

Project Ginkgo

Contributor Guidelines

Version 1 01 July 2026

Table of Contents

Overview.....	3
Project Terminology and Definitions.....	3
Task Instructions	4
Step 1 — Listen Before You Type	4
Step 2 — Spoken-Form Transcription Rules.....	5
Step 3 — Written-Form Transcription Rules.....	8
Step 4 — Save State	11
Step 5 — Number of Identified Speakers	12
Step 6 — Speaker Details: Gender and Nativity	13
Quality Standards	14
Accuracy	14
Completeness.....	14
Reliability	14
Edge Cases and Escalation.....	14
Decision Rules	14
Acronyms and Initialisms.....	15
Non-Target Languages.....	16
Song Lyrics	16
Appendix — Tag Quick Reference.....	17
Backchannel Words.....	18
Spoken-Form vs. Written-Form Quick Examples	19

Overview

In this project you will listen to audio recordings and produce accurate text transcriptions. These transcriptions help evaluate and improve Automatic Speech Recognition (ASR) models. Every task has two required formats: a spoken-form transcription (lowercase, no punctuation) and a written-form transcription (standard capitalization, punctuation, and digits). Accurate, consistent transcriptions across both formats improve model quality. Every sound in the audio must be accounted for — transcribed as words or captured with a tag; nothing should be left out.

Each task requires you to complete the following six steps:

1. Listen to the audio file in full before typing anything.
2. Provide a spoken-form transcription following the tag and formatting rules.
3. Provide a written-form transcription of the same audio in the separate field.
4. Select the correct save state — Good or Discard.
5. Indicate the number of speakers in the audio.
6. Indicate gender and nativity for each identified speaker, then submit.

Project Terminology and Definitions

The following terms are used throughout these guidelines.

Label	Definition	Examples
Spoken-Form	Lowercase-only transcription with no punctuation except apostrophes and hyphens. Uses words for all numbers. Currency and symbols (\$, %, &) are also written as words.	play music please set alarm for five minutes it costs fifty dollars
Written-Form	Same content and tags as spoken-form but with standard capitalization, punctuation, and numbers as digits.	Play music, please. Set alarm for 5 minutes. It costs \$50.
Backchannel	A brief listener response (mhm, uh-huh, hmm, oh) that signals attention without taking a full turn. Transcribed as words, not as [fp].	mhm uh-huh mm-mm nuh-huh
Filled Pause [fp]	A non-word hesitation sound not covered by the backchannel table. Placed exactly where the pause occurs.	can you [fp] play some music i was [fp] thinking
Unintelligible (())	Speech that cannot be understood or transcribed. Guessed content goes inside: ((guessed word)). Multiple unintelligible words can be grouped in a single (()). Mark any word requiring repeated slow listens to confirm.	play (()) music play ((songs)) repeat after me (())

Cross Talk	Unintelligible, overlapping speech from multiple speakers. Use <ct> tag.	play <ct> don't listen to her
Catalog Entity	A titled or branded item the tool could search for: songs, albums, artists, apps, movies, books, cities. Transcribe per official spelling; normalize non-pronounceable symbols.	panic at the disco t-pain oral-b u. two in n. out burger
Non-Target Language	Any language other than the expected locale language. Attempt to transcribe; use (()) for unintelligible portions. Discard only if the entire utterance is unintelligible non-target language.	quiero taco bell wie geht es dir (()) (fully unintelligible)
Good (Save State)	Select when the file contains at least some intelligible target-language speech from the main speaker, or background speech only, or a non-target language you were able to transcribe.	Partial intelligible speech non-target language fully transcribed only non-lexical vocal sounds
Discard (Save State)	Select when none of the main speaker's speech is usable or the audio has a technical failure. Seven discard reasons — see Step 4.	Garbled audio no signal entirely unintelligible humming only sensitive content no speech non-target language that cannot be transcribed

Task Instructions

Step 1 — Listen Before You Type

Audio Transcription

Provide a **spoken-form** and **written-form** transcription for each audio file. Follow the conventions exactly and use the Tag Reference panel on the right.

1

Listen to Audio

Replay as many times as needed — do not use ASR predictions.

▶ 0:00 / 0:12

Replay

3s

Play

3s

Figure 1: Step 1 — Listen before you type.

What you'll see: an audio player with Replay, back 3s, Play, and forward 3s controls.

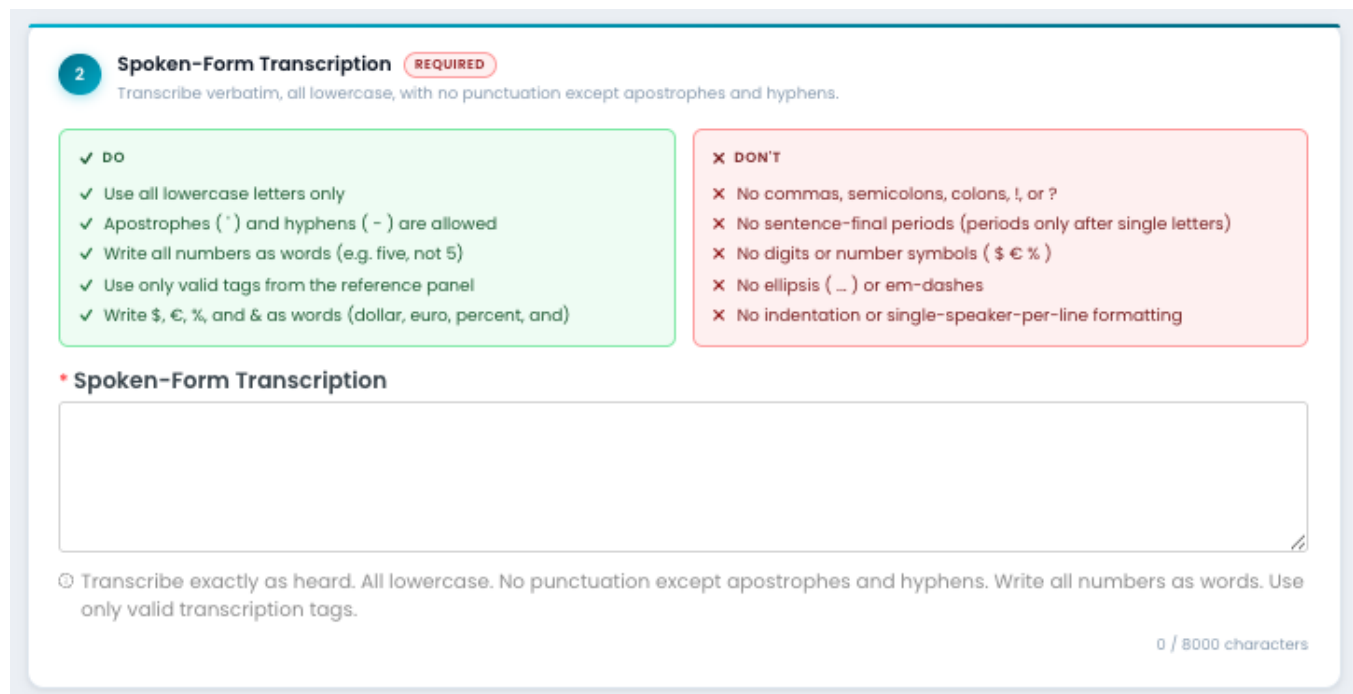
What you do:

- Play the audio file in full before typing anything.
- Replay as many times as needed.
- **Do NOT** use any ASR (Automatic Speech Recognition) or auto-transcription predictions to produce your answer.

Edge cases / what to skip

- If the audio fails to load or you cannot process it for any reason, select Discard in Step 4 and choose the appropriate discard reason.
- If the content is sensitive and you choose not to process it, select Discard and choose the 'Sensitive content' reason.

Step 2 — Spoken-Form Transcription Rules



2 Spoken-Form Transcription REQUIRED

Transcribe verbatim, all lowercase, with no punctuation except apostrophes and hyphens.

✓ **DO**

- ✓ Use all lowercase letters only
- ✓ Apostrophes (') and hyphens (-) are allowed
- ✓ Write all numbers as words (e.g. five, not 5)
- ✓ Use only valid tags from the reference panel
- ✓ Write \$, €, %, and & as words (dollar, euro, percent, and)

✗ **DON'T**

- ✗ No commas, semicolons, colons, !, or ?
- ✗ No sentence-final periods (periods only after single letters)
- ✗ No digits or number symbols (\$ € %)
- ✗ No ellipsis (...) or em-dashes
- ✗ No indentation or single-speaker-per-line formatting

*** Spoken-Form Transcription**

0 / 8000 characters

Figure 2: Step 2 — Spoken-form transcription.

What you'll see: a Spoken-Form Transcription box with Do/Don't reminder lists above it and a character counter below.

What you do: type your spoken-form transcription in the box, following the rules and tags below.

Quick reference: Lowercase only | Numbers as words | Apostrophes & hyphens only | Tags where they occur | Speech exactly as heard | Every sound tagged or transcribed

Topic	Rule
Capitalization	Use lowercase only. No capitals anywhere.
Punctuation	Omit all punctuation except apostrophes (in contractions and possessives) and hyphens. Use hyphens in: Truncations (cut-off words): elev- Dictionary-hyphenated words: well-known, state-of-the-art Person's names: jean-luc picard Place names: moreton-in-marsh Single-letter + word brand names: t-pain, oral-b Made-up words in branded titles: ob-la-di
Numbers	Write all numbers as words: five, eighty-seven, fifth, three hundred and six.
Currency & symbols	Write currency symbols (\$, €, %), the ampersand (&), and all other special symbols as words: fifty dollars, twenty percent, and.
URLs & emails	Transcribe URLs and email addresses as individual spoken words: 'dot' for periods, 'at' for @, 'slash' for /. E.g., w. w. w. dot amazon dot com; janesmith at gmail dot com.
Speech fidelity	Transcribe speech exactly as spoken, including grammatical errors, repetitions, and non-standard grammar (e.g., 'me name is William' stays 'me name is William').
Mispronunciations	Normalize mispronounced words to the standard dictionary spelling when the intended word is clear (e.g., mushic becomes music, liberry becomes library, expresso becomes espresso).
Ambiguous intended word	If you can hear the sounds clearly but are genuinely unsure what standard word was intended (i.e., more than one normalization is plausible), transcribe exactly what you hear rather than guessing a normalization. This differs from Unintelligible (()), which is for sounds you cannot make out at all.
Extended syllables	Do NOT add extra letters for extended syllables (e.g., 'haaaaate' becomes 'hate').
Tags	Insert tags from the Tag Reference exactly where each feature occurs in the audio.
Informal contractions	Informal contractions are transcribed as written, not expanded: gimme, gonna, gotta, lemme, wanna, whatcha (plus any listed in the American Heritage Dictionary, e.g., c'mon, 'cause). Any informal contraction not on this list should be transcribed as standard words (e.g., "whaddya" → what do you).

Edge cases / what to skip

- Transcribe backchannels (mhm, uh-huh, mm-mm, hmm, oh, ooh, uh, um) as words, not as [fp].
- Always transcribe 'OK' or 'okay' as ok (lowercase, no space, no period).
- Catalog entities are always lowercase in spoken-form, regardless of official branding. E.g., SZA → sza, T-Pain → t-pain, PewDiePie → pewdiepie. Official casing is applied in written-form only.
- Double contractions (three-word combinations like 'I'd've' or 'mustn't've'): expand only the final contraction. E.g., 'I'd've done better' → i'd have done better; 'you mustn't've heard me' → you mustn't have heard me.

- **Acronyms:** For rules on acronyms and initialisms (e.g., NASA vs. F-B-I), see [Acronyms and Initialisms](#) in [Edge Cases and Escalation](#).

Tag Reference: Use only the tags listed in the Appendix. For rules on tag placement and usage, see below; for the full tag list, see [Appendix — Tag Quick Reference](#).

Topic	Rule
No new tags	Do not invent new tags.
Tag placement	Place each tag exactly where the feature occurs in the audio (not at the end, unless specified).
Tag spacing	Separate tags from adjacent text and from each other with a single space.
Multiple speakers	For multiple speakers: assign <s1>, <s2>, etc. in order of first appearance. Always retag speakers inside blocks.
Overlapping speech tags	Use for intelligible overlapping speech; use <ct> for unintelligible overlap.
Standalone tags	Tags that are the sole content of a transcription (<ga>, <na>, <ns>, <ct>, <hum>) require no other text. A lone period (.) is used the same way for an extremely truncated word that is the entire transcription.
End-placement tags	End-placement tags (e.g., <pws>) are inserted only once per audio file, at the very end. If multiple end-placement tags apply, insert them in alphabetical order (e.g., <hum> before <pws>).
Tag Reference panel	The Tag Reference panel on the right side of the task interface is clickable — select any tag to insert it at your cursor position. Keyboard shortcuts are also available (e.g., Alt+L for [laughter], Alt+F for [fp], Alt+U for (())).

More edge cases for tags

- Multiple consecutive non-lexical sounds of the same type are transcribed with a single tag — e.g., two coughs in a row = [cough] once. This applies to all non-lexical sound tags, not just [fp].
- Previously identified speakers: if a speaker you have already identified becomes temporarily unintelligible, keep their speaker tag — do not switch to [bg]. Only use [bg] when you genuinely cannot determine who is speaking.
- If you cannot determine whether a noise came from the main speaker or another speaker, assume it came from the main speaker.
- If background speech includes non-lexical sounds from unidentifiable speakers, use [bg] [hn] together.

Step 3 — Written-Form Transcription Rules

3 Written-Form Transcription REQUIRED

Same content as spoken-form, with capitalization, punctuation, and digits applied.

✓ **DO**

- ✓ Capitalize the first word of each sentence and all proper nouns
- ✓ Apply standard punctuation: periods, question marks, commas
- ✓ Write all numbers as digits (e.g. 5, not five)
- ✓ Use currency symbols with digits (\$5, €10, 60¢)
- ✓ Carry over every tag in its identical position
- ✓ Always capitalize the pronoun "I"

✗ **DON'T**

- ✗ Do not add, remove, or clean up any speech content
- ✗ No exclamation marks (!)
- ✗ Do not change tag types or positions

↓ **Copy from spoken-form** *Press the button to copy the spoken-form text here, then add capitalization and punctuation.*

*** Written-Form Transcription**

○ Apply capitalization and punctuation to the spoken-form. Write numbers as digits. Keep all tags in identical positions. Do not alter the speech content.

0 / 8000 characters

Figure 3: Step 3 — Written-form transcription.

What you'll see: a Written-Form Transcription box with its own Do/Don't lists and a "Copy from spoken-form" button above it.

What you do: in written-form, use the "Copy from spoken-form" button to pull in your text, then apply the formatting rules below. Do not change or clean the speech itself.

Quick reference: Capitalize sentences & proper nouns | Numbers as digits | Standard punctuation | Tags in identical positions | Do not alter speech content

Addresses, Phone & Names

Format	Rule
Non-English titles	Spell out in full in both forms (e.g., Madame, not Mme). Do not abbreviate.
Honorifics	Use the standard abbreviation with a period (e.g., mister → Mr., missus → Mrs., doctor → Dr., miss → Ms.)

Catalog entities	Some catalog entities use the abbreviated form (e.g., “Mr. Mister” the band), others use the full word. Follow the entity’s official spelling rather than the general rule.
Addresses	Comma after street address, comma after city, no comma between state and ZIP. Use USPS abbreviations (Ave, St, Blvd, Dr).
Phone numbers	555-1234 (7-digit) or (555) 555-1234 (10-digit). ZIP codes: no spaces (e.g., 90210).

Numbers & Symbols

Type	Rule
Measurements	Metric: standard abbreviations (80 km, 80 m). Imperial: spell out (5 miles, 3 pounds). No plural s on metric abbreviations.
Large numbers	Use commas as thousands separators: 1,234 / 12,500 / 1,200,000.
Cents	\$0.XX by default. Use ¢ only when multiple cents-only amounts appear in the same sentence. When dollar amounts also appear, use \$0.XX.
Currency & symbols	Use \$, €, %, standard address abbreviations with digits.
URLs & emails	Collapse to standard format: w. w. w. dot amazon dot com → www.amazon.com; janesmith at gmail dot com → janesmith@gmail.com.
Letter sequences	Collapse to uppercase no-space: a. b. c. → ABC; o. p. e. c. → OPEC.
Arithmetic	Use standard symbols: five plus five → 5 + 5; ten minus three → 10 – 3.
Fractions	Numeric form: one half → 1/2; three quarters → 3/4; two thirds → 2/3.
Numbers	Write as digits: 5, 5th, 87, \$50, 50%, 3:30, 7 A.M. Space between digit and A.M./P.M.
Letter “o” vs. digit 0	In a numeric or code sequence, transcribe a spoken “o” as the digit 0: five o five → 505. In a spelled-out name, transcribe it as the letter O: j o h n → J O H N.
Decimals	Write decimal numbers with a period between the whole and fractional parts, digits as spoken: forty-two point six three → 42.63.

Punctuation

Topic	Rule
"And" comma patterns	(a) comma before "and" when joining two independent clauses — I went to the store, and she stayed home. (b) comma after "and" for a parenthetical — I called John and, surprisingly, he answered. (c) no comma when a single subject has two verbs — Play music and set a timer.
Commas before <i>	Place a comma before each <i> tag in written-form. E.g., set alarm for monday <i> wednesday and friday mornings → Set alarm for Monday, <i> Wednesday, and Friday mornings.
Backchannels	Set backchannels off with commas when they appear mid-sentence. Capitalize and follow with a comma when they begin a sentence. E.g., mm-mm that's not what i said → Mm-mm, that's not what I said. / uh-huh i'll do it → Uh-huh, I'll do it.
Double quotes	Use only when the utterance itself signals a direct quote (e.g., 'he said,' 'she told me,' 'and I was like'). Capitalize the first word inside the quote if it begins a full sentence. Place periods and commas inside the closing quote mark. If it is ambiguous whether the speaker is quoting, do not use double quotes.
Colons	Colons are permitted for introducing lists, time notation, and explanations — e.g., Here's what I need: eggs, milk, and bread.
General	Add standard English punctuation: periods, question marks (only for syntactic questions), commas. Do not use exclamation marks in spoken-form or written-form — transcribe excited or emphatic speech with a period, regardless of the speaker's intonation.

Capitalization

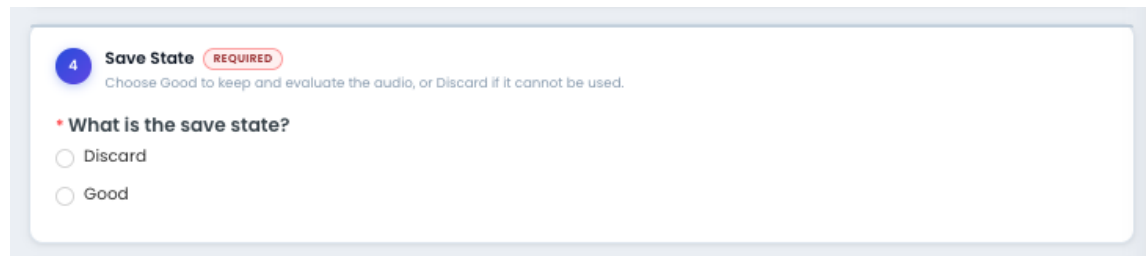
Topic	Rule
Pronoun 'I'	Always capitalize the pronoun 'I', regardless of its position in the sentence.
Sentences & proper nouns	Capitalize the first word of every sentence and all proper nouns. Catalog entities use official branded casing (SZA, PewDiePie, T-Pain).

Edge cases / what to skip

- Truncated numbers that you cannot write as digits stay in spoken-form (e.g., 'elev-' stays as 'elev-').
- Tags do not absorb punctuation — place the period or comma on the speech side of any tag.
- Capitalize the first word after every speaker tag. Use . . . only when another speaker begins before the current speaker finishes a grammatically complete sentence, or the sentence is syntactically incomplete at the point of overlap. Do NOT use . . . when the speaker completes a full thought before the next speaker begins.
- When 'you know' is a filler, use a comma — not a question mark. When it is a genuine question, use a question mark.

- Standalone tag-only transcriptions (<ga>, <na>, <ns>, <ct>) are identical in both forms — no punctuation added.
- When a period (.) is the entire transcription content — used for an extremely truncated word — it is written identically in spoken and written form. No capitalization or punctuation is added.

Step 4 — Save State



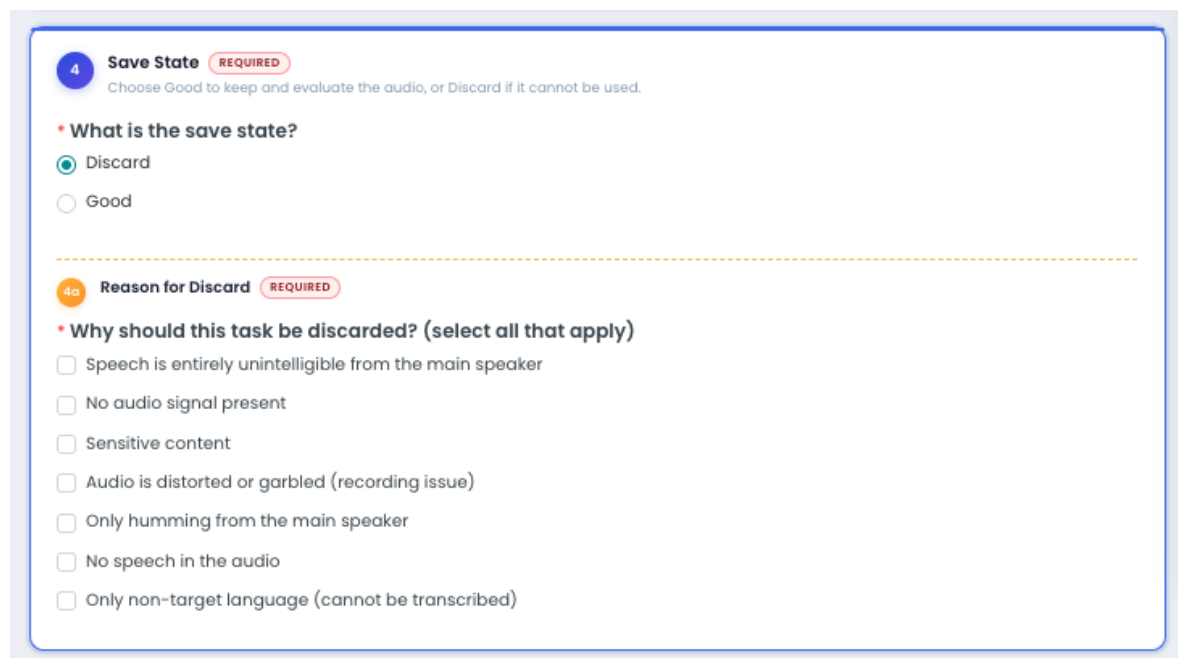
The screenshot shows a task interface for 'Step 4 — Save State'. It includes a sub-header '4 Save State' with a 'REQUIRED' tag and a instruction 'Choose Good to keep and evaluate the audio, or Discard if it cannot be used.' Below this is a question 'What is the save state?' with two radio button options: 'Discard' and 'Good'.

Figure 4: Step 4 — Save State (1 of 2).

What you'll see: Discard and Good as your two choices; picking Discard opens a checklist of reasons where you can select more than one.

What you do: choose Good or Discard based on the audio; if you discard, select all reasons that apply.

- **Good:** select when the file contains at least some intelligible target-language speech from the main speaker, or background speech only, or non-lexical vocal sounds only.



The screenshot shows the continuation of the 'Step 4 — Save State' interface. It features the same '4 Save State' header and instruction. Below the 'What is the save state?' question, the 'Discard' radio button is selected. A dashed line separates this from the next section, '4a Reason for Discard', which also has a 'REQUIRED' tag. This section asks 'Why should this task be discarded? (select all that apply)' and lists seven checkbox options: 'Speech is entirely unintelligible from the main speaker', 'No audio signal present', 'Sensitive content', 'Audio is distorted or garbled (recording issue)', 'Only humming from the main speaker', 'No speech in the audio', and 'Only non-target language (cannot be transcribed)'.

Figure 5: Step 4 — Save State (2 of 2).

- **Discard:** when selected, select all reasons that apply from the table below.

Discard Reason	When to Select	Transcription Field
Speech entirely unintelligible	Main speaker's speech cannot be understood at all	(())
No audio signal	No audio signal is present in the file	<na>
Sensitive content	You choose not to process the content	Transcribe if possible, or write DISCARD
Audio distorted / garbled	Recording issue makes transcription impossible	<ga>
Humming only	Audio contains only humming from the main speaker	<hum>
No speech	Audio exists but contains no human speech	<ns>
Non-target language	Entire utterance is in a language you cannot transcribe	(())

Step 5 — Number of Identified Speakers

5

Number of Identified Speakers
REQUIRED

Use <s1> <s2> for 2 or more speakers. Do NOT use <s1> for single-speaker audio. Select 0 if no speaker can be identified (or when discarding).

*** How many identified speakers does the audio contain?**

☐ 0
 ☐ 1
 ☐ 2
 ☐ 3
 ☐ 4 or more

Figure 6: Step 5 — speaker count.

What you'll see: a choice of 0, 1, 2, 3, or 4 or more speakers.

What you do: Select the number of speakers you can identify in the audio.

- Select **1** for single-speaker audio — even though you do not use the <s1> tag in the transcription.
- Select **0** only when no speaker can be identified at all, or when discarding. Selecting 0 suppresses the gender and nativity fields.
- Assign speakers in order of first appearance: the first speaker heard is Speaker 1, the second is Speaker 2, and so on.

Step 6 — Speaker Details: Gender and Nativity

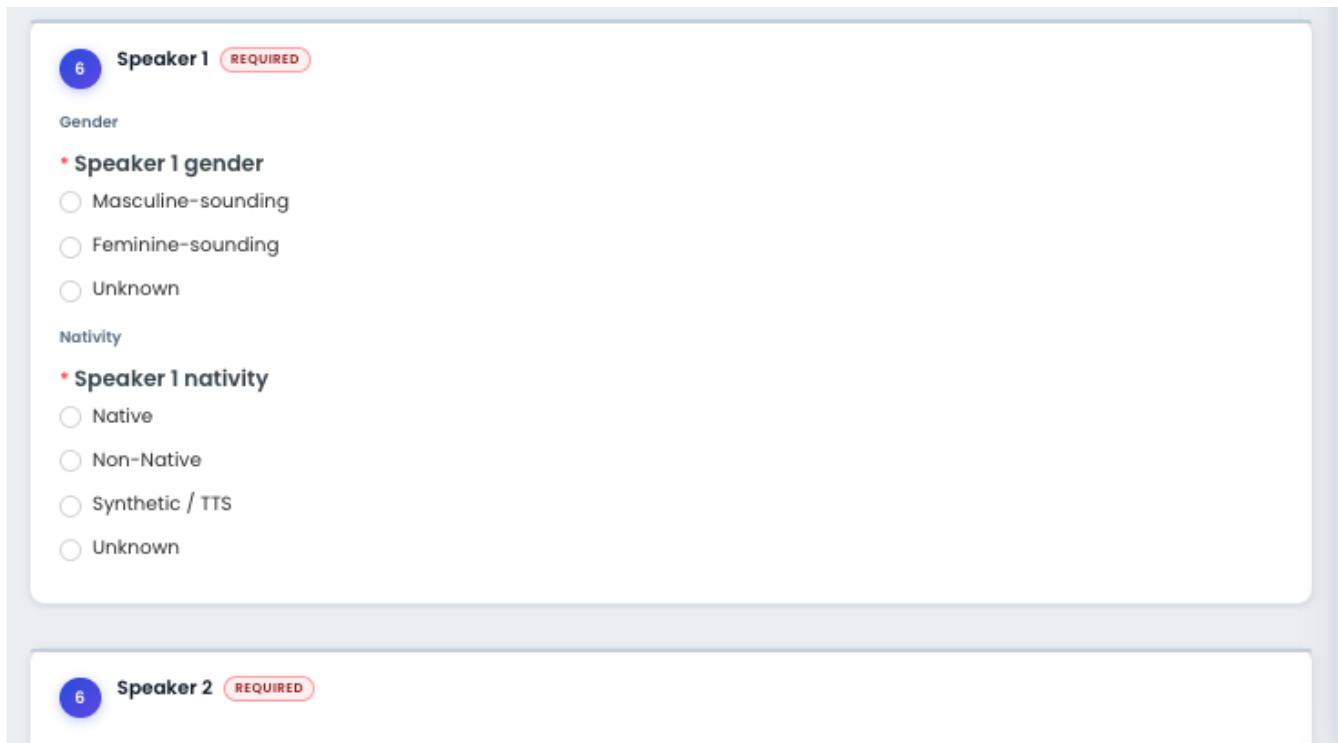


Figure 7: Step 6 — speaker gender and nativity.

What you’ll see: one gender-and-nativity card per speaker you identified — or one shared card for “Speaker 4 (and beyond)” if you selected 4 or more.

What you do:

- **Gender:** select Masculine-Sounding, Feminine-Sounding, or Unknown for each identified speaker. Select Unknown for toddlers or voice-distorted audio.
- **Nativity:** select Native, Non-Native, Synthetic / Text-to-Speech (TTS), or Unknown for each identified speaker. Select Unknown when nativity cannot be determined or when both types are present.
- **Do NOT** complete Gender or Nativity fields for Discarded tasks.

Edge cases / what to skip

- A speaker who produces only non-lexical sounds (no intelligible speech) gets Unknown nativity.
- Computer-generated or TTS agent utterances: select Synthetic / TTS for nativity.
- If you selected “4 or more” speakers, the Speaker 4 card covers every speaker from #4 onward: use a specific gender or nativity label only if all of them share it. If they differ, select Unknown.

Quality Standards

Accuracy

Transcribe every word the main speaker says exactly as spoken, with correct tag placement. Normalize mispronounced words to dictionary spelling. Catalog entities use official spelling. Spoken-form uses lowercase only; written-form uses correct capitalization, punctuation, and digits. Adding words, dropping words, and substituting words are all flagged as errors, along with a missed tag or incorrect save state.

Completeness

Both the spoken-form and written-form fields must be completed for every task. All identified speakers must be tagged. Every sound in the audio must be accounted for: if it is transcribable speech, write it as words; if it is not, it must be captured with the correct tag from the Appendix. No audible sound — speech, non-lexical vocal sound, or background noise — should be left out of the transcription.

Reliability

Apply rules identically across all tasks. The same utterance must receive the same transcription from any contributor. Backchannel spellings, tag placement, and catalog entity normalization must be consistent. Inter-annotator agreement is monitored; tasks that repeatedly diverge from the calibration set will be flagged for review.

Edge Cases and Escalation

Decision Rules

Topic	Rule
Words needing repeated listens	If a word requires multiple slow listens to confirm, mark it as unintelligible (()) rather than guessing — unless you have a reasonable guess, in which case use ((guessed word)).
Overlapping speech turns unintelligible	If overlapping speech is intelligible in places and unintelligible in others, close the tag at the point where it becomes unintelligible and follow with <ct>.
Cross-language homophones	If a cross-language homophone appears in isolation, assume it is the target language.
Catalog entities with symbol letters	If a catalog entity contains a symbol that replaces a letter (e.g., Ke\$ha), normalize to the letter in spoken-form (kesha) and keep the symbol in written-form (Ke\$ha).

Topic	Rule
Numbers in catalog entity names	Numbers inside catalog entity names stay as words in spoken-form: “I Would Die 4 U” → i would die four u. / “U2” → u. two.
Unpronounceable branded forms	Unpronounceable branded forms normalize to pronunciation: “CHVRCHES” → churches.
‘Lil’	‘Lil’ as a common adjective normalizes to ‘little’ (e.g., ‘you’re a lil silly’ → you’re a little silly). When ‘Lil’ is part of a catalog entity name, preserve it as-is (e.g., ‘Lil Wayne’ → lil wayne in spoken-form).
Question vs. statement intonation	If you are unsure whether intonation makes something a question, rely on syntax not intonation: subject-verb inversion or a question word = question mark; command or statement = period.
Uncertain comma placement	When comma placement is uncertain, ask: ‘Would a comma appear here in typical written communication of this content?’ If yes, use one.
Ambiguous units (currency vs. other)	When a unit is ambiguous between currency and another meaning (e.g., “pounds” as weight vs. sterling), use the interpretation that fits the speaker’s evident context. E.g., “she weighs ten thousand pounds” → 10,000 pounds (imperial unit, not \$10,000).

Acronyms and Initialisms

The treatment of letter-spelled sequences depends on how the speaker pronounces them:

- Spelled out as individual letters (F-B-I, C-E-O): transcribe with periods and spaces in spoken-form → f. b. i.
- Pronounced as a single word (NASA, OPEC, SCUBA): transcribe as one lowercase word in spoken-form → nasa, opec, scuba.
- In written-form, collapsed letter sequences become uppercase with no spaces or periods → FBI, OPEC.
- To make an initialism plural or possessive, attach the suffix directly after the final period in spoken-form: c. d.s | i. b. m.'s president.

Audio / Pronunciation	Spoken-Form	Written-Form
FBI (letters F-B-I)	f. b. i.	FBI
NASA (pronounced 'nah-suh')	nasa	NASA
OPEC (pronounced 'oh-peck')	opec	OPEC
OPEC (spelled out: O-P-E-C)	o. p. e. c.	OPEC
CDs (plural)	c. d.s	CDs

IBM's president	i. b. m.'s president	IBM's president
-----------------	----------------------	-----------------

Non-Target Languages

If a speaker uses a language other than the expected locale language, apply the following rules:

Topic	Rule
Attempting to transcribe	Attempt to transcribe. If you understand the language, transcribe what you hear. Do not add any special tag for non-target language speech.
Fictional/invented languages	Do not treat fictional or invented languages (e.g., Elvish, Klingon, Na'vi) as non-target languages — apply regular transcription rules, including (()) for unintelligible portions.
Unclear words	If a word is unclear, use the target language's regular unintelligible tag: (()) or ((closest guess)).
Diacritics	Remove diacritics that do not exist in English (e.g., medellín → medellin).
Background speech	Do not transcribe non-target language background speech — use [bg] instead.
Fully unintelligible utterance	If you cannot transcribe ANY of a non-target language utterance, insert (()) for the entire unintelligible section.
Entire utterance unintelligible	If the ENTIRE main-speaker utterance is unintelligible non-target language and you cannot transcribe any of it, select Discard with reason Non-target language.

Scenario	Spoken-Form	Action
Speaker says 'quiero taco bell' in Spanish (confident)	quiero taco bell	Transcribe — no tag needed
Speaker says a German phrase you recognize	wie geht es dir	Transcribe — no tag needed
Speaker says a word you can't confirm after research	((siza))	Guess in (())
Speaker says something entirely unintelligible in another language	(())	Discard — Non-target language
Cross-language homophone in isolation (e.g., 'home')	home	Treat as target language

Song Lyrics

Transcribe song lyrics as the user speaks or sings them, even if the lyrics are incorrect. Follow the same transcription rules as for regular speech:

- Use the `<sg>` `</sg>` tags around any live singing from an identified speaker.
- Transcribe nonsense syllables as logically as possible. Do not do outside research to determine spelling of nonsense lyrics.
- If lyrics are unintelligible, use () inside the `<sg>` tags: it goes like `<sg> ( ) </sg>`.

- Spoken-form song lyrics follow all regular spoken-form rules (lowercase, numbers as words). Written-form lyrics follow all regular written-form rules (digits, capitalization).

Appendix — Tag Quick Reference

Use only the tags below. Every tag must be spaced from adjacent text and other tags by a single space. Tags placed at the end of the transcription (e.g., `<pws>`) are inserted only once per audio file, in alphabetical order if multiple apply.

Tag	Meaning	Position / Notes
<code><s1>, <s2>, ...</code>	Multiple speakers	Start of each speaker's turn. Assign in order of first appearance.
<code> </code>	Overlapping speech — intelligible	Around the entire overlap block. Retag all speakers inside.
<code><ct></code>	Cross talk — unintelligible overlap	Where it occurs. Use Discard (Speech entirely unintelligible) if this is the entire audio.
<code><ws> </ws></code>	Whisper	Around the affected speech.
<code><lv> </lv></code>	Low voice (intentionally quiet, not mic distance)	Around the affected speech. Do not tag naturally soft voices throughout.
<code><pws></code>	Partial whisper — convenience marker indicating the audio contains some whispered speech	End of transcribed text only. Placed once. Use alphabetical order if multiple end-tags apply.
<code><sg> </sg></code>	Singing (live, identified speaker)	Around the affected speech.
<code><hum></code>	Humming	Where it occurs. Select Discard (Humming only) if humming is the entire audio.
<code><i></code>	List item separator — used between sequential items not joined by a conjunction (and / or)	Between items. Example: add beans [cough] <code><i></code> pears <code><i></code> cheese and bread to my list
<code>[fp]</code>	Filled pause (any hesitation not in the backchannel table)	Where it occurs. Transcribe once even for consecutive pauses.
<code>[hn]</code>	Non-lexical vocal sound (yawn, cry, or any other not listed below)	Where it occurs.
<code>[laughter]</code>	Laughter	Where it occurs.
<code>[cough]</code>	Cough	Where it occurs.

[sigh]	Sigh	Where it occurs.
[throat-clear]	Throat clear	Where it occurs.
[sneeze]	Sneeze	Where it occurs.
[lip-smack]	Lip smack	Where it occurs.
[bg]	Background speech (unattributable)	Where it occurs. One tag for multiple simultaneous unidentifiable voices.
[artifact]	Brief audio artifact (pop, click, static burst)	Where it occurs. Not for persistent noise.
(())	Unintelligible speech. Group multiple consecutive unintelligible words in a single tag. Use ((guessed word)) or ((phrase)) for a reasonable guess.	Where it occurs.
<ga>	Garbled audio — technical distortion makes transcription impossible	Entire transcription only. Select Discard.
<na>	No audio signal	Entire transcription only. Select Discard.
<ns>	No speech present (audio exists but no human speech)	Entire transcription only. Select Discard.

Backchannel Words

Transcribe the following as written words, not as [fp].

Transcription	Meaning
mhm	Agreeing / listening — mouth closed
uh-huh	Agreeing / listening — mouth open
hmm	Thinking or considering
uh	Short hesitation — open vowel
um	Hesitation — closed lip / nasal
mm-mm	Disagreeing — mouth closed
nuh-huh	Disagreeing — mouth open
oh	Surprise or realization (rhymes with 'owe')

ooh	Excitement or impression (rhymes with 'food')
-----	---

Spoken-Form vs. Written-Form Quick Examples

Spoken-Form	Written-Form
play music please	Play music, please.
what time is it	What time is it?
set alarm for five minutes	Set alarm for 5 minutes.
eighty-seven	87
it costs fifty dollars	It costs \$50.
it costs sixty cents	It costs \$0.60.
the gum is twenty-five cents and the mint is fifty cents	The gum is 25¢ and the mint is 50¢.
call mister johnson	Call Mr. Johnson.
play songs by the band mister mister	Play songs by the band Mr. Mister.
play songs by sza	Play songs by SZA.
play the song that goes twinkle twinkle little star	Play the song that goes twinkle twinkle little star.
play <sg> twinkle twinkle little star </sg>	Play <sg> Twinkle, twinkle, little star. </sg>
<s1> play <s2> songs	<s1> Play... <s2> Songs.
set alarm for elev-	Set alarm for elev- (truncated — fallback to spoken-form)
mm-mm that's not what i said	Mm-mm, that's not what I said.
well i think so	Well, I think so.
she turned to me and said nice to meet you	She turned to me and said, "Nice to meet you."
my zip code is nine zero two one zero	My zip code is 90210.
my number is five five five one two three four	My number is 555-1234.
a. b. c. one two three	ABC123 (when prefaced by 'my code is')
w. w. w. dot example dot com	www.example.com

eighty kilometers	80 km
nasa	NASA
f. b. i.	FBI
<s1> do you want to come <s1> along <s2> yes	<s1> Do you want to come... <s1> Along? <s2> Yes.