Amazon Transcribe Toxicity Detection

AWS AI Service Cards

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS AI Service Cards: Amazon Transcribe Toxicity Detection

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Amazon Transcribe Toxicity Detection

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 release of Amazon Transcribe Toxicity Detection that is current as of November 27, 2023.

Overview

Amazon Transcribe Toxicity Detection is a feature of Amazon Transcribe that enables customers to detect audio content that may be considered toxic by human moderators, based on the scores assigned to an audio input. This AI Service card describes considerations for responsibly detecting potential toxicity within human speech for use cases that include voice chat in online gaming, social media and peer to peer dialogue platforms where customers have a need to augment their human content moderation. The feature is configured through the Transcribe::StartTranscriptionJob API for Automatic Speech Recognition (ASR) and by enabling the ToxicityDetection parameter. Transcribe Toxicity Detection uses both voice-based and text-based cues to identify overall toxicity, and toxicity across seven categories irrespective of whether it is coded by words or tone. For responsible use of Automatic Speech Recognition (ASR) see the Amazon Transcribe - Batch (English US) AI Service card.

Transcribe Toxicity Detection returns scores indicating the confidence that an audio input contains toxic content. The minimum confidence score for overall and categorical toxicity is 0.0, implying no likelihood of toxicity and the maximum score is 1.0, implying the highest likelihood of toxicity. The seven categories of toxic content that Transcribe Toxicity Detection classifies are: (1) Profanity, which refers to speech that contains words, phrases, or acronyms that are impolite, vulgar, or offensive. (2) Hate speech, which refers to speech that criticizes, denounces, or dehumanizes a person or group on the basis of an identity (such as race, ethnicity, gender, religion, sexual orientation, ability, and national origin). (3) Sexual, which refers to speech that indicates sexual interest, activity, or arousal using direct or indirect references to body parts, physical traits, or sex.

(4) Insults, which refers to speech that includes demeaning, humiliating, mocking, insulting, or belittling language. (5) Violence or threat, which refers to speech that includes threats seeking to inflict physical, emotional or psychological harm toward a person or group. (6) Graphic, which refers to speech that uses visually descriptive and unpleasantly vivid imagery. This type of language is often intentionally verbose to amplify a recipient's discomfort. (7) Harassment or abuse, which refers to interactions intended to affect the psychological well-being of the recipient, including demeaning and objectifying terms. Our categories are not mutually exclusive and overlaps are possible. An example of an overlap would be content that has specific language that is simultaneously classified as profanity, insults, and hate speech.

We assess the quality of Transcribe Toxicity Detection by measuring the accuracy of the system in detecting non-toxic instances (referred to as “non-toxic recall”) and toxic instances (referred to as “toxic recall”). When a speaker says something that contains examples of toxic and non-toxic content, as subjectively judged by a trained human content moderator, we expect the toxic content in the speech to have a high toxic recall and the non-toxic content to have a high non-toxic recall.

Classification of toxicity is subjective and it is possible for different individuals to have different perceptions of toxicity, especially when taking several modalities into account such as tone of voice, emotion, physical gestures, and facial expression. Toxicity is also context-dependent and may vary based on social setting and historical background. Transcribe Toxicity Detection operates on the human speech present in audio signals to detect the instances of toxicity where a majority of human content moderators would agree. It is designed to differentiate between the types of variations in the audio signals that help distinguish non-toxic and toxic intent (intrinsic) and the variations that should be ignored (confounding). Some examples of intrinsic variation include differences in (1) the use of vocabulary and slang; (2) grammar and pronunciation; (3) tone of speech. Some examples of confounding variation include differences in (1) dialects and accents; (2) background noise and echo; (3) recording devices; (4) overlapping speech. The system is trained on data specifically collected with these variations in mind to improve robustness.

Intended use cases and limitations

Transcribe Toxicity Detection enables content moderation across many applications including voice chat functionality. The service operates on US English speech contained in non-real-time audio files (batch mode). Transcribe Toxicity Detection cannot detect all types of toxic content. For example, categories like stereotype or child sexual abuse material (CSAM) are not supported. In our evaluations we calibrated confidence scores based on our test datasets, which included examples for all of the categories of toxicity that the system supports.

- **Voice Chat use case**: Voice chat applications use toxicity detection on single or multiple speaker audio to identify toxic content. Customers can use a threshold, which is setting a value between 0 and 1 to act as a decision point for toxicity if the confidence score exceeds the set value. This allows customers to create filters for overall toxicity and for each category. In this use case some customers may want to permit friendly banter in their voice chat applications, where context specific jargon and mild trash talk are distinguished from toxic content by setting thresholds at a level where users can speak more freely. Other customer applications may focus on maintaining a more inclusive environment by setting higher thresholds. In this use case there is high variation among speakers and context, so customers should carefully consider the appropriate thresholds for their application. Additional guidance on how to experiment with thresholds is included in the Workflow design section below.

Design of Transcribe Toxicity Detection

Machine learning

Transcribe Toxicity Detection is built using ML and ASR technology. It works as follows: (1) Segment the audio input and for each audio segment, extract the relevant acoustic features. (2) Generate a transcription associated with the audio input, and extract relevant text-based features. (3) Combine the acoustic and text-based features, alongside conversational context, to generate the confidence score for overall toxicity, and the confidence scores for the toxic content categories for all the segments. For more information, see Detecting toxic speech in the *Amazon Transcribe Developer Guide* and the Amazon Transcribe API Reference.

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 toxicity detection applications A and B. Application A enables voice chat for a popular online game that has multiple voices speaking at different levels of excitement per recording channel, variation in the quality of the microphones and loud background noise. Application B enables voice chat in a mobile social media app that has two speakers per channel with high quality microphones and negligible background noise. Because A and B have differing kinds of inputs, they will likely have differing accuracy rates, even assuming that each application is deployed perfectly using Transcribe Toxicity Detection. For model performance updates, customers can expect consistency and improvements with respect to accuracy metrics for their threshold choices.

Test-driven methodology

We use multiple 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. We measure toxicity detection performance by testing with evaluation datasets containing both toxic and non-toxic audio recordings. Groups in the dataset are comprised of speakers and their audio samples which can be defined by acoustic features (such as pitch, tone, intonation), demographic attributes (such as dialect, gender, age and ancestry), confounding variables (such as recording equipment varieties, the distance of each speaker from recording equipment, post-processing and background noises), 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 performance for Transcribe Toxicity Detection over our different evaluation datasets, takes steps to increase accuracy for groups on which the service performed least well, works to improve the effectiveness of the evaluation datasets to identify the performance of diverse speaker groups, and then iterates.

Fairness and bias

Our goal is for Transcribe Toxicity Detection to accurately identify toxicity in US English speech (a) that is sourced from a diverse range of speakers and (b) that targets a diverse range of identity groups. We consider speaker groups defined by regional dialects, such as Lowland South or New York City, language learning method such as native or non-native, and demographic groups such as ancestry, age, and gender. We test for content toxic towards women, people with mental and physical disabilities, and national origins, among other identity groups. 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, and human validated toxicity labels. As an example, on one dataset of non-toxic natural speech, comprised of unique speakers from 65 demographic groups, we find that the system correctly labels 94% or more of the examples as non-toxic, for every group and every intersection (e.g., Male + Asian ancestry) of speakers. We evaluate toxic speech on a synthetic dataset of toxic and non-toxic statements mentioning 13 identity groups. The justifications for using a synthesized dataset were to (1) strictly control the intrinsic variation of the spoken statements and; (2) avoid subjecting humans to the potential harm of reading and speaking highly offensive statements. On this synthesized dataset with

a threshold set at 0.6, we find that the system correctly labels the toxicity in 61% or more of individual statements (avg. statement length=20 words), for every identity group being mentioned. In a typical content moderation use case, this means that a speaker who uses toxic language within five conversational turns will be detected with a probability of 90%, while the probability of penalizing a benign speaker in the same number of turns remains low (23%); this assumes similar turn length in words to the length of statements from our test set. Because results depend on Transcribe Toxicity Detection, the customer workflow and the evaluation dataset, we recommend that customers test on their own content, calibrate thresholds for that content and validate the results with human content moderators.

Explainability

Transcribe Toxicity Detection returns start and end timestamps for the segments of the transcription that are detected as toxic content. Customers can use these start and end timestamps to listen to the segments of the input audio and verify the detection of toxicity.

Robustness

We rely on recall analysis on a wide variety of datasets to test the robustness of the system, evaluating how likely the non-toxic speech is identified as toxic. The feature is trained to be resilient under various acoustic environments, such as recording quality, background noise and room reverberation. By operating directly on audio, the system is designed to be robust to tone of voice and conversational context.

Privacy and security

Transcribe Toxicity Detection only processes audio input data. Audio inputs are never included in the output returned by the service. Audio inputs and service outputs are never shared between customers. Customers can opt out of training on customer content via AWS Organizations or other opt out mechanisms we may provide. 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.

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 Transcribe Toxicity Detection 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 Transcribe customers, who define the workflow where audio is captured from end users. Transcribe provides features for customers to optimize their recognition performance within the API. Confidence thresholds, 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. **Confidence Thresholds**: We recommend that customers experiment with performance for their content by starting with a threshold of 0.5 and incrementing or decrementing by 0.05. Customers with labeled evaluation datasets can calibrate thresholds for their specific use cases.
3. **Human oversight**: Human review should be incorporated into the application workflow where appropriate. An ASR system with toxic content classification is intended to serve as a tool to reduce the effort incurred by fully manual solutions, and to allow human moderators to expeditiously review and assess the audio in order to make a moderation decision.
4. **Consistency:** Customers should set and enforce policies for the kinds of audio inputs permitted, and for how humans use their own judgment to assess toxicity detection outputs. These policies should be consistent across all demographic groups. Inconsistently modifying audio inputs could result in unfair outcomes for different demographic groups.
5. **Performance drift:** Updates to the models powering the feature may lead to different outputs over time. To address these changes, customers should consider periodically retesting the performance of Toxicity Detection, and adjusting their workflow if necessary.

Further information

- For service documentation, see Transcribe Toxicity Detection.

- 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.