BAEmpEval: A Framework for Probing Large Audio Language Models on Multidimensional Empathetic Response Assessment

Jingxi Liang[1,*] **Pengze Guo**[1,*] **Zhiwen Xie**[1,2,†]
Qifeng Wang[1] **Yujia Huo**[3] **Derek F. Wong**[1,†]

[1]NLP[2]CT Lab, Department of Artificial Intelligence, University of Macau
[2]School of Computer Science, Central China Normal University
[3]School of Data Science and Information Engineering,
Guizhou Minzu University, Guizhou, China
nlp2ct.{jingxi1,pengze,qifeng}@gmail.com
zwxie@ccnu.edu.cn, huo.yujia@gzmu.edu.cn, derekfw@um.edu.mo

Abstract

Large Audio Language Models (LALMs) are increasingly applied in affective and conversational AI, yet their ability to align with human empathetic judgments in spoken dialogue remains poorly understood. We introduce **BAEmpEval**, a multimodal evaluation framework for analyzing empathetic response judgment in instruction-following LALMs under subjective conversational settings. The framework contains 5,436 dialogue pairs collected from three speech sources, where models evaluate listener responses across six fine-grained dimensions spanning both vocal perception (Vocal Warmth, Perceived Genuineness, Prosodic Attunement) and conversational response strategies (Exploration, Interpretation, Emotional Reaction). Experiments reveal a challenging setting: experiments across recent LALMs reveal uneven alignment with human relational judgments across empathy dimensions and speech conditions. Our findings suggest that empathetic response judgment remains a challenging multidimensional perceptual task for current LALMs and highlight the importance of evaluation frameworks that account for subjective relational variability. We release benchmark annotations, source identifiers, and evaluation code to support future research in human-centric AI.[1]

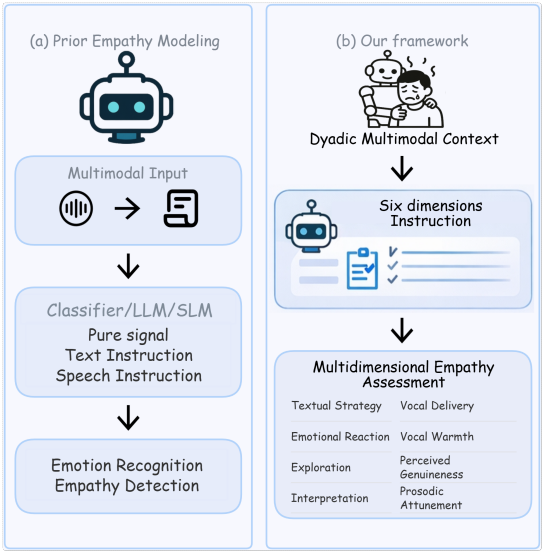


Figure 1: Comparison of previous empathy evaluation formulations and our six-dimensional listener-response assessment framework. BAEmpEval decomposes empathetic response judgment into three vocal-delivery dimensions — Vocal Warmth, Perceived Genuineness, and Prosodic Attunement — and three conversational dimensions — Emotional Reaction, Exploration, and Interpretation.

1 Introduction

Large Audio Language Models (LALMs) are rapidly improving in spoken dialogue understanding, multimodal interaction, and audio-based reasoning (Wang et al., 2025a; Cui et al., 2025; Yang et al., 2024). As these systems become increasingly integrated into conversational assistants, social agents, and spoken support applications, an important open question is whether they can reliably judge empathetic listener responses in spoken interaction. Unlike text-only settings, spoken empathetic responses are shaped not only by lexical content, but also by vocal delivery cues such as prosody, pacing, tone, and emotional regulation (Natale, 1978; Welkowitz and Kuc, 1973).

Existing empathy evaluation settings commonly operationalize empathy through holistic response scores, affective labels, or text-centered conversational evaluation (Sharma et al., 2020; Xie et al., 2024; Luo et al., 2026). However, listener responses may express supportive behavior through

[*]Equal contribution. [†]Corresponding author.
[1]Link to the dataset: https://github.com/NLP2CT/BaeEmpEval

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2377–2403
November 6-10, 2026 ©2026 Association for Computational Linguistics

multiple distinct conversational and vocal properties. As illustrated in Figure 1, responses with similar overall empathetic impressions may differ substantially in warmth, emotional acknowledgment, interpretive engagement, or prosodic attunement. Consequently, coarse-grained empathy labels provide limited visibility into how LALMs judge different perceptual and conversational aspects of spoken listener behavior.

Empathetic listener-response judgment is also inherently subjective. Perceptual attributes such as warmth, sincerity, or emotional attunement may admit multiple plausible interpretations across annotators, particularly in spoken interaction where vocal delivery can alter perceived intent. Prior work in subjective speech perception suggests that annotator disagreement may reflect meaningful perceptual variation rather than simple annotation noise (Chou et al., 2025). Empathetic response evaluation should therefore be treated as an ambiguity-sensitive perceptual judgment task rather than a fully deterministic classification problem.

To study how LALMs behave under these ambiguity-sensitive conditions, we introduce **BAEmpEval**, a framework for behavioral evaluation of empathetic listener-response judgment in spoken dialogue. Rather than treating empathy as a single global construct, BAEmpEval decomposes listener-response evaluation into six fine-grained dimensions spanning Vocal Delivery — Vocal Warmth, Perceived Genuineness, and Prosodic Attunement — and Conversational Behavior — Emotional Reaction, Exploration, and Interpretation. The framework contains 5,436 multimodal dialogue-response pairs collected from real-world, scripted, and synthetic speech settings, together with a human-validated gold subset designed for ambiguity-aware evaluation. Example annotations from multiple sources are shown in Table 1.

Using this framework, we evaluate recent frontier and open-source LALMs under zero-shot multimodal conversational settings. Our experiments reveal substantial variation in empathetic judgment capability across current systems. Stronger multimodal models exhibit meaningful sensitivity toward several vocal-delivery dimensions, while conversational dimensions requiring contextual interpretation remain substantially more difficult. Models also overpredict supportive or highly empathetic judgments, suggesting persistent calibration biases toward socially positive interpretations.

Our contributions are threefold. We introduce **BAEmpEval**, a behavioral framework for evaluating LALM judgments of empathetic listener responses across six vocal and conversational dimensions. We release a reproducible benchmark of 5,436 real-world, scripted, and synthetic dialogue-response pairs, including a 1,007-sample human-validated gold subset. Experiments across recent LALMs reveal substantial differences in human alignment, score calibration, and sensitivity to speech conditions and annotation ambiguity.

2 Empathetic Response Judgment

BAEmpEval evaluates how Large Audio Language Models (LALMs) judge spoken empathetic listener responses under multimodal conversational settings. Rather than treating empathy as a single holistic score, it evaluates complementary conversational and vocal dimensions. We use *behavioral evaluation* to denote observable model judgments and their alignment with human annotations, without assuming genuine empathy, emotional experience, or human-like internal understanding.

2.1 Empathy Perception and Generation

Empathy perception and empathy generation represent related but distinct capabilities. Generation-oriented evaluation asks whether a model can produce an appropriate empathetic response, whereas perception-oriented evaluation asks whether a model can recognize and differentiate the qualities that make an existing response more or less empathetic. BAEmpEval focuses on the latter.

We do not assume that strong empathy perception necessarily implies strong empathetic generation, nor that a model capable of producing empathetic responses will necessarily provide well-calibrated judgments of empathy. Establishing the relationship between these two capabilities remains an open empirical question. We therefore isolate perception as an independently measurable capability before studying how it interacts with generation.

This perception-oriented formulation also has value beyond diagnostic benchmarking. Reliable empathetic response judgment may support automatic evaluation of spoken dialogue systems, scalable annotation of multimodal conversational data, and calibration of automatically generated labels. More broadly, BAEmpEval provides a controlled setting for future studies investigating whether improving models' sensitivity to empathetic qualities leads to better empathetic response generation.

Source	Dialogue Example	Annotations (0–2)	Explanation
MESC	**Listener:** "You wanna write everybody's story for them?" **Speaker:** "I really think I'd be a shitty writer. I'd want to give everybody a happy ending." **Listener:** "Paul, I'm very sorry about your patient."	E:2 Ex:0 I:1 G:2 P:0 W:2	Sympathetic text, warm, emotionally congruent delivery, but lacks exploration; flat prosody.
Anno-MI	**Listener:** "Do you have a plan as to where you might want to go?" **Speaker:** "Um, I'm not sure." **Listener:** "So, kind of keeping your options open."	E:0 Ex:0 I:1 G:1 P:2 W:1	Paraphrases uncertainty; neutral prosody, but no emotional engagement.
AvaMERG	**Listener:** "Did you call the police?" **Speaker:** "No but if it happens again I will. I ordered a camera for my porch. I couldnt get any sleep that night." **Listener:** "Do you live in a house or an apartment?"	E:0 Ex:1 I:0 G:1 P:1 W:0	Factual clarification, detached, minimally expressive delivery and warmth.

Table 1: Example annotated responses from three dataset sources. Each listener response is evaluated on six empathy dimensions (0–2 scale). E=Emotional Reaction, Ex=Exploration, I=Interpretation, G=Perceived Genuineness, P=Prosodic Attunement, W=Vocal Warmth.

Framework	Modality	Scale	Dataset Size	Dimensions	Judgments
Epitome (Sharma et al., 2020)	Text	0–2	10,143	3	30k
MEDIC (Zhu et al., 2023)	Video+Text	0–2	771	3	2,313
EMPSPEECH (Srikanth et al., 2025)	Video	0–1	1,718	1	1,718
SpeechEmpathy (Chen et al., 2024a)	Speech	0–1	346	1	346
BAEmpEval (Ours)	Speech+Text	0–2	1,007 + 5,436 silver	6	6,042

Table 2: **BAEmpEval in context.** Existing empathy benchmarks typically focus on either conversational behavior or coarse affective perception. BAEmpEval jointly evaluates conversational and vocal empathetic judgment across six dimensions. The benchmark contains 1,007 gold interactions annotated across six perceptual dimensions (6,000 total judgments), together with 5,436 silver-annotated samples.

2.2 Empathetic Response Dimension Design

Existing empathy evaluation frameworks often capture only part of spoken empathetic interaction. Text-based approaches primarily evaluate supportive conversational behaviors expressed through language (Sharma et al., 2020), while speech- and video-based settings frequently reduce empathy to coarse affective or binary judgments (Chen et al., 2024a; Srikanth et al., 2025). Consequently, existing formulations provide limited visibility into how models judge distinct conversational and vocal properties of spoken listener responses. BAEmpEval extends response-centered empathy evaluation into spoken multimodal interaction through six dimensions spanning *vocal delivery* and *conversational behavior*. These dimensions are not intended to constitute an exhaustive taxonomy of human empathy. Rather, they provide an interpretable operationalization of spoken empathetic response perception that combines established conversational empathy strategies with relational qualities of vocal delivery.

Vocal dimensions. Our vocal dimensions are motivated by person-centered therapy principles (Rogers, 1957), which emphasize warmth, genuineness, and empathic attunement as central relational qualities of supportive communication. Prior work further shows that listeners rely on paralinguistic cues to infer warmth, attentiveness, and interpersonal engagement (Natale, 1978; Welkowitz and Kuc, 1973), while authentic and acted emotional speech can produce different perceptual impressions (Drolet et al., 2012). Based on these principles, BAEmpEval evaluates three vocal dimensions. **Vocal Warmth** captures whether the delivery is perceived as emotionally supportive, comforting, or cold and detached. **Perceived Genuineness** captures whether the response sounds sincere and emotionally congruent rather than performative, sarcastic, or artificial. **Prosodic Attunement** captures whether pacing, intonation, expressiveness, and other aspects of delivery are perceived as appropriate to the conversational context. Importantly, Prosodic Attunement does not require explicit acoustic matching between the speaker and listener. These vocal dimensions should be interpreted as human-perceptual relational constructs, rather than independent acoustic primitives. Perceived warmth or genuineness may therefore be influenced jointly by prosody, lexical content, speaker characteristics, speech naturalness, recording conditions, and conversational context. Such influences are not necessarily annotation errors be-

cause they may themselves contribute to human perception of the response. Accordingly, BAEmpEval does not attempt to causally isolate individual acoustic cues; instead, it evaluates whether LALM judgments align with holistic human perception under realistic multimodal conditions.

Conversational dimensions. For conversational behavior, we adopt the established framework of (Sharma et al., 2020), consisting of **Exploration**, **Interpretation**, and **Emotional Reaction**. Exploration captures whether the listener actively investigates the speaker's situation or emotional state; Interpretation captures whether the response demonstrates understanding or reflection of the speaker's experience; and Emotional Reaction captures whether the listener explicitly or implicitly acknowledges the speaker's emotional state. The six dimensions are evaluated independently rather than combined into a single additive empathy score. A response may, for example, be vocally warm while offering little exploration, or demonstrate strong interpretation while sounding emotionally detached. This multidimensional formulation therefore enables analysis of which specific aspects of spoken empathetic behavior current LALMs can or cannot reliably judge. Detailed operational definitions, annotation instructions, and score-level criteria (0–2) for all six dimensions are provided in Appendix B.

2.3 Annotation and Measurement Design

Spoken empathetic listener-response judgment is inherently subjective, particularly for perceptual properties such as warmth, genuineness, and prosodic attunement. Different annotators may interpret the same response differently because of variation in perceived sincerity, conversational appropriateness, or vocal delivery. Prior work in subjective speech perception further suggests that disagreement may reflect meaningful listener-dependent variation rather than purely erroneous labeling (Chou et al., 2025). BAEmpEval treats empathetic response judgment as an ambiguity-sensitive perceptual evaluation task rather than a fully deterministic classification problem.

To support consistent comparison across conversational and vocal dimensions, we adopt a shared three-level ordinal scale (0–2). Binary labels cannot distinguish intermediate or ambiguous response behaviors, whereas substantially finer-grained scales may introduce distinctions that annotators cannot reliably separate, particularly for perceptual dimensions such as genuineness (Kolden et al., 2019). Table 2 situates BAEmpEval relative to existing empathy evaluation frameworks. We report both Exact Accuracy and Within-1 (± 1) Accuracy. Exact Accuracy evaluates whether model predictions match the majority human judgment at the same ordinal level, while Within-1 Accuracy measures whether predictions remain within a neighboring category under a more tolerant ordinal criterion. However, Within-1 is not intended to serve as a standalone measure of model quality. On a three-point scale, high Within-1 scores can be obtained even by models whose predictions collapse toward a dominant category. We therefore interpret it jointly with Exact Accuracy, prediction distributions, score-conditioned confusion matrices, and ambiguity-aware analyses. Human Within-1 agreement exceeds 90% across all dimensions (Appendix D), indicating that most human disagreement occurs between adjacent categories rather than incompatible interpretations.

2.4 Ambiguity-Aware Evaluation

Rather than reducing the three human annotations to majority labels alone, BAEmpEval retains their disagreement patterns and categorizes samples into four ambiguity situations. **Full Consensus** refers to cases where all annotators assign the same label. **Adjacent Deviation** captures disagreement restricted to neighboring ordinal labels (e.g., 2-2-1), reflecting mild category-boundary uncertainty. **Polarized Deviation** refers to cases containing distant disagreement (e.g., 2-2-0), indicating stronger interpretive divergence. Finally, **Non Consensus** describes cases where all three annotators assign different labels (e.g., 0-1-2), suggesting substantially different interpretations. These categories do not redefine the benchmark ground truth; majority voting remains the standardized reference used for aggregate model comparison. Instead, they contextualize model performance according to the degree and form of human uncertainty. This allows us to distinguish model errors on relatively unambiguous examples from disagreements arising in cases where human listeners themselves differ, rather than treating all annotation variability uniformly as noise.

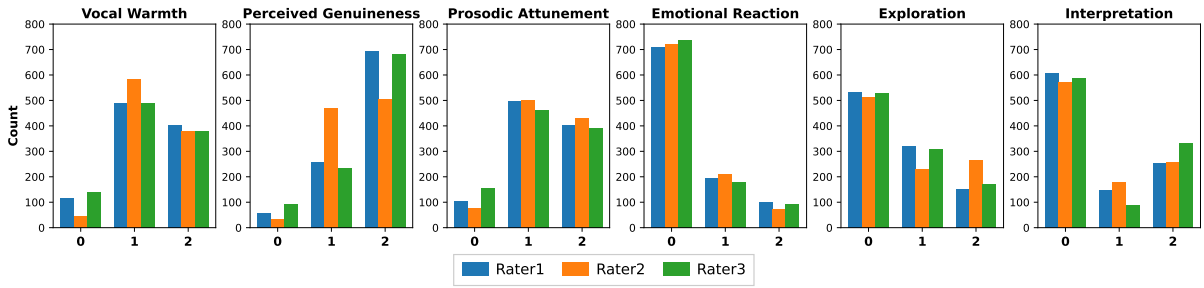


Figure 2: Distribution of human ratings in the gold subset of BAEmpEval across the six empathy dimensions. Ratings are annotated on a three-point ordinal scale (0–2). The distributions differ markedly across dimensions, reflecting behaviors in natural multimodal interactions.

Dim.	R1–R2	R1–R3	R2–R3
VW	0.330	0.656	0.512
PG	0.380	0.645	0.466
PA	0.435	0.587	0.653
ER	0.670	0.836	0.757
EX	0.761	0.870	0.840
INT	0.662	0.791	0.791

Table 3: Pairwise Quadratic Weighted Kappa (QWK) across the full annotation set. Abbreviations: VW = Vocal Warmth, PG = Perceived Genuineness, PA = Prosodic Attunement, ER = Emotional Reaction, EX = Exploration, INT = Interpretation.

3 Dataset Construction

3.1 Task Formulation

We formulate empathetic response judgment as a turn-level multimodal ordinal prediction task on the listener response. For each dialogue turn, the model receives (i) preceding dialogue context as text, (ii) the listener's textual response, and (iii) listener-response audio, and predicts ordinal scores (0–2) on six empathetic response dimensions.

The framework is designed to probe perceptual properties of the listener response rather than full dyadic interaction modeling. While some dimensions (e.g., Vocal Warmth, Perceived Genuineness) are largely judgeable from the response itself, others (e.g., Exploration, Interpretation) require conversational context. We therefore provide textual context while treating the listener response as the primary analysis target. Providing both listener text and audio prevents conversational-strategy judgments from being dominated by transcription errors while enabling evaluation of vocal delivery.

Speaker context is provided as text only in the primary setting. Current LALM interfaces provide limited standardized support for multi-speaker audio conditioning, and exploratory sequential-audio prompting produced inconsistent speaker attribution across models. We therefore exclude speaker audio from the primary setting to maintain comparable conditions across model families. Speaker-audio analyses are reported in Appendix I.

3.2 Dataset Construction

Dataset Sources. To evaluate empathetic response judgment across diverse conversational and acoustic conditions, we construct BAEmpEval using three complementary spoken-dialogue settings. First, we use real-world motivational interviewing conversations from Anno-MI (Wu et al., 2022), a counseling corpus widely used in empathy-related research (Kumar et al., 2024; Cohen et al., 2024; Kim et al., 2025). Anno-MI serves as the primary conversational source due to its relatively natural listener-support interactions and clinically grounded conversational structure. Second, we incorporate scripted therapeutic dialogue from MESC (Chu et al., 2025). Third, we use the synthetic training split released for the AvaMERG Challenge (Zhang et al., 2025), derived from EmpatheticDialogues (Rashkin et al., 2019). Its TTS-generated speech provides a complementary condition for examining model behavior under reduced vocal naturalness. Together, these datasets provide variation in conversational realism, speaking style, and acoustic naturalness while remaining centered on supportive listener-response evaluation.

Data Selection. All samples are drawn from previously released research resources rather than newly collected conversations. For Anno-MI, we retain interactions spanning both higher- and lower-quality listener behavior. For MESC, we filter heavily dramatized or sensitive scenarios outside the intended scope of ordinary supportive interaction. For AvaMERG, we retain samples from the synthetic challenge training split with valid listener

Statistic	Anno-MI	MESC	AvaMERG
Silver	3,858	733	845
Gold	627	132	248
σF_0 (Hz)	69.9 (57.9)	47.7 (37.9)	39.7 (42.8)

Table 4: Dataset composition and annotation statistics. Silver annotations were generated by Gemini-2.5-Pro; the gold subset was independently annotated by three researchers. Within-utterance pitch variability (σF_0) summarizes prosodic variability across datasets; values are reported as Gold (Silver).

audio. Across all sources, samples with missing, inconsistent, or unusable audio are removed before annotation and evaluation. Because the resulting data remain empathy-oriented rather than class-balanced, we complement aggregate accuracy with prediction-distribution, class-conditional, and ambiguity-aware analyses. To support reproducibility, the exact sample list used in BAEmpEval, together with source identifiers and filtering metadata, will be released in our public repository.

3.3 Annotation Methodology

Three researchers with experience in affective computing and conversational AI independently annotated the human-validated gold subset following structured calibration involving guideline study, pilot annotation, disagreement discussion, and iterative refinement. The dimensions concern perceived properties of listener responses rather than therapeutic effectiveness or clinical diagnosis. The resulting gold subset contains 1,007 spoken dialogue-response pairs with listener audio, corresponding to 18,126 individual dimension-level judgments. Pairwise agreement is generally stronger for conversational dimensions than for the more subjective vocal dimensions, while human Within-1 agreement exceeds 90% across all dimensions (Appendix D), indicating that most disagreements occur between neighboring categories. Inter-dimension Spearman correlations suggest related but non-redundant constructs: among majority-vote labels, the three vocal dimensions are moderately correlated ($\rho = .46$–$.58$), whereas correlations between vocal and conversational dimensions are weaker ($\rho = -.18$–$.32$; Appendix E). Three annotators provide a practical compromise between benchmark scale and multimodal annotation cost. We therefore retain individual ratings and separately analyze consensus and disagreement patterns rather than treating majority vote as absolute ground truth; larger annotator pools remain for future work.

Silver Annotation. To support larger-scale behavioral analysis, we additionally generate silver annotations for all 5,436 dialogue turns using Gemini-2.5-Pro (Comanici et al., 2025). Gemini was selected because it showed comparatively stable multidimensional scoring and strong agreement with human annotators on the gold subset.

The silver annotations are not intended to replace human ground truth. Instead, Gemini provides scalable paired scores and rationales that can serve as an initial annotation layer for later human inspection and correction. The gold subset remains the primary evaluation reference, while the silver set supports large-scale behavioral analysis and future annotation expansion. Human–Gemini agreement, label distributions, and inter-dimension correlations are reported in Appendix D.

3.4 Evaluation Metrics

Model predictions are evaluated primarily against majority-vote labels derived from the gold annotations. Majority voting provides a reference for aggregate comparison, while individual annotations are retained for ambiguity-aware analysis.

Exact Accuracy. Measures the proportion of predictions that exactly match the majority-vote label:

$$\mathbf{Acc}^{(k)}_{exact} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i^{(k)} = y_i^{(k)}),$$

where $y_i^{(k)} \in \{0, 1, 2\}$ denotes the gold score for instance i on empathy dimension k, $\hat{y}_i^{(k)}$ is the corresponding model prediction, and $\mathbf{1}(\cdot)$ is the indicator function.

Within-1 (One-Point-Off) Accuracy. Because the empathy scale is ordinal and neighboring categories may represent perceptually similar judgments, we additionally report Within-1 Accuracy:

$$\mathbf{Acc}^{(k)}_{\pm 1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|\hat{y}_i^{(k)} - y_i^{(k)}| \leq 1\right).$$

Within-1 is interpreted as a complementary measure of ordinal proximity rather than a standalone performance metric. On a three-point scale it is permissive: when the gold label is 1, every possible prediction falls within one point, while for labels 0 or 2, two of three predictions do so. Its expected random performance depends on the label distribution.

Complementary Metric Interpretation. Exact Accuracy captures strict agreement, whereas Within-1 captures broader ordinal proximity. We therefore interpret both jointly with prediction distributions, score-conditioned confusion matrices (Appendix G.1), and ambiguity-aware analyses, distinguishing genuine score discrimination from performance driven by label imbalance, prediction collapse, or human uncertainty. Together they reveal whether apparently strong aggregate performance reflects genuine discrimination among empathy levels or instead arises from label imbalance, prediction collapse, or human uncertainty.

4 Experimental Setup

Models. We evaluate seven recent large audio-language models (LALMs), spanning both proprietary and open-weight systems with diverse architectures and training domains. Our evaluation includes Qwen-3-Omni-Flash and Qwen-2.5-Omni (Xu et al., 2025; Chu et al., 2024), the Qwen-2-Audio-Instruct (Chu et al., 2024), and multimodal open-weight models including Audio-Flamingo-3 (Goel et al., 2025), and Music-Flamingo (Ghosh et al., 2025). GPT-4o-Audio-Preview (OpenAI et al., 2024) and Gemini-2.5-pro serves as a commercial baseline. All models are evaluated in a zero-shot setting using a standardized prompt requesting ordinal scores (0–2) for each dimensions that models can stably produce valid format. In case of failure the data is rerun with the same prompt. The final size for evaluation after filtering out empty prediction in some models are 929. Prompt templates are provided in Appendix F. We adopt a zero-shot setting to examine whether structured empathetic judgment emerges without task-specific calibration. While few-shot prompting may improve instruction adherence for some models, it can also introduce prompt-dependent calibration effects in ambiguity-sensitive perceptual evaluation tasks. We therefore use zero-shot prompting as the primary setting and report additional few-shot ablations in the appendix I.

5 Results and Analysis

Overall Empathetic Judgment Performance. Table 5 reports Exact Accuracy and Within-1 Accuracy on the human-validated gold set, with bootstrap confidence intervals provided in Table 21. Gemini-2.5-Pro achieves the strongest overall alignment, with Exact Accuracy ranging from 75.5% to 87.5% and Within-1 Accuracy from 94.7% to 98.5% across dimensions. Its Exact Accuracy numerically exceeds the leave-one-out human agreement reference on all six dimensions. However, because models are evaluated against majority-vote labels whereas human performance reflects leave-one-out inter-annotator agreement, we interpret this as strong alignment with the aggregate human reference rather than evidence of superhuman empathetic perception.

Performance among the remaining models is more uneven. Audio-Flamingo-3 performs strongly on Emotional Reaction, while Music-Flamingo achieves high Prosodic Attunement accuracy but substantially weaker Perceived Genuineness and Vocal Warmth. These contrasting profiles suggest that current LALMs capture different components of empathetic response judgment rather than exhibiting a uniform empathy capability.

Importantly, high Within-1 Accuracy does not necessarily indicate reliable discrimination among levels. Qwen2-Audio obtains 98–100% Within-1 Accuracy across several dimensions despite extremely low Exact Accuracy, illustrating how a permissive ordinal metric can remain high under collapsed predictions. We therefore interpret aggregate metrics jointly with the analyses below.

Prediction Calibration and Label Use. Figure 4 reveals substantial differences in how models use the three-point scale. Several models exhibit a strong preference for the highest empathy score, with Qwen2-Audio predicting score 2 for almost all samples across dimensions. Such collapse explains why Within-1 Accuracy can remain superficially high despite poor exact alignment.

In contrast, Gemini-2.5-Pro and GPT-4o-Audio-Preview make substantially broader use of the available scoring range, with prediction distributions varying across empathy dimensions. These results suggest that stronger models are not only more accurate but also better calibrated to differences in perceived empathy intensity. The prediction-distribution analysis complements aggregate accuracy by distinguishing genuine score discrimination from performance driven by majority-label bias.

Variation Across Conversational and Speech Conditions. Figure 5 compares model performance across Anno-MI, MESC, and AvaMERG. Gemini-2.5-Pro remains the strongest and most consistent model across the three sources, while other systems exhibit considerably greater source-

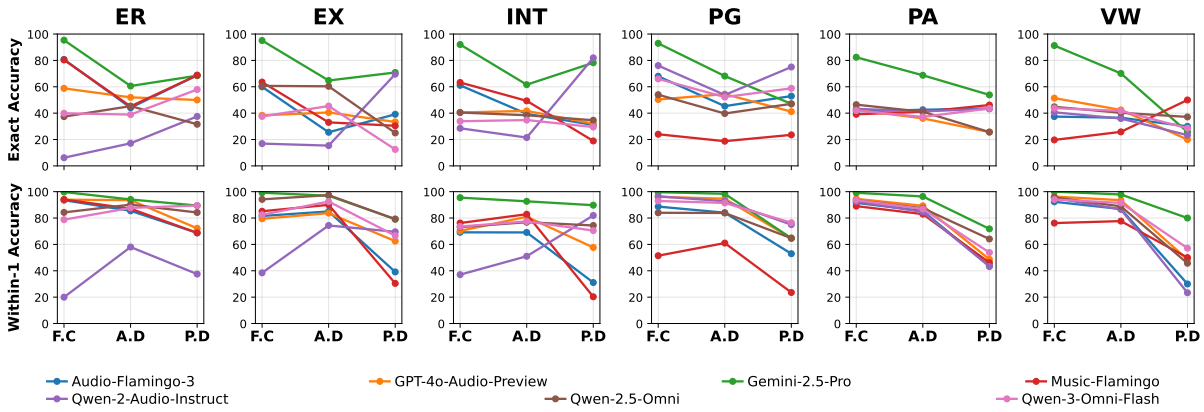


Figure 3: Models prediction trends across 3 ambiguity situation on six empathy dimensions. Each point corresponds to one evaluation dimension, and values indicate dimension-wise exact accuracy or within-one accuracy. VW = Vocal Warmth, PG = Perceived Genuineness, PA = Prosodic Attunement, ER = Emotional Reaction, EX = Exploration, and INT = Interpretation.

	Emotion		Exploration		Interpretation		Genuine		Prosody		Warmth	
Model	**E**	**W1**	**E**	**W1**	**E**	**W1**	**E**	**W1**	**E**	**W1**	**E**	**W1**
AudioFlamingo	72.8	91.6	52.1	81.3	53.3	67.2	56.9	86.3	41.1	87.6	36.3	88.1
Gemini-2.5-Pro	87.3	98.4	87.5	98.2	83.7	94.7	80.0	98.5	75.5	97.1	79.3	98.4
GPT4o	57.1	93.5	38.7	80.0	39.2	72.1	51.3	95.0	38.9	91.0	46.3	93.5
MusicFlamingo	73.7	95.8	69.9	96.0	70.8	92.2	27.4	81.9	82.1	99.0	24.8	77.0
Qwen2.5	33.6	81.4	49.8	86.7	44.6	77.5	53.0	95.2	30.3	84.2	62.8	95.5
Qwen2-Audio	22.1	98.9	20.2	98.4	17.2	98.3	0.9	99.9	1.1	99.0	0.3	99.8
Qwen3.0	50.5	86.0	49.7	81.6	42.3	68.6	47.0	89.7	39.4	88.7	53.8	89.9
Human	84.5	98.5	83.2	98.3	80.5	93.7	68.3	98.2	71.8	96.8	68.7	97.3

Table 5: Exact Accuracy (E) and Within-1 Accuracy (W1) on the full evaluation set using majority-vote labels(%). Human performance corresponds to average leave-one-out inter-annotator agreement.

and dimension-specific variation. AvaMERG provides a particularly informative contrast because its responses are synthesized with TTS. Across several models, Emotional Reaction and other content-dependent dimensions remain comparatively robust, whereas Vocal Warmth and Perceived Genuineness show larger degradation relative to natural or scripted speech. This pattern suggests that explicit supportive content may remain accessible under synthetic speech while vocal qualities related to naturalness and perceived sincerity are more sensitive to the characteristics of speech generation. These results further motivate evaluating conversational and vocal empathy dimensions separately.

Sensitivity to Human Annotation Ambiguity. Figure 3 examines model performance under Full Consensus (F.C.), Adjacent Deviation (A.D.), and Polarized Deviation (P.D.) among human annotators. Exact Accuracy is generally highest on samples with full human consensus and declines as annotator disagreement increases. This pattern is more clear for Gemini-2.5-Pro, whose strong performance on consensus examples decreases under disagreement, especially for the vocal dimensions.

Within-1 Accuracy is more stable between Full Consensus and Adjacent Deviation, consistent with adjacent human disagreements primarily reflecting uncertainty near ordinal category boundaries. Under Polarized Deviation, however, performance decreases more substantially across models and dimensions, indicating that cases producing qualitatively different human interpretations also remain difficult for LALMs. Because polarized cases constitute only a small portion of the gold set, it can be noisy and are interpreted primarily as diagnostic trends rather than strict monotonic relationships. Overall, 66.1% of gold annotations exhibit full consensus, 28.9% adjacent disagreement, and only 3.4% polarized disagreement (Table 6). Together with the complementary metric analysis, these results indicate that model–human disagreement reflects both model calibration failures and genuine ambiguity in human perception, motivating evaluation that considers not only aggregate accuracy but also label usage and annotator uncertainty.

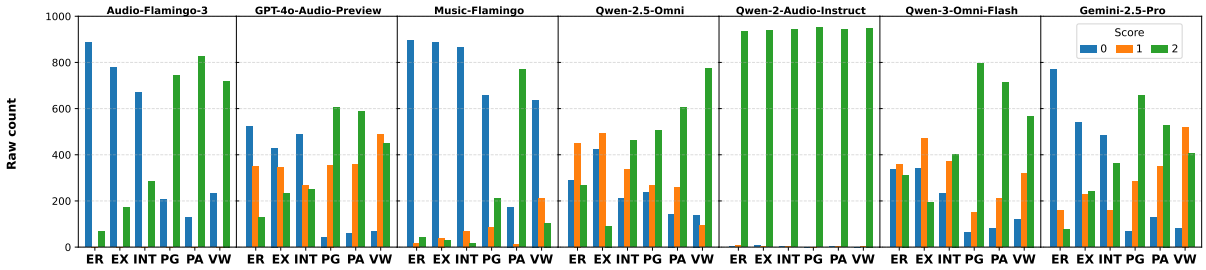


Figure 4: Models prediction distribution on six empathy dimensions. Each spoke corresponds to one evaluation dimension.VW = Vocal Warmth, PG = Perceived Genuineness, PA = Prosodic Attunement, ER = Emotional Reaction, EX = Exploration, and INT = Interpretation.

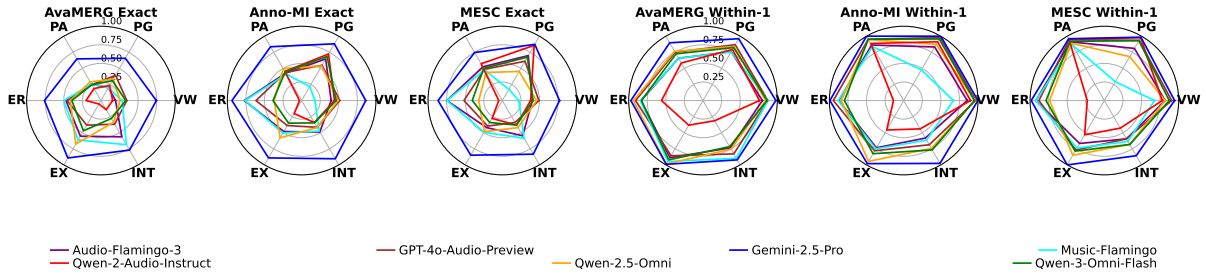


Figure 5: Comparing model performance across six empathy dimensions on three datasets (Anno-MI, MESC, and AvaMERG). Each spoke corresponds to one evaluation dimension, and values indicate dimension-wise exact accuracy. VW = Vocal Warmth, PG = Perceived Genuineness, PA = Prosodic Attunement, ER = Emotional Reaction, EX = Exploration, and INT = Interpretation.

Dimension	F.C.	A.D.	P.D.	N.C.
Emotional Reaction	78.7	19.2	1.5	0.6
Exploration	76.7	20.7	2.4	0.2
Interpretation	74.1	15.2	7.4	3.4
Perceived Genuineness	54.7	41.5	1.8	2.0
Prosodic Attunement	59.4	34.6	4.1	2.0
Vocal Warmth	53.4	42.3	3.2	1.2
Overall	66.1	28.9	3.4	1.6

Table 6: Distribution of annotation ambiguity types. F.C.=Full Consensus, A.D=Adjacent Deviation,P.D.= Polarized Disagreement, N.C.= No Consensus.

6 Conclusion

We introduced BAEmpEval, a framework for probing empathetic response judgment in spoken dialogue through six dimensions spanning vocal delivery and conversational behavior. Evaluating seven recent LALMs reveals uneven alignment with human relational judgments: models show partial sensitivity to vocal cues, but substantially weaker performance on conversational dimensions requiring contextual interpretation and calibrated empathy assessment. Across models, we observe prediction bias toward highly empathetic ratings and reduced robustness under synthetic speech conditions.

Our findings highlight empathetic response assessment as an ambiguity-sensitive and multidimensional perceptual task rather than a single holistic score. Human disagreement frequently reflects neighboring or plausible relational interpretations, motivating evaluation settings that account for ordinal uncertainty rather than treating all disagreement as annotation noise. We hope BAEmpEval provides a useful framework for studying multimodal conversational systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62266013, 62402194), the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST, MYRG-GRG2026-00018-FIC, MYRG-CRG2026-00007-FIC), the Dr. Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the Guizhou Provincial Basic Research Program (Grant No. MS[2026]276), the Macao Young Scholars Program (Grant No. AM2024015), and the China Postdoctoral Science Foundation under Grant Number 2024M751062.

Limitations

This work focuses on English-language spoken dialogue, and the proposed dimensions and observed model behaviors may not generalize across languages, cultures, or conversational norms. Broader multilingual and cross-cultural validation remains necessary.

The vocal dimensions in BAEmpEval are intended as perceptual relational constructs rather than isolated acoustic primitives. Human judgments of Vocal Warmth, Perceived Genuineness, and Prosodic Attunement may therefore be influenced jointly by lexical content, speaker characteristics, recording quality, speech naturalness, and conversational context. Our analyses show that the dimensions are empirically distinguishable, but we do not claim to causally disentangle the contribution of individual acoustic factors.

The gold subset is annotated by three researchers. Although this yields 18,126 dimension-level judgments and enables explicit analysis of consensus and disagreement, three annotators cannot fully characterize the population distribution of subjective empathetic perception. Future extensions should include larger and more diverse annotator pools.

Current LALM interfaces also provide limited standardized support for multi-speaker audio conditioning. We therefore evaluate listener audio with textual dialogue context in the primary setting and report exploratory speaker-audio prompting in Appendix I. Richer dyadic audio conditioning may reveal additional interactional cues.

Finally, the source datasets are not class-balanced and differ in conversational realism and speech naturalness. In particular, the synthetic AvaMERG Challenge training split may introduce source-specific artifacts. Although we use prediction-distribution, class-conditional, and ambiguity-aware analyses to reduce misleading interpretations, broader datasets with more balanced and naturalistic interaction conditions would strengthen future evaluation.

Ethical Considerations

Our original code and annotations are released under the MIT and CC BY 4.0 licenses, respectively. Source audio and transcripts are not redistributed and remain governed by the access conditions of the original datasets; released identifiers are intended solely to facilitate reconstruction for research use. This study does not collect new conversational data from human participants. All dialogue samples are drawn from previously released research datasets, and the additional gold annotations were produced by members of the research team on existing samples. We do not attempt to identify speakers or infer sensitive personal attributes beyond the perceptual labels required by the benchmark.

Because empathetic and counseling-related dialogue may contain sensitive content, we preserve the privacy protections and usage conditions of the original datasets and do not redistribute source materials beyond what is permitted by their respective licenses. Automated empathy ratings may mischaracterize speakers, particularly across cultural, linguistic, clinical, or synthetic-speech conditions, and may privilege majority perceptions over minority interpretations. BAEmpEval is intended solely for research evaluation and should not be used to diagnose individuals, assess therapists or patients, make clinical decisions, or rank people without domain-specific validation and human oversight.

References

Elham Aghakhani, Lu Wang, Karla T. Washington, George Demiris, Jina Huh-Yoo, and Rezvaneh Rezapour. 2025. From conversation to automation: Leveraging LLMs for problem-solving therapy analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25189–25207, Vienna, Austria. Association for Computational Linguistics.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems*, volume 37, pages 56898–56918. Curran Associates, Inc.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. EmpCRL: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, Turin, Italy. ELRA and ICCL.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and EngSiong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. In *The Thirteenth International Conference on Learning Representations*.

Run Chen, Haozhe Chen, Anushka Kulkarni, Eleanor Lin, Linda Pang, Divya Tadimeti, Jun Shin, and Julia Hirschberg. 2024a. Detecting empathy in speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1080–1084.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024b. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. 2025. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*.

Cheng-Han Chiang, Hung-yi Lee, and Michal Lukasik. 2025. TRACT: Regression-aware fine-tuning meets chain-of-thought reasoning for LLM-as-a-judge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2952, Vienna, Austria. Association for Computational Linguistics.

Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N. Salman, Chi-Chun Lee, and Carlos Busso. 2025. Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule. *IEEE Transactions on Affective Computing*, 16(1):41–55.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio technical report. *Preprint*, arXiv:2407.10759.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*, 27:8276–8287.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italy. ELRA and ICCL.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. VoxEval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16735–16753, Vienna, Austria. Association for Computational Linguistics.

Matthis Drolet, Ricarda I. Schubotz, and Julia Fischer. 2012. Authenticity affects the recognition of emotions in speech: behavioral and fMRI evidence. *Cognitive, Affective, & Behavioral Neuroscience*, 12(1):140–150.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71, Bangkok, Thailand. Association for Computational Linguistics.

Aya E. Fouda, Abdelrahamn A. Hassan, Radwa J. Hanafy, and Mohammed E. Fouda. 2025. PsychiatryBench: A Multi-Task Benchmark for LLMs in Psychiatry. *Preprint*, arXiv:2509.09711.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784, Vienna, Austria. Association for Computational Linguistics.

Charles J. Gelso and Jean A. Carter. 1994. Components of the psychotherapy relationship: Their interaction and unfolding during treatment. *Journal of Counseling Psychology*, 41(3):296–306.

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, and 4 others. 2025. Osum-echat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *Preprint*, arXiv:2508.09600.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Music flamingo: Scaling music understanding in audio language models. *Preprint*, arXiv:2511.10289.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Md Rakibul Hasan, Md Zakir Hossain, Shreya Ghosh, Aneesh Krishna, and Tom Gedeon. 2025. Empathy

detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods. *IEEE Transactions on Affective Computing*, 16(4):2507–2525.

Md Rakibul Hasan, Md Zakir Hossain, Aneesh Krishna, Shafin Rahman, and Tom Gedeon. 2024. Thesis proposal: Detecting empathy using multimodal language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 338–349, St. Julian's, Malta. Association for Computational Linguistics.

Yuna Hong, Bonhwa Ku, and Hanseok Ko. 2025. Performance evaluation metrics for empathetic LLMs. *Information*, 16(11):977.

Guanyu Hou, Jiaming He, Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, and Wenbo Jiang. 2025. Evaluating robustness of large audio language models to audio injection: An empirical study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25671–25687, Suzhou, China. Association for Computational Linguistics.

Yifan Hu, Rui Liu, Yi Ren, Xiang Yin, and Haizhou Li. 2025. Chain-talker: Chain understanding and rendering for empathetic conversational speech synthesis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1988–2003, Vienna, Austria. Association for Computational Linguistics.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. Apathetic or empathetic? evaluating LLMs' emotional alignments with humans. In *Advances in Neural Information Processing Systems*, volume 37.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7B. *Preprint*, arXiv:2310.06825.

Liting Jiang, Di Wu, Bohui Mao, Yanbing Li, and Wushour Slamu. 2023b. Empathy intent drives empathy detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Singapore. Association for Computational Linguistics.

Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. 2025. KMI: A dataset of Korean motivational interviewing dialogues for psychotherapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10803–10828, Albuquerque, New Mexico. Association for Computational Linguistics.

Gregory G. Kolden, Chia-Chih Wang, Sarah B. Austin, Ying Chang, and Marjorie H. Klein. 2019. Congruence/genuineness. In John C. Norcross and Michael J. Lambert, editors, *Psychotherapy Relationships That Work: Evidence-Based Therapist Contributions*, 3 edition, pages 323–350. Oxford University Press, New York, NY.

Vivek Kumar, Pushpraj Singh Rajwat, Giacomo Medda, Eirini Ntoutsi, and Diego Reforgiato Recupero. 2024. Unlocking LLMs: Addressing scarce data and bias challenges in mental health and therapeutic counselling. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 238–251, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

Minzhi Li, William Held, Michael J. Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, and Diyi Yang. 2025. Mind the gap: Static and interactive evaluations of large audio models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8749–8766, Vienna, Austria. Association for Computational Linguistics.

Xuan Luo, Lewei Yao, Libo Zhao, Lanqing Hong, Kai Chen, Dehua Tao, Daxin Tan, Ruifeng Xu, and Jing Li. 2026. Aeq-bench: Measuring empathy of omnimodal large models. *Preprint*, arXiv:2601.10513.

Muhammad Arslan Manzoor, Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Can machines resonate with humans? evaluating the emotional and empathic comprehension of LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14683–14701, Miami, Florida, USA. Association for Computational Linguistics.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, and 57 others. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *Preprint*, arXiv:2503.01743.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2025. Speech-integrated modeling for behavioral coding in counseling. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 152–158, Avignon, France. Association for Computational Linguistics.

Michael Natale. 1978. Perceived empathy, warmth and genuineness as affected by interviewer timing of speech in a telephone interview. *Psychotherapy: Theory, Research & Practice*, 15(2):145–152.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Carl R. Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4663–4677, Online. Association for Computational Linguistics.

Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Park. 2024. EmpathicStories++: A multimodal dataset for empathy towards personal experiences. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4525–4536, Bangkok, Thailand. Association for Computational Linguistics.

Maya Srikanth, Run Chen, and Julia Hirschberg. 2025. Mixed signals: Understanding model disagreement in multimodal empathy detection. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1978–1991, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Trang Tran, Yufeng Yin, Leili Tavabi, Joannalyn Delacruz, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2023. Multimodal analysis and assessment of therapist empathy in motivational interviews. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 406–415, New York, NY, USA. Association for Computing Machinery.

Chenwei Wan, Matthieu Labeau, and Chloé Clavel. 2025. EmoDynamiX: Emotional support dialogue strategy prediction by modelling MiXed emotions and discourse dynamics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1678–1695, Albuquerque, New Mexico. Association for Computational Linguistics.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoyu Wang, Guangyan Zhang, Jiale Chen, Jingyu Li, Yuehai Wang, and Yiwen Guo. 2025b. Empathy Omni: Enabling Empathetic Speech Response Generation through Large Language Models. *Preprint*, arXiv:2508.18655.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025c. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23588–23609, Vienna, Austria. Association for Computational Linguistics.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, and Chao Zhang. 2025d. Enabling auditory large language models for automatic speech quality evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Joan Welkowitz and Marta Kuc. 1973. Interrelationships among warmth, genuineness, empathy, and temporal speech patterns in interpersonal interaction. *Journal of Consulting and Clinical Psychology*, 41(3):472–473.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938, Abu Dhabi, UAE. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-MI: A dataset of expert-annotated counselling dialogues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Henry J. Xie, Jinghan Zhang, Xinhao Zhang, and Kunpeng Liu. 2024. Scoring with large language models: A study on measuring empathy of responses in dialogues. In *Proceedings of the 2024 IEEE International Conference on Big Data*, pages 7433–7437.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting

He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-Omni technical report. *Preprint*, arXiv:2509.17765.

Haoqiu Yan, Yongxin Zhu, Kai Zheng, Bing Liu, Haoyu Cao, Deqiang Jiang, and Linli Xu. 2024. Talk with human-like agents: Empathetic dialogue through perceptible acoustic reception and reaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15009–15022, Bangkok, Thailand. Association for Computational Linguistics.

Chih-Kai Yang, Neo S. Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10181, Suzhou, China. Association for Computational Linguistics.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

Tengfei Yu, Siyu Pan, Caoyun Fan, Siyang Luo, Yaohui Jin, and Binglei Zhao. 2025. Can large language models exhibit cognitive and affective empathy as humans? *Computers in Human Behavior: Artificial Humans*, 6:100233.

Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. 2025. Kardia-R1: Unleashing LLMs to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. *arXiv preprint*. ArXiv:2512.01282.

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM Web Conference 2025*, pages 2872–2881.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. 2026. Echomind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In *The Fourteenth International Conference on Learning Representations*.

Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. MEDIC: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6054–6062, Ottawa, Canada. Association for Computing Machinery.

A Related Work

Computational Empathy The constructs of empathy, warmth, and genuineness originate in client-centered therapy as core relational conditions (Rogers, 1957; Gelso and Carter, 1994). However, their computational treatment has largely reduced empathy to a textual classification task (Sharma et al., 2020; Zhu et al., 2023). Newer resources, even when multimodal, often rely on similar holistic or self-reported labels (Shen et al., 2024), or estimate empathetic similarity (Manzoor et al., 2024), perpetuating this gap in qualitative assessment.

On the other hand, research across NLP and speech has shown that LLMs can support a range of therapeutic tasks (Jiang et al., 2023b; Wan et al., 2025; Aghakhani et al., 2025; Chen et al., 2024b; Wu et al., 2025; Cai et al., 2024; Fouda et al., 2025). These advances collectively demonstrate the viability of using LLMs for therapeutic and evaluative purposes, including empathetic generation (Yu et al., 2025; Huang et al., 2024; Yuan et al., 2025; Zhou et al., 2026; Geng et al., 2025). However, these efforts primarily focus on perception or generation quality rather than explicitly evaluating whether a spoken response itself constitutes an empathetic act across interpretable relational dimensions, or only as a unimodal textual task. Speech-processing work highlights correlations between acoustic cues and perceived empathy (Chen et al., 2024a). Recent audio dialogue benchmarks further evaluate conversational quality using metrics such as naturalness, coherence, supportiveness, and delivery (Luo et al., 2026). While these metrics capture aspects of response quality and vocal expression, they do not explicitly decompose empathy into interpretable relational dimensions. Consequently, current multimodal approaches still treat speech largely as an external feature (Hasan et al., 2024).

Large Audio Language Models The field's first priority has been to establish rigorous evaluations for LALMs (Yang et al., 2025), resulting in a suite of benchmarks covering general audio intelligence and spoken dialogue understanding (Wang et al., 2025a; Gao et al., 2025; Hou et al., 2025; Ao et al., 2024; Cheng et al., 2025), knowledge and reasoning (Cui et al., 2025; Li et al., 2025), competent evaluators (Chen et al., 2025; Wang et al., 2025d,c) and Instruction following ability (Yang et al., 2024).It show that these models can judge objective speech quality metrics and articulate de-

tailed reasoning, showing LALMs are capable perceptual judges.

Parallel to these developments, a growing body of work explores multimodality for emotion and empathy modeling in speech (Hasan et al., 2025; Tran et al., 2023; Min et al., 2025). Other studies focus on generating empathetic speech responses or expressive conversational agents (Hu et al., 2025; Wang et al., 2025b; Yan et al., 2024; Zhang et al., 2025; Fei et al., 2024). However, such generation-oriented approaches often operationalize empathy narrowly, emphasizing stylistic cues such as warmth, prosodic variation, or controlled tempo (Hu et al., 2025; Hong et al., 2025). This perspective overlooks the relational nature of empathic communication, where vocal quality and textual content interact to shape how a response is interpreted.

These research threads converge on a critical gap: Despite advances in LALM evaluation and empathetic speech generation, existing benchmarks primarily evaluate perception, generation quality, or acoustic attributes, rather than whether models can judge the quality of spoken supportive responses across interpretable relational dimensions. This limits systematic progress in empathetic speech systems.

B Annotation Guideline

The following annotation guidelines were provided verbatim to annotators. They are included in full to ensure reproducibility.
Speech-Based Dimensions:
Vocal Warmth
Score 0: The tone is cold, irritated, dismissive, or emotionally mismatched with the speaker's context;
Score 1: The tone is neutral or professional; not warm but also not harmful;
Score 2: The tone is warm, soothing, compassionate, and clearly conveys emotional support.
Perceived Genuineness Score 0: The tone sounds sarcastic, mocking, performative, or emotionally contradictory;
Score 1: The tone is neutral or detached but not deceptive;
Score 2: The tone sounds authentic, grounded, and emotionally congruent.
Prosodic Attunement Score 0: The delivery is monotone, robotic, or has awkward pacing and unnatural phrasing;
Score 1: The delivery sounds natural but is minimally expressive; like reading from a textbook;
Score 2: The delivery is expressive and emotionally attuned; it has good pacing, natural pauses, and dynamic prosody.
Text-Based Dimensions:
Emotional Reaction
Score 0: No emotional reaction is expressed. The response contains only advice, information, or questions;
Score 1: An implicit or weak emotional reaction is present;
Score 2: An explicit emotional reaction is present.
Interpretation
Score 0: No understanding of the speaker's feelings or experience is shown;
Score 1: A weak or generic understanding is shown;
Score 2: A specific or reflective understanding is shown; an accurate inference or appropriate shared experience is expressed.
Exploration
Score 0: No attempt is made to explore the speaker's feelings or the situation;
Score 1: A generic exploration question is asked;
Score 2: A specific exploration is made, tied directly to the speaker's emotion or context.

Implementation Details. Evaluation is conducted on the 1007-sample gold test set using official APIs or locally deployed checkpoints. Inference settings follow the respective documentation. The model version is the latest as of 2026 May 25. Each models is set to temperature 0 for deterministic output, unless otherwise specify, every other parameters use the default value in the original study, Phi-4-multimodal is additionally used in appendix for ablation analysis (Microsoft et al., 2025)

Dialogue length varies substantially across the three source datasets (Table 7), ranging from short exchanges in AvaMERG to long-form counseling and scripted therapy sessions in AnnoMI and MESC. During evaluation, models are provided with the local dialogue context surrounding the target response, while the scoring task focuses on judging the empathy of the response itself.

C Mappings to Everyday Concepts

To help readers connect our technical dimensions to intuitive concepts in human interaction, we provide

Dataset	Mean Turns	Median	Range
AvaMERG	1.18	1	1–3
AnnoMI	79.05	50	7–598
MESC	152.9	158.5	134–170

Table 7: Dialogue length statistics across the three source datasets used in BAEmpEval.

Cmp.	VW	PG	PA	ER	EX	INT
R1–R2 (E)	.596	.610	.645	.801	.775	.755
R1–R2 (W)	.969	.987	.967	.976	.974	.918
R1–R3 (E)	.714	.771	.711	.887	.895	.826
R1–R3 (W)	.993	.985	.971	.991	.987	.947
R2–R3 (E)	.750	.669	.799	.847	.826	.834
R2–R3 (W)	.957	.975	.965	.987	.989	.945
R1–Gem (E)	.612	.680	.597	.812	.805	.741
R1–Gem (W)	.978	.973	.953	.973	.965	.893
R2–Gem (E)	.803	.814	.803	.674	.702	.656
R2–Gem (W)	.983	.990	.979	.931	.934	.912
R3–Gem (E)	.783	.801	.746	.871	.867	.855
R3–Gem (W)	.981	.971	.969	.983	.982	.941

Table 8: Exact-match (E) and within-one (W1) agreement across annotators and the Gemini baseline. VW = Vocal Warmth, PG = Perceived Genuineness, PA = Prosodic Attunement, ER = Emotional Reaction, EX = Exploration, INT = Interpretation.

the following mappings:

Everyday Concept	Related Dimensions
Sympathy / Compassion	ER + VW
Supportiveness	ER + EX
Interpersonal Warmth	VW + PA
Sincerity / Authenticity	PG
Empathic Understanding	IN + ER
Active Listening	EX + IN

Table 9: Conceptual relationships between everyday empathy-related concepts and the operational dimensions used in our dataset. Abbreviations: ER = Emotional Reaction, EX = Exploration, IN = Interpretation, VW = Vocal Warmth, PG = Perceived Genuineness, PA = Prosodic Attunement. Everyday concepts often correspond to combinations of multiple dimensions rather than a single measure.

While no single dimension fully captures the complexity of human empathy, their combination—particularly the integration of prosodic attunement, exploration, and interpretation—approximates what laypeople recognize as empathic understanding. Emotional Reaction captures the related but distinct concept of sympathy, while Vocal Warmth and Perceived Genuineness correspond to interpersonal qualities that modulate how empathy is perceived in spoken interaction.

D Annotation Quality and Rater Reliability

D.1 Pilot Study and Gold Set Expansion

The initial 379-sample pilot subset contained several disproportionately challenging interaction patterns, including emotionally volatile MESC sessions involving therapist-patient conflict, highly flattened prosody in portions of AvaMERG, and diagnostically oriented AnnoMI interactions where therapists adopted instructional rather than overtly supportive conversational styles. While these cases remain important components of empathetic evaluation, their relative concentration in the smaller pilot subset increased annotation ambiguity and inter-rater variability. Expanding the human-annotated gold set to 1007 randomly sampled instances from the full 5,436 pool introduced a broader distribution of interaction styles, including calmer therapeutic exchanges, more naturally supportive listener responses, and speech segments with more human-like prosodic variability. This broader sampling reduced the dominance of highly ambiguous edge cases and produced more stable agreement and correlation estimates while preserving the benchmark's overall diversity and subjectivity characteristics.

D.2 Gold and Silver Validity

To evaluate annotation reliability, we measure pairwise exact-match (E) and within-one (W1) agreement among three trained human annotators. For diagnostic comparison, we also report agreement between human annotators and the Gemini-2.5-Pro model used during the silver annotation stage. Results are shown in Table 8.

Exact agreement is consistently higher for conversational dimensions such as Emotional Reaction and Exploration than for speech-based dimensions including Vocal Warmth, Perceived Genuineness, and Prosodic Attunement. This difference is expected, as vocal judgments rely on subtler and more subjective prosodic cues. Nevertheless, within-one agreement remains consistently high across all dimensions and rater pairs, typically exceeding 90%, indicating that most disagreements occur between adjacent ordinal categories rather than reflecting fundamentally incompatible interpretations.

Human–LLM agreement is broadly comparable to human–human agreement across multiple dimensions, suggesting that the Gemini silver annotations fall within the natural variability of human

annotation rather than behaving as a systematic outlier. Importantly, although within-one agreement among human annotators frequently exceeds 90% and high Within-1 agreement is also observed for several LALMs, as discussed in the main paper, this metric can remain high under collapsed predictions and is therefore interpreted jointly with Exact Accuracy and class-conditional analyses.

Overall, these results suggest that the annotation task is subjective but not arbitrarily noisy: disagreements are common near category boundaries, yet strongly polarized judgments remain comparatively rare. This supports the use of ambiguity-aware evaluation for fine-grained spoken empathetic response understanding.

E Dataset Statistics and Label Distributions

This section reports descriptive statistics of the datasets used in our experiments, including label distributions and inter-dimension correlations for both the silver-annotated training set and the gold set.

Silver set. The silver-annotated dataset is used for calibrator training. Figure 7 shows the correlation structure across the six empathy dimensions. Speech-based dimensions exhibit moderate positive correlations, while text-based dimensions form a distinct cluster, suggesting partial construct separation between vocal perception and linguistic empathy strategies.

Figure 8 presents the label distribution for each dimension. Labels are defined on a three-point ordinal scale (0–2). The distributions are imbalanced and vary substantially across dimensions. This imbalance reflects the relative rarity of strong empathetic behaviors and motivates the use of a calibrated evaluation framework.

Gold set. Figure 6 illustrates the inter-dimension correlation structure of the gold subset. Compared to the silver set, correlations are generally weaker and more heterogeneous, indicating reduced coupling between dimensions. Spearman correlations show moderate associations among vocal dimensions and substantially weaker coupling between vocal and conversational dimensions, supporting related but non-redundant constructs.

F Prompt Templates for Evaluation

This section documents the prompt format used to elicit dimension-level empathy judgments from all evaluated models. Detailed definitions and annotation criteria for each empathy dimension are provided in Appendix B and are not repeated here. Example is shown in 9.

All dimensions are evaluated on a three-point ordinal scale (0–2). A holistic empathy judgment is additionally collected as an auxiliary signal for exploratory analysis annotated by Gemini-2.5-Pro; it is not independently validated by multiple human annotators and is therefore excluded from primary analysis. The prompt template shown below was selected to ensure consistent formatting and comparability across models.

G Extra Model Result Details

Model	Human	Model Prediction (%)		
		0	1	2
Phi-4	0	0.6	40.4	59.0
Phi-4	1	0.3	25.8	73.9
Phi-4	2	0.2	27.4	72.4
Qwen2.5	0	8.2	4.2	87.6
Qwen2.5	1	2.0	8.1	89.9
Qwen2.5	2	0.6	4.7	94.6
Qwen3	0	8.2	9.5	82.3
Qwen3	1	2.0	8.0	90.0
Qwen3	2	0.9	3.0	96.1
Qwen2-Audio	0	1.3	11.5	87.2
Qwen2-Audio	1	0.2	7.0	92.8
Qwen2-Audio	2	1.9	8.6	89.5
Flamingo	0	0.1	4.5	95.4
Flamingo	1	0.4	14.6	84.9
Flamingo	2	0.0	29.8	70.2
Music-Flamingo	0	0.4	36.0	63.7
Music-Flamingo	1	1.9	35.1	63.0
Music-Flamingo	2	0.6	37.0	62.4
GPT-4o	0	**42.6**	28.8	28.6
GPT-4o	1	8.3	**33.7**	58.0
GPT-4o	2	4.3	28.5	**67.2**

Table 10: Complete confusion matrices for all seven models on the pilot set. Values are row percentages (sum to 100% across each human label row). GPT-4o is the only model that utilizes the full scoring range when human annotations are low (0).Model abbreviations: GPT-4o = GPT-4o-Audio-Preview; Flamingo = Audio-Flamingo-3; Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Qwen2-Audio = Qwen-2-Audio-Instruct; Phi-4 = Phi-4-Multimodal.

G.1 Full Confusion Matrices and correlation

Table 10 presents the complete confusion matrices for all seven models, showing the distribution of model predictions conditional on human annotations. Values are percentages (row-normalized),

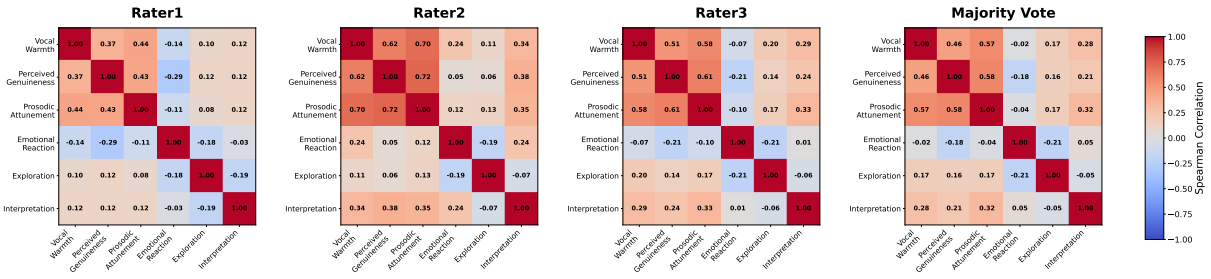


Figure 6: Inter-dimension correlation matrix for the gold subset. Compared to the silver set, correlations are weaker and more heterogeneous, suggesting greater separation between evaluated empathy dimensions.

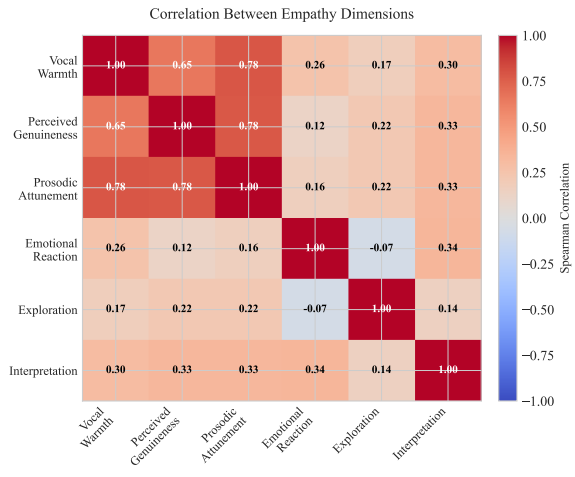


Figure 7: Inter-dimension correlation heatmap for the silver-annotated dataset. Speech-based dimensions show moderate positive correlations, while text-based dimensions exhibit a distinct correlation pattern, suggesting differences in annotation behavior across modalities.

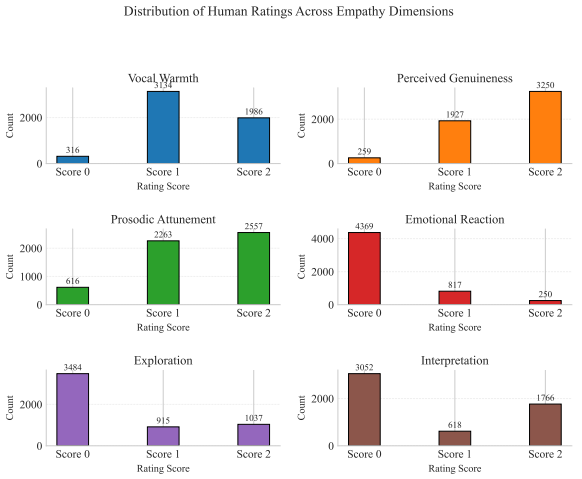


Figure 8: Label distribution of the silver-annotated dataset across the six empathy dimensions. Labels are defined on a three-point ordinal scale (0–2). Distributions are imbalanced and vary across dimensions, with higher scores occurring less frequently.

with the true label (human judgment) as rows and model predictions as columns.

The patterns reveal a clear dichotomy: all models except GPT-4o-Audio-Preview exhibit systematic bias toward over-predicting score 2, with many models assigning the highest score to over 80% of instances regardless of human judgment. This bias is most extreme in Flamingo, which predicts score 2 for 95% of instances where humans assigned 0 or 1. GPT-4o alone demonstrates reasonable calibration, correctly predicting score 2 only 28.6% of the time when humans assigned 0, and showing more balanced predictions across all categories.

Table 11 presents the internal correlation matrices for each model across the six empathy dimensions. Most models exhibit weak correlations between dimensions, with values generally close to zero. In contrast, GPT-4o-Audio-Preview shows several moderate positive correlations between related dimensions, indicating stronger internal consistency in its predictions. These results provide a detailed view of how different models distribute scores across empathy dimensions.

H Qualitative Case Study Analysis

To illustrate the challenges of multimodal empathy judgment and diagnose specific failure modes of current LALMs, we examine three contrasting examples from our gold evaluation set. These cases demonstrate systematic patterns in model behavior that quantitative metrics alone cannot reveal.

H.1 Example 1: AvaMERG

Analysis. In this example, the speaker expresses frustration through tense prosody and emphatic stress on “hate,” resulting in high gold scores across the vocal dimensions (VW=2, PG=2, PA=2). However, the textual content remains self-focused (“I hate being late”) and does not attempt to interpret or

Model	Example 1			Example 2			Example 3		
	ER	IT	EX	ER	IT	EX	ER	IT	EX
Gold	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**
Phi-4	2	1	2	1	1	2	2	2	2
Qwen2.5	2	2	2	2	2	2	2	2	2
Qwen3	2	2	1	2	2	2	2	2	2
Qwen2-Audio	2	1	2	1	2	2	2	2	2
Flamingo	2	2	2	2	2	2	2	2	2
Music-Flamingo	1	2	2	0	1	2	0	2	2
GPT-4o	2	2	1	0	1	2	2	2	1

Table 11: Conversational-dimension predictions for the three qualitative examples. Cells shaded in gray indicate predictions matching the gold annotation. ER = Emotional Reaction, IT = Interpretation, EX = Exploration.

Dataset / ID	Context	Response	Gold Scores
AvaMERG			
Conv. 1394	Speaker is frustrated about being late due to carpooling	"I hate being late"	VW=2, PG=2, PA=2 ER=1, IT=0, EX=0
Anno-MI			
Conv. 21	Therapeutic discussion about college plans	"So, kind of keeping your options open."	VW=1, PG=1, PA=2 ER=0, IT=1, EX=0
MESC			
Conv. s01e40_387	Philosophical discussion about writers controlling characters' fates	"You wanna write everybody's story for them?"	VW=1, PG=2, PA=1 ER=0, IT=1, EX=0

Table 12: Qualitative examples illustrating different empathy configurations across datasets. Example 1 shows strong vocal empathy with limited textual empathy strategies. Example 2 reflects professional therapeutic reflection with neutral emotional content. Example 3 demonstrates philosophical inquiry that may be mistaken for emotional engagement despite limited affective connection.

Prompt Template (Vocal Warmth)

You are evaluating empathetic response quality.
Speaker previous utterance: "context"
Listener response: "response"
Evaluate ONLY the listener's VOCAL WARMTH from the audio.

VOCAL WARMTH scoring rules:
0 = Cold, irritated, dismissive, or emotionally mismatched
1 = Neutral/professional; not warm but not harmful
2 = Warm, soothing, compassionate; conveys emotional support

Return ONLY Score: [0/1/2]

Figure 9: Prompt template for evaluating LALMs in the Vocal Warmth dimension.

explore the other speaker's experience, leading to low textual empathy scores (ER=1, IT=0, EX=0).

Model predictions show mixed performance on the vocal dimensions. Perceived genuineness and prosodic attunement are frequently predicted correctly, while vocal warmth is less consistent across models.

In contrast, textual dimensions are systematically overestimated. Nearly all models assign high scores to Emotional Reaction, Interpretation, and Exploration, despite the absence of these strategies in the utterance. No model correctly predicts the gold values for Interpretation or Exploration.

This example illustrates a common tendency for models to infer additional empathic engagement from a short expressive utterance, even when the text itself does not contain explicit empathic strategies.

H.2 Example 2: Anno-MI

Analysis. The therapist's response is delivered in a soft voice with a rising pitch on "kind of," producing a reflective conversational tone. The prosody therefore receives a high gold score (PA=2), while the overall warmth and genuineness remain moderate (VW=1, PG=1).

Textually, the utterance paraphrases the client's

situation ("keeping your options open"), which supports a score of Interpretation=1, but it does not explicitly express emotional reaction or encourage further elaboration (ER=0, EX=0).

Most models assign higher scores to Emotional Reaction and Exploration, suggesting that reflective paraphrases are frequently interpreted as stronger empathic engagement than the annotation guidelines specify. Music-Flamingo and GPT-4o correctly identify the absence of emotional reaction (ER=0), while most other models assign higher ER scores.

H.3 Example 3: MESC

Analysis. This response is delivered as a question with a rising pitch contour ("You wanna write everybody's story for them?"). The tone conveys engagement with the topic but relatively limited warmth, which corresponds to moderate vocal scores in the gold annotation (VW=1, PG=2, PA=1).

Textually, the utterance challenges the idea under discussion rather than responding to an emotional experience. As a result, the gold annotation assigns Interpretation=1 while Emotional Reaction and Exploration remain at 0.

Most models assign substantially higher scores to Emotional Reaction and Exploration than the gold annotation specifies. Only Music-Flamingo correctly predicts ER=0, while no model correctly predicts Exploration=0. This pattern suggests that questions engaging with an idea may be interpreted by models as stronger empathic responses than intended by the annotation criteria.

H.4 Synthesis of Observed Patterns

The three examples reveal several consistent tendencies in model predictions across different conversational settings.

- **Systematic overestimation of textual empathy.** Across all three cases, models frequently assign higher scores to textual dimensions than the gold annotations specify. In Example 1, every model over-predicts both Interpretation and Exploration for a short self-focused utterance. A similar pattern appears in Example 3, where Exploration is overestimated for a question that engages with an idea rather than the speaker's emotional experience.
- **Robust recognition of prosodic engagement.** Prosodic attunement (PA) is often predicted correctly across models, particularly in Examples 1 and 2 where the vocal delivery is clearly expressive or reflective. However, other vocal dimensions such as warmth show greater variability, suggesting that models capture general prosodic engagement more reliably than nuanced vocal affect.
- **Challenges separating conversational engagement from empathy.** Responses that actively engage with the topic—such as reflective paraphrases in Example 2 or rhetorical questions in Example 3—are frequently interpreted as stronger empathic responses than intended by the annotation guidelines. This leads to inflated scores on Emotional Reaction or Exploration even when these strategies are absent.

Together, these cases illustrate how conversational style and vocal delivery can influence model judgments across multiple empathy dimensions. They also highlight the importance of evaluating vocal and textual components separately, as the two modalities may convey different aspects of empathic behavior.

I Reasoning-Aligned Evaluation Framework (Exploratory)

Large Audio Language Models (LALMs) can be sensitive to prompt formatting and instruction-following variations, sometimes producing outputs that deviate from the required response format (Yang et al., 2025). In subjective tasks such as empathy evaluation, these inconsistencies make it difficult to directly recover a model's intended ordinal judgment from its textual response.

Exploratory Prompt and Context Sensitivity Analysis Here we report the effect of different prompt designs effects on LALMs. We only chooses models that can handle long instruction and can produce formatted response. Models include Qwen-2.5-Omni, Qwen-3-Omni-Flash and Gemini-2.5-Pro. GPT-4o-Audio-Preview is excluded due to occasional error of "I cant listen to audio so i can not provide help". Under this settings we instruct models to score all dimensions in a single output instead of separately to save computational cost. Since this ablation contains multiturn and few shot and long speaker audio

as extra context, long instruction following ability is a must, which motivate our decision of using combine prompt for this ablation. From table 13, 14, 15 we observe no strong differences in results across different condition, this is expected because our task focus on listener response evaluation that rely little on context understanding.

Instruction-Following Limitations. We also report results across 3 prompt condition(Output every dimension at once, output each dimension separately, reason first before output score). All evaluated models occasionally fail to produce the required score+rationale format, generating only scores, only rationales, or malformed responses (Table 18), while output every dimension at once works for the selected models, they still has occasional failure and output null value, overall only output each dimension separately can ensure high validity output, we also observe that these two condition has small difference in most dimensions but has significant variation in emotional reaction for Qwen-series 16. Phi-4-Multimodal shows the highest recoverability rate (90.2%), while other models exhibit more frequent formatting inconsistencies (e.g., Qwen-2.5-Omni: 16.1%). Prior work addresses such issues by using multiple prompt variants (Wang et al., 2025a); instead, we apply calibration to normalize outputs.

This calibration step substantially increases the number of usable samples by recovering predictions that would otherwise be discarded. Importantly, as shown in Table 19 and Table 20, calibration does not materially change performance trends across models or empathy dimensions. Rather, it acts as a normalization step that preserves model behavior while improving evaluation coverage.

To study this issue, we explore a reasoning-aligned evaluation framework that attempts to recover structured scores from model outputs through a two-stage process (Figure 10). This framework is presented as an exploratory analysis and is not used for the primary evaluation reported in the main paper, where results are computed directly from normalized model outputs. Instead, it examines how model reasoning could be aligned with structured scoring in future multimodal evaluation.

The framework consists of two sequential stages: output normalization and calibration.

First, we perform output normalization and recoverability analysis to extract any directly recoverable numeric signal independent of accompanying rationales. Let $s_i^{(k)}$ denote the raw output for empathy dimension k on instance i. We define a deterministic normalization function $\mathcal{N}_{\text{raw}} : \mathcal{S} \to \{0, 1, 2, \text{NaN}\}$ that operates as follows. Here, NaN denotes an unrecoverable output, corresponding to cases where no unique ordinal score can be deterministically extracted from the model response. For discrete inputs where $s_i^{(k)} \in \{0, 1, 2\}$, we preserve the value: $\mathcal{N}_{\text{raw}}(s_i^{(k)}) = s_i^{(k)}$. For continuous numeric inputs, we discretize using fixed thresholds:

$$\mathcal{N}_{\text{raw}}(s_i^{(k)}) = \mathbf{1}(s_i^{(k)} \geq 0) + \mathbf{1}(s_i^{(k)} \geq 1.5),$$

where $\mathbf{1}(\cdot)$ is the indicator function, and values greater than 2 are clipped to 2. For textual inputs, we search for numeric tokens in $\{0, 1, 2\}$: if exactly one such token n is found, we set $\mathcal{N}_{\text{raw}}(s_i^{(k)}) = n$; otherwise, $\mathcal{N}_{\text{raw}}(s_i^{(k)}) = \text{NaN}$.

The proportion of non-NaN outputs after this step defines the recoverability rate, which measures basic instruction-following success. Importantly, this normalized score is used only for recoverability analysis, not for final performance evaluation.

Calibration To enable consistent comparison across models, we fine-tune Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a) as a calibration model that converts model reasoning into structured empathy scores. Similar to approaches that train LLMs as judges with reasoning alignment (Chiang et al., 2025), we adopt a training-based method rather than relying on API-based evaluators (Yang et al., 2024). This allows learning domain-specific mappings between reasoning patterns and ordinal scores while ensuring full evaluation reproducibility.

We define a calibration function:

$$\mathcal{C}_\theta(s_i^{(k)}) = f_\theta(t_i^{(k)}) \in \{0, 1, 2\},$$

where $t_i^{(k)}$ denotes the textual content of $s_i^{(k)}$, and f_θ is a fine-tuned Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a) that maps from a model's full textual output (including any scores or rationales) to a coherent ordinal score. We train f_θ on 5,436 silver-standard samples (annotated by Gemini-2.5-Pro), achieving more than 95% accuracy on a held-out validation set, confirming its reliability in recovering intended judgments. Training uses default model hyperparameters.

Model	Speech-Based Dimensions						Text-Based Dimensions					
	Vocal Warmth		Genuineness		Prosody		Emotional		Interpretation		Exploration	
	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1
Qwen2.5(Audio)	0.227	0.748	0.257	0.515	0.227	0.727	0.612	0.952	0.467	0.817	0.511	0.825
Qwen2.5(Audio+Text)	0.476	0.945	0.732	0.940	0.497	0.937	0.092	0.319	0.358	0.573	0.316	0.916
Qwen3(Audio)	0.071	0.534	0.068	0.248	0.090	0.524	0.812	0.947	0.547	0.663	0.566	0.789
Qwen3(Audio+Text)	0.544	0.973	0.643	0.944	0.522	0.963	0.128	0.592	0.367	0.733	0.524	0.901
Gemini(Audio)	0.501	0.959	0.681	0.917	0.501	0.944	0.370	0.847	0.358	0.832	0.560	0.869
Gemini(Audio+Text)	0.584	0.978	0.732	0.959	0.560	0.961	0.423	0.854	0.465	0.803	0.683	0.890

Table 13: Dimension-wise accuracy across six empathy dimensions and different listener ablation. Dimensions are grouped into speech-level (Vocal Warmth, Perceived Genuineness, Prosodic Attunement) and text-level dimensions (Exploration, Interpretation, Emotional Reaction). Abbreviations: Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Gemini=Gemini-2.5-Pro; Genuineness=Perceived Genuineness; Prosody=Prosodic Attunement

Model	Speech-Based Dimensions						Text-Based Dimensions					
	Vocal Warmth		Genuineness		Prosody		Emotional		Interpretation		Exploration	
	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1
Qwen3(None)	0.532	0.976	0.653	0.942	0.532	0.963	0.180	0.750	0.366	0.755	0.520	0.888
Qwen3(Text)	0.544	0.973	0.643	0.944	0.522	0.963	0.128	0.592	0.367	0.733	0.524	0.901
Qwen3(Audio)	0.511	0.971	0.649	0.929	0.518	0.960	0.159	0.732	0.402	0.759	0.500	0.880
Qwen3(Audio+Text)	0.582	0.990	0.702	0.949	0.543	0.961	0.157	0.748	0.416	0.819	0.543	0.883
Qwen2.5(None)	0.490	0.936	0.745	0.936	0.493	0.921	0.069	0.260	0.347	0.517	0.323	0.907
Qwen2.5(Text)	0.477	0.939	0.647	0.909	0.466	0.918	0.121	0.455	0.320	0.656	0.438	0.934
Qwen2.5(Audio)	0.521	0.972	0.741	0.932	0.518	0.962	0.096	0.500	0.357	0.705	0.416	0.900
Qwen2.5(Audio+Text)	0.579	0.979	0.699	0.944	0.581	0.938	0.105	0.545	0.397	0.786	0.479	0.917
Gemini(None)	0.615	0.902	0.767	0.954	0.586	0.964	0.372	0.839	0.414	0.752	0.689	0.925
Gemini(Text)	0.422	0.878	0.222	0.767	0.289	0.811	0.422	0.844	0.333	0.856	0.889	1
Gemini(Audio)	0.623	0.981	0.754	0.961	0.563	0.978	0.345	0.842	0.429	0.767	0.653	0.903
Gemini(Audio+Text)	0.614	0.979	0.757	0.963	0.578	0.968	0.412	0.876	0.468	0.782	0.648	0.884

Table 14: Dimension-wise accuracy across six empathy dimensions and different speaker content. Dimensions are grouped into speech-level (Vocal Warmth, Perceived Genuineness, Prosodic Attunement) and text-level dimensions (Exploration, Interpretation, Emotional Reaction). Abbreviations: Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Gemini=Gemini-2.5-Pro; Genuineness=Perceived Genuineness; Prosody=Prosodic Attunement

Models Reasoning qualitative analysis. The qualitative examples(17) suggest that stronger LALMs produce more dimension-sensitive and contextually grounded rationales compared to weaker models, which frequently rely on surface-level acoustic descriptions. However, several strong models still exhibited upward calibration bias, assigning highly positive empathy scores even when their own explanations described only moderate warmth or expressiveness. These observations are consistent with our quantitative findings on partial relational sensitivity alongside persistent score inflation.

Model	Speech-Based Dimensions						Text-Based Dimensions					
	Vocal Warmth		Genuineness		Prosody		Emotional		Interpretation		Exploration	
	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1
Qwen2.5(Singleturn)	0.476	0.945	0.732	0.940	0.497	0.937	0.092	0.319	0.358	0.573	0.316	0.916
Qwen2.5(Multiturn)	0.493	0.942	0.706	0.932	0.503	0.926	0.090	0.346	0.352	0.567	0.288	0.745
Qwen3(Singleturn)	0.544	0.973	0.643	0.944	0.522	0.963	0.128	0.592	0.367	0.733	0.524	0.901
Qwen3(Multiturn)	0.581	0.973	0.674	0.939	0.552	0.966	0.164	0.662	0.397	0.711	0.506	0.911
Gemini(Singleturn)	0.584	0.978	0.732	0.959	0.560	0.961	0.423	0.854	0.465	0.803	0.683	0.890
Gemini(Multiturn)	0.567	0.973	0.762	0.954	0.526	0.951	0.327	0.831	0.428	0.717	0.675	0.893

Table 15: Dimension-wise accuracy across six empathy dimensions and different conversation turn ablation. Dimensions are grouped into speech-level (Vocal Warmth, Perceived Genuineness, Prosodic Attunement) and text-level dimensions (Exploration, Interpretation, Emotional Reaction). Abbreviations: Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Gemini=Gemini-2.5-Pro; Genuineness=Perceived Genuineness; Prosody=Prosodic Attunement

Model	Speech-Based Dimensions						Text-Based Dimensions					
	Vocal Warmth		Genuineness		Prosody		Emotional		Interpretation		Exploration	
	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1	Exact	±1
Qwen2.5(Combined)	0.476	0.945	0.732	0.940	0.497	0.937	0.092	0.319	0.358	0.573	0.316	0.916
Qwen2.5(Separate)	0.456	0.920	0.505	0.809	0.473	0.915	0.349	0.816	0.447	0.779	0.553	0.930
Qwen3(Combined)	0.544	0.973	0.643	0.944	0.522	0.963	0.128	0.592	0.367	0.733	0.524	0.901
Qwen3(Separated)	0.414	0.912	0.635	0.902	0.436	0.919	0.343	0.752	0.329	0.750	0.338	0.813

Table 16: Dimension-wise accuracy across six empathy dimensions across two prompting methods. Dimensions are grouped into speech-level (Vocal Warmth, Perceived Genuineness, Prosodic Attunement) and text-level dimensions (Exploration, Interpretation, Emotional Reaction). Abbreviations: Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Gemini=Gemini-2.5-Pro; Genuineness=Perceived Genuineness; Prosody=Prosodic Attunement

Model	Dimension	Generated Rationale
Audio Flamingo	Prosodic Attunement	"The speaker's prosody is very monotone, robotic, and awkward, with a very slow pace and unnatural phrasing. The delivery lacks any expressive inflection, making it sound like a textbook reading rather than a natural conversation."
Qwen3-Omni	Vocal Warmth	"The listener's tone is light and amused, with a soft, conversational quality that suggests genuine engagement. Their vocal warmth is warm and soothing, conveying empathy and shared amusement without being overly effusive."
Qwen-Audio	Perceived Genuineness	"The speaker's voice is perceived as genuine. The delivery seems spontaneous and earnest, reflecting a heartfelt and sincere expression of fear and surprise. There's no evidence of sarcasm, mocking, or performative speech."
Qwen2.5-Omni	Prosodic Attunement	"The speaker's prosody is characterized by a natural yet slightly hesitant pace, with moderate vocal expressiveness that conveys reassurance mixed with concern." Despite this restrained characterization, the model assigned the maximum score of 2.

Table 17: Representative free-form rationales generated by different LALMs for speech-based empathy dimensions. Weaker models often focused on surface acoustic properties (e.g., monotonicity or robotic delivery), while stronger models produced more relational and context-sensitive explanations. However, stronger models also occasionally exhibited calibration inflation by assigning maximal empathy scores despite relatively moderate qualitative descriptions.

	Speech Dimensions			Text Dimensions			Sample
Model	**Warmth**	**Genuine.**	**Attune.**	**Exploration**	**Interpret.**	**Emotion**	**Min**
Qwen-3-Omni-Flash	171	296	318	338	366	326	171
Qwen-2.5-Omni	305	347	315	372	375	61	61
Qwen-2-Audio-Instruct	155	331	345	359	313	367	155
Audio-Flamingo-3	355	358	121	354	360	363	121
Music-Flamingo	378	378	378	378	378	378	378
Phi-4-Multimodal	342	353	339	376	379	379	342
GPT-4o-Audio-Preview	298	301	302	304	304	304	298

Table 18: Number of recoverable evaluation samples per model and empathy dimension on a 379 gold subset. Counts reflect successful extraction of dimension-level scores under the standardized prompt format. The *Min* column reports the minimum number of recoverable samples across dimensions for each model.

Dimension	Phi-4	Qwen2.5	Qwen3	Qwen2-Audio	Flamingo
Vocal Warmth	0.515 → 0.515	0.279 → 0.338	0.281 → 0.266	0.234 → 0.237	0.225 → 0.243
Perceived Genuineness	0.480 → 0.398	0.569 → 0.575	0.490 → 0.567	0.548 → 0.551	0.402 → 0.264
Prosodic Attunement	0.156 → 0.158	0.197 → 0.211	0.174 → 0.158	0.154 → 0.174	0.587 → 0.338
Emotional Reaction	0.187 → 0.288	0.137 → 0.158	0.147 → 0.187	0.161 → 0.145	0.140 → 0.164
Exploration	0.053 → 0.058	0.191 → 0.190	0.228 → 0.222	0.043 → 0.042	0.037 → 0.079
Interpretation	0.153 → 0.164	0.115 → 0.127	0.123 → 0.134	0.148 → 0.137	0.119 → 0.172

Table 19: Exact accuracy before and after calibration (shown as before → after). Calibration slightly adjusts model predictions but does not substantially alter the relative performance trends across dimensions. In particular, perceived genuineness and vocal warmth remain the strongest dimensions, while exploration and interpretation remain consistently difficult for all models. Model abbreviations: GPT-4o = GPT-4o-Audio-Preview; Flamingo = Audio-Flamingo-3; Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Qwen2-Audio = Qwen-2-Audio-Instruct; Phi-4 = Phi-4-Multimodal. Genuineness=Perceived Genuineness

Dimension	Phi-4	Qwen2.5	Qwen3	Qwen2-Audio	Flamingo
Vocal Warmth	0.918 → 0.910	0.836 → 0.860	0.784 → 0.850	0.827 → 0.826	0.825 → 0.826
Perceived Genuineness	0.930 → 0.844	0.957 → 0.955	0.933 → 0.950	0.939 → 0.950	0.966 → 0.649
Prosodic Attunement	0.876 → 0.865	0.886 → 0.897	0.854 → 0.868	0.868 → 0.873	0.992 → 0.882
Emotional Reaction	0.707 → 0.710	0.403 → 0.420	0.497 → 0.525	0.435 → 0.451	0.377 → 0.397
Exploration	0.521 → 0.522	0.644 → 0.646	0.645 → 0.644	0.508 → 0.512	0.511 → 0.541
Interpretation	0.728 → 0.728	0.279 → 0.298	0.303 → 0.340	0.499 → 0.491	0.294 → 0.324

Table 20: One-off accuracy before and after calibration. Similar to exact accuracy, calibration does not substantially change the overall trend: models remain consistently close to the human ratings within one scale point, while the relative difficulty ordering of the six empathy dimensions remains largely unchanged. Model abbreviations: GPT-4o = GPT-4o-Audio-Preview; Flamingo = Audio-Flamingo-3; Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Qwen2-Audio = Qwen-2-Audio-Instruct; Phi-4 = Phi-4-Multimodal. Genuineness=Perceived Genuineness

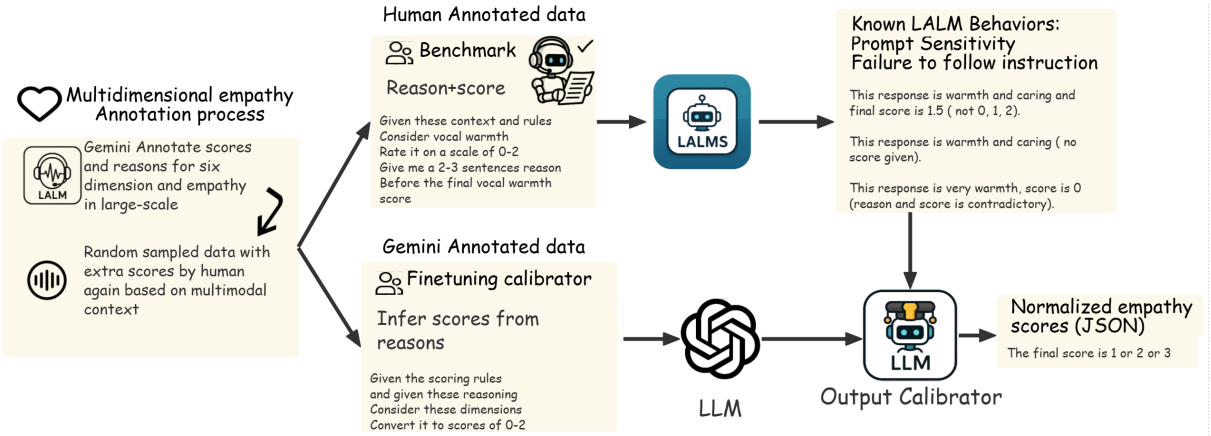


Figure 10: Overview of the multidimensional empathy annotation and reasoning-aligned evaluation pipeline. Large-scale silver annotations are generated by Gemini-2.5-Pro with scores and rationales, while a human-annotated subset forms a gold dataset. A calibrator model is fine-tuned to infer empathy scores from rationales, enabling normalization and correction of LALM outputs that exhibit prompt sensitivity or instruction-following errors.

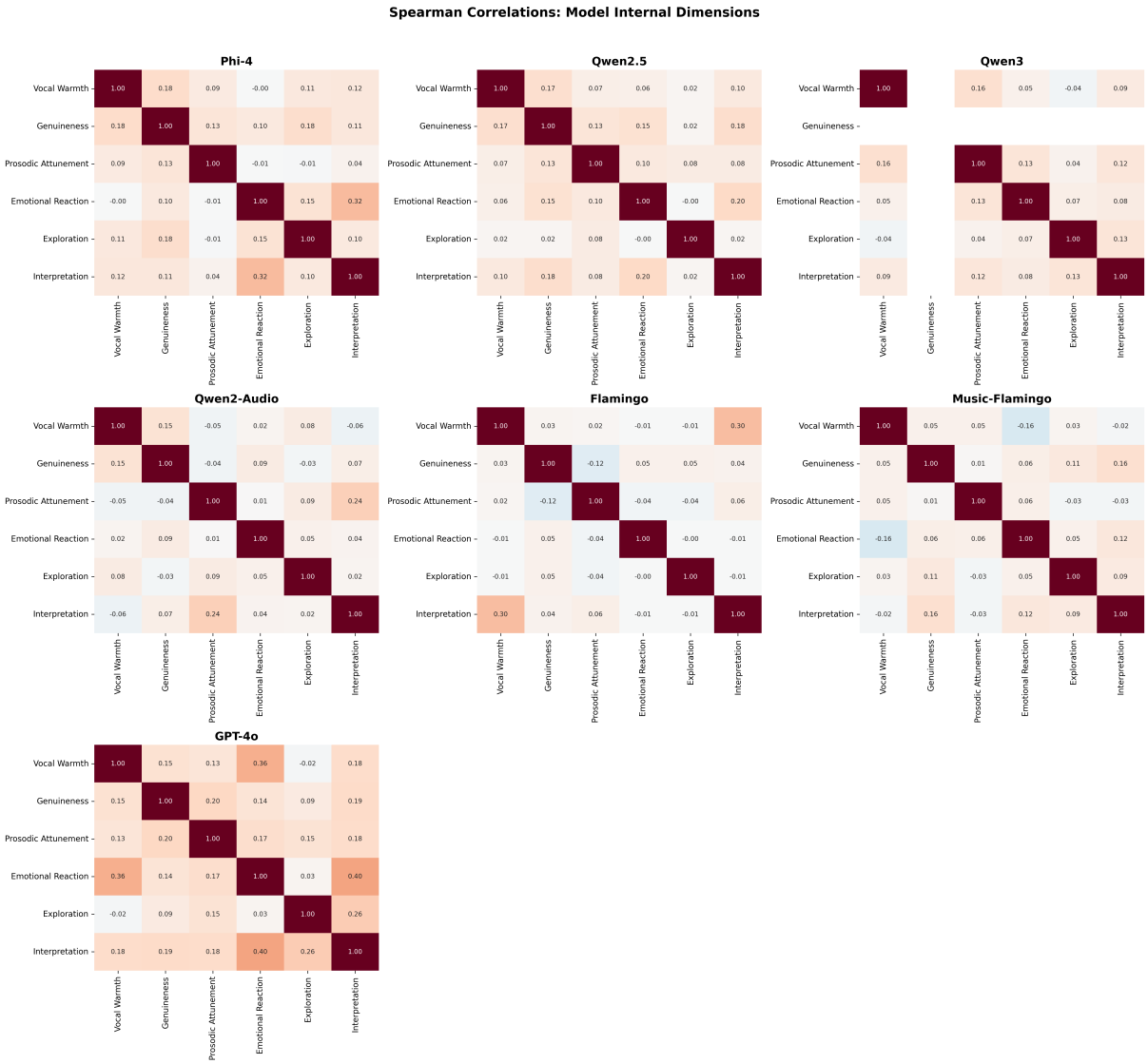


Figure 11: Internal Model Correlation. Model abbreviations: GPT-4o = GPT-4o-Audio-Preview; Flamingo = Audio-Flamingo-3; Qwen3 = Qwen-3-Omni-Flash; Qwen2.5 = Qwen-2.5-Omni; Qwen2-Audio = Qwen-2-Audio-Instruct; Phi-4 = Phi-4-Multimodal. Genuineness=Perceived Genuineness.

Model	Dimension	Exact Accuracy	Within-1 Accuracy
AudioFlamingo	Emotional Reaction	$72.8_{[70.0,75.7]}$	$91.6_{[89.7,93.4]}$
	Exploration	$52.1_{[49.0,55.2]}$	$81.3_{[78.9,83.7]}$
	Interpretation	$53.3_{[50.0,56.4]}$	$67.2_{[64.5,70.2]}$
	Perceived Genuineness	$56.9_{[53.8,60.1]}$	$86.3_{[84.1,88.7]}$
	Prosodic Attunement	$41.1_{[37.7,44.2]}$	$87.6_{[85.4,89.8]}$
	Vocal Warmth	$36.3_{[33.2,39.2]}$	$88.1_{[86.0,90.1]}$
GPT4o	Emotional Reaction	$57.1_{[53.7,60.1]}$	$93.5_{[91.9,94.9]}$
	Exploration	$38.7_{[35.6,41.8]}$	$80.0_{[77.4,82.5]}$
	Interpretation	$39.2_{[36.2,42.1]}$	$72.1_{[69.3,74.8]}$
	Perceived Genuineness	$51.3_{[48.3,54.4]}$	$95.0_{[93.6,96.2]}$
	Prosodic Attunement	$38.9_{[35.9,42.1]}$	$91.0_{[89.2,92.8]}$
	Vocal Warmth	$46.3_{[43.2,49.4]}$	$93.5_{[92.0,95.0]}$
MusicFlamingo	Emotional Reaction	$73.7_{[70.8,76.5]}$	$95.8_{[94.3,97.1]}$
	Exploration	$69.9_{[66.9,72.9]}$	$96.0_{[94.7,97.2]}$
	Interpretation	$70.8_{[67.9,73.8]}$	$92.2_{[90.5,93.9]}$
	Perceived Genuineness	$27.4_{[24.6,30.3]}$	$81.9_{[79.4,84.2]}$
	Prosodic Attunement	$82.1_{[79.7,84.6]}$	$99.0_{[98.3,99.5]}$
	Vocal Warmth	$24.8_{[22.1,27.5]}$	$77.0_{[74.4,79.7]}$
Qwen2.5	Emotional Reaction	$33.6_{[30.5,36.6]}$	$81.4_{[79.0,83.8]}$
	Exploration	$49.8_{[46.7,53.0]}$	$86.7_{[84.5,88.7]}$
	Interpretation	$44.6_{[41.3,47.7]}$	$77.5_{[74.8,80.1]}$
	Perceived Genuineness	$53.0_{[49.9,56.0]}$	$95.2_{[93.8,96.5]}$
	Prosodic Attunement	$30.3_{[27.3,33.2]}$	$84.2_{[81.9,86.4]}$
	Vocal Warmth	$62.8_{[59.8,65.9]}$	$95.5_{[94.1,96.8]}$
Qwen2-Audio	Emotional Reaction	$22.1_{[19.4,24.7]}$	$98.9_{[98.2,99.5]}$
	Exploration	$20.2_{[17.7,22.8]}$	$98.4_{[97.5,99.1]}$
	Interpretation	$17.2_{[14.8,19.6]}$	$98.3_{[97.4,99.0]}$
	Perceived Genuineness	$0.9_{[0.3,1.5]}$	$99.9_{[99.7,100.0]}$
	Prosodic Attunement	$1.1_{[0.4,1.9]}$	$99.0_{[98.4,99.6]}$
	Vocal Warmth	$0.3_{[0.0,0.7]}$	$99.8_{[99.5,100.0]}$
Qwen3.0	Emotional Reaction	$50.5_{[47.3,53.7]}$	$86.0_{[83.8,88.2]}$
	Exploration	$49.7_{[46.6,52.8]}$	$81.6_{[79.2,84.1]}$
	Interpretation	$42.3_{[39.1,45.4]}$	$68.6_{[65.7,71.5]}$
	Perceived Genuineness	$47.0_{[43.8,50.1]}$	$89.7_{[87.7,91.5]}$
	Prosodic Attunement	$39.4_{[36.4,42.5]}$	$88.7_{[86.6,90.7]}$
	Vocal Warmth	$53.8_{[50.6,56.9]}$	$89.9_{[87.9,91.7]}$
Gemini-2.5-Pro	Emotional Reaction	$87.3_{[85.4,89.2]}$	$98.4_{[97.6,99.2]}$
	Exploration	$87.5_{[85.4,89.5]}$	$98.2_{[97.4,99.0]}$
	Interpretation	$83.7_{[81.4,85.9]}$	$94.7_{[93.4,96.0]}$
	Perceived Genuineness	$80.0_{[77.9,82.5]}$	$98.5_{[97.8,99.2]}$
	Prosodic Attunement	$75.5_{[73.0,78.1]}$	$97.1_{[96.0,98.1]}$
	Vocal Warmth	$79.3_{[76.9,81.8]}$	$98.4_{[97.6,99.1]}$

Table 21: Exact Accuracy and Within-1 Accuracy on the full evaluation set using majority-vote labels. Values in brackets denote 95% bootstrap confidence intervals estimated from 1000 resamples.

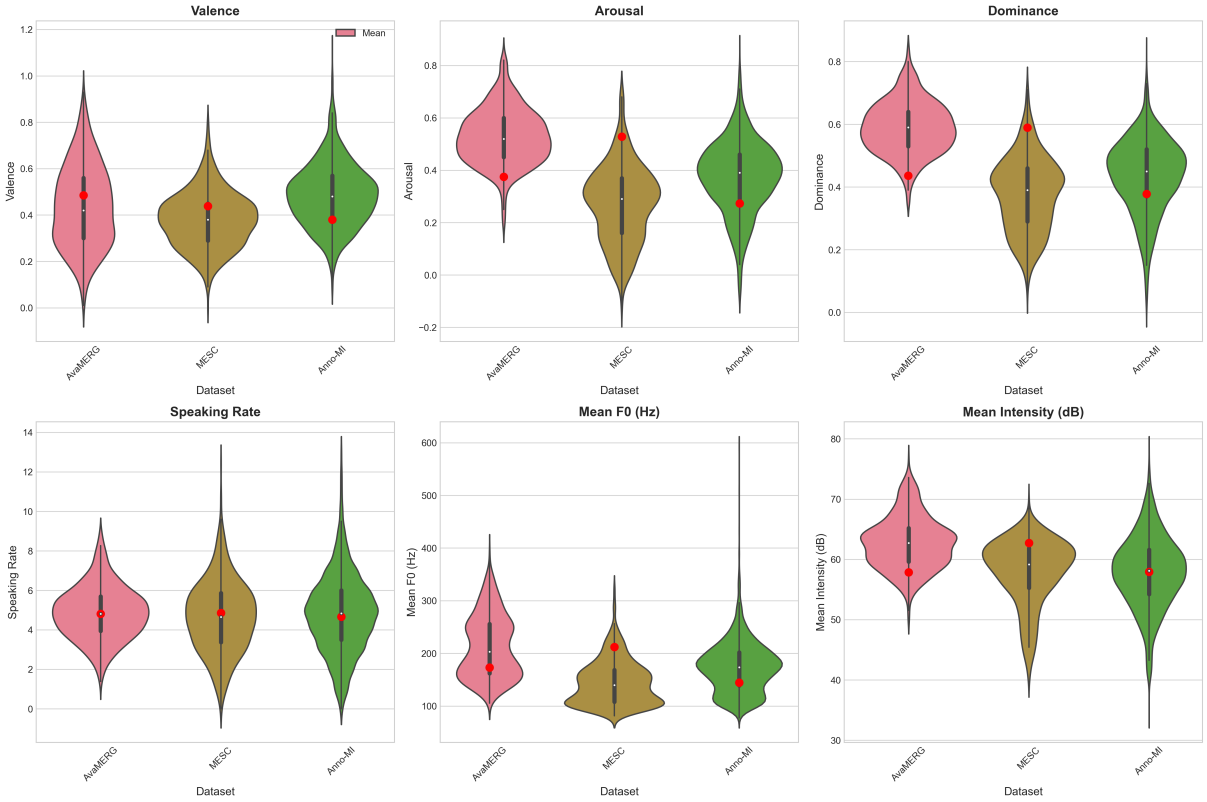


Figure 12: Distribution of acoustic features across human-annotated vocal empathy scores (0–2) on the a portion of the gold evaluation set ($n = 379$). Violin plots illustrate how prosodic properties such as pitch variability, intensity, and speaking rate vary with perceived Vocal Warmth, Perceived Genuineness, and Prosodic Attunement. Observable distributional shifts across score levels indicate that the annotated vocal empathy dimensions correspond to measurable acoustic patterns rather than arbitrary labels.