Detailed Use Case Description: ASR Evaluation Within Customer Boundary

Use Case

Axon's Automatic Speech Recognition (ASR) technology converts body-worn camera and other evidence audio into written transcripts that officers review and, when needed, correct before submitting them as legal evidence. This use case describes a privacy-protected process that uses those officer corrections, together with short matching audio clips, to improve Axon's speech-to-text models entirely within the customer's secure environment.

Improving speech-to-text for real body-camera audio requires learning from accurate, real-world examples. Longer-term programs that carefully prepare and annotate recordings remain an important path for building large training datasets. This use case complements those programs by comparing performance of release candidate models on customer — without asking customers for a separate annotation project, without Axon personnel reviewing transcripts or listening to audio, and without moving raw audio or raw transcript text outside the customer environment for this evaluation process.

This process is eyes-off and ears-off. Eyes-off means no Axon individual views customer transcript text. Ears-off means no Axon individual listens to customer evidence audio. Only automated systems process that content inside the customer environment.

Customer Benefit

This use case helps Axon improve speech-to-text quality in the real conditions where transcription is hardest — for example noisy body-camera audio — by identifying differences in transcripts and using models to classify whether those differences improve the transcript or not. Customers benefit from progress toward more accurate transcripts while remaining eyes-off and ears-off: raw evidence audio and full transcript content stay inside the customer environment for this process, are never reviewed or listened to by Axon personnel, and do not require a separate labeling project for each improvement cycle.

Data Accessed

For eligible use of this feature, Axon's automated systems may access the following data within the customer environment. No Axon individual reviews transcript text or listens to evidence audio as part of this use case: Audio and transcripts for evidence accessed by automation only, used solely to cut short audio clips and measure basic audio quality (for example how noisy the clip is). No individual listens to this audio.Separate evaluation materials that Axon places into the customer environment only to measure whether an improved model is more accurate.

How Axon Uses the Data

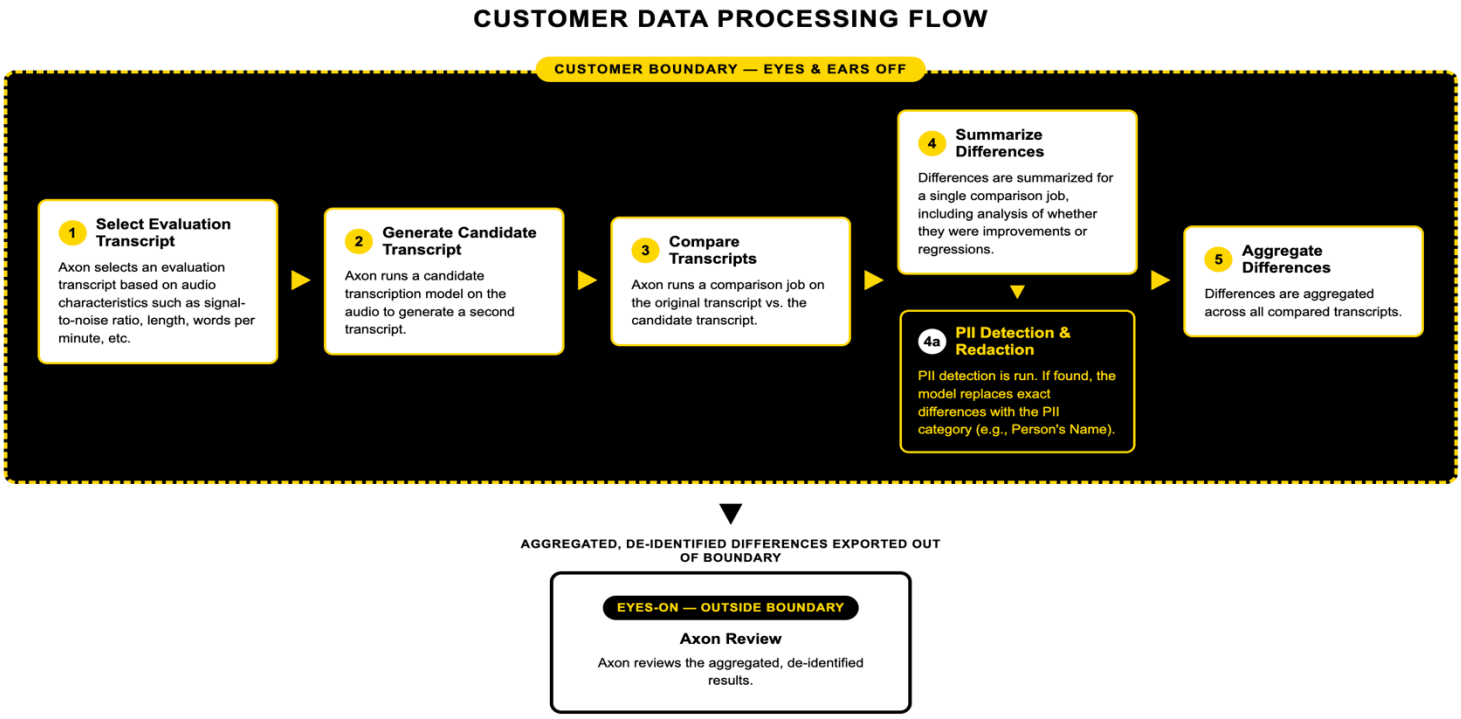
Axon uses this information to run an improvement process entirely within the customer environment. At every stage below, customer transcript text and customer evidence audio are handled only by automation; Axon personnel do not view or listen to that content:

- **Finding eligible examples**. Starting from evidences where a transcript was created and calculate simple quality numbers such as how noisy the audio is and how confident the original transcription was. No individual inspects transcript text or audio during this step. The result is (1) a restricted internal record used later by automation for evaluation and (2) a de-identified summary of numbers only — with no organization or evidence identifiers, no transcript text, and no audio — for Axon personnel to review.
- **Setting filters**. Axon personnel review only those de-identified numbers (for example audio noisiness, transcription confidence, clip length, and word counts) and choose simple limits such as how noisy a clip may be, allowed word-count ranges, and a maximum confidence level. Personnel do not see organization or evidence identifiers, do not see transcript text, and do not listen to audio.
- **Evaluating the candidate model**. Automation then uses the evidence audio to generate a second candidate transcript. Automation identifies areas for which the original and candidate transcript differs and classifies those differences in categories and provides an indication whether the difference should be considered an improvement or not. Those are then aggregated to identify themes which can be used to assess what areas the team should focus training efforts on in the future to further improve model performance.

Privacy-Preserving Technique

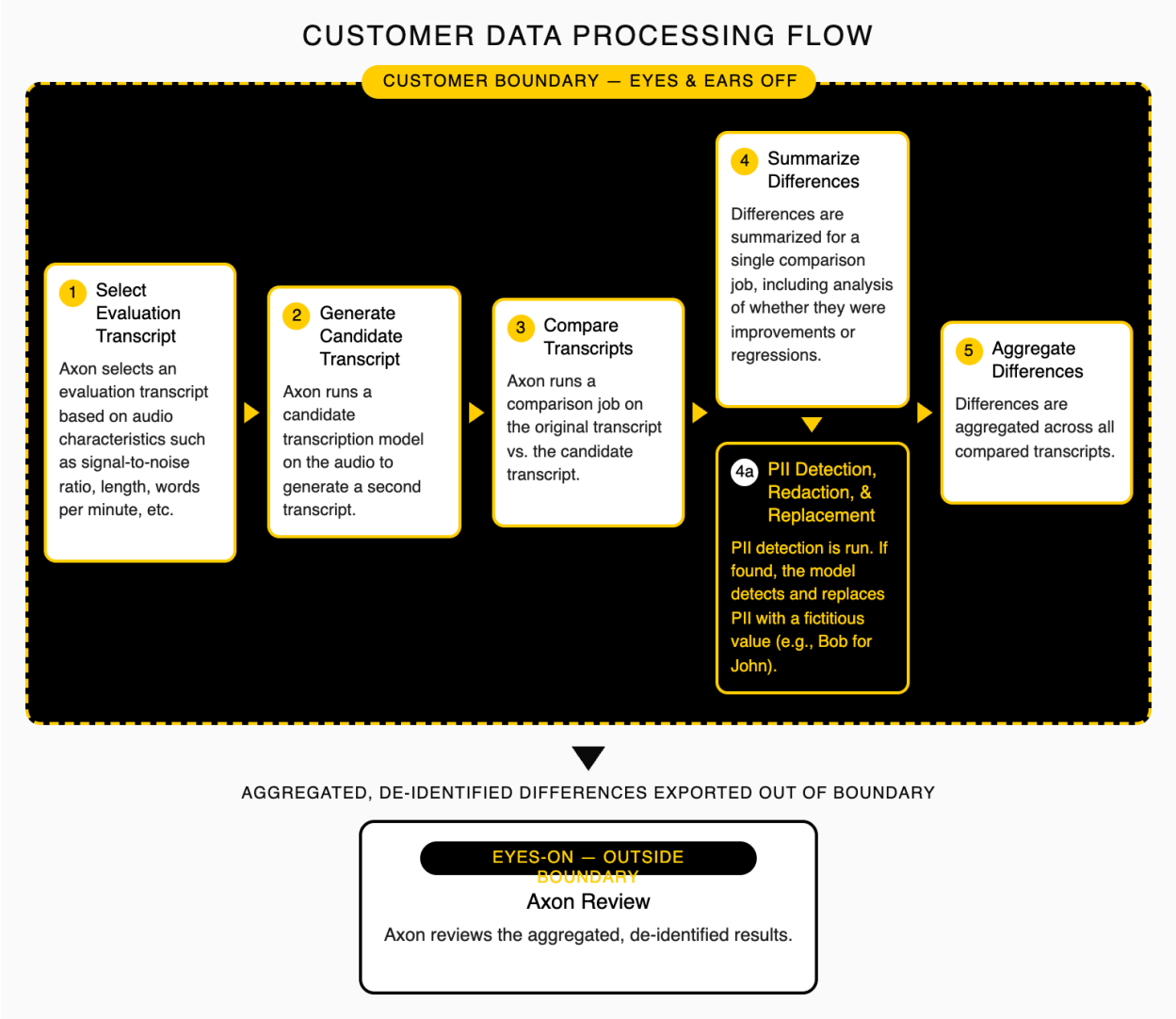
- All processing for this use case runs within the customer's secure Axon environment under ACEIP on an eyes-off, ears-off basis. Raw customer transcripts and evidence audio are accessed only by automated systems during discovery and evaluation. No Axon individual views raw transcript content or listens to audio at any.
- Audio and text clips used for evaluation are kept only briefly (for example up to 24 hours*), with no human access, are used only by automated processes, and are automatically deleted afterward whether evaluation succeeds or not.

- *Addendum 10/7/26: for example up storing transcripts for up to 14 days
- During evaluation text is automatically checked for personal information such as names, phone numbers, social security numbers, dates of birth, email addresses, physical addresses, badge numbers, case numbers, license plates, vehicle identification numbers, and other sensitive identifiers. If any personal information is detected the evaluation is then instructed to categorize the difference as the type of PII and not include the exact value. For example, if the difference was on 'John Doe' vs 'Jane Doe' the output would be Difference Category Type = Name. These are then further aggregated across all results to further remove a chance of privacy concern. *
- *Addednum 11/3/26: Any detected PII will first be redacted as above and then replaced with a fictitious new value, not found in the source text, e.g., 'John Doe' might be replaced with 'Bob Smith'. Aggregation then proceeds as above, examples in the aggregated report that have detected PII will be dropped.
- What may leave the customer environment for this use case is limited to (1) the de-identified numbers used to understand the evaluation mix and (2) evaluation results and aggregated findings on candidate generated transcripts (for



example overall agreement scores, list and frequency of types of differences between original transcripts and candidate transcript).

*Addendum 10/7/26:



Data Schema

The following categories of data are relevant to this use case. Only the de-identified numbers and evaluation results described below may leave the customer environment for Axon personnel review.

Restricted internal record used by automation (inside the customer environment; not accessible to individuals):

- **Organization ID:** an identifier for the customer organization associated with the evidence.
- **Evidence ID:** an identifier for the evidence recording.
- **Average transcription confidence:** a number reflecting how confident the original transcription was.
- **Number of speakers:** how many speakers appear in the transcript.
- ~~**Number of words edited:** how many words were corrected by a person.~~
- ~~**Number of segments:** how many short eligible portions contain corrections.~~
- **Words per minute:** a speaking-rate number for the evidence.
- ~~**Segment details:**~~
  - ~~start and end times~~
  - ~~duration~~
  - ~~word count~~

- ~~number of corrected words~~
- ~~the numbered index of the corrected word in this segment~~
- ~~a speaker label such as Speaker 0 or Speaker 1, not the actual speaker identity~~
- ~~an audio-quality number (how noisy the clip is)~~
- ~~average confidence score from the current production ASR model~~
- ~~the words in the clip with confidence score, correction flags and timing.~~

De-identified summary for Axon personnel (leave the customer environment; no identifiers, no transcript text, no audio):

- **Average transcription confidence:** a confidence number from the current production ASR model
- **Number of speakers:** a speaker-count number.
- ~~**Number of words edited:** a correction-count number.~~
- ~~**Number of segments:** a segment-count number.~~
- **Words per minute:** a speaking-rate number.
- ~~**Segment statistics:** length, word count, number of corrected words, edit type, audio-quality number, and confidence scores of the corrected words~~
- **Comparative statistics:** distribution and types of differences found within aggregated compared data. This includes categories that were programmatically identified and classified as improvements or regressions.

*Addednum 11/3/26: Number of words edited, number of segments and segment statistics will no longer be accessed and exported.

Temporary evaluation clips (inside the customer environment; short retention; not accessible to individuals):

- **Audio clip:** an audio portion used to generate a second candidate transcript.
- **Original Transcript:** the transcript which was originally generated to be compared against the candidate transcript.

Evaluation outputs:

- **Evaluation results:** accuracy and related scores computed. Algorithmic outputs of aggregated categories of error categories, counts, and statistics. Algorithm analysis of differences in regard to quality improvements or regressions. These may be shared with Axon personnel and do not include customer transcript text or customer evidence audio.

Examples of What May or May Not Be Extracted

See example below:

Original Customer Data	**De-identified Numbers for Human Review**

{
"organization_id":
"550e8400-e29b-41d4-a716-446655440000",
"evidence_id":
"123e4567-e89b-12d3-a456-426614174000",
"average_confidence": 0.82,
"number_of_speakers": 2,
~~"number_of_words_edited": 7,~~
~~"number_of_segments": 3,~~
"words_per_minute": 142,
~~"segments": [~~
~~{~~
~~"segment_id": "0",~~
~~"start_time_seconds": 1.51,~~
~~"end_time_seconds": 14.8,~~
~~"duration_seconds": 13.29,~~
~~"word_count": 18,~~
~~"edited_word_index": [2, 3],~~
~~"speaker_label": "0",~~
~~"audio_quality": 15.3,~~
~~"average_confidence": 0.71,~~
~~"modified_words_confidence":[0.82, 0.61],~~
	{
	"average_confidence": 0.82,
	"number_of_speakers": 2,
	~~"number_of_words_edited": 7,~~
	~~"number_of_segments": 3,~~
	"words_per_minute": 142,
	~~"segments": [~~
	~~{~~
	~~"segment_id": "0",~~
	~~"start_time_seconds": 1.51,~~
	~~"end_time_seconds": 14.8,~~
	~~"duration_seconds": 13.29,~~
	~~"word_count": 18,~~
	~~"edited_word_index": [2, 3],~~
	~~"audio_quality": 15.3,~~
	~~"average_confidence": 0.71,~~
	~~"modified_words_confidence":[0.82, 0.61],~~
	~~}~~
	~~]~~
	}

~~"words": [{ "text": "while", "corrected": true, "start_time_seconds": 5.5, "end_time_seconds": 5.6 }, { "text": "the", "corrected": false, "start_time_seconds": 5.7, "end_time_seconds": 5.8 }] }~~] }	

*Addednum 11/3/26: Number of words edited, number of segments and segment statistics will no longer be accessed and exported.

Comparison Data:

% of transcript exact agreement: 92%
% of transcript disagreement: 8%

Categorical Analysis:
10% of regressed differences are a result from non-inclusion of filler words (example: um, uhs, like, you know, etc.)
12% of regressed differences are a result from spoken numbers or acronyms (example: "twenty twenty six" vs "2026" or "A.S.R." vs "ASR").
20% of regressed differences are from informal contractions (example: "gonna" vs "going to", "wanna" vs "want to", "'cause" vs "because")
7% of regressed differences are from incorrect speaker identification based on context (example: misassigning turns during rapid cross-talk or background chatter)
5% of regressed differences are from incorrect word selection based on context (example: choosing homophones like "their" vs "there" or misinterpreting domain jargon)
3% of regressed differences are from incorrect Punctuation and Formatting (example: capitalization, comma placement, or sentence delimiters)

40% of the improved differences are from correct word selection based on context (example: properly distinguishing specialized terminology and phonetic equivalents in low-volume audio)
23% improvement on Temporal Alignment and Timestamping. (example: Differences in word-level timestamping logic, which can misalign segment boundaries)
17% improvement of Named Entity Recognition. (example: identifying proper nouns, technical product names, or specific locations)
14% improvement of Multilingual cases. (example: when speakers alternate between languages or dialects)
5% of the improved differences are from sections with overlap.
0% presence of Hallucinations and Repetitive Loops**. (**example: repetitive text patterns or insert 'phantom' words in low-confidence silent intervals where no speech exists.)

18% of the improved difference are from improved transcriptions of slang
The largest contributing factor to errors is when there are multiple speakers.
The second largest contributing factor to errors is when there are multiple speakers.

*Addednum 11/3/26: added a more detailed example of what is exported and what is not

120 recordings compared · 10,000 differences · average word error rate 3.8%
MODEL_2 finds more speakers than the reference in 82 of 120 recordings. Punctuation is missing from 20.8% of words and capital letters from 15.4%.

Types of differences

Type	Count	Share	Example
Filler words	3,300	33.0%	MODEL_1=[uh] MODEL_2=[(none)]
Speaker turns merged	1,570	15.7%	*none*
Non-speech sounds	810	8.1%	MODEL_1=[laugh] MODEL_2=[(none)]

Speaker turns split	600	6.0%	*none*
Number written differently	600	6.0%	MODEL_1=[one two] MODEL_2=[twelve]

Every example is checked for PII. A flagged example is **dropped whole**; if all examples for a type are dropped, the type shows none. Counts and shares are never affected.

What the example contained	Example	In the report?
A person's name	MODEL_1=[siobhan wieczorek] MODEL_2=[shivon vechorek]	**Dropped**
A street address	MODEL_1=[two one seven brookhaven] MODEL_2=[seventeen brook haven]	**Dropped**
Part of a phone number	MODEL_1=[three zero three five] MODEL_2=[(none)]	**Dropped**
A date of birth	MODEL_1=[three one nine eight] MODEL_2=[third nineteen eighty]	**Dropped**
Only a number	MODEL_1=[five] MODEL_2=[fifty]	Kept
Only everyday words	MODEL_1=[into] MODEL_2=[in to]	Kept

Full list of allowed Type of differences:

- **People's names**: a person's name is written differently.
- **Place names**: a street, city or business name is written differently.
- **Specialist terms**: a technical, medical or public-safety term is wrong.
- **Sound-alike words**: a word that sounds the same replaces the right one ("their" and "there").
- **Small connecting words**: an article, preposition or pronoun differs ("on" and "in").
- **Word endings**: the same word with a different tense, plural or possessive.
- **Missed speech**: spoken words in the reference are absent from one model.
- **Extra words**: a model has words the reference does not.
- **Filler words**: "uh" or "um" appears in only one.
- **Short responses**: a brief reply such as "uh-huh" appears in only one.
- **Non-speech sounds**: laughter, crosstalk or similar is marked in only one.
- **Informal wording**: slang on one side and standard wording on the other.
- **Contractions**: "don't" on one side and "do not" on the other.
- **Wrong number**: the value itself differs ("fourteen" and "forty").
- **Number written differently**: the same value is written differently ("one two" and "twelve").
- **Dates, times, money**: a date, time, amount or unit is written differently.
- **One word split in two**: "backup" and "back up".
- **Two words joined**: "a lot" and "alot".
- **Punctuation missing**: the reference has punctuation that the other leaves out.
- **Capital letters missing**: the reference has capitals that the other leaves out.
- **Speaker turns merged**: separate turns in the reference are joined into one in the other.
- **Speaker turns split**: one turn in the reference is broken into several in the other.
- **Number of speakers differs**: the two find a different number of speakers.
- **Right words, wrong speaker**: the same words are given to a different speaker.

Retention Period

Restricted internal records used by automation are kept only as long as needed to select examples and run evaluation (up to 30 days so enough eligible examples can accumulate, and filter limits can be set before evaluation runs), then automatically deleted. Those records remain eyes-off and ears-off for their entire lifetime — no individual access. Temporary audio and transcripts used for evaluation are kept only for the short period needed to finish an evaluation run (for example up to 24 hours*), likewise with no human access, and are automatically deleted afterward whether evaluation succeeds or not. De-identified numbers and evaluation results from evaluation may be kept for a limited period no longer than necessary to support model improvement for

this use case. Customer evidence audio and customer transcript text are not kept outside the customer environment for this process and are never kept in a form available for human viewing or listening.

- *Addendum 10/7/26: for example up storing transcripts for up to 14 days

Preservation of Original Content

This use case does not modify the customer's original evidence, source transcripts, or live product behavior. Original evidence and transcripts in customer systems remain unchanged. Automated systems read transcript and audio content temporarily within the customer environment for finding examples, generating a second candidate transcript and comparison; this use case does not change how transcription works for officers in production. Turning on an improved model for end users is outside the scope of this description.

Withdrawal

Customers participating in this use case may request withdrawal. After withdrawal, Axon would stop collecting new data for this use case from that customer and would stop including that customer's evidence in this improvement process. Previously generated de-identified numbers, evaluation results, and model improvements that no longer identify the customer, user, or original evidence may continue to be retained where permitted.

Summary

This use case helps Axon improve ASR models by evaluating candidate model transcripts on evidence audio, entirely within the customer environment under an eyes-off, ears-off model. Automated systems access transcript and audio content only as needed to generate a candidate transcript, select clips, calculate quality numbers, and run privacy preserving evaluation — with no Axon individual viewing transcript text or listening to customer audio. Axon personnel review only de-identified numbers and evaluation results from Axon's own evaluation materials. Raw customer evidence audio and customer transcript text used for evaluation do not leave the customer environment; candidate transcripts are short-lived, inaccessible to individuals, and automatically deleted; and this use case does not modify original customer evidence or live transcription behavior.