-grained access control of identity-based policies, resource-based policies and service-specific policy condition keys. Amazon Transcribe is integrated with AWS CloudTrail (CloudTrail), which logs all API calls made to the service. Amazon Transcribe integrates with Amazon CloudWatch (CloudWatch) to provide real-time monitoring of transcription jobs.

Amazon Transcribe has undergone application security review. The model is restricted, and its parameters cannot be inferred through external testing. Access controls, encryption, and monitoring mechanisms are implemented to protect against unauthorized access to the service. Customer data is isolated; inputs and outputs are never shared between customers. Amazon Transcribe has achieved ISO/IEC 42001:2023 Artificial Intelligence Management System accredited certification. For a complete list of compliance programs for which Amazon Transcribe is in scope, see AWS Services in Scope by Compliance Program.

Customers operating in regulated industries should review the specific compliance programs relevant to their industry and jurisdiction to confirm that Amazon Transcribe is in scope. Customers who are required to document their use of AI services for audit or regulatory purposes can reference this Service Card, the supporting documentation listed in the Further Information section, and may request additional detail through their AWS account team.

Transparency

Amazon Transcribe provides information to customers in the following locations: this Service Card, AWS documentation, AWS educational channels (for example, blogs, developer classes), and the AWS Console. We accept feedback through customer support mechanisms such as account managers. Where appropriate for their use case, customers who incorporate Amazon Transcribe in their workflow should consider disclosing their use of ML and ASR technology to end users and other individuals impacted by the application, and customers should give their end users the ability to provide feedback to improve workflows. In their documentation, customers can also reference this Service Card. Amazon Transcribe is a transcription service, not a generative AI system. Its outputs represent the service's best-guess transcription of the audio input provided. Where customers surface transcription outputs to end users (for example, as closed captions or meeting notes), we recommend clearly indicating that the text was generated by automated speech recognition, so that users can assess the output with appropriate expectations.

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, best practice development, dissemination, and training.

As with all AWS AI services, numerous stakeholders are involved in Amazon Transcribe's model lifecycle, including security, engineering, data science, product, data, responsible AI, and legal teams. Each model update undergoes quality readiness review before deployment, which includes evaluation of accuracy, fairness, robustness, and other responsible AI dimensions. For details on model lifecycle management, change control, ongoing monitoring, and update communication, see Model Updates under Deployment and Performance Optimization Best Practices below.

Customers who are required to validate AI or ML models as part of their governance processes may find the following information relevant. Amazon Transcribe performs speech-to-text conversion and does not generate novel content. The model assumes that input audio contains naturally occurring human speech in a supported language, recorded at a supported sample rate and format. Customers can use custom vocabularies, PII redaction, vocabulary filtering, confidence score thresholds, and human oversight as compensating controls to address model limitations for their specific use case. Word-level confidence scores are recalibrated during each model update cycle to ensure that scores accurately reflect transcription correctness and that confidence score distributions remain consistent, supporting the stability of downstream workflows that depend on specific thresholds. Amazon Transcribe is monitored by automated canary systems that periodically evaluate accuracy, latency, and other metrics against reference datasets, with alerts triggered on significant drift and rollback mechanisms in place. (See Model Updates and Ongoing Monitoring for details.) Customers should implement their own monitoring appropriate to their use case, including periodic retesting on representative data, tracking of output quality metrics, and review of edge cases. (See Performance Drift under Deployment and Performance Optimization Best Practices.)

Deployment and performance optimization best practices

We encourage customers to build and operate their applications responsibly, as described in AWS Responsible Use of AI Guide. This includes implementing Responsible AI practices to address key

dimensions including controllability, safety, fairness, veracity, robustness, explainability, privacy, security, transparency, and governance.

Workflow Design

The performance of any application using Amazon Transcribe depends on the design of the customer workflow, including the factors discussed below:

- **Effectiveness Criteria:** Customers should define and enforce criteria for the kinds of use cases they will implement, and, for each use case, further define criteria for the inputs and outputs permitted, and for how humans should employ their own judgment to determine the final results. These criteria should systematically address controllability, safety, fairness, and the key dimensions listed above.
- **Configuration:** Amazon Transcribe provides various configuration parameters to help customers achieve the best results for each feature:
  - **Sample rate:** Customers can specify the sample rate of their input audio. Amazon Transcribe supports audio from 8kHz (typical for telephony) to 48kHz (high-fidelity recording). Higher sample rates generally provide better audio quality and may improve transcription accuracy.
  - **Speaker diarization:** Customers can enable speaker diarization and optionally specify the expected number of speakers (up to 30). Specifying the correct number of speakers can improve diarization accuracy. If the number of speakers is not specified, the service will estimate it automatically.
  - **Language identification:** Customers can specify a set of candidate language codes for automatic language identification. Providing a smaller, more targeted set of candidate languages generally improves identification accuracy.
  - **PII redaction:** Customers can select which PII entity types to identify and redact, and choose whether to receive both the original and redacted transcripts (batch only) or only the redacted transcript. Customers should review the list of supported PII types for their language and mode to ensure coverage of the PII categories relevant to their use case.
  - **Custom vocabularies:** Customers can provide domain-specific terms and text to improve recognition accuracy for their use case. These should be tested and refined iteratively based on transcription results.
  - **Output formatting:** For supported languages, Amazon Transcribe automatically provides punctuation, capitalization, and digit formatting. Customers should review output formatting for their specific language to confirm it meets their requirements.

- **Recording Conditions:** Workflows should include steps to address variation in recording conditions, such as speaking far from the microphone or in noisy conditions. If variation is high, consider providing help and instructions that are accessible to all end users, and monitor recording quality by periodically and randomly sampling inputs.
- **Vocabulary Filtering and PII Redaction:** These optimizations can improve the security and privacy of the language produced in transcriptions. Vocabulary filtering enables customers to mask or remove words that are sensitive or unsuitable for their audience from transcription results, based on a customer-defined list. PII redaction enables customers to generate a transcript where PII has been removed, based on PII types that Amazon Transcribe identifies.

- **Human Oversight:** If a customer's application workflow involves a high risk or sensitive use case, such as a decision that impacts an individual's rights or access to essential services, human review should be incorporated into the application workflow where appropriate. ASR systems can serve as tools to reduce the effort incurred by fully manual solutions, and to allow humans to expeditiously review and assess audio content. Customers should define clear escalation paths for cases where transcription confidence is low or where errors could have significant consequences.
- **Consistency:** Customers should set and enforce policies for the kinds of workflow customization and audio inputs permitted, and for how humans use their own judgment to assess Amazon Transcribe outputs. These policies should be applied consistently, especially when being applied across demographic groups. Inconsistently modifying audio inputs could result in unfair outcomes for different demographic groups.
- **Performance Drift:** A change in the kinds of audio that a customer submits to Amazon Transcribe, or a change to the service, may lead to different outputs. For example, switching from high-quality studio recordings to noisy telephony audio, or expanding to a new language, may result in different accuracy characteristics. Because Amazon Transcribe models are continuously improved as part of the managed service, transcription output for the same audio may change over time even without a specific update notification. We recommend establishing a regular testing cadence using a representative dataset, and tracking key metrics (such as WER or F1) over time to detect any changes that may affect your workflow.

Model Updates and Ongoing Monitoring

As a managed AI service, Amazon Transcribe models are continuously improved by AWS to enhance accuracy, expand language support, and add new capabilities. Routine model improvements – such as incremental accuracy gains, robustness enhancements, and training data expansions – are deployed seamlessly as part of the managed service, and do not require customer action or

notification. For major updates that materially impact service behavior – such as the addition of new features, new language support, or fundamental architectural changes – AWS notifies customers and provides time to evaluate and adapt their workflows. Customers should consider periodically retesting the performance of Amazon Transcribe on their use cases, regardless of whether a specific update has been communicated, as cumulative improvements may result in changes to transcription output over time.

- **Model lifecycle management.** New models are developed, tested, validated, and deployed end-to-end by AWS. Customers do not need to manage model versions, training, or deployment themselves. AWS maintains backward compatibility in the API, so model updates do not require customer-side code changes. Customers continue using the same API endpoints and receive improved transcriptions seamlessly.
- **Change control.** All deployments follow standard AWS best practices, which include automated testing and rollbacks in the event of failures. Deployments are staggered across stages and regions, with changes validated at each stage before proceeding. One-box (single instance) deployments complete first, followed by integration tests and bake time before broader rollout. Functional tests after each stage block deployment on any failure, and auto-rollback mechanisms are configured to revert changes automatically if service availability or health metrics degrade. Extended bake times ensure sufficient data is collected before deciding to proceed to the next stage. Model updates go through a dedicated release workflow that validates accuracy and latency do not regress beyond defined thresholds.
- **Model update communication.** For major updates to Amazon Transcribe — such as the addition of new features, new language support, or architectural changes that materially impact service behavior — AWS communicates via launch announcements on the AWS What's New blog and in the Amazon Transcribe documentation history. New language launches, updated service quotas, and feature additions are also reflected in the Amazon Transcribe Developer Guide and the Supported Languages page. Routine model improvements that enhance accuracy without materially changing service behavior are deployed as part of the managed service and are not individually communicated.
- **Ongoing monitoring.** Amazon Transcribe follows a structured ongoing monitoring approach that includes regular internal evaluations for any models planned for launch, ensuring quality and performance standards are met before deployment. Automated canary systems continuously monitor service behavior, measuring accuracy and performance metrics against reference datasets and alerting on any significant drift. Amazon Transcribe maintains a published Service Level Agreement (SLA) that defines uptime and availability commitments, which is

actively monitored. AWS employs automated testing and rollback mechanisms to detect and remediate issues in production.

- **Responding to issues.** If model limitations or assumptions are found to no longer be valid — for example, through customer feedback, internal evaluation, or monitoring alerts — AWS may evaluate, test, train, validate, and re-deploy the model to remedy the identified limitations. For escalations, customers can reach out to AWS Support or to their account manager.

Further information

- For service documentation, see Amazon Transcribe Developer Guide.
- For a list of supported languages, see Supported Languages.
- For an example of a Contact Center Analytics workflow design, see Amazon Transcribe Call Analytics.
- 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 system behavior to reflect system design goals

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: Embedding best practices within the AI supply chain, including providers and deployers

Word Error Rate (WER): The industry standard metric for measuring speech-to-text accuracy. WER counts the number of incorrect words (insertions, deletions, and substitutions) identified during recognition, then divides by the total number of words in the reference transcript.

F1 (Word Recognition): A metric that evenly balances precision (the percentage of predicted items that are correct) against recall (the percentage of correct items that are included in the prediction). Values range from 0% to 100%, with higher values indicating better performance. Amazon Transcribe uses F1 scores to evaluate the accuracy of specific output elements such as custom vocabulary terms, named entities, digit transcription, punctuation, and capitalization.

Confidence Score: A value assigned to each word in a transcript indicating the service's confidence that the word was correctly recognized. Higher scores indicate greater confidence.

Speaker Diarization: The process of differentiating speakers in an audio input based on voice characteristics and annotating the transcript with speaker index labels.

Custom Vocabulary: A customer-provided list of domain-specific terms, brand names, or acronyms used to bias the model toward improved recognition accuracy for those words. This is distinct from vocabulary filtering, which masks or removes specified words from transcription output.

PII Redaction: The process of identifying and removing personally identifiable information from transcription output.

Compensating Controls: Mechanisms available to customers (such as custom vocabularies, confidence thresholds, PII redaction, vocabulary filtering, and human review) to address known model limitations for their specific use case.

Calibration: In the context of Amazon Transcribe, the process of ensuring that word-level confidence scores accurately reflect transcription correctness, and that confidence score distributions remain consistent across model updates.