MEmoS2S-Bench: An Open-Ended Benchmark for Emotional Understanding, Reasoning, and Empathetic Spoken Response Generation in Multilingual Speech-LLMs

Anonymous authors
Paper under double-blind review

Abstract

Speech large language models (SpeechLLMs) are deployed as spoken conversational partners, a role that requires hearing how something was said, reasoning about why the speaker feels that way, and answering aloud. Existing benchmarks test these abilities in isolation, so they cannot show whether the stages work together inside one interaction. We present **MEmoS2S-Bench**, which elicits an emotion, a rationale, a text reply and a spoken reply within a single interaction, intervenes on each intermediate stage, and pairs every intervention with controls. It comprises **MELD-ST-Reason** (3,180 utterances in English, German and Japanese) and **MDET-Reason** (11,951 utterances in English and Chinese); we evaluate twelve SpeechLLMs over 3,720 scored cells. The controls expose three gaps that stage-wise evaluation cannot see. A *hearing gap*: from speech alone the models recover less emotion than a text-only LLM reading the words (macro-F1 0.358 against 0.512 on English), and every model improves with the transcript. A *credit gap*: a judge shown the gold emotion credits gold intermediates with gains that shrink several-fold under a label-blind judge, and a rationale helps only through the emotion it carries: one borrowed from another item with the same emotion works as well as the item's own. With delivery held fixed, the correctness of a model's own emotion makes no reliable difference to a blind judge. A *saying gap*: several models write the expected language but their speech is identified as another; for Japanese, five models' speech is identified as Chinese with high confidence. MEmoS2S-Bench makes these gaps measurable and attributable to a stage.

1 Introduction

Speech large language models (SpeechLLMs) are increasingly deployed as spoken conversational partners. A partner asked to console someone must do several things within a single exchange: hear that the speaker is upset rather than merely read what they said, form an account of why, choose words that answer that account, and say those words aloud in the right language. A growing family of end-to-end systems now attempts exactly this, accepting speech and returning speech without a text bottleneck (Zeng et al., 2024; Xu et al., 2025; Qwen Team, 2025; StepFun Audio Team, 2025; Kimi Team, 2025; Fu et al., 2025).

Current evaluation, however, measures these abilities one stage at a time and in isolation. Speech emotion recognition benchmarks score a categorical label (Busso et al., 2008; Poria et al., 2019; Lotfian & Busso, 2019); audio reasoning models are scored on their final answers, not on whether the rationale they emit helped produce them (Xie et al., 2025); empathetic dialogue benchmarks score written replies, usually from transcripts that discard the acoustic signal under study (Rashkin et al., 2019); and speech synthesis is scored separately for intelligibility and naturalness, on text it is given rather than on the model's own reply (Figure 1a). Each measures a real capability, and a model can score respectably on all of them. None, however, evaluates them together in a real interaction, where the stages must

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

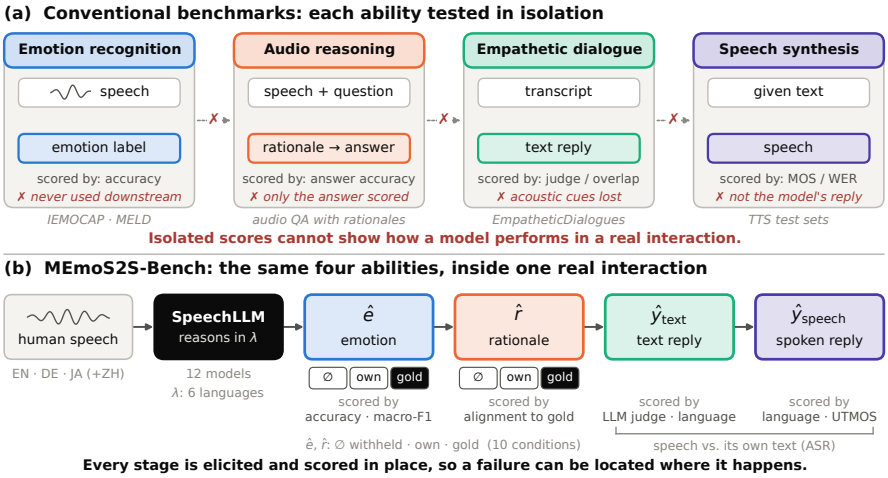

Figure 1: Conventional evaluation versus MEmoS2S-Bench. (a) Existing benchmarks test each ability of an empathetic spoken exchange on its own data and input: emotion recognition scores a label (Busso et al., 2008; Poria et al., 2019), audio reasoning scores only the final answer (Xie et al., 2025), empathetic dialogue scores written replies to transcripts (Rashkin et al., 2019), and speech synthesis scores speech rendered from given text. No stage's output reaches the next, so how the model performs in a real interaction is never observed. (b) MEmoS2S-Bench elicits the same four abilities (same colours) within one interaction: from human speech, a SpeechLLM reasoning in a controlled language λ produces an emotion $\hat{e}$, a rationale $\hat{r}$, a text reply $\hat{y}_{\text{text}}$, and a spoken reply $\hat{y}_{\text{speech}}$. The emotion and rationale can each be withheld, taken from the model itself, or replaced with ground truth (ten conditions); every stage is scored in place, and the spoken reply is checked against the model's own text, so a failure can be attributed to the stage where it occurs.

compose: where the emotion a model perceives should be the emotion its reply is built on, and the reply it writes should be the reply it speaks.

This leaves failures unattributable. When an empathetic reply is judged poor, existing instruments cannot tell whether the model misheard the emotion, reasoned badly from a correct perception, wrote a generic reply, or spoke something other than what it wrote—four failures that call for four different remedies, which a single quality score cannot distinguish. Existing work is also largely monolingual and text-side. It cannot ask in which language a model should reason about speech it heard in another, and it cannot detect a spoken reply that diverges from the text the model produced.

We therefore propose **MEmoS2S-Bench**, an open-ended benchmark that elicits emotional understanding, emotion-grounded reasoning, and empathetic text and spoken responses within a single interaction (Figure 1b). We release two datasets for the task: **MELD-ST-Reason**, 3,180 utterances drawn from parallel English, German and Japanese renditions of the same television dialogue, and **MDET-Reason**, 11,951 English and Chinese utterances from two independently collected corpora, each annotated with a gold emotion and a gold rationale. Given a human-recorded conversational utterance, a model produces an emotion, a rationale grounded in the audio, a text response, and a spoken response under one protocol. The benchmark then intervenes on the interaction: ten conditions independently withhold, elicit, or replace with ground truth each intermediate stage, so that the contributions of perception, reasoning, and composition are separated rather than inferred. It is multilingual by design, treating the language of the rationale as a controlled variable crossed against the language of the speech, and it scores the spoken response against the model's own text, so that a divergence between what a model writes and what it says becomes measurable. Because every stage is elicited and scored in place, a failure can be located rather than merely detected. Because conclusions drawn from judged interventions depend on what the judge is told, every intervention is paired with a label-blind judge and

Table 1: Comparison with representative benchmarks for audio-language models and spoken dialogue systems. ✓, △, and × indicate full, partial, and absent coverage. *Ctx.* denotes contextual dialogue; *Multi.* denotes an explicit multilingual evaluation; *S2S* denotes evaluation of speech output; *T–S* denotes consistency between generated text and speech; *Linked* denotes evaluation of multiple stages on the same interaction; *X-Ling.* denotes controlled cross-lingual reasoning; *Transfer* denotes cross-stage dependence analysis; *Interv.* denotes gold-information interventions; and *Joint* denotes instance-level compositional success.

	Interaction Setting			Evaluated Capability				Compositional Evaluation				
Benchmark	**Ctx.**	**Multi.**	**S2S**	**Emotion**	**Reason.**	**Response**	**T–S**	**Linked**	**X-Ling.**	**Transfer**	**Interv.**	**Joint**
AudioBench (Wang et al., 2025a)	×	△	×	✓	×	×	×	×	×	×	×	×
Dynamic-SUPERB (Huang et al., 2024)	×	△	×	✓	△	×	×	×	×	×	×	×
AIR-Bench (Yang et al., 2024)	×	△	×	△	△	△	×	×	×	×	×	×
Audio Entailment (Deshmukh et al., 2025)	×	×	×	×	✓	×	×	×	×	×	×	×
MMAU (Sakshi et al., 2025)	×	×	×	△	✓	×	×	×	×	×	×	×
MMSU (Wang et al., 2026)	×	△	×	✓	✓	×	×	×	×	×	×	×
ADU-Bench (Gao et al., 2025)	✓	✓	×	△	△	△	×	×	×	×	×	×
SD-Eval (Ao et al., 2024)	△	×	×	✓	×	✓	×	×	×	×	×	×
VoxDialogue (Cheng et al., 2025)	✓	×	✓	✓	×	✓	×	×	×	×	×	×
EChat-eval (Geng et al., 2025)	✓	×	✓	✓	△	✓	△	△	×	×	×	×
URO-Bench (Yan et al., 2025)	✓	✓	✓	△	✓	✓	△	×	×	×	×	×
EchoMind (Zhou et al., 2026)	△	×	✓	✓	✓	✓	△	✓	×	△	△	×
Ours	△	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

with mismatched-information controls, and every recognition score with majority, acoustic and text-only references. Together, these separate what a model hears, what it is credited for, and what it says.

2 Related Work

Audio-based benchmarks. Universal speech benchmarks probe representations task by task (Yang et al., 2021; Shi et al., 2023), and audio-understanding benchmarks for Speech-LLMs score perception and reasoning from audio (Wang et al., 2025a; Huang et al., 2024; Yang et al., 2024; Sakshi et al., 2025; Wang et al., 2026; Deshmukh et al., 2025). Emotion benchmarks score a categorical label, on single corpora (Busso et al., 2008; Poria et al., 2019; Lotfian & Busso, 2019) or across corpora and languages (Ma et al., 2024a), and emotion captioning describes the cues behind it (Xu et al., 2024). Their shared limitation is that each stops at a single stage: a correct label, caption or answer is the end point, so they cannot show whether what a model perceives is what its reply is built on. Rationales from audio reasoning models (Xie et al., 2025) are judged, if at all, by the answer they accompany, although a written chain of thought need not reflect what drives the answer (Turpin et al., 2023). They also rarely place a SpeechLLM against the single-modality systems it should subsume, an acoustic recogniser and a text model reading the words, on the same items. A benchmark for empathetic interaction must instead elicit perception, reasoning and reply together, intervene on the stages in between, and compare each stage with such references, which is what MEmoS2S-Bench does.

Multilingual empathetic spoken dialogue. Voice-assistant and spoken-dialogue benchmarks score replies to speech (Chen et al., 2024b; Gao et al., 2025; Ao et al., 2024; Cheng et al., 2025; Yan et al., 2025), and empathetic dialogue benchmarks score written replies or emotional understanding from text (Rashkin et al., 2019; Liu et al., 2021; Sabour et al., 2024), discarding the acoustic signal. The closest work links several stages: EchoMind (Zhou et al., 2026) for understanding, reasoning and response, EChat-eval (Geng et al., 2025) for empathetic spoken replies, and URO-Bench (Yan et al., 2025) for the agreement of spoken and written replies. These evaluations share three limitations. First, they are largely confined to English and Chinese; multilingual speech benchmarks evaluate recognition and translation rather than affect (Conneau et al., 2023; Shi et al., 2023), and although text LLMs are known to answer in an unintended language (Marchisio et al., 2024), no spoken benchmark checks the language of the speech itself. Second, they do not intervene on the intermediate stages of one interaction, so a good reply cannot be attributed to a correct perception. Third, they score the reply with judges or references that are given the answer, although LLM judges are sensitive to what they are shown (Zheng et al., 2023; Wang et al.,

2024). MEmoS2S-Bench addresses all three, building on the parallel English, German and Japanese versions of MELD-ST (Chen et al., 2024a) (Table 1).

3 MEmoS2S-Bench

3.1 Dataset construction

We release two datasets for the task. **MELD-ST-Reason** builds on MELD-ST (Chen et al., 2024a), which pairs English sitcom utterances with German and Japanese versions. Its English and Japanese partitions are the same 1,008 utterances; the German partition is 1,164 utterances from the English–German split (3,180 in total). Emotion labels are MELD's seven classes ($\mathcal{E} = \{$anger, disgust, fear, joy, neutral, sadness, surprise$\}$), annotated on the English original with video and dialogue context and shared by the German and Japanese versions, and each item carries the next line of the original dialogue as a reference reply. **MDET-Reason** combines EmotionTalk (Sun et al., 2026) (3,173 Mandarin utterances) and MultiDialog (Park et al., 2024) (8,778 English utterances), two independently collected corpora of performed speech with their own speakers and label protocols (11,951 in total), and serves as a cross-corpus test.

Every item in both datasets receives a **gold rationale**: an explanation of the speaker's emotion grounded in audible cues. Rationales were generated by an audio reasoning model (Xie et al., 2025) from the audio and the gold emotion (when its own prediction disagreed with the label, it was asked to explain why the speaker expresses the gold emotion), rewritten for fluency with GPT-4.1, and screened so that they describe audible properties without naming the label. They therefore encode the label implicitly, which §5.3 quantifies. We verify them with standard acoustic analysis: from every clip we extract fundamental frequency (YIN), RMS energy, speaking rate and a smoothed intonation slope, and match them with what each rationale claims. Claims about level cues agree with the measurements: rationales that describe a high pitch go with higher measured fundamental frequency than those describing a low pitch ($d = +0.43$, $p = 3 \times 10^{-7}$, agreement 0.59), and loud with higher energy than quiet ($d = +0.51$, $p = 0.01$, agreement 0.72). Claims about temporal cues do not: speaking pace ($d = -0.14$) and rising versus falling intonation ($d = +0.09$) are at chance, so they should not be relied on. An audio-capable model (gpt-audio) that hears each clip confirms the grounding on a random sample of 191 items: 81% of gold rationales receive 4 or 5 of 5 for acoustic accuracy, above a rationale taken from a different clip with a different emotion (4.04 against 3.44; difference +0.60, 95% CI [+0.40, +0.80]). Machine translations into five further languages support varying the reasoning language (Appendix J).

3.2 Task and interventions

Given a human-recorded utterance x in interaction language ℓ, a model produces, within one interaction,

$$(\hat{e}, \hat{r}, \hat{y}_{\text{text}}, \hat{y}_{\text{speech}}) = M(x, \pi_c, \lambda), \tag{1}$$

an emotion, a rationale, a text reply and a spoken reply, where π_c is the prompt for condition c and λ the language requested for the rationale (English unless stated). Both replies are retained, so their agreement is measured rather than assumed.

Ten conditions intervene on the two intermediate stages (Appendix C): four *elicit* conditions (reply only; emotion only; rationale only; emotion and rationale), three *predictive* conditions in which the model states its own emotion, rationale, or both before replying, and three *oracle* conditions in which the gold emotion, rationale, or both are inserted into the prompt. Predictive conditions are generated in a single pass, so a predictive and an oracle prompt differ in form as well as in the correctness of the intermediate value. We therefore pair every comparison on identical items and add controls that isolate what an intermediate contributes: the model's own intermediate inserted through the oracle template, a *mismatched rationale* borrowed from another item with the same or a different gold emotion, and judges that are or are not told the gold emotion (§3.3).

3.3 Evaluation metrics

Each stage is scored where it occurs; Appendix A gives definitions and implementation details. **Emotion**: accuracy and macro-F1 over the seven classes, with macro-F1 primary because a constant *neutral* already reaches accuracy 0.475. **Rationale**: BERTScore (Zhang et al., 2020) and LaBSE similarity (Feng et al., 2022) to the gold rationale, Jensen–Shannon similarity of the emotion distributions of the two texts (Lin, 1991), and agreement of the rationale's dominant emotion with the gold label. **Text reply**: BERTScore, Sentence-BERT similarity (Reimers & Gurevych, 2019), bidirectional NLI contradiction (Bowman et al., 2015), affective similarity and emotion agreement against the reference reply, and the rate of replies in the expected language. **Spoken reply**: word or character error of a Whisper large-v3 (Radford et al., 2023) transcript against the model's own text, spoken-language accuracy, UTMOS (Saeki et al., 2022), and emotion2vec (Ma et al., 2024b) similarity to the reference speech. **Joint success**: the share of items on which every stage passes, with a soft geometric-mean variant. **LLM judge**: GPT-4.1 rates each reply from 1 to 5 on emotional appropriateness, empathy, naturalness, relevance and overall quality, blind to model and condition, on 60 items per model and language drawn identically across conditions, in two variants: *label-visible*, told the gold emotion as in standard rubric-based judging, and *label-blind*, given only the transcript and the reply.

4 Experimental Setup

Models. We evaluate twelve SpeechLLMs at their released weights with default decoding and no fine-tuning, in three groups. *General SpeechLLMs*: GLM-4-Voice (Zeng et al., 2024), Kimi-Audio (Kimi Team, 2025), Qwen2.5-Omni (Xu et al., 2025), VITA-1.5 (Fu et al., 2025), Baichuan-Omni-1.5 (Li et al., 2025) and Fun-Audio-Chat-8B (Chen et al., 2025). *Speech-LLMs with reasoning capabilities*: Qwen3-Omni (Qwen Team, 2025), Step-Audio-2 (StepFun Audio Team, 2025) and MiMo-Audio-7B (Xiaomi LLM-Core Team, 2025). *Emotion-focused SpeechLLMs*: OSUM-EChat (Geng et al., 2025), FreezeEmpath (Hong et al., 2026) and OpenS2S-1.5 (Wang et al., 2025b).

Prompt setup. Every condition uses one prompt template assembled from blocks (Appendix C): an instruction to listen to the utterance, a rationale block that asks for a 50–200-word explanation grounded in prosodic, acoustic and non-verbal cues (English unless the rationale language is varied), an emotion block with the closed label set, the gold emotion or rationale in tags for oracle conditions, and a reply block asking for one or two spoken sentences in the interaction language. The model returns all requested fields in one JSON object, from which the text reply is read and the spoken reply is synthesised by the model's own speech decoder. Speech-only replies are instructed not to presume the words, so that they rely on what is heard.

Audio input setup. All conditions are run on the speech alone, and the emotion-producing conditions also on the speech paired with its gold transcript, which isolates how much a model relies on the words. MELD-ST-Reason supplies three interaction languages over the same material (English, German and Japanese); MDET-Reason supplies performed speech from independent corpora as a cross-corpus test (Appendix E).

5 Results

5.1 Stage-level results

Table 2 reports every stage for the twelve models. No model leads every stage: Qwen2.5-Omni has the best emotion scores, Qwen3-Omni the best reasoning, speech and text–speech averages, and OpenS2S-1.5 the best text average despite the lowest emotion accuracy, because the text average does not penalise the reply language. Instance-level success is rare: even Qwen3-Omni completes all four stages on only 28.5% of items (lowest 6.6%).

Table 2: Main results on MELD-ST-Reason (mean over English, German and Japanese; one pass, English rationale). Best value per column in bold. Metrics and task averages (Avg.) are defined in Appendix A; per-language results are in Appendix B.

(a) Emotional understanding, reasoning, and text generation

	Emotion Understanding			Reasoning					Text Response Generation					
				Lex.-Sem.		Aff.			Lex.-Sem.			Aff.		
Model	Acc. ↑	F1 ↑	Avg. ↑	BS ↑	LaBSE ↑	JS ↑	EM ↑	Avg. ↑	BS ↑	SBERT ↑	$1-C$ ↑	JS ↑	EM ↑	Avg. ↑
General SpeechLLMs														
GLM-4-Voice	0.464	0.163	0.314	0.884	0.744	**0.985**	0.477	0.772	0.856	0.315	0.604	**0.476**	0.405	0.531
Kimi-Audio	0.353	0.308	0.331	0.868	0.538	0.681	0.422	0.627	0.851	0.335	0.645	0.425	0.388	0.529
Qwen2.5-Omni	**0.500**	**0.359**	**0.430**	0.886	0.716	0.938	**0.479**	0.755	0.865	0.336	0.595	0.455	0.389	0.528
VITA-1.5	0.311	0.195	0.253	0.881	0.672	0.927	0.465	0.736	0.870	0.327	0.663	0.447	0.381	0.538
Baichuan-Omni-1.5	0.334	0.269	0.302	0.864	0.432	0.857	0.445	0.649	0.851	0.328	0.655	0.438	0.391	0.533
Fun-Audio-Chat-8B	0.426	0.358	0.392	0.862	0.564	0.768	0.474	0.667	0.870	0.350	0.640	0.409	0.416	0.537
SpeechLLMs with Reasoning Capabilities														
Qwen3-Omni	0.366	0.308	0.337	**0.889**	**0.790**	0.981	0.473	**0.783**	0.875	**0.373**	0.633	0.390	0.355	0.525
Step-Audio-2	0.414	0.298	0.356	0.886	0.699	0.934	0.474	0.748	0.869	0.350	0.611	0.390	0.350	0.514
MiMo-Audio-7B	0.441	0.309	0.375	0.869	0.614	0.885	0.471	0.710	0.865	0.341	0.654	0.445	**0.425**	0.546
Emotion-Focused SpeechLLMs														
OSUM-EChat	0.490	0.250	0.370	0.875	0.575	0.913	0.470	0.708	0.860	0.332	0.589	0.451	0.417	0.530
FreezeEmpath	0.358	0.231	0.295	0.880	0.707	0.975	0.472	0.759	**0.876**	0.342	0.644	0.435	0.384	0.536
OpenS2S-1.5	0.173	0.165	0.169	0.873	0.716	0.654	0.306	0.637	0.857	0.325	**0.759**	0.456	0.386	**0.557**

(b) Speech generation, cross-modal consistency, and end-to-end success

	Speech Response Generation						Text–Speech Consistency					End-to-End	
	Linguistic		Quality	Aff.		Avg.	Lex.-Sem.			Aff.	Avg.		
Model	Err. ↓	Lang. ↑	UTMOS ↑	E2V ↑	JS ↑	↑	BS ↑	LaBSE ↑	$1-C$ ↑	JS ↑	↑	J_{strict} ↑	J_{soft} ↑
General SpeechLLMs													
GLM-4-Voice	0.586	0.963	3.130	0.799	0.359	0.613	0.932	0.743	0.848	0.416	0.735	0.264	**0.381**
Kimi-Audio	5.341	0.382	2.617	**0.816**	0.306	0.382	0.855	0.457	0.755	0.298	0.591	0.069	0.242
Qwen2.5-Omni	1.033	0.426	3.254	0.727	0.349	0.413	0.894	0.478	0.753	0.234	0.590	0.113	0.354
VITA-1.5	0.629	0.352	3.125	0.795	0.356	0.481	0.910	0.566	0.800	0.371	0.662	0.121	0.232
Baichuan-Omni-1.5	0.472	0.609	3.216	0.758	0.404	0.571	0.929	0.713	0.825	0.428	0.724	0.153	0.254
Fun-Audio-Chat-8B	0.374	**0.984**	3.296	0.797	0.365	0.669	0.944	0.734	0.863	0.490	0.758	0.272	0.338
SpeechLLMs with Reasoning Capabilities													
Qwen3-Omni	**0.149**	0.930	3.418	0.807	0.390	**0.716**	**0.980**	**0.928**	**0.961**	0.369	**0.810**	**0.285**	0.297
Step-Audio-2	0.201	0.673	**3.616**	0.633	0.396	0.631	0.971	0.914	0.942	0.350	0.794	0.248	0.327
MiMo-Audio-7B	0.480	0.740	2.707	0.814	0.395	0.579	0.931	0.730	0.850	0.398	0.727	0.217	0.343
Emotion-Focused SpeechLLMs													
OSUM-EChat	0.553	0.312	3.175	0.793	0.304	0.480	0.939	0.728	0.911	**0.640**	0.804	0.097	0.372
FreezeEmpath	0.713	0.343	3.066	0.725	**0.425**	0.459	0.907	0.524	0.811	0.227	0.617	0.134	0.252
OpenS2S-1.5	0.343	0.338	2.976	0.802	0.329	0.524	0.956	0.852	0.941	0.332	0.770	0.066	0.130

Table 3: Emotion recognition on MELD-ST-Reason against reference systems. English and Japanese are the same 1,008 utterances; German is the 1,164-utterance English–German split. SpeechLLM figures pool the four emotion-producing conditions. Accuracy is close to the majority-class rate, so macro-F1 is the informative column.

	English		German		Japanese	
System	Acc.	F1	Acc.	F1	Acc.	F1
Majority class (*neutral*)	0.475	0.092	0.473	0.092	0.475	0.092
emotion2vec+ (speech only)	0.577	0.437	0.310	0.210	0.240	0.174
GPT-4.1 (transcript only)	0.605	0.512	0.541	0.397	0.537	0.429
SpeechLLMs, speech (mean)	0.478	0.358	0.377	0.233	0.351	0.227
SpeechLLMs, speech (best)	0.575	0.450	0.504	0.361	0.491	0.349
SpeechLLMs, speech + transcript (mean)	0.558	0.447	0.463	0.325	0.450	0.339

5.2 The hearing gap: how much emotion do the models hear?

Table 3 compares the SpeechLLMs with three references on the same items. On English speech their mean macro-F1 is 0.358; the best, Kimi-Audio (0.450), is on par with a dedicated acoustic recogniser, emotion2vec+ (Ma et al., 2024b) (0.437), and all are below GPT-4.1 reading only the gold transcript (0.512). The labels are thus largely recoverable from the words. Supplying the transcript alongside the speech raises macro-F1 for 12/12 models (mean +0.096). The accuracy gain is similar across languages in absolute terms (+8.0, +8.7 and +9.8 points for English, German and Japanese), so the larger relative gains for German and Japanese mostly reflect their lower baselines.

Table 4: Cross-stage interventions under two GPT-4.1 judges: one shown the gold emotion (as in standard rubric-based judging) and one that must infer the emotion from the transcript. Each entry is the paired change in the overall score (1–5) against Speech → Response on the same item (12 models × 3 languages × 60 items). Brackets are 95% bootstrap intervals clustered by model; * marks intervals that exclude zero.

	Judge sees gold emotion		Label-blind judge	
Condition	Δ	95% CI	Δ	95% CI
+ own emotion	−0.081	[-0.16, +0.00]	−0.069	[-0.16, +0.03]
+ own rationale	−0.135	[-0.26, +0.03]	+0.051	[-0.07, +0.18]
+ own emotion & rationale	−0.087	[-0.27, +0.14]	+0.079	[-0.06, +0.24]
+ gold emotion	+0.551*	[+0.46, +0.65]	+0.072	[-0.02, +0.17]
+ gold rationale	+0.306*	[+0.22, +0.39]	+0.119*	[+0.03, +0.22]
+ gold emotion & rationale	+0.551*	[+0.46, +0.65]	+0.150*	[+0.04, +0.28]

Table 5: Correctness versus delivery: change against Speech → Response for the own intermediate generated in the same pass or inserted, and the gold one inserted. Correctness is gold minus own (both inserted); delivery is inserted minus single-pass own. Brackets: 95% intervals.

Judge	Stage	Own (1-pass)	Own (ins.)	Gold (ins.)	Correctness	Delivery
Label-visible	Emotion	−0.079	+0.129	+0.557	+0.428 [+0.33, +0.53]	+0.208 [+0.12, +0.29]
	Rationale	−0.169	+0.078	+0.179	+0.101 [+0.00, +0.20]	+0.247 [+0.15, +0.35]
	Both	−0.092	+0.050	+0.421	+0.371 [+0.26, +0.48]	+0.142 [+0.04, +0.24]
Label-blind	Emotion	−0.043	+0.026	+0.083	+0.057 [-0.02, +0.14]	+0.069 [-0.00, +0.15]
	Rationale	−0.044	+0.072	+0.035	−0.037 [-0.12, +0.05]	+0.117 [+0.03, +0.20]
	Both	+0.034	+0.023	+0.079	+0.056 [-0.04, +0.15]	−0.011 [-0.10, +0.08]

Across groups, SpeechLLMs with reasoning capabilities recognise emotion best (mean macro-F1 0.305), general SpeechLLMs next (0.276) and emotion-focused SpeechLLMs last (0.215): models built for empathetic dialogue are not better at hearing the emotion (Table 2).

5.3 The credit gap: what do intermediate stages contribute?

Table 4 reports the interventions under both judges. The label-visible judge gives a stark picture: the gold emotion adds +0.551, the gold rationale +0.306, and a model's own intermediates subtract up to −0.135. The label-blind judge does not reproduce it. The gold emotion adds +0.072 (interval including zero), the gold rationale +0.119 and both together +0.150, while a model's own intermediates range from −0.069 to +0.079 with every interval spanning zero. Most of the label-visible effect is thus the judge rewarding agreement with the label it was shown.

Correctness versus delivery. Because predictive conditions are single-pass, the own-versus-gold contrast of Table 4 also changes how the intermediate reaches the reply. We separate the two by inserting each model's own emotion and rationale, taken from its elicit conditions, through the oracle template (Table 5; 4 models, 696 items per condition). With delivery held fixed, the label-visible judge attributes large effects to correctness (+0.428 for the emotion, +0.371 for emotion and rationale), but the label-blind judge finds none that is reliable (+0.057, −0.037 and +0.056, all intervals including zero); its only reliable effect is a small benefit of inserting the model's own rationale rather than generating it in the same pass (+0.117). For these models, the conclusion that correct perception drives reply quality is largely a product of judging with the label.

Gold rationales help through the emotion they encode. Although the gold rationales describe their own audio (§3), what they contribute to the reply is their emotion. GPT-4.1 recovers the gold label from the gold rationale alone for 91.3–92.4% of items, despite the screening for label words. A mismatched-rationale control on six models shows that the item-specific content adds nothing measurable (Table 6): a gold rationale borrowed from another item *with the same emotion* changes the reply score by +0.297 (label-visible)

Table 6: Mismatched-rationale control (six models: GLM-4-Voice, Kimi-Audio, Qwen2.5-Omni, VITA-1.5, Qwen3-Omni and Step-Audio-2; 60 judged items per model and language). The gold rationale of the item is replaced by the gold rationale of a different item whose gold emotion is the same or different. Entries are paired changes in the overall score against Speech → Response, with 95% item-bootstrap intervals.

Rationale supplied	Judge sees gold emotion	Label-blind judge
Gold rationale of this item	+0.227 [+0.16, +0.30]	+0.082 [+0.02, +0.15]
Gold rationale of another item, same emotion	+0.297 [+0.22, +0.38]	+0.094 [+0.03, +0.17]
Gold rationale of another item, different emotion	−0.183 [-0.26, -0.10]	+0.023 [-0.05, +0.10]

Table 7: The spoken channel is not the text channel. TextLang and SpeechLang are the rates at which the written and the spoken response are in the expected language; Agree is the rate at which the two agree; ASR-Err is the character error between a transcription of the generated audio and the model's own generated text, averaged over the three interaction languages. SpeechLang uses Whisper large-v3 language identification; ASR is forced to the expected language. Several models write the expected language while their speech is not identified as that language, which no text-side metric detects.

Model	TextLang	SpeechLang	Agree	ASR-Err ↓	UTMOS
GLM-4-Voice	0.938	0.963	0.926	0.398	3.130
Kimi-Audio	0.958	0.382	0.398	1.231	2.617
Qwen2.5-Omni	0.995	0.426	0.423	0.532	3.254
VITA-1.5	0.997	0.352	0.354	0.468	3.125
Baichuan-Omni-1.5	0.995	0.609	0.614	0.362	3.216
Fun-Audio-Chat-8B	1.000	0.984	0.984	0.286	3.296
Qwen3-Omni	0.999	0.930	0.931	0.047	3.418
Step-Audio-2	0.906	0.673	0.753	0.148	3.616
MiMo-Audio-7B	0.995	0.740	0.739	0.307	2.707
OSUM-EChat	0.731	0.312	0.393	0.338	3.175
FreezeEmpath	1.000	0.343	0.343	0.591	3.066
OpenS2S-1.5	0.600	0.338	0.713	0.302	2.976
Mean	**0.926**	**0.587**	**0.631**	**0.417**	**3.133**

and +0.094 (blind), against +0.227 and +0.082 for the item's own rationale, while one borrowed from an item with a *different* emotion gives −0.183 and +0.023. Models' own rationales imply the correct emotion for only 32.1% of items, and replies conditioned on one that does score +0.31 higher under the blind judge (+1.24 label-visible). Intermediates help through the emotion they carry, substantially so only for a judge that knows the answer, and self-generated ones rarely carry the right one.

5.4 The saying gap: the spoken channel is not the text channel

Table 7 separates what a model writes from what it says. Ten of the twelve models write the expected language at least 90.6% of the time, but across all twelve the speech is identified (Whisper large-v3) as the expected language only 0.587 of the time on average. The gap is sharpest where the text is flawless: FreezeEmpath writes the expected language every time yet its speech is identified as that language 34.3% of the time, and VITA-1.5, Kimi-Audio and Qwen2.5-Omni reach 0.352, 0.382 and 0.426 in speech against 99.7%, 95.8% and 99.5% in text. Character error against the model's own text rises from 0.134 in English to 0.411 in German and 0.707 in Japanese; the spoken reply fails while every text-side metric scores the interaction as clean.

The failures are systematic rather than noisy (Table 8). For Japanese input, the speech of 5 models (Kimi-Audio, Qwen2.5-Omni, VITA-1.5, OSUM-EChat, FreezeEmpath) is identified as Chinese for most items, with mean confidence 0.91, consistent with speech decoders that read the kanji of a Japanese reply with Mandarin pronunciation. For German input, the speech of 8 models is identified mostly as English, with lower confidence (0.71), yet it still transcribes to the model's German text with median character error 0.43: accented, largely

Table 8: Language of the reply: share of written (Text) and spoken (Speech, Whisper language ID) replies in the expected language, and the most frequent other spoken language (share, mean confidence).

	German input			Japanese input		
Model	Text	Speech	Other (share, conf.)	Text	Speech	Other (share, conf.)
GLM-4-Voice	0.88	0.95	English (0.03, 0.62)	0.94	0.94	Chinese (0.05, 0.81)
Kimi-Audio	0.95	0.17	English (0.54, 0.74)	0.92	0.08	Chinese (0.88, 0.90)
Qwen2.5-Omni	0.99	0.23	English (0.44, 0.57)	1.00	0.05	Chinese (0.92, 0.92)
VITA-1.5	0.99	0.00	English (0.97, 0.79)	1.00	0.05	Chinese (0.91, 0.90)
Baichuan-Omni-1.5	0.99	0.01	English (0.88, 0.74)	1.00	0.82	Chinese (0.14, 0.72)
Fun-Audio-Chat-8B	1.00	0.96	English (0.03, 0.57)	1.00	0.99	Korean (0.01, 0.47)
Qwen3-Omni	1.00	0.86	English (0.08, 0.56)	1.00	0.93	Chinese (0.07, 0.63)
Step-Audio-2	0.73	0.04	English (0.92, 0.75)	0.98	0.98	Chinese (0.01, 0.52)
MiMo-Audio-7B	0.99	0.36	English (0.55, 0.63)	0.99	0.86	Chinese (0.07, 0.62)
OSUM-EChat	0.77	0.05	English (0.68, 0.69)	0.47	0.00	Chinese (0.98, 0.93)
FreezeEmpath	1.00	0.03	English (0.94, 0.71)	1.00	0.00	Chinese (0.97, 0.88)
OpenS2S-1.5	0.76	0.02	English (0.79, 0.61)	0.04	0.00	English (0.95, 0.91)

intelligible German. The Chinese-identified Japanese speech does not (median 0.97 against the Japanese text). On the human input utterances the identifier errs only on very short clips.

6 Conclusion

MEmoS2S-Bench elicits emotion, rationale, text and speech within one interaction and intervenes on each intermediate stage with matched controls. Across twelve SpeechLLMs and two released datasets, it exposes three gaps: the models hear less emotion than single-modality references, with reasoning-capable models best and emotion-focused ones worst (*hearing*); a judge told the gold emotion inflates the value of intermediates several-fold, and a blind judge finds no reliable effect of whether a model's own intermediate is correct (*credit*); and the speech channel fails outside English (*saying*). Evaluations of empathetic SpeechLLMs should report macro-F1 against majority and acoustic baselines, judge replies without revealing the reference emotion, and score the spoken reply itself.

AI use statement

Large language models were used to polish the writing of this paper, improving grammar, wording and readability. The authors reviewed all text and take full responsibility for the content of this work.

Ethics statement

The benchmark builds on existing corpora (MELD-ST, EmotionTalk, MultiDialog) under their original licences; MELD-ST derives from a copyrighted television series and is distributed for research use, so we release rationales, prompts, model outputs and scripts, and direct users to the original distributors for the audio. No new human-subject data were collected. Automatic emotion recognition can be misused for surveillance or manipulation; we evaluate it only as a component of empathetic dialogue, and our results show that current systems are unreliable at it, especially outside English. The emotion taxonomy and the English-centred labels may not transfer across cultures, which we flag where it affects conclusions.

Reproducibility statement

Section 3, Section 4 and Appendices A, C and F specify the data, conditions, prompts, judge rubrics, metrics and control experiments. All tables and every number in the text are generated by scripts from the released model outputs; the evaluation harnesses, parsing rules, scoring and table generation code, and the judge outputs will be released with the benchmark (anonymised supplementary material during review).

References

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Proceedings of NeurIPS (Datasets and Benchmarks Track)*, 2024.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pp. 632–642, 2015.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335–359, 2008.

Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. Fun-Audio-Chat technical report. *arXiv preprint arXiv:2512.20156*, 2025.

Sirou Chen, Sakiko Yahata, Shuichiro Shimizu, Zhengdong Yang, Yihang Li, Chenhui Chu, and Sadao Kurohashi. MELD-ST: An emotion-aware speech translation dataset. In *Findings of ACL*, 2024a.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024b.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. VoxDialogue: Can spoken dialogue systems understand information beyond words? In *Proceedings of ICLR*, 2025.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2023.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. Audio entailment: Assessing deductive reasoning for audio understanding. In *Proceedings of AAAI*, 2025.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and Khiet P Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2): 190–202, 2016.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of ACL*, pp. 878–891, 2022.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of ACL*, 2025.

Xuelong Geng, Qijie Shao, Hongfei Xue, Shuiyuan Wang, Hanke Xie, Zhao Guo, Yi Zhao, Guojian Li, Wenjie Tian, Chengyou Wang, Zhixian Zhao, Kangxiang Xia, Ziyu Zhang, Zhennan Lin, Tianlun Zuo, Mingchen Shao, Yuang Cao, Guobin Ma, Longhao Li, Yuhang Dai, Dehui Gao, Dake Guo, and Lei Xie. OSUM-EChat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *arXiv preprint arXiv:2508.09600*, 2025.

Yun Hong, Yan Zhou, and Yang Feng. FreezeEmpath: Efficient training for empathetic spoken chatbots with frozen LLMs. In *Findings of ACL*, 2026.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-yi Lee. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *Proceedings of ICASSP*, 2024.

Kimi Team. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of ACL*, 2021.

Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proceedings of Interspeech*, 2024a.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of ACL*, 2024b.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In *Proceedings of EMNLP*, 2024.

Se Jin Park, Chae Won Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeong Hun Yeo, and Yong Man Ro. Let's Go Real Talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of ACL*, 2024.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, 2019.

Qwen Team. Qwen3-Omni technical report. *arXiv preprint*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, 2023.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*, 2019.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of ACL*, 2024.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022. In *Proceedings of Interspeech*, pp. 4521–4525, 2022.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics. In *Proceedings of Interspeech*, 2024.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *Proceedings of ICLR*, 2025.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and Shinji Watanabe. ML-SUPERB: Multilingual speech universal PERformance benchmark. In *Proceedings of Interspeech*, 2023.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

StepFun Audio Team. Step-Audio 2 technical report. *arXiv preprint*, 2025.

Haoqin Sun, Jinghua Zhao, Xuechen Wang, Shiwan Zhao, Jiaming Zhou, Hui Wang, Xi Yang, Yequan Wang, and Yonghua Lin. EmotionTalk: An interactive Chinese multimodal emotion dataset with rich annotations. In *Findings of ACL*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *Proceedings of NAACL*, 2025a.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and Jiajun Zhang. OpenS2S: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of EMNLP: System Demonstrations*, 2025b.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *Proceedings of ICLR*, 2026.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of ACL*, 2024.

Xiaomi LLM-Core Team. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-Reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech emotion captioning with large language model. In *Proceedings of AAAI*, 2024.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. URO-Bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. In *Findings of EMNLP*, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of ACL*, 2024.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. SUPERB: Speech processing universal PERformance benchmark. In *Proceedings of Interspeech*, 2021.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR*, 2020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. EchoMind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In *Proceedings of ICLR*, 2026.

A Evaluation Metrics

We evaluate the models along four stages of an emotionally intelligent spoken interaction: emotional understanding, emotion-grounded reasoning, textual response generation, and spoken response generation. We additionally measure cross-modal text–speech consistency and end-to-end compositional success. Unless otherwise stated, higher values indicate better performance. For the multilingual main results (Table 2), we first calculate each metric separately for English, German, and Japanese inputs and then report the unweighted macro-average across the three languages, ensuring that each language contributes equally. The LLM-judged response quality used in the cross-stage analysis is described in Section 3.

A.1 Emotional Understanding

Emotional understanding is evaluated by comparing the predicted emotion label with the gold annotation. We report strict classification accuracy and macro-averaged F1. Macro-F1 calculates F1 independently for each class and then assigns equal weight to all classes, making it more informative than accuracy alone under an imbalanced emotion distribution (Sokolova & Lapalme, 2009). Predictions that cannot be mapped to the benchmark emotion inventory are counted as incorrect. We additionally report the valid-prediction rate as a diagnostic measure of instruction-following reliability.

A.2 Emotion-Grounded Reasoning

We evaluate the generated acoustic-emotional reasoning against the reference rationale (Section 3). Because valid explanations can differ substantially in surface form, the principal reasoning metrics emphasize semantic and affective agreement.

Semantic similarity. We report BERTScore F1, which aligns candidate and reference tokens using contextual representations (Zhang et al., 2020), and LaBSE cosine similarity, which compares sentence-level representations in a shared multilingual embedding space (Feng et al., 2022). BERTScore captures fine-grained contextual correspondence, whereas LaBSE evaluates global cross-lingual semantic similarity.

Affective agreement. We apply a multilingual text-emotion classifier to the generated and reference reasoning. We compare its emotion posterior distributions using Jensen–Shannon divergence (Lin, 1991). For interpretability, we convert divergence into a bounded similarity:

$$S_{\mathrm{JS}}(P, Q) = 1 - \mathrm{JSD}(P \| Q), \tag{2}$$

where P and Q denote the reference and generated emotion distributions. Higher values indicate stronger affective agreement. We also report whether the dominant emotion of the generated reasoning agrees with the gold emotion of the input utterance (EM in Table 2), and additionally compute affective cosine similarity, dominant-emotion agreement with the reference, and cosine similarity between the hidden emotion-sensitive text representations.

Reasoning-language compliance. Automatic language identification is applied to the generated reasoning. Reasoning-language accuracy is

$$\mathrm{Acc}_{\text{R-LID}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\widehat{\ell}_i^R = \ell_{\text{target}}^R\right]. \tag{3}$$

In the multilingual main condition, English is the required reasoning language for English, German, and Japanese inputs. This measure is an application of language identification to controlled chain-of-reasoning generation and is therefore treated as a benchmark-specific compliance metric.

A.3 Textual Response Generation

Semantic and lexical similarity. We evaluate semantic similarity to the reference response using BERTScore (Zhang et al., 2020) and Sentence-BERT cosine similarity (Reimers & Gurevych, 2019). Surface-form correspondence is evaluated using ROUGE-L (Lin, 2004) and chrF++ (Popović, 2015).

Logical consistency. We use a multilingual natural-language inference model to estimate entailment and contradiction. NLI formalizes whether a hypothesis follows from, is neutral to, or contradicts a premise (Bowman et al., 2015). We evaluate both directions and report

$$E_{\mathrm{bi}} = \frac{p(R_{\mathrm{ref}} \Rightarrow R_{\mathrm{gen}}) + p(R_{\mathrm{gen}} \Rightarrow R_{\mathrm{ref}})}{2}, \tag{4}$$

$$C_{\mathrm{bi}} = \max\left\{p(R_{\mathrm{ref}} \perp R_{\mathrm{gen}}), p(R_{\mathrm{gen}} \perp R_{\mathrm{ref}})\right\}. \tag{5}$$

Higher entailment and lower contradiction indicate greater logical compatibility; Table 2 reports $1 - C_{\mathrm{bi}}$ so that higher is better throughout. Bidirectional aggregation is our evaluation design rather than a metric proposed in the original NLI literature.

Affective appropriateness. The generated and reference responses are compared using affective cosine similarity, Jensen–Shannon emotion-distribution similarity (Lin, 1991), dominant-emotion agreement, and emotion-sensitive embedding cosine similarity. We also measure whether the emotion expressed by the generated text agrees with the emotion of the input utterance.

Language accuracy. Text-language accuracy measures whether the detected language of the response matches the input language:

$$\mathrm{Acc}_{\text{T-LID}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\widehat{\ell}_i^{T} = \ell_i^{\text{input}}\right]. \tag{6}$$

A.4 Spoken Response Generation

Perceptual quality. We estimate speech quality using UTMOS, a non-intrusive MOS-prediction model developed for synthetic and converted speech (Saeki et al., 2022). UTMOS provides an automatic proxy for perceived quality and should not be interpreted as a replacement for human listening tests. We additionally compute duration, root-mean-square energy, clipping ratio, and silence ratio as diagnostic acoustic statistics.

Intelligibility and content preservation. Generated speech is transcribed using automatic speech recognition, and the transcript is compared with the model's own text response. The error rate is

$$\mathrm{Err} = \frac{S + D + I}{N}, \tag{7}$$

where S, D, and I denote substitutions, deletions, and insertions, and N is the number of words in the text response for English and German (word error rate, WER) or the number of characters for Japanese and Chinese (character error rate, CER). This formulation follows the standard edit-distance principle introduced by Levenshtein (1966). Lower values indicate more faithful realization of the intended response; because insertions are counted, the rate can exceed one.

We additionally compare the ASR transcript with the reference response using BERTScore (Zhang et al., 2020) and Sentence-BERT similarity (Reimers & Gurevych, 2019). When same-split reference audio is available, we compute SpeechBERTScore, which extends contextual token-alignment evaluation to self-supervised speech representations (Saeki et al., 2024).

Emotional speech alignment. Generated and reference speech are encoded using emotion2vec, a self-supervised representation model developed for general speech-emotion analysis (Ma et al., 2024b). We report cosine similarity between their emotion2vec representations and Jensen–Shannon similarity between predicted speech-emotion distributions, and additionally compute dominant-emotion agreement and the similarity of standardized eGeMAPS acoustic descriptors, which capture prosodic, spectral, energy, and voice-quality characteristics relevant to affective speech (Eyben et al., 2016).

Speech-language accuracy. A speech language-identification model is applied to each generated waveform. Speech-language accuracy measures whether the detected language matches the target response language. Because automatic speech LID can be unreliable for short or accented recordings, we interpret it together with the text-language accuracy of the same response rather than in isolation.

A.5 Text–Speech Consistency

Text–speech consistency is treated as a separate benchmark dimension because semantic or emotional quality in one modality does not guarantee that the other modality conveys the same response.

Semantic consistency. We compare the generated text with the ASR transcript of the generated speech using BERTScore (Zhang et al., 2020), LaBSE similarity (Feng et al., 2022), and bidirectional NLI (Bowman et al., 2015). These metrics detect omissions, semantic alterations, and contradictions introduced during speech generation.

Language consistency. We define text–speech language agreement as

$$A_{\text{lang}}^{T,S} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\widehat{\ell}_i^{T} = \widehat{\ell}_i^{S}\right]. \tag{8}$$

Unlike the individual text and speech language accuracies, this benchmark-specific measure directly tests whether both generated modalities use the same language.

Affective consistency. We compare the text- and speech-emotion posterior distributions using Jensen–Shannon similarity (Lin, 1991), cross-modal affective cosine similarity, and dominant-emotion agreement. These benchmark-specific metrics identify responses whose text is emotionally appropriate but whose vocal realization conveys a conflicting affect.

A.6 End-to-End Compositional Success

The following joint measures are introduced in this work. Unlike the stage-specific metrics inherited from prior studies, they are designed to evaluate whether all stages of an emotional interaction succeed on the same example.

Let E_i, R_i, T_i, and S_i denote the normalized emotional understanding, reasoning, textual response, and spoken response scores, respectively. Reasoning, text, and speech are considered successful when their stage scores exceed 0.6; emotion understanding requires an exactly correct label. A stage score is defined only when at least 50% of its constituent metrics are available.

Strict joint success is

$$J_{\text{strict},i} = \mathbb{I}\left[E_i^{\text{pass}} \wedge R_i^{\text{pass}} \wedge T_i^{\text{pass}} \wedge S_i^{\text{pass}}\right]. \tag{9}$$

Reasoning success additionally requires the requested reasoning language, text success requires the target response language, and speech success requires a valid waveform, correct speech language, and text–speech language agreement. Missing required stages are counted as end-to-end failures, while joint coverage is reported separately.

We further introduce a soft joint score:

$$J_{\text{soft},i} = \left(E_i R_i T_i S_i\right)^{1/4}. \tag{10}$$

The geometric mean follows the general principle of conjunctive aggregation: a low-performing stage strongly reduces the overall score without forcing all partially successful examples to zero.

A.7 Task Averages

Emotion Avg. is the mean of accuracy and macro-F1. Reasoning Avg. is the mean of BERTScore, LaBSE, affective JS similarity, and input-emotion agreement (EM). Text Avg. is the mean of BERTScore, Sentence-BERT, $1 - \text{NLI contradiction}$, affective JS similarity, and input-emotion agreement. Speech Avg. uses normalised UTMOS, $(\text{UTMOS} - 1)/4$; ASR fidelity, $1 - \min(\text{Err.}, 1)$, where Err. is the word error rate for English and German and the character error rate for Japanese, measured against the model's own text response; speech-language accuracy; emotion2vec similarity; and speech-emotion JS similarity. Text–speech Avg. is the mean of BERTScore, LaBSE, $1 - \text{NLI contradiction}$, and affective JS similarity.

B Results by Input Language

Tables 9–11 break the main results of Table 2 down by the language of the input speech.

C Conditions, Prompts and Judge Rubrics

Prompt construction. Each condition's prompt is assembled from blocks: an instruction to listen to the utterance; a rationale block (rationale-eliciting conditions only), which asks

Table 9: Results on the English partition of MELD-ST-Reason. Columns, task averages, and model groups are as in Table 2; WER is measured against the model's own text response. The best value in each column is in bold.

(a) Emotional understanding, reasoning, and text generation

Model	Emotion Understanding			Reasoning					Text Response Generation					
	Acc. ↑	F1 ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. LaBSE ↑	Aff. JS ↑	Aff. EM ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. SBERT ↑	Lex.-Sem. $1-C$ ↑	Aff. JS ↑	Aff. EM ↑	Avg. ↑
General SpeechLLMs														
GLM-4-Voice	0.514	0.272	0.393	0.883	0.739	0.965	0.484	0.768	0.862	0.260	0.613	0.422	0.412	0.514
Kimi-Audio	0.468	0.417	0.442	0.870	0.558	0.719	0.473	0.655	0.847	0.251	0.684	0.436	0.436	0.531
Qwen2.5-Omni	0.546	0.414	0.480	0.886	0.717	0.935	0.494	0.758	0.842	0.234	0.548	0.449	0.303	0.475
VITA-1.5	0.426	0.288	0.357	0.879	0.651	0.919	0.476	0.731	0.862	0.246	0.672	0.440	0.447	0.533
Baichuan-Omni-1.5	0.403	0.358	0.381	0.865	0.433	0.848	0.466	0.653	0.849	0.245	0.681	0.424	0.429	0.526
Fun-Audio-Chat-8B	0.492	**0.445**	0.468	0.864	0.575	0.769	0.491	0.675	0.864	0.272	0.641	0.414	0.465	0.531
SpeechLLMs with Reasoning Capabilities														
Qwen3-Omni	0.405	0.362	0.383	0.889	**0.793**	0.978	0.473	**0.783**	**0.869**	**0.279**	0.650	0.404	0.399	0.520
Step-Audio-2	0.527	0.380	0.454	0.887	0.716	0.940	0.484	0.757	0.866	0.270	0.644	0.408	0.380	0.514
MiMo-Audio-7B	0.518	0.425	0.472	0.871	0.621	0.870	**0.507**	0.717	0.863	0.258	0.640	0.423	0.485	0.534
Emotion-Focused SpeechLLMs														
OSUM-EChat	**0.587**	0.433	**0.510**	0.873	0.563	0.874	0.491	0.700	0.865	0.272	0.577	0.426	0.480	0.524
FreezeEmpath	0.467	0.289	0.378	**0.891**	0.760	**0.984**	0.477	0.778	0.867	0.264	0.671	**0.462**	**0.502**	**0.553**
OpenS2S-1.5	0.238	0.231	0.235	0.872	0.715	0.716	0.376	0.670	0.861	0.261	**0.724**	0.427	0.440	0.543

(b) Speech generation, cross-modal consistency, and end-to-end success

Model	Speech Response Generation						Text–Speech Consistency					End-to-End	
	Linguistic WER ↓	Linguistic Lang. ↑	Quality UTMOS ↑	Aff. E2V ↑	Aff. JS ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. LaBSE ↑	Lex.-Sem. $1-C$ ↑	Aff. JS ↑	Avg. ↑	J_{strict} ↑	J_{soft} ↑
General SpeechLLMs													
GLM-4-Voice	0.270	0.995	3.292	0.770	0.337	0.681	0.987	0.961	0.984	0.512	0.861	0.401	0.434
Kimi-Audio	5.445	0.887	2.605	0.811	0.310	0.482	0.901	0.635	0.861	0.289	0.672	0.181	0.352
Qwen2.5-Omni	0.948	0.995	3.444	0.728	0.271	0.532	0.921	0.692	0.873	0.274	0.690	0.292	0.430
VITA-1.5	0.082	**1.000**	3.412	0.777	0.333	0.726	0.984	0.959	0.959	0.448	0.837	0.358	0.356
Baichuan-Omni-1.5	0.157	**1.000**	3.334	0.724	0.358	0.702	0.980	0.932	0.969	0.499	0.845	0.302	0.328
Fun-Audio-Chat-8B	0.110	**1.000**	3.513	0.752	**0.402**	0.735	0.990	0.964	0.994	0.617	**0.891**	0.383	0.404
SpeechLLMs with Reasoning Capabilities													
Qwen3-Omni	**0.049**	0.996	3.569	0.792	0.355	**0.747**	**0.992**	0.978	**0.995**	0.526	0.873	0.326	0.334
Step-Audio-2	0.166	0.998	**3.765**	0.646	0.277	0.689	0.986	0.955	0.988	0.426	0.839	**0.433**	0.437
MiMo-Audio-7B	0.087	**1.000**	3.136	**0.828**	0.348	0.725	0.988	0.966	0.985	0.558	0.874	0.420	0.430
Emotion-Focused SpeechLLMs													
OSUM-EChat	0.191	0.885	3.330	0.814	0.345	0.687	0.982	0.952	0.960	**0.617**	0.878	0.284	**0.459**
FreezeEmpath	0.204	**1.000**	3.199	0.716	0.315	0.675	0.983	**0.979**	0.994	0.300	0.814	0.402	0.387
OpenS2S-1.5	0.221	0.993	3.171	0.763	0.327	0.681	0.980	0.957	0.983	0.421	0.835	0.198	0.194

Table 10: Results on the German partition of MELD-ST-Reason. Columns, task averages, and model groups are as in Table 2; WER is measured against the model's own text response. The best value in each column is in bold.

(a) Emotional understanding, reasoning, and text generation

Model	Emotion Understanding			Reasoning					Text Response Generation					
	Acc. ↑	F1 ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. LaBSE ↑	Aff. JS ↑	Aff. EM ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. SBERT ↑	Lex.-Sem. $1-C$ ↑	Aff. JS ↑	Aff. EM ↑	Avg. ↑
General SpeechLLMs														
GLM-4-Voice	0.454	0.116	0.285	0.884	0.748	**0.994**	0.471	0.774	0.854	0.334	0.576	**0.436**	0.356	0.511
Kimi-Audio	0.327	0.286	0.307	0.868	0.537	0.680	0.402	0.622	0.852	0.360	0.651	0.390	0.375	0.526
Qwen2.5-Omni	**0.476**	**0.313**	**0.395**	0.886	0.719	0.943	0.467	0.754	0.865	0.357	0.650	0.421	0.408	0.540
VITA-1.5	0.283	0.161	0.222	0.881	0.675	0.925	0.455	0.734	0.868	0.335	0.662	0.402	0.325	0.519
Baichuan-Omni-1.5	0.282	0.217	0.250	0.863	0.426	0.844	0.430	0.641	0.861	0.343	0.686	0.396	0.335	0.524
Fun-Audio-Chat-8B	0.398	0.306	0.352	0.862	0.557	0.758	**0.473**	0.663	0.868	0.362	0.654	0.395	**0.409**	0.538
SpeechLLMs with Reasoning Capabilities														
Qwen3-Omni	0.364	0.287	0.326	**0.889**	**0.791**	0.982	0.472	**0.783**	0.869	**0.382**	0.635	0.353	0.332	0.514
Step-Audio-2	0.342	0.230	0.286	0.886	0.707	0.936	0.465	0.749	0.859	0.352	0.606	0.360	0.301	0.496
MiMo-Audio-7B	0.389	0.219	0.304	0.868	0.606	0.892	0.446	0.703	0.867	0.352	0.652	0.436	0.374	0.536
Emotion-Focused SpeechLLMs														
OSUM-EChat	0.437	0.177	0.307	0.876	0.594	0.922	0.457	0.712	0.849	0.316	0.631	0.403	0.355	0.511
FreezeEmpath	0.330	0.199	0.265	0.862	0.630	0.972	0.464	0.732	**0.871**	0.355	0.673	0.384	0.296	0.516
OpenS2S-1.5	0.145	0.144	0.145	0.873	0.718	0.670	0.309	0.642	0.866	0.353	**0.777**	0.419	0.342	**0.552**

(b) Speech generation, cross-modal consistency, and end-to-end success

Model	Speech Response Generation						Text–Speech Consistency					End-to-End	
	Linguistic WER ↓	Linguistic Lang. ↑	Quality UTMOS ↑	Aff. E2V ↑	Aff. JS ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. LaBSE ↑	Lex.-Sem. $1-C$ ↑	Aff. JS ↑	Avg. ↑	J_{strict} ↑	J_{soft} ↑
General SpeechLLMs													
GLM-4-Voice	0.590	0.955	3.048	0.844	0.389	0.622	0.946	0.859	0.894	0.432	0.783	0.290	**0.373**
Kimi-Audio	6.979	0.174	2.417	0.813	0.320	0.332	0.841	0.424	0.729	0.286	0.570	0.020	0.206
Qwen2.5-Omni	0.997	0.231	3.275	0.746	0.496	0.409	0.880	0.472	0.745	0.228	0.581	0.041	0.326
VITA-1.5	0.676	0.003	3.303	0.845	0.499	0.449	0.890	0.526	0.707	0.312	0.609	0.000	0.191
Baichuan-Omni-1.5	0.698	0.009	3.201	0.784	0.507	0.430	0.899	0.609	0.766	0.431	0.676	0.000	0.189
Fun-Audio-Chat-8B	**0.181**	**0.960**	3.263	0.830	0.443	**0.724**	0.966	**0.905**	0.929	0.441	**0.810**	**0.291**	0.319
SpeechLLMs with Reasoning Capabilities													
Qwen3-Omni	0.284	0.864	3.376	0.834	0.522	0.706	**0.971**	0.896	**0.956**	0.285	0.777	0.252	0.290
Step-Audio-2	0.290	0.039	**3.615**	0.591	0.546	0.508	0.955	0.862	0.908	0.359	0.771	0.014	0.245
MiMo-Audio-7B	0.662	0.363	2.451	**0.884**	0.488	0.487	0.905	0.596	0.770	0.423	0.673	0.075	0.281
Emotion-Focused SpeechLLMs													
OSUM-EChat	0.661	0.049	3.134	0.840	0.379	0.428	0.936	0.779	0.901	**0.514**	0.782	0.006	0.308
FreezeEmpath	0.736	0.027	3.132	0.684	**0.601**	0.422	0.883	0.456	0.760	0.179	0.570	0.000	0.201
OpenS2S-1.5	0.589	0.018	2.551	0.866	0.466	0.430	0.917	0.674	0.859	0.364	0.704	0.000	0.099

Table 11: Results on the Japanese partition of MELD-ST-Reason. Columns, task averages, and model groups are as in Table 2; CER is measured against the model's own text response. The best value in each column is in bold.

(a) Emotional understanding, reasoning, and text generation

Model	Emotion Understanding			Reasoning					Text Response Generation					
	Acc. ↑	F1 ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. LaBSE ↑	Aff. JS ↑	Aff. EM ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. SBERT ↑	Lex.-Sem. $1-C$ ↑	Aff. JS ↑	Aff. EM ↑	Avg. ↑
General SpeechLLMs														
GLM-4-Voice	0.425	0.102	0.263	0.884	0.745	**0.997**	0.475	0.775	0.853	0.351	0.623	**0.569**	0.446	0.568
Kimi-Audio	0.265	0.221	0.243	0.866	0.520	0.643	0.390	0.604	0.853	0.393	0.599	0.449	0.353	0.530
Qwen2.5-Omni	**0.479**	**0.350**	**0.414**	0.885	0.710	0.936	**0.476**	0.752	0.887	0.415	0.585	0.494	**0.457**	0.568
VITA-1.5	0.225	0.136	0.181	0.882	0.689	0.936	0.464	0.743	0.880	0.401	0.656	0.498	0.370	0.561
Baichuan-Omni-1.5	0.316	0.233	0.275	0.865	0.437	0.879	0.438	0.655	0.845	0.397	0.598	0.495	0.411	0.549
Fun-Audio-Chat-8B	0.386	0.324	0.355	0.861	0.562	0.775	0.459	0.664	0.879	0.416	0.623	0.417	0.375	0.542
SpeechLLMs with Reasoning Capabilities														
Qwen3-Omni	0.330	0.275	0.303	**0.889**	**0.786**	0.984	0.474	**0.783**	0.888	**0.458**	0.613	0.415	0.336	0.542
Step-Audio-2	0.373	0.283	0.328	0.884	0.675	0.926	0.474	0.740	0.882	0.427	0.583	0.400	0.370	0.532
MiMo-Audio-7B	0.416	0.282	0.349	0.867	0.616	0.892	0.460	0.709	0.866	0.414	0.669	0.477	0.417	0.569
Emotion-Focused SpeechLLMs														
OSUM-EChat	0.447	0.139	0.293	0.875	0.569	0.943	0.462	0.712	0.865	0.408	0.560	0.524	0.418	0.555
FreezeEmpath	0.278	0.205	0.241	0.888	0.732	0.969	**0.476**	0.766	**0.889**	0.405	0.587	0.458	0.353	0.539
OpenS2S-1.5	0.137	0.120	0.128	0.874	0.716	0.576	0.234	0.600	0.844	0.362	**0.775**	0.521	0.375	**0.575**

(b) Speech generation, cross-modal consistency, and end-to-end success

Model	Speech Response Generation						Text–Speech Consistency					End-to-End	
	Linguistic CER ↓	Linguistic Lang. ↑	Quality UTMOS ↑	Aff. E2V ↑	Aff. JS ↑	Avg. ↑	Lex.-Sem. BS ↑	Lex.-Sem. LaBSE ↑	Lex.-Sem. $1-C$ ↑	Aff. JS ↑	Avg. ↑	J_{strict} ↑	J_{soft} ↑
General SpeechLLMs													
GLM-4-Voice	0.897	0.938	3.051	0.784	0.351	0.538	0.862	0.409	0.666	0.302	0.560	0.103	0.335
Kimi-Audio	3.599	0.084	2.830	**0.825**	0.287	0.331	0.823	0.311	0.674	0.320	0.532	0.006	0.167
Qwen2.5-Omni	1.153	0.051	3.042	0.706	0.280	0.310	0.880	0.271	0.640	0.201	0.498	0.005	0.307
VITA-1.5	1.131	0.053	2.659	0.764	0.236	0.293	0.855	0.212	0.735	0.354	0.539	0.004	0.148
Baichuan-Omni-1.5	0.562	0.819	3.113	0.766	0.349	0.580	0.908	0.597	0.740	0.355	0.650	0.158	0.244
Fun-Audio-Chat-8B	0.832	**0.991**	3.112	0.810	0.250	0.549	0.876	0.334	0.664	0.412	0.571	0.141	0.290
SpeechLLMs with Reasoning Capabilities													
Qwen3-Omni	**0.113**	0.930	3.308	0.795	0.292	**0.696**	**0.977**	0.910	0.933	0.296	**0.779**	0.276	0.266
Step-Audio-2	0.147	0.983	**3.469**	0.661	**0.363**	0.695	0.974	**0.923**	0.932	0.264	0.773	**0.296**	0.298
MiMo-Audio-7B	0.690	0.856	2.535	0.731	0.348	0.526	0.900	0.629	0.794	0.213	0.634	0.157	0.319
Emotion-Focused SpeechLLMs													
OSUM-EChat	0.807	0.002	3.061	0.726	0.188	0.325	0.900	0.453	0.871	**0.788**	0.753	0.000	**0.348**
FreezeEmpath	1.198	0.001	2.867	0.775	0.358	0.320	0.855	0.139	0.680	0.200	0.468	0.001	0.169
OpenS2S-1.5	0.220	0.002	3.207	0.776	0.193	0.461	0.972	0.923	**0.980**	0.211	0.771	0.000	0.096

Table 12: The ten conditions. *Elicit* conditions score the intermediate stages on their own; *predictive* conditions feed the model's own intermediate output into its response; *oracle* conditions substitute ground truth for that same intermediate value, (the gold value is inserted into the prompt). Predictive conditions are generated in a single pass.

Group	Condition	Emotion input	Rationale input
Elicit	`response_only`	—	—
	`emotion_only`	—	—
	`reasoning_only`	—	—
	`emotion_reasoning`	—	—
Predictive	`emotion_to_response`	own	—
	`reasoning_to_response`	—	own
	`emotion_reasoning_to_response`	own	own
Oracle	`gold_emotion_to_response`	gold	—
	`gold_reasoning_to_response`	—	gold
	`gold_emotion_reasoning_to_response`	gold	gold

for 50–200 words in the requested language grounded in prosodic, acoustic and non-verbal cues and states "Assume that you do not understand the transcript meaning, and avoid using transcript content to justify your reasoning"; an emotion block with the closed label set; the gold emotion or gold rationale in tags (oracle conditions only); and a reply block asking for one or two natural spoken sentences in the interaction language, which states "If no transcript is provided, do not pretend to know the exact semantic content. Make the reply generic but appropriate to the perceived affect." The model returns all requested fields in one JSON object. Two consequences follow for interpretation: the rationale is instructed to ignore lexical content, and speech-only replies are instructed to stay generic, which a judge that penalises generic replies scores down in every speech-only condition alike.

Label-visible judge (system prompt).

```
You are evaluating a spoken dialogue system's empathetic reply.

You are given what the speaker said, the speaker's true emotional state, and a
candidate reply the system produced.

Score the reply on four dimensions, each an integer 1-5:
- emotional_appropriateness: does the reply respond suitably to the speaker's actual
  emotion?
- empathy: does it convey understanding and care, rather than deflecting or lecturing?
- naturalness: would a fluent speaker plausibly say this out loud in conversation?
- relevance: does it engage with what the speaker actually said?

Also give overall: an integer 1-5 for how good a spoken empathetic reply this is.

Penalise: replies that describe the emotion analytically instead of responding
to it; replies in the wrong language; generic filler that would fit any input.

Return ONLY a JSON object with keys emotional_appropriateness, empathy,
naturalness, relevance, overall. No other text.
```

The user message gives the transcript, "Speaker's true emotion: <gold>", the expected reply language and the candidate reply.

Label-blind judge (system prompt). Identical except for the two sentences below; the user message omits the emotion line.

```
You are given a transcript of what the speaker said (you cannot hear the audio) and a
candidate reply the system produced. You are NOT told the speaker's emotion.
- emotional_appropriateness: does the reply respond suitably to the emotion the speaker
  is most plausibly expressing?
```

Table 13: Emotion accuracy from speech alone versus speech paired with the gold transcript, averaged over the four emotion-producing conditions and the three interaction languages (ENG/DEU/JPN). Every model improves when the words are supplied, which isolates how much of the reported "emotion understanding" is carried by lexical content rather than by the speech signal.

Model	Speech	Speech+Transcript	Δ	Rel. Δ (%)
Qwen2.5-Omni	0.522	0.564	+0.042	+8.0
Step-Audio-2	0.493	0.566	+0.072	+14.6
GLM-4-Voice	0.472	0.510	+0.038	+8.1
Qwen3-Omni	0.456	0.507	+0.052	+11.3
OSUM-EChat	0.435	0.457	+0.022	+5.1
Kimi-Audio	0.391	0.440	+0.049	+12.5
FreezeEmpath	0.388	0.522	+0.133	+34.4
VITA-1.5	0.370	0.511	+0.141	+38.1
Fun-Audio-Chat-8B	0.368	0.462	+0.094	+25.6
Baichuan-Omni-1.5	0.337	0.483	+0.146	+43.3
MiMo-Audio-7B	0.327	0.443	+0.116	+35.5
OpenS2S-1.5	0.267	0.423	+0.156	+58.2
Mean	0.402	0.491	**+0.088**	**+24.6**

Table 14: Emotion accuracy from speech alone and with the gold transcript, by corpus (accuracy, as in earlier versions; macro-F1 results are in §5.2). MELD-ST-Reason German and Japanese are dubbed; the MDET corpora are performed speech with their own label protocols, so the rows are not directly comparable.

Corpus	Condition	Speech	+Transcript	Rel. Δ	Improved
MELD-ST (English, original)	within	0.478	0.559	+18.4%	12/12
MELD-ST (German, dubbed)	within	0.377	0.463	+26.8%	12/12
MELD-ST (Japanese, dubbed)	within	0.352	0.450	+32.0%	12/12
MDET (Mandarin, spontaneous)	cross	0.402	0.431	+8.5%	9/12
MDET (English, spontaneous)	cross	0.516	0.530	+4.3%	6/12

D Perception References

Reference systems. emotion2vec+ large (Ma et al., 2024b) is applied to every clip at utterance level, with its labels mapped onto the seven benchmark classes and the argmax taken over those seven. GPT-4.1 receives only the gold transcript and returns one of the seven labels. English and Japanese are the same utterances, so their comparison is paired; the German dubs are paired with their own English originals from the English–German split.

Tables 13 and 14 report accuracy, as in earlier versions of this analysis. Because *neutral* is close to half of MELD-ST-Reason, accuracy alone can reward a constant prediction: GLM-4-Voice predicts *neutral* for most speech-only items, which places it third by accuracy but lower by macro-F1. The transcript effect survives on macro-F1 for every model (§5.2). The MDET corpora consist of performed rather than spontaneous speech, and on MultiDialog the transcript effect is small in accuracy but positive for more models in macro-F1, so we do not draw conclusions from its accuracy row.

E MDET-Reason Results

On the cross-corpus set (Table 15; per corpus in Table 16), speech-only macro-F1 is again modest, 0.402 on EmotionTalk and 0.256 on MultiDialog at best (Qwen3-Omni on both), and the transcript raises it for 11/12 and 12/12 models, by +0.029 and +0.027 on average: the words help less on these performed corpora than on MELD-ST-Reason. Across all three corpora the model groups order the same way. SpeechLLMs with reasoning capabilities

Table 15: Results on MDET-Reason, averaged over EmotionTalk (Mandarin) and MultiDialog (English). Emotion is pooled over the four emotion-producing conditions; the rationale is scored against the gold rationale (BERTScore F1 and sentence-embedding cosine) in the full-interaction condition. Per-corpus results are in Table 16. Best value per column in bold.

Model	Acc.	F1	F1 (+transcript)
General SpeechLLMs			
GLM-4-Voice	0.510	0.217	0.257
Kimi-Audio	0.512	0.310	0.327
Qwen2.5-Omni	**0.569**	0.288	0.299
VITA-1.5	0.497	0.228	0.263
Baichuan-Omni-1.5	0.416	0.244	0.280
Fun-Audio-Chat-8B	0.346	0.272	0.296
SpeechLLMs with Reasoning Capabilities			
Qwen3-Omni	0.497	**0.329**	**0.349**
Step-Audio-2	0.535	0.310	0.331
MiMo-Audio-7B	0.424	0.290	0.310
Emotion-Focused SpeechLLMs			
OSUM-EChat	0.472	0.235	0.247
FreezeEmpath	0.373	0.205	0.273
OpenS2S-1.5	0.386	0.231	0.266
Majority class	0.571	0.103	

Table 16: Emotion recognition on MDET-Reason by corpus (macro-F1), from speech alone and with the gold transcript, pooled over the four emotion-producing conditions. Best value per column in bold.

	EmotionTalk (Mandarin)		MultiDialog (English)	
Model	Speech	+Transcript	Speech	+Transcript
General SpeechLLMs				
GLM-4-Voice	0.249	0.305	0.185	0.208
Kimi-Audio	0.382	0.405	0.238	0.248
Qwen2.5-Omni	0.363	0.363	0.212	0.234
VITA-1.5	0.275	0.307	0.182	0.218
Baichuan-Omni-1.5	0.291	0.321	0.197	0.238
Fun-Audio-Chat-8B	0.354	0.371	0.190	0.222
SpeechLLMs with Reasoning Capabilities				
Qwen3-Omni	**0.402**	**0.425**	**0.256**	**0.273**
Step-Audio-2	0.395	0.401	0.226	0.261
MiMo-Audio-7B	0.368	0.396	0.213	0.224
Emotion-Focused SpeechLLMs				
OSUM-EChat	0.265	0.271	0.204	0.223
FreezeEmpath	0.234	0.310	0.176	0.235
OpenS2S-1.5	0.271	0.319	0.190	0.214
Majority class	0.087	0.087	0.118	0.118

recognise emotion best (mean macro-F1 0.388 on EmotionTalk, 0.232 on MultiDialog, 0.305 on MELD-ST-Reason), general SpeechLLMs next (0.319, 0.201, 0.276), and the emotion-focused SpeechLLMs last (0.257, 0.190, 0.215): models built for empathetic dialogue are not better at hearing the emotion.

F Judge Validity and Intervention Controls

Reliability. Re-judging a stratified 720 judged replies with the same judge and prompt reproduces the overall score exactly for 91.4% of them (quadratic-weighted $\kappa = 0.98$). A

Table 17: Three judges on the same replies (20 items per model and language): GPT-4.1 told the gold emotion, GPT-4.1 given only the transcript, and gpt-audio hearing the speech. Change against Speech → Response; 95% intervals clustered by model; * excludes zero.

	Told the gold emotion		Transcript only		Hears the speech	
Condition	Δ	95% CI	Δ	95% CI	Δ	95% CI
+ own emotion	−0.084	[-0.22, +0.07]	−0.064	[-0.18, +0.06]	−0.025	[-0.11, +0.06]
+ own rationale	−0.084	[-0.23, +0.10]	+0.130*	[+0.01, +0.26]	+0.111*	[+0.01, +0.23]
+ own emotion & rationale	−0.053	[-0.25, +0.18]	+0.147*	[+0.00, +0.32]	+0.139*	[+0.01, +0.30]
+ gold emotion	+0.608*	[+0.47, +0.75]	+0.128	[-0.01, +0.27]	+0.172*	[+0.07, +0.28]
+ gold rationale	+0.338*	[+0.22, +0.47]	+0.167*	[+0.04, +0.31]	+0.138*	[+0.04, +0.26]
+ gold emotion & rationale	+0.620*	[+0.49, +0.76]	+0.223*	[+0.08, +0.38]	+0.223*	[+0.11, +0.35]

second judge, GPT-4o, scores 4,320 replies (20 items per model and language in all six conditions) with Spearman $\rho = 0.91$ against GPT-4.1 on the overall score, and reproduces the direction and approximate size of every label-visible intervention effect. Because both are shown the gold emotion, this establishes reliability, not validity.

Sensitivity controls. On 720 judged replies, substituting the reply of another item (same model, language and condition) lowers the overall score by −1.12; keeping the reply but changing the stated emotion to a different label lowers emotional appropriateness by −1.01. The judge therefore reads the reply, and it also reads, and relies on, the label it is given.

Correct versus incorrect perception. In the own-emotion condition, replies whose emotion prediction was correct score +1.51 higher on emotional appropriateness under the label-visible judge and +0.33 under the blind judge. Comparing each item with its own reply-only baseline gives a difference-in-differences of +0.73 and +0.17 respectively. The unpaired figure of earlier versions (+1.5) is confounded by item difficulty and by the judge's access to the label.

A judge that hears the speech. gpt-audio receives the speaker's audio, its transcript and the candidate reply, is not told the emotion, and returns the emotion it perceives with the five scores. On the 5,032 replies of Table 17 it perceives the gold emotion for 0.620, 0.569 and 0.564 of English, German and Japanese items, and its intervention effects match the transcript-only judge rather than the judge told the label. As a stand-alone recogniser on the clips it answers (it declines very short ones) it reaches macro-F1 0.442, 0.382 and 0.390. Where own intermediates hurt under the blind judge, it is in Japanese (down to −0.18).

MDET-Reason. The transcript raises macro-F1 by +0.029 on EmotionTalk (11/12 models) and +0.027 on MultiDialog (12/12); on MultiDialog a constant *neutral* scores accuracy 0.704, so accuracy is uninformative there.

Mismatched rationales. For six models (GLM-4-Voice, Kimi-Audio, Qwen2.5-Omni, VITA-1.5, Qwen3-Omni and Step-Audio-2) and the 60 judged items per model and language, the gold rationale was replaced by the gold rationale of a randomly chosen other item with the same gold emotion, or with a different gold emotion, and the replies were generated and judged exactly as in the oracle condition (Table 6).

What the speech channel produces. Table 8 lists, for German and Japanese input, the language the spoken reply is identified as. To check the identifier, Whisper large-v3 was run on the human input utterances, whose language is known: it identifies 99.8%, 88.7% and 71.4% of the English, German and Japanese clips correctly. Its errors are concentrated on very short clips (a few characters), where German is confused mainly with English and Japanese mainly with Korean; on utterances of sentence length it is above 93% in both languages. Spoken replies are one or two sentences long, so the Chinese identifications of Japanese replies, made with high confidence, are not an identifier artefact; the English identifications of German replies, made with lower confidence, may include heavily accented German.

Text cascade. GPT-4.1 receives the gold transcript of each judged item and is asked for a one- or two-sentence empathetic spoken reply in the interaction language; its replies are scored by the label-blind judge exactly as the SpeechLLMs' replies are. It exceeds the SpeechLLMs' speech-only replies by +1.88 (GPT-4.1 judge) and +1.67 (GPT-4o judge), and their replies with the transcript supplied by +1.31; supplying the transcript to the SpeechLLMs raises their own replies by +0.56. With Whisper large-v3 transcripts instead of gold ones, the cascade reaches emotion macro-F1 0.478, 0.398 and 0.409 for English, German and Japanese (gold transcripts: 0.512, 0.397, 0.429), and its replies score 4.44. The cascade uses a much larger text model, so the comparison measures headroom rather than a modality effect.

G Reference Metrics versus the Judge

Reference-based overlap inverts the central comparison. chrF++ ranks gold-conditioned responses *below* predicted-conditioned ones, the opposite of the judge's verdict, because open-ended empathetic replies sit near the floor of n-gram overlap against a single human reference (ROUGE-L ≈ 0.10, token Jaccard ≈ 0.07). A gold rationale moves a response toward being appropriate, not toward reproducing the particular words the human happened to use. We therefore report the judge as the primary response metric and overlap metrics as secondary, and we regard the disagreement itself as a finding: the instruments most commonly used to score empathetic generation would have reported the reverse of what we observe.

H Benchmark Composition

Table 18: Composition of the two released datasets.

Dataset	Split	Items	Languages
MELD-ST-Reason	English	1,008	original production
	German	1,164	dubbed
	Japanese	1,008	dubbed
	Total	**3,180**	parallel, within-corpus
MDET-Reason	EmotionTalk	3,173	Mandarin, performed
	MultiDialog	8,778	English, performed
	Total	**11,951**	cross-corpus

The emotion label distribution is naturally imbalanced, as conversational corpora are: the English MELD-ST-Reason split contains 479 neutral, 186 joy, 121 surprise, 85 anger, 73 sadness, 39 disgust, and 25 fear. We report macro-F_1 alongside accuracy throughout for this reason.

I Audio–Text Consistency by Model and Language

Table 19 gives the per-model breakdown behind the speech-channel result in §5.

J Reasoning Language

The benchmark can also vary the language λ in which the rationale is requested: six values (English, German, Japanese, Chinese, Malay, Spanish) crossed with the three interaction languages of MELD-ST-Reason. λ appears only in the prompts of conditions that elicit a rationale; the conditions that do not are identical across arms and act as a placebo. Gold rationales are machine-translated into each λ and audited in Appendix K.

Restricted to the two conditions whose prompt contains the rationale language (the other two, a placebo, are flat across arms: -0.001), matching the rationale to the speech language

Table 19: Consistency between the generated audio and the model's own generated text on the unified task, by interaction language. CER is the character error rate of a Whisper-large-v3 transcription against that text; F is the share of utterances within 0.3 WER of it; WER splits on whitespace, so F is not meaningful for Japanese, and CER is the comparable figure across the three scripts. ASR is forced to the expected language, and no kana normalisation is applied.

	ENG		DEU		JPN	
Model	CER ↓	F ↑	CER ↓	F ↑	CER ↓	F ↑
GLM-4-Voice	0.074	0.93	0.274	0.68	0.847	0.01
Kimi-Audio	1.124	0.24	1.197	0.10	1.372	0.00
Qwen2.5-Omni	0.041	0.96	0.581	0.06	0.972	0.00
VITA-1.5	0.022	0.98	0.430	0.08	0.951	0.00
Baichuan-Omni-1.5	0.093	0.87	0.470	0.12	0.522	0.03
Fun-Audio-Chat-8B	0.024	0.98	0.077	0.88	0.757	0.00
Qwen3-Omni	0.005	0.99	0.057	0.92	0.079	0.55
Step-Audio-2	0.113	0.92	0.187	0.57	0.144	0.38
MiMo-Audio-7B	0.022	0.98	0.415	0.14	0.484	0.05
OSUM-EChat	0.011	0.97	0.247	0.51	0.756	0.01
FreezeEmpath	0.026	0.99	0.567	0.03	1.179	0.00
OpenS2S-1.5	0.054	0.90	0.431	0.29	0.420	0.51
Mean CER	**0.134**		**0.411**		**0.707**	

Table 20: Emotion accuracy by requested rationale language (rows) and speech language (columns), restricted to the two conditions whose prompt contains the rationale language, and excluding cells in which fewer than half of the rows could be scored. n is the number of models per cell (of twelve). Matched cells are underlined.

Rationale language	ENG	DEU	JPN	Mean
English	<u>0.447</u> (12)	0.342 (12)	0.311 (11)	0.367
German	0.428 (11)	<u>0.329</u> (11)	0.315 (11)	0.358
Japanese	0.377 (11)	0.277 (10)	<u>0.274</u> (12)	0.309
Chinese	0.420 (10)	0.316 (10)	0.301 (10)	0.346
Malay	0.393 (12)	0.297 (11)	0.269 (11)	0.320
Spanish	0.434 (12)	0.334 (11)	0.337 (12)	0.368

gives no advantage (+0.003, $p = 0.64$), and no relatedness hypothesis (shared script, language family, cultural sphere) survives correction for nine tests (smallest Holm-adjusted $p = 0.12$; Table 21); rationales in English, German or Spanish beat those in Chinese, Japanese or Malay in 10/12 models (+0.037, uncorrected $p = 0.014$). The rationale language does affect the reply: a matched rationale raises the share of replies in the speech's language by +0.06, +0.13 and +0.17 for Japanese, German and English speech, and among the 51 of 60 model–language pairs that follow the instruction, a foreign rationale lowers it by 13 points relative to an English one.

Table 20 is the grid underlying the contrasts of Table 21. It uses only the two rationale-eliciting conditions, because the rationale language appears only in their prompts; in the other two emotion-producing conditions the prompt is identical across arms and accuracy agrees to within 0.003, which we use as a placebo. Cells scored on fewer than half their rows are excluded; an earlier version of this grid averaged all four conditions and included a German–Spanish FreezeEmpath cell scored on a single row (read as accuracy 1.000), which had produced a spurious "Spanish is best for German speech" result. The English-rationale row is the highest for German speech (0.342 against 0.334 for Spanish), and the matched Japanese-rationale cell is the lowest for Japanese speech (0.274).

Rationale and reply language. The language-control figures in Section 5 identify languages with a rule-based detector (script for Japanese, Chinese and Korean; function-word profiles for English, German, Spanish and Malay) that scores 99.9–100% on gold rationales.

Table 21: Does linguistic relatedness between the rationale language and the speech predict emotion accuracy? Each row is a paired per-model contrast over the two rationale-eliciting conditions (the only ones whose prompt contains the rationale language), excluding cells scored on fewer than half their rows. p is a one-sample t-test on the per-model contrasts; Holm corrects across the nine rows. No contrast survives the correction.

	Contrast	Speech	Δ acc. [95% CI]	Models	p	Holm
H1	Chinese (Han) rationale vs Latin-script rationale	JPN	−0.025 [−0.065, +0.007]	6/11	0.219	0.66
H2	Han-script rationale vs Latin-script rationale	JPN	−0.030 [−0.059, −0.003]	4/12	0.066	0.40
H3	Eastern rationale vs Western rationale	JPN	−0.045 [−0.080, −0.016]	1/12	0.021	0.17
H4	Western rationale vs Eastern rationale	ENG, DEU	+0.034 [+0.010, +0.062]	10/12	0.033	0.23
H5	Germanic rationale vs non-Germanic rationale	ENG, DEU	+0.028 [+0.002, +0.057]	8/12	0.087	0.44
H6	Latin-script rationale vs Han-script rationale	all	+0.021 [−0.001, +0.044]	6/12	0.105	0.44
H7	rationale = speech language vs all other languages	all	+0.003 [−0.008, +0.015]	8/12	0.640	0.98
H8	Western rationale vs Eastern rationale	all	+0.037 [+0.015, +0.061]	10/12	0.014	0.12
H9	Western gain on Western speech vs on Eastern speech	all	−0.011 [−0.046, +0.014]	6/12	0.492	0.98

Table 22: Objective quality of the translated rationale corpus, averaged over the three interaction splits ($n = 400$ aligned items each). Matched is the cross-lingual cosine between a rationale and its own translation; Mismatch is the same against *other* items' translations, and Margin is the separation between them. LangID verifies the output is actually in the target language. Leak is the rate at which a rationale names an emotion label, which the protocol forbids. The control line is the key to reading this table: two *different* English rationales are already 0.817 similar, so the corpus is formulaic and the margin, not the absolute similarity, is the informative quantity.

Target	Matched	Mismatch	Margin	LangID	Leak	Acoustic	Len. ratio
Chinese	0.836	0.739	0.097	99.9%	0.3%	7.5	0.26
Japanese	0.856	0.755	0.101	99.9%	0.0%	7.7	0.41
German	0.896	0.786	0.110	100.0%	0.1%	6.5	1.14
Spanish	0.924	0.789	0.135	100.0%	0.2%	6.6	1.05
Malay	0.865	0.751	0.114	100.0%	0.7%	5.0	1.02
English source	—	0.817	—	—	0.2%	9.0	1.00

Compliance is the share of rationales in the requested language in the reasoning-only condition, averaged over the three speech languages; reply language is measured in the reasoning-to-response condition, where the reply follows the model's own rationale, over the replies the detector can place (very short replies such as "Okay." carry no function words). The leakage contrast pairs each model with itself under an English rationale and is restricted to cases where the model both followed the foreign instruction and replied in the speech's language when reasoning in English ($\geq 90\%$ each), so that it measures contamination rather than disobedience or a model's general failure to answer in the speaker's language.

K Quality of the Translated Rationale Corpus

The five non-English reasoning arms are produced by translating the English rationale, so a translation that silently degraded would confound every cross-lingual conclusion in §5. The pipeline's own `translation_status` field reports `ok` for 1,008/1,008 items in every language, which establishes only that the translation call returned. We therefore measure the output directly on five axes (Table 22), on 400 aligned items per interaction split.

Reading the table requires the control. Cross-lingual cosine similarity between a rationale and its own translation is 0.836–0.924. Taken alone that number is uninterpretable, because two *different* English rationales are already 0.817 similar to each other. The gold rationales are formulaic: they describe the same acoustic dimensions—pitch, loudness, tempo, pauses, stress, timbre—in largely the same order. The informative quantity is therefore the *margin* between matched and mismatched pairs, which is positive for every language and every split (+0.097 to +0.135), and matched similarity exceeds the same-language different-

item ceiling throughout. Translations carry item-specific content, not merely the shared template.

Target-language identity. Every translation is in its target language (99.9–100%). For Chinese and Japanese this is verified by script; for German, Spanish and Malay, which share the Latin script with the source, a script test cannot distinguish a translation from an untouched passthrough, so we instead require lexical divergence from the English original. No language shows passthrough.

Emotion-label leakage. The protocol requires a rationale to cite audible properties without naming the emotion. Leakage is 0.0–0.7% against a source baseline of 0.2%, so translation does not introduce the answer.

This figure required care. An initial pass reported 34.8% leakage for Spanish and 21.8% for Malay. The Spanish figure was a substring artefact: the emotion term *ira* (anger) matches *respira* (breathes), which is an *acoustic* term, and disappears under word-boundary matching. The Malay figure was real as a string match but not as leakage: *terkejut* renders English "gasp" (192 of 195 co-occurrences) inside the boilerplate acoustic checklist, and it appears *less* often when the gold label is surprise (11.6%) than when it is not (20.1%). We report the verified figure and note the failed measurement, because both errors would have inflated a reported defect by roughly two orders of magnitude.

Length and acoustic content. Length ratios behave as the writing systems predict: 0.26 for Chinese and 0.41 for Japanese, which are denser per character, and 1.02–1.14 for the Latin-script targets. Acoustic-term counts fall from 9.0 in the source to 5.0–7.7 in translation, with Malay lowest. We treat this axis as indicative rather than decisive, because it depends on per-language lexicon coverage that we wrote by hand and that is certainly less complete for Malay than for German.

One defect found. Every translated row carries a duplicated `<REASONING>` opening tag (1,008/1,008 rows for Chinese, Japanese, Spanish and Malay; 1,003/1,008 for German), introduced when the translation was wrapped. The tag is stripped before scoring and before the rationale is fed to a model, so it does not affect the reported numbers, but it is present in the released files and we flag it rather than silently repairing it after the fact.

What this does and does not establish. It establishes that the translations are in the right language, preserve item-specific content, and do not leak the label. It does not establish that they are fluent or idiomatic; that would need human raters, and it is part of the same outstanding validation as the rationales themselves. The formulaicity of the source rationales, surfaced by the control, is a separate and more consequential finding, and we treat it as a limitation in its own right.