AEQ-Bench: Measuring Empathy of Omni-Modal Large Models

Xuan Luo[1,2], Lewei Yao[3], Lanqing Hong[3], Kai Chen[4], Dehua Tao[3], Daxin Tan[3], Yukun Deng[2], Ruifeng Xu[2,5], Jing Li[1]

[1]The Hong Kong Polytechnic University, Hong Kong
[2]The Harbin Institute of Technology, Shenzhen
[3]Huawei, Hong Kong
[4]Hong Kong University of Science and Technology, Hong Kong
[5]Shenzhen Loop Area Institute, Shenzhen

Correspondence: jing-amelia.li@polyu.edu.hk

Abstract

While the automatic evaluation of omni-modal large models (OLMs) is essential, assessing empathy remains a significant challenge due to its inherent affectivity. To investigate this challenge, we introduce **AEQ-BENCH** (Audio Empathy Quotient Benchmark), a novel benchmark to systematically assess two core empathetic capabilities of OLMs: (i) **generating empathetic responses** by comprehending affective cues from multi-modal inputs (audio + text), and (ii) **judging the empathy of audio responses** without relying on text transcription. Compared to existing benchmarks, AEQ-BENCH incorporates two novel settings that vary in context specificity and speech tone. Comprehensive assessment across linguistic and paralinguistic metrics reveals that (1) OLMs trained with audio output capabilities generally outperformed models with text-only outputs, and (2) while OLMs align with human judgments for coarse-grained quality assessment, they remain unreliable for evaluating fine-grained paralinguistic expressiveness.

🤗 AEQ-Bench

1 Introduction

Empathy is the ability to understand, share, and respond to the feelings and experiences of another person by taking their perspective (Szalita, 1976)—essentially, to "*put oneself in another's shoes*". It is crucial for NLP models to gain such capabilities for providing positive user experiences during interactions with humans. This need is becoming more pressing yet challenging as omni-modal large models (OLMs) become the popular backbone, which integrate modalities such as audio, vision, and text to enable more human-like interactions (Wu et al., 2023; Yin et al., 2024).

While OLMs grow in complexity, the automatic evaluation of empathy becomes crucial for training paradigms like reinforcement learning. However, this task is non-trivial, as empathy is inherently affective. Consequently, establishing a benchmark to quantify nuanced emotional resonance across audio and textual modalities remains a significant open challenge. Most existing benchmarks focus on cognitive abilities, such as knowledge retrieval, complex reasoning, and instruction following (Yue et al., 2024; Zhang et al., 2025c) (see Fig. 1), largely overlooking empathy evaluation. As OLMs become increasingly human-like, we aim to benchmark their capacity to *generate empathetic responses* and *judge empathy* in diverse contexts.

Human communication conveys empathy through not only the linguistic content of *what* is said but also the paralinguistic tone of *how* it is said. For instance, the utterance "*It was a curious coincidence*" implies genuine surprise or delight when delivered with warmth. However, a cool, sarcastic delivery of the same words suggests scepticism or criticism. To be truly effective, an OLM should understand and generate both the right words and the appropriate acoustic cues.

Previous NLP research has prioritised *linguistic* features (e.g., lexical choice and explicit emotional statements) for empathy, while *paralinguistic* cues (e.g., prosody and acoustic delivery) are equally critical for audio-capable OLMs (Aziz-Zadeh et al., 2010). However, a significant research gap persists in their joint examination. Recent studies reduced audio to mere transcriptions (Wang et al., 2025) in order to rely on established text-based metrics (Zhang et al., 2025a), ignoring the nuances of vocal delivery. This limitation presents a fundamental question: *Can OLMs automatically deliver and judge empathy by accounting for both linguistic and paralinguistic cues like human*?

To answer this question, we introduce AEQ-BENCH, a benchmark designed to assess the **empathic responsiveness and judgment** capabilities of OLMs. It comprises 1,885 English instances, each pairing an audio utterance with a concise tex-

Findings of the Association for Computational Linguistics: ACL 2026, pages 36390–36407
July 2-7, 2026

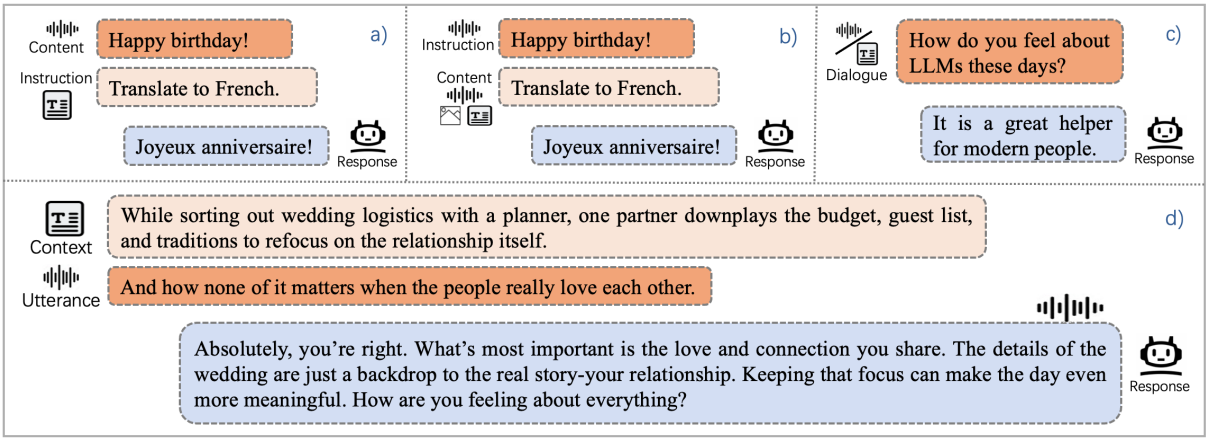

Figure 1: Overview of omni-modal tasks: instruction following (a & b) and conversation (c & d). The input configurations are: (a) audio content with textual instruction; (b) audio instruction with text, audio, or image content; (c) unimodal dialogue; and (d) mixed-modality dialogue (demonstrated by AEQ-BENCH on the MELD subset). In the mixed-modality setting, the text provides context (e.g., chat history summary) for the current audio utterance, requiring the model to respond accordingly. Outputs across all tasks may be text, audio, or both.

tual context. It overcomes the limitations of existing empathetic dialogue evaluations, which lack multimodality (see Fig. 1) and parallel comparisons (over varying contexts and acoustic tones). There exhibits two novel design strategies: (**i**) *Contextual Variation*, where distinct background contexts are applied to the same utterance to alter its pragmatic meaning (Fig. 2); and (**ii**) *Tonal Variation*, where identical contexts are paired with the same utterance delivered in varying emotional tones to shift user intent (Fig. 3). To *the best of our knowledge,* AEQ-BENCH *is the first benchmark to jointly examine linguistic and paralinguistic empathy via parallel contexts and acoustic tonal variations*.

We then conducted an extensive evaluation on AEQ-BENCH, benchmarking current state-of-the-art OLMs on their ability to *generate empathetic responses* and *judge generations from other models* compared to human evaluators. This assessment covered seven dimensions designed to comprehensively measure both linguistic and paralinguistic features: *modality reliance*, *naturalness*, *coherence*, *supportiveness*, *discrimination*, and *delivery*.

Experimental results demonstrate that OLMs trained with audio output capabilities generally outperform those limited to text-only output. GPT produced the most human-like and empathetic narratives, followed closely by Qwen-Omni. These findings underscore the value of integrating text and audio for learning empathy. However, while OLMs align with human judgment on coarse-grained tasks, they remain unreliable for fine-grained emotional evaluation. Specifically, the models lack paralinguistic empathy, often delivering flat speech, and OLM-based judges diverge significantly from human perception regarding emotional prosody. Consequently, generating and evaluating expressive empathetic tone remains an open challenge.

2 Related Work

Empathy and Emotional Intelligence. Empathy is widely defined as the ability to perceive, understand, and respond to another individual's emotional states. It is fundamental for satisfactory and effective conversational communication (Szalita, 1976). Empathy in dialogue has been studied in text through datasets and frameworks that annotate empathic responses and strategies. EMPATHETICDIALOGUES introduced 25k conversations grounded in emotional situations for empathic response generation (Rashkin et al., 2019). Work in mental-health support provides theory-grounded taxonomies and labelled corpora to detect and analyse empathy mechanisms in text (Sharma et al., 2020). With the development of LLMs, automatic evaluation of response empathy is available (Zhang et al., 2025a). For generated audio responses, it remains confined to the text modality empathy evaluation, converting the audio content into text (Wang et al., 2025). In contrast, we evaluate the potential of OLMs as automatic judges directly on the audio response, in order to study empathy from both linguistic and paralinguistic perspectives.

Paralinguistic Research and Audio Cues. Human language is generally split into the verbal chan-

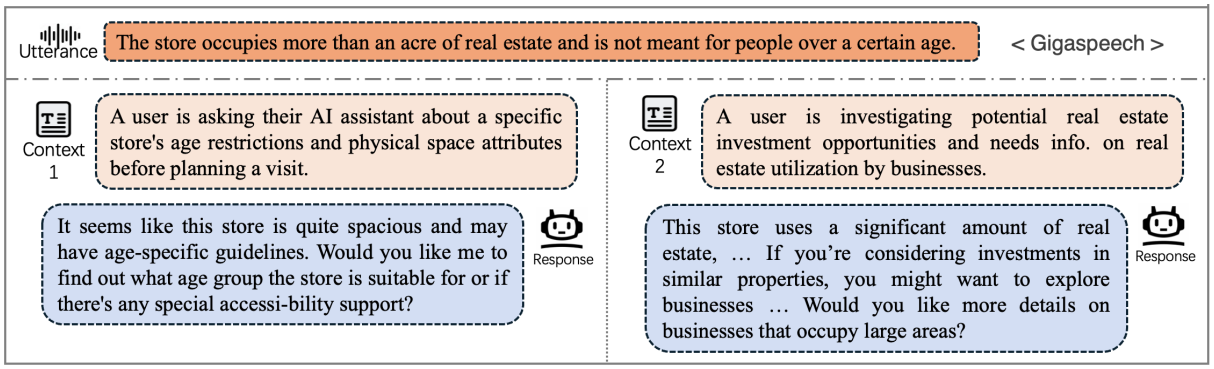

Figure 2: The constructed GIGASPEECH subset of AEQ-BENCH featuring *context variation*. For each utterance, we construct two plausible contexts, each associated with a corresponding reference response. (Appx. D)

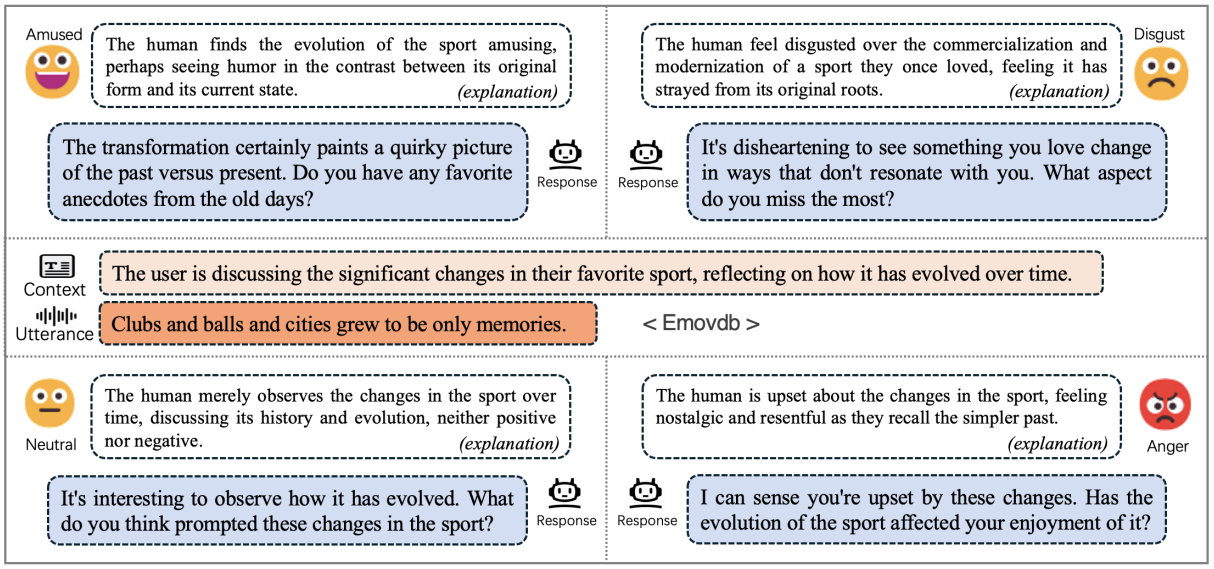

Figure 3: The constructed Emovdb subset of AEQ-BENCH with *tone variation*. The middle part is the constructed context and the original audio. The utterances with different emotions/tones share the same context. Each audio tone (amused, disgusted, angry, and neutral) has a plausible explanation for its context and a corresponding response.

nel (the semantic content or chosen words) and the vocal channel (the non-semantic delivery (Trager, 1958). Paralinguistic features belong to this vocal channel, conveying emotional meaning through prosody, which is variations in pitch, loudness, timbre, speech rate, and pauses. These paralinguistic cues are indispensable for empathetic communication (Loveys et al., 2021) because they can drastically alter the meaning of verbal content, such as in the case of sarcasm, or reveal emotional states that text transcription alone cannot capture. MFCCs (Mel-Frequency Cepstral Coefficients) (Davis and Mermelstein, 1990), for instance, capture the characteristics of the sound that are most discernible to the human ear. MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008) are widely used benchmark for SER. EMOV-DB (Adigwe et al., 2018) is an emotional speech dataset used for synthesis and generation purposes, offering varied content in different tones delivered by both male and female voices. Furthermore, specialised datasets, such as VocalSound (Gong et al., 2022), provide collections of non-lexical human vocalisations, including laughter and sighs, to train and evaluate models across a broader range of human-voice acoustics. For multimodal dialogue, the evaluation on generated audio quality is conducted by human listeners rating the naturalness or quality of the speech (Zhang et al., 2025b; Xu et al., 2025a; Du et al., 2024), using Mean Opinion Score (MOS) (Viswanathan and Viswanathan, 2005). However, OLMs' sensitivity to paralinguistic perception, such as tone and emotions, remains underexplored. AEQ-BENCH fills the gaps by setting the dialogue with both text and human audio, thereby evaluating the multi-modal empathetic con-

Subset	Source	Language	Core Contribution	Length
MELD	TV Series	Colloquial	Daily Conversational Emotion	18, 30
EMOV-DB	Studio recording	Literary	Explicit Emotional Range	10, 20
GIGASPEECH	Streaming content	Formal	Domain Diversity, Acoustic Variability	22, 17

Table 1: Contribution of each data source to AEQ-BENCH. EMOV-DB and MELD primarily contain audios with clear emotions. Length is the average words of the utterances and generated contexts in AEQ-BENCH.

Source	#Audio	#Context	Total
MELD	281	(×3) 843	843
GIGASPEECH	303	(×2) 606	606
EMOV-DB	(×4) 436	109	436
Total	1,020	1,558	1,885

Table 2: Statistics of AEQ-BENCH. There are 1,885 English instances, each pairing an audio utterance with a textual context. ($\times n$) means the number at the cell is n times to the neighbouring number at the same row.

versation ability and investigating the feasibility of automatic speech evaluation using OLMs.

3 AEQ-BENCH Benchmark

3.1 Benchmark Design

Empathy broadly refers to our reaction to the experiences of others. It has two major components: 1) **Cognitive empathy**, defined as the process of understanding another person's perspective[1] and 2) **Affective empathy**, also called emotional empathy, defined as an observer's emotional response to the affective state of others (Davis, 1983).

AEQ-BENCH for evaluating OLMs' empathy is constructed from three complementary datasets: MELD (Poria et al., 2019), GIGASPEECH (Chen et al., 2021), and EMOV-DB (Adigwe et al., 2018). They span the critical dimensions of real-world communication: contextual dialogue, domain diversity, and explicit emotional expression.[2] The contexts are constructed in reverse according to the Utterance. The features of each data source are present in Table 1 and the composition is listed in Table 2. It is classified into two types (985 chatting + 900 asking for help) and 6 topics (Appx. Table 8).

AEQ-BENCH covers two complementary evaluation tasks, as illustrated in Fig. 2 and 3:

1) Same utterance, different contexts. The first type of data focuses on the same utterance audio used in different contexts (**A–Cs**), where background framing shifts its interpretation. For example, the statement "*Then just give him some money*" could occur in a discussion about supporting a struggling relative, implying personal aid, or in a conversation about backing a political candidate, suggesting financial sponsorship. Despite the identical wording, the empathic strategies required are very different. (See the example in Fig. 5).

2) Same context, different utterances. The second type of data involves the same context and utterance *audio* in different emotional tones (**C-As**). Although the literal words remain identical, the affective delivery changes the underlying meaning. For instance, an angry "*Sometimes you really need to quarrel to solve problems*" conveys frustration and a need for validation, whereas the same sentence spoken calmly suggests a pragmatic and constructive view of conflict. Models should recognise the emotional signal and adjust their response style accordingly. (See the example in Fig. 6).

3.2 Benchmark Construction

Our benchmark is constructed in three steps:

1) Audio Filtering. Because our benchmark focused on empathy evaluation, we filtered out scenarios of (i) Physical actions, (ii) Mathematical reasoning or complex task completion, and (iii) Targeting real people. More details are in Appx. F.

2) Contexts Generation. We adopted GPT to frame different contexts for the retained data.

- **A–Cs** evaluate the ability to associate a specific context. MELD and GIGASPEECH both have one specific audio for each utterance. MELD are daily short colloquial conversations that could fit in various contexts. GIGASPEECH are more domain-specific and formal ones, restricting their flexibility. Therefore, we set 3 contexts for each MELD audio and 2 contexts for each GIGASPEECH audio.
- **C–As** evaluate the hearing ability of models. Different tones reflect the speaker's different at-

[1] In real human interaction, most daily conversations are informational or neutral. In these settings, empathy is expressed not through emotional mirroring but through helpfulness, perspective-taking, alignment with the interlocutor's goals, and context-appropriate support.

[2] We exclusively employ human speech to ensure natural paralinguistic variability (e.g., accents, prosody), a complexity absent in stable synthesised voices.

titude towards the same context, which requires models to infer the speaker's role or standpoint given a specific context. EMOV-DB recorded utterances with different explicit emotions. Here, we set only one context for all the different tones.

3) Quality Validation. To further validate the data quality of AEQ-BENCH, we first prompted GPT to generate a plausible explanation for why the context and utterance are coherent and provide a reference response. Then, human annotators manually verified that each constructed context is coherent.[3] Specifically, they checked: (i) The context plausibly leads to the audio. (ii) For A–Cs, the contexts are semantically different; For C-As, all the emotions are explainable in such contexts.

3.3 Evaluation Metrics

We evaluate from two aspects: **linguistic** and **paralinguistic**. *modality reliance*, *naturalness*, *coherence*, *supportiveness*, and *discrimination* are evaluated as linguistic features. For paralinguistic, we focus on *delivery*.[4] Model judges are shown in Table 3 and more details in Appx. G and H.

Judge	Modality	Metrics
GPT-5	Text	M.R., Nat., Disc.
OLMs	Audio	Coh., Sup., Delivery

Table 3: The automatic judges for evaluation metrics.

• **Modality Reliance** (M.R.): *Which modality does the model rely on to generate the response?*

• **Coherence** (Coh.): *Is the response logically consistent to the context and utterance?* (semantic check for (Good, Fair, Poor)) (Ickes et al., 2000).

• **Naturalness** (Nat.): *How natural is the response?* More human-like responses have higher scores with the range (1-4) (Kühne et al., 2020).

• **Discrimination** (Disc.): *Across different contexts/tones, do the model responses vary?* Higher scores are for those tailored for context/tone with the range (NA, 1–6) (Ickes et al., 2000).

• **Supportiveness** (Sup.) : *Imagine yourself in this situation. How supportive does the response feel?* Here, we consider acknowledgment of feelings, perspective-taking, supportive intent (comfort/encouragement/help), and non-judgmental language for (Good, Fair, Poor) (Rogers, 1957).

• **Delivery**: *How supportive the response sounds?* Here we consider the tone of voice, pacing, pauses, etc., for (Good, Fair, Poor) (Burleson and Kunkel, 2009; Loveys et al., 2021).

[3]More annotation details are discussed in Appx. F.

[4]We also attempted to evaluate paralinguistic *emotions*, yet it cannot work. We discussed details in the Limitation section.

4 Experimental Settings

Baselines. 1) Qwen3-Omni (Xu et al., 2025b), 2) Qwen2.5-Omni (Xu et al., 2025a), 3) Qwen2-Audio (Chu et al., 2024), 4) Qwen-audio (Chu et al., 2023), 5) SALMONN (Tang et al.), 6) LLaMA (Fang et al., 2025), 7) Flamingo (Ghosh and Duraiswami), 8) Baichuan (Li et al., 2025), and 9) GPT (Hurst et al., 2024). (See Appx. B).

Human Annotation. For empathy judgement, we compared OLMs' results to human's. Audio responses were sampled evenly from each OLM. For each model pair, their responses to the same instance were evaluated by the same annotator to ensure consistent judgment; also, inter-annotator consistency was computed based on models' ranking within each pair. Details are described in Appx. I.

Consistency Measurement. Because empathetic scores are relatively subjective, for the *Inter-annotator consistency*, annotations are considered consistent if any of the following conditions hold: (i) the paired models receive different scores but the same rank order from both annotators; (ii) the paired models are tied in score by both annotators; or (iii) one annotator records a tie, and the other reports a score difference ≤ 1. For *Human–OLM judge consistency*, each model response is first assigned the average score of its two human annotators. The resulting average rank for each model pair is then compared against the OLM-judge ranking, following the aforementioned conditions.

Model	T	A	T+A	F	Norm
GPT	0	18	82	0	**0.91**
LLaMA	1	36	62	1	0.81
Baichuan	6	34	59	1	0.79
Qwen3-Omni	1	24	74	0	0.87
Qwen2.5-Omni	0	25	74	0	0.87
Qwen2-Audio	4	18	52	25	0.63
Qwen-Audio	5	18	65	12	0.77
SALMONN	13	27	57	3	0.77
Flamingo	10	36	47	7	0.70

Table 4: Evaluation on input modality reliance. The values indicate the proportion (%) of responses based on text only (T), audio only (A), both text and audio (T+A), or failure to respond (F). Normalised score (Norm) is the sum of the value T+A and the average of T and A.

Audio Output	Model	MELD		GIGASPEECH		EMOV-DB		Overall		Normalised Scores	
		Nat.	Disc.	Nat.	Disc.	Nat.	Disc.	Nat.	Disc.	Nat. (Rank)	Disc. (Rank)
✓	GPT	**3.95**	4.85	**3.93**	4.38	**3.97**	2.17	**3.95**	4.22	**0.98** (1)	0.64 (6)
✓	LLaMA	3.88	4.81	3.77	4.79	3.88	2.70	3.85	4.47	0.95 (4)	0.69 (4)
✓	Baichuan	3.90	**5.15**	3.81	4.95	3.94	3.08	3.88	4.74	0.96 (2)	0.75 (2)
✓	Qwen3-Omni	3.61	5.14	3.88	**5.02**	3.93	2.77	3.77	4.71	0.92 (5)	0.74 (3)
✓	Qwen2.5-Omni	3.87	4.38	3.87	4.63	3.94	2.34	3.89	4.17	0.96 (2)	0.63 (7)
-	Qwen2-Audio	3.01	4.63	2.32	3.85	3.18	2.93	2.82	4.00	0.61 (7)	0.60 (8)
-	Qwen-Audio	2.25	4.78	2.46	5.01	2.48	**4.63**	2.37	**4.86**	0.46 (8)	**0.77** (1)
-	SALMONN	3.25	4.61	3.18	4.66	3.31	3.09	3.24	4.39	0.75 (6)	0.68 (5)
-	Flamingo	2.40	4.10	2.31	3.79	1.95	2.66	2.27	3.73	0.42 (9)	0.55 (9)

Table 5: Naturalness (Nat.) measures how human-like the responses are (ranges 1-4 and the higher the more human-like). Discrimination (Disc.) assesses the model's ability to distinguish responses across varying contexts (MELD & GIGASPEECH) or audio tone (EMOV-DB) (ranges 1-6, and the higher the better). They both are measured on text and evaluated by GPT-5-mini (with manual checking) following the text evaluation practice (Fu et al., 2024).

5 Results on Omni-Models as Responders

We first discuss model responses with empathy. Table 4 shows *Modality Reliance* of baselines, Table 5 their *Naturalness and Discrimination*, and Table 7 *Delivery* evaluated by human and OLM-judges. We observe that: (1) OLMs with audio outputs outperform text-only-output models. GPT and Qwen's Omni series represent the top-tier audio synthesis quality. These indicate the benefits of multimodal integration in learning empathy. (2) Models' performance varies on Discrimination and Naturalness, indicating they may be decoupled abilities.

• **Modality Reliance.** As shown in Table 4, all OLMs primarily rely on both audio and text modalities (over 50%). GPT demonstrates the best multimodal integration (normalised to 0.91), followed by Qwen's omni series (0.87). Earlier audio analysis models exhibit higher failure rates and weaker multimodal grounding, i.e., Qwen's audio series.

• **Naturalness.** Among OLMs, GPT achieves the highest overall Naturalness (normalised to 0.98), closely followed by Baichuan and Qwen2.5-Omni (both 0.96). It suggests that recent OLMs can produce responses better resembling human conversational style. In contrast, Flamingo and the Qwen-Audio series exhibit substantially lower Naturalness, often providing analytical or descriptive content rather than natural, conversational replies.

• **Discrimination.** In contrast, Qwen-Audio attains the highest Discrimination score (normalised to 0.77), particularly on the EMOV-DB subset. It indicates strong sensitivity to varying emotional tones of input audio. However, this sensitivity is primarily due to analytical variation (not communicating), as evidenced by its poor Naturalness.

• **Delivery.** Only OLMs with audio outputs are evaluated for the paralinguistic metric. Both Human and OLM-judges consistently rank GPT, Qwen3-Omni, and Qwen2.5-Omni as the top-tier models in paralinguistic quality, indicating superior control over prosody and vocal delivery compared to LLaMa and Baichuan.

6 Results on Omni-Models as Evaluators

We then discuss how OLMs evaluate empathy from generated audio. Tables 7 and 6 show the results.[5]

Alignment with Human Judgment. In Table 6, GPT exhibits the highest overall consistency with human annotators across coherence (84.9%) and delivery (88.7%).[6] It closely approaches human-human agreements. Qwen2.5-Omni shows competitive performance in terms of supportiveness (average 88.8%, comparable to human-human interactions). These imply that the best OLMs have the potential for coarse-grained automatic empathy evaluation from audio. It again highlights the benefits of more generic, multimodal capabilities in understanding empathy. In contrast, Qwen2-Audio yields substantially lower agreement across all dimensions, confirming that earlier audio-analysis–oriented models are not effective here.

OLMs' Evaluation Consistency Table 7 shows the ratings from OLM- and human-judges over delivery. Qwen2-Audio (Q-A) exhibits polarised scoring and a marked bias toward Qwen3-O, hindering

[5]The Coherence and Supportiveness scores are in Appx. C.

[6]For Delivery, level-samples are provided in prompts, which suggests GPT has stronger in-context-learning ability.

Judge	MELD (%)			GIGASPEECH (%)			EMOV-DB (%)			Average (%)		
	Coh.	Sup.	Para.	Coh.	Sup.	Para.	Coh.	Sup.	Para.	Coh.	Sup.	Para.
Human	89.7	89.7	87.9	**95.0**	90.0	80.0	**82.5**	**88.8**	**88.8**	**87.6**	**89.3**	86.5
GPT	**94.1**	85.3	**88.2**	87.5	83.3	**100.0**	77.1	75.0	83.3	84.9	80.2	**88.7**
Q-Omni	87.9	**91.4**	86.2	92.5	**95.0**	82.5	57.5	83.8	75.0	75.3	88.8	80.3
Q-Audio	58.6	60.3	75.9	57.5	65.0	72.5	57.5	51.3	56.3	57.9	57.3	66.3

Table 6: The consistency between human annotators and model judges on coherence (Coh.), supportiveness (Sup.), and Delivery (Para.), respectively. Model judges are GPT-4o-audio-preview (GPT), Qwen2.5-omni (Q-omni), and Qwen2-Audio (Q-Audio). The row of Human is the annotation consistency between two human annotators.

Model	MELD					GIGASPEECH					EMOV-DB				
	GPT	Q3	Q-O	Q-A	H	GPT	Q3	Q-O	Q-A	H	GPT	Q3	Q-O	Q-A	H
GPT	**0.65**	**0.99**	**0.96**	0.14	0.77	**0.56**	**0.98**	**0.93**	0.06	0.81	**0.59**	**1.00**	0.96	0.09	0.55
Qwen3-omni	0.56	0.84	0.81	0.95	**0.94**	0.53	0.95	**0.93**	0.97	**0.85**	0.56	0.96	**0.96**	0.94	**0.86**
Qwen2.5-Omni	0.61	0.95	0.92	0.14	0.93	0.55	0.97	0.88	0.09	**0.85**	0.57	0.97	0.95	0.10	0.80
LLaMA	0.54	0.84	0.78	0.13	0.67	0.51	0.84	0.75	0.09	0.52	0.53	0.86	0.82	0.06	0.34
Baichuan	0.49	0.74	0.72	0.15	0.64	0.51	0.73	0.67	0.08	0.61	0.51	0.76	0.79	0.08	0.53

Table 7: Evaluation on paralinguistic features: *Delivery*. For easy reading, we present normalised scores from a 3-point Likert scale (the higher, the better). The columns show the judges evaluating models in rows. Judges include GPT-4o-audio-preview (GPT), Qwen3-Omni (Q3-O), Qwen2.5-Omni (Q-O), Qwen2-Audio (Q-A), and Human (H). The boldface numbers indicate the best ones according to the original scores, listed in Table 12, Appx. C.

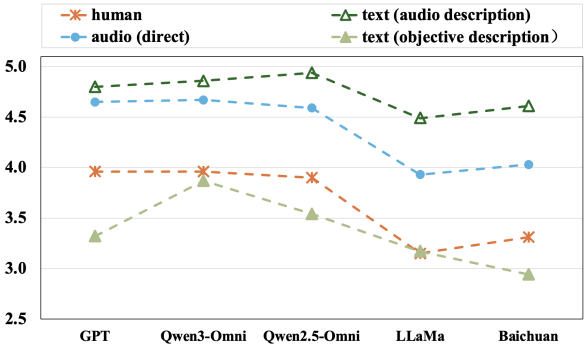

Figure 4: Average 5-point paralinguistic (delivery) evaluation (y-axis) for model on x-axis. Higher is better. For judges, stars indicate human evaluation, while circles indicate OLM-judges using direct audio. Empty and filled triangles denote OLM-judges using additional contexts of the original and objective audio caption, respectively.

reliable rank differentiation for scores exceeding 2. In contrast, other OLM-judges consistently distinguish Qwen-Omni models and GPT from lower-tier models, aligning well with human judgment. This demonstrates the potential of OLM-judges to effectively evaluate coarse-grained delivery. However, regarding fine-grained delivery, a divergence emerges: while human judges prefer the Qwen-Omni series over GPT, OLM-judges favour the opposite. These findings suggest that the ability of OLMs to discern fine-grained superiority in delivery is not yet aligned with human preference.

7 Further Discussion

Quantitative Analyses of Paralinguistics. We have highlighted the challenges in fine-grained evaluation of paralinguistic delivery. To provide further insight, we investigate whether textual descriptions of audio (captions) can aid in assessing paralinguistic empathy. We utilised Qwen3-Omni for audio captioning and conducted a fine-grained evaluation, expanding the scoring metric from a 3-point to a 5-point scale to capture greater nuance. Results are shown in Fig. 4 with details in Appx. J.

Direct audio judging (circles) aligns with the general ranking order of human (stars) but achieves only 43% consistency with human judgements. Incorporating the original audio captions (empty triangles) exacerbates the issue, as the captions tend to be biased and overly positive regarding machine delivery (e.g., "*The speaker's voice is calm, gentle, and empathetic...*"). Furthermore, although prompting the OLM to emphasise objective descriptions (filled triangles) brings the results closer to human evaluation, it tends to introduce a negative bias. It implies that captions can partially help, yet their effectiveness is highly sensitive to their quality.

Moreover, these results demonstrate that textual audio captions, whether prompted for objectivity or not, are not faithful proxies for auditory perception. It is because nuanced tones and implicit cues can

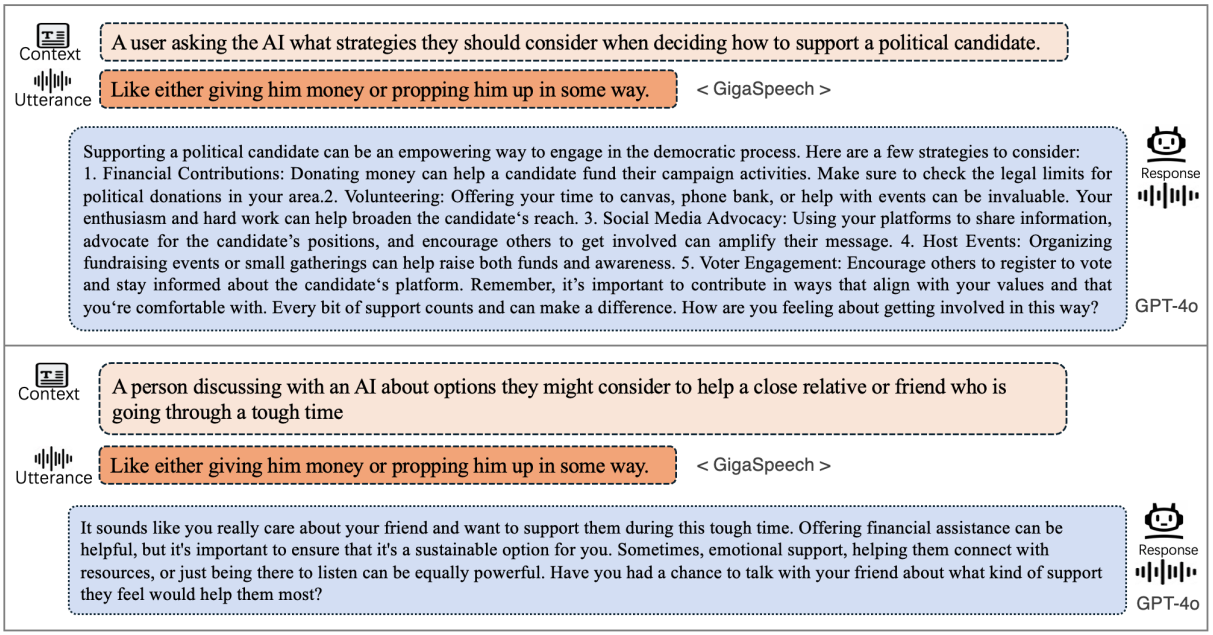

Figure 5: The example of context-variant responses generated by *GPT-4o*. Its responses are typically long and adhere to a standardised structural format: first, a situation analysis, then suggestions, and finally a question.

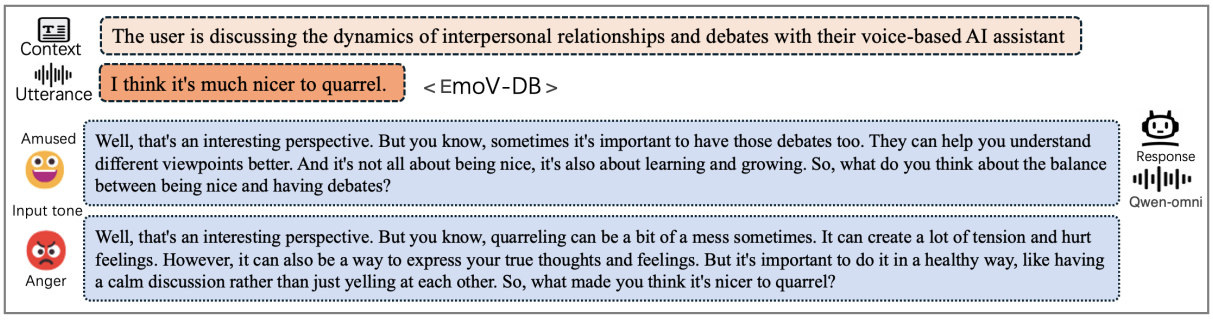

Figure 6: The example of tone-variant responses generated by *Qwen2.5-Omni*. The model shows a distinct stylistic pattern, frequently starting with "*well*" and adopting a reflective tone marked by phrases such as "*but you know*".

be easily overlooked and excluded in captions. As a result, captions cannot substitute for direct audio evaluation, particularly when assessing emotional delivery. It highlights a critical limitation of current OLM evaluators and underscores the necessity for a more dedicated audio paralinguistic modelling.

Qualitative Analyses. We further interviewed the annotators and found GPT-4 to be perceived as the most human-like system, followed by Qwen-Omni. While providing empathetic content, Qwen-Omni tends to append template-style sentences, resulting in less discrimination. Baichuan shows more stable conversational behaviour than LLaMA, whereas LLaMA occasionally generates sarcastic responses rather than uniformly supportive ones.

Finally, we present a qualitative analysis of the generated samples. While generally lacking strong expressive prosody, interestingly, the models display distinct stylistic characteristics. GPT-4o offers fast, structured responses with smooth, standardised delivery (see Fig. 5). Qwen2.5-Omni adopts a reflective tone, often initiating with "*well*" at a moderate pace (see Fig. 6). While Qwen3-Omni occasionally generates vivid, emotional audio, it suffers from instability. Baichuan shows lower consistency, characterised by evident fluctuations in accent. These findings highlight a persistent gap in achieving truly human-like, empathetic generation.

8 Conclusion

We have presented AEQ-BENCH to evaluate empathetic response generation and judgments in OLMs. By employing novel context- and tone-variant settings, we found that although top-tier OLMs exhibit strong linguistic empathy, they still face significant challenges in mastering fine-grained paralinguistic delivery. Moreover, our results show that textual

audio captions cannot replace direct audio analysis, underscoring the urgent need for audio-native evaluators to develop genuinely empathetic AI systems.

Limitations

Our work provides an initial exploration into the EQ of OLMs, but it is subject to several limitations that suggest future research directions:

Language and Cultural Scope: AEQ-BENCH is constructed exclusively in English. Empathy is highly language- and culture-dependent, meaning the emotional cues, appropriate supportiveness strategies, and even the naturalness of delivery can vary drastically across different linguistic groups.

Coarse-Grained Metric: Our metrics use a coarse-grained categorical rating ("Good," "Fair," "Poor"). While evaluating audio quality using OLM judges has high consistency with coarse-grained metrics, the complexity of assessing the nuanced emotional intent and consistency of the model-generated audio output remains challenging.

Emotion evaluation: The lack of expressive variation in the synthesized audio results in mostly neutral emotion in responses. The GPT/Qwen models primarily rely on textual analysis to judge emotion, rating various emotion distributions. So far, the acoustic quality limitation of the synthesized speech prevents the evaluation results from reliably showing a disparity in emotional rendering between the tested systems.

Ethical Considerations

The dual use of highly empathetic OLM capabilities presents the primary risk: the generated supportive responses could be exploited for manipulation, phishing, or disinformation. Furthermore, the study raises concerns about fairness and bias due to the potential for emotional overgeneralization or bias confirmation stemming from the underlying data. Finally, processing a user's sensitive emotional state from audio highlights significant privacy risks. Our work is confined to research, with future deployment requiring strong safeguarding mechanisms and adherence to privacy-by-design principles.

Additionally, to ensure ethical annotation data collection, all human annotators (student helpers) involved in model response evaluation were compensated fairly. Annotators were paid at a standard university rate of 16 USD per hour.

Acknowledgments

This work is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.PolyU/25200821 and Project No. T41-517/25-N), the Innovation and Technology Fund (Project No. PRP/047/22FX), and a gift fund from Huawei (N-ZGM3).

This work was supported by the National Natural Science Foundation of China 62576120.

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Lisa Aziz-Zadeh, Tong Sheng, and Anahita Gheytanchi. 2010. Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. *PloS one*, 5(1):e8759.

Brant R Burleson and Adrianne Kunkel. 2009. Revisiting the different cultures thesis: An assessment of sex differences and similarities in supportive communication. In *Sex differences and similarities in communication*, pages 133–154. Routledge.

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

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 4376–4380. International Speech Communication Association.

Shun Chen, Faith Liao, David Murphy, and Stephen Joseph. 2023. Development and validation of a 12-item version of the barrett-lennard relationship inventory (bl ri: mini) using item response theory. *Current Psychology*, 42(13):10566–10580.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *CoRR*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Mark H Davis. 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Steven B. Davis and Paul Mermelstein. 1990. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, page 65–74. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Sreyan Ghosh and Ramani Duraiswami. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *TTIC Summer Workshop on Foundations of Speech and Audio Foundation Models 2025*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

William Ickes, ANN Buysse, HAO Pham, Kerri Rivers, James R Erickson, Melanie Hancock, Joli Kelleher, and Paul R Gesn. 2000. On the difficulty of distinguishing "good" and "poor" perceivers: A social relations analysis of empathic accuracy data. *Personal Relationships*, 7(2):219–234.

Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics*, 14:593732.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025. Baichuan-omni-1.5 technical report. *CoRR*.

Kate Loveys, Mark Sagar, Xueyuan Zhang, Gregory Fricchione, Elizabeth Broadbent, and 1 others. 2021. Effects of emotional expressiveness of a female digital human on loneliness, stress, perceived support, and closeness across genders: randomized controlled trial. *Journal of medical Internet research*, 23(11):e30624.

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

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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.

Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95.

Ashish Sharma, Adam Miner, David 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 5263–5276.

Alberta B Szalita. 1976. Some thoughts on empathy: The eighteenth annual frieda fromm-reichmann memorial lecture. *Psychiatry*, 39(2):142–152.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

George L. Trager. 1958. Paralanguage: A first approximation. *Studies in Linguistics*, 13:1–12.

Mahesh Viswanathan and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer speech & language*, 19(1):55–83.

Haoyu Wang, Guangyan Zhang, Jiale Chen, Jingyu Li, Yuehai Wang, and Yiwen Guo. 2025. Emotion omni: Enabling empathetic speech response generation through large language models. *arXiv preprint arXiv:2508.18655*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

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. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, and 1 others. 2025a. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*.

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

Yiman Zhang, Ziheng Luo, Qiangyu Yan, Wei He, Borui Jiang, Xinghao Chen, and Kai Han. 2025c. Omnieval: A benchmark for evaluating omni-modal models with visual, auditory, and textual inputs. *arXiv preprint arXiv:2506.20960*.

A Details of Utilized Benchmark

The comparisons are listed in Table 9 and 10.

1) MELD (Multimodal EmotionLines Dataset), sourced from TV series Friends and was labeled for Emotion Recognition in Conversation. The dialogues are daily communication among family, friends, and neighbourhood rather than academic or knowledge.

2) GIGASPEECH, compiled from diverse YouTube and Podcast content, tests the system's ability to handle relatively formal and terminology across topics, thereby broadening the EQ assessment beyond simple daily conversation.

3) EMOV-DB (Emotional Voices Database) provides the acoustically controlled emotional ground truth. EMOV-DB 's studio-recorded speech offers clean, high-fidelity examples of explicit core emotions. This inclusion is crucial for isolating pure paralinguistic cues and ensuring the system can accurately discriminate core acoustic-emotional features, independent of the complexities introduced by dialogue context.

The topic distribution are listed in Table 8.

Topic	#
Personal Life / Feelings	821
Career/Work	448
Relationship	205
Educational / Study / Life Skill	174
Entertainment / Art	89
Others	148
sum	1,885

Table 8: Statistics of AEQ-BENCH. For MELD and GIGASPEECH, each audio has three and two different contexts, respectively. For EMOV-DB, each context has four audios with different tones. Each *(audio, context)* pair is considered one data instance, a total of 1,885.

B Details of Models

Tested baseline models are listed in Table 11.

1) **Qwen3-Omni-30B-A3B-Instruct** (Xu et al., 2025b) and 2) **Qwen2.5-Omni-7B** (Xu et al., 2025a) have an end-to-end architecture that processes text, images, audio, and video inputs and generates simultaneous text and speech outputs. It is designed for streaming, low-latency, and deployment on edge devices.

3) **Qwen2-Audio-7B-Instruct** (Chu et al., 2024) and 4) **Qwen-audio-chat** (Chu et al., 2023) are similar series, both designed for audio chat (with text output only) and audio analysis, such as music analysis or sound recognition.

5) **SALMONN-7B** (Tang et al.) is capable of emotion recognition, speaker verification, and music and audio captioning beyond traditional tasks.

6) **LLaMA-Omni2-7B** (Fang et al., 2025) focuses on creating a real-time spoken chatbot by streaming speech synthesis. Although it can only process audio and text modalities, the "Omni" suggests broader multimodal capabilities in the LLaMA series.

7) **audio-flamingo-3** (Ghosh and Duraiswami) advances understanding and chain-of-thought-type reasoning across speech, along with multi-turn, multi-audio chat ability. Code for Flamingo Streaming-TTS pipeline has not been released yet.

8) **Baichuan-Omni-1d5** (Li et al., 2025) is another comprehensive model capable of processing text, image, audio, and video inputs and generating text and voice output. It decodes the text and audio simultaneously, which is likely to be slightly different literally but semantically the same. While TTS models and those providing the audio transcription have identical contents in audio and text.

9) **gpt-4o-audio-preview** (Hurst et al., 2024) is a version of GPT-4o with expanded support for audio inputs and the ability to generate text and audio responses.

Benchmark	Task	In	Out	Size
GigaSpeech (Chen et al., 2021)	ASR (En, multi-domain)	A	T	∼40k h (10k h labeled)
MELD (Poria et al., 2019)	Emotion recognition (multi-party)	A+V+T	T	∼1.4k dialogs; 13k utt.
Emov-DB (Adigwe et al., 2018)	Emotional TTS (controllable)	T	A	4 speakers; 5 emotions
AEQ-BENCH (Ours)	**Empathy response generation**	**A+T**	**A/T**	**1,885 instances**

Table 9: Multi-modality benchmarks. A: Audio; V: Video; T: Text.

Dataset	Original Modality	Contextual Depth
MELD	Video with audio, Text	High: multi-turn, multi-party conversations
GIGASPEECH	Audio, Text (Transcription)	Low to Medium (domain focus)
EMOV-DB	Audio, Text	Low (explicit emotions without context)

Table 10: Benchmark features. EMOV-DB and MELD primarily contain audios with clear emotions.

NO.	Model	Input	Output	ASR	Multiturn	Ins.Fo.	Size(B)	Judge
1	Qwen3-Omni	A, T, I, V	A, T	✓	✓	✓	35	✓
2	Qwen2.5-Omni	A, T, I, V	A, T	✓	✓	✓	10.7	✓
3	Qwen2-Audio	A, T	T	✓	✓	✓	8.4	✓
4	Qwen-Audio	A, T	T	✓	✓	✓	8.4	×
5	SALMONN	A, T	T	✓	×	✓	~7.0	×
6	Flamingo	A, T	T, TTS	✓	✓	✓	~7.0	×
7	LLaMA-Omni	A, T	T, TTS	×	✓	×	9.0	×
8	Baichuan-Omni	A, T, I, V	A, T	✓	✓	✓	11.0	×
9	GPT	A, T	A, T	✓	✓	✓	-	✓

Table 11: For Input and Output modalities, there are Audio(A, TTS), Text(T), Image(I), and Video(V). For model capabilities, ASR, Multiturn, and Ins.Fo. stand for automatic speech recognition, multi-turn diglogues, and text instruction following, respectively. The model sizes come from HuggingFace (~ denotes approximation). The model versions are Qwen3-Omni-30B-A3B-Instruct, Qwen2.5-Omni-7B, Qwen2-Audio-7B-Instruct, Qwen-audio-chat, SALMONN-7B, LLaMA-Omni2-7B, audio-flamingo-3, Baichuan-Omni-1d5, and gpt-4o-audio-preview, in order. The Judge column indicates whether they are evaluated as a judge model.

Models	MELD					GIGASPEECH					EMOV-DB				
	GPT	Q3-O	Q-O	Q-A	H	GPT	Q3-O	Q-O	Q-A	H	GPT	Q3-O	Q-O	Q-A	H
GPT	**1.71**	**1.01**	**1.09**	2.72	1.46	**1.88**	**1.05**	**1.14**	2.87	1.38	**1.83**	**1.00**	1.09	2.83	1.90
Qwen3-O	1.89	1.31	1.37	1.11	**1.13**	1.94	1.11	**1.14**	1.06	**1.31**	1.88	1.07	**1.08**	1.11	**1.28**
Qwen2.5-O	1.79	1.10	1.16	2.72	1.15	1.90	1.05	1.24	2.83	**1.31**	1.86	1.07	1.11	2.81	1.41
LLaMA	1.92	1.32	1.45	2.74	1.67	1.98	1.32	1.51	2.83	1.97	1.95	1.28	1.37	2.88	2.32
Baichuan	2.02	1.53	1.57	2.71	1.73	1.98	1.53	1.67	2.85	1.78	1.99	1.48	1.43	2.84	1.95

Table 12: Evaluation on paralinguistic features: *Delivery*. The columns show the judges evaluating audio from models in rows. Judges include GPT-4o-audio-preview (GPT), Qwen3-Omni (Q3-O), Qwen2.5-Omni (Q-O), Qwen2-Audio (Q-A), and Human (H). The level ranges are 1: Good, 2: Fair, 3: Poor. The smaller the value, the better the delivery. Q3-O rates GPT 1.04545 and Qwen2.5-Omni 1.0512, respectively.

Model	MELD		GIGASPEECH		EMOV-DB		MELD		GIGASPEECH		EMOV-DB	
	Coh.	Sup.	Coh.	Sup.	Coh.	Sup.	Coh.	Sup.	Coh.	Sup.	Coh.	Sup.
GPT	**1.01**	**1.01**	**1.02**	**1.00**	**1.01**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**
Qwen2.5-Omni	1.07	1.03	1.03	1.03	1.02	1.02	0.97	0.99	0.99	0.99	0.99	0.99
Qwen3-Omni	1.38	1.28	1.29	1.03	1.38	1.05	0.81	0.86	0.86	0.99	0.81	0.97
LLaMA-Omni	1.12	1.15	1.17	1.16	1.12	1.13	0.94	0.93	0.92	0.92	0.94	0.94
Baichuan-Omni	1.07	1.27	1.11	1.34	1.07	1.25	0.97	0.87	0.95	0.83	0.97	0.88

Table 13: The average scores of Coherence (Coh.) by GPT and Supportiveness (Sup.) by Qwen2.5-Omni. The level ranges are 1: Good, 2: Fair, 3: Poor. The smaller the value, the better. The right part is the normalised scores.

C Evaluation Results

The results of Delivery are in Table 12. The results of Coherence and Supportiveness are in Table 13.

D Explanation for Figure 2

In this example:

1. **Correctly infer the user's perspective and intention.**

- (text) Go to the store in person.
- (text) Get investment info.

2. **Interpret the underlying concerns.**

- (audio) people over a certain age → potential wheelchair accessibility concerns.
- (audio) more than an acre of → scale of commercial usage.

3. **Adapt the response to be appropriate, relevant, and supportive to address the user's actual concern.**

- (response) Special accessibility support.
- (response) Invest in similar properties that occupy a large area.

E Difference from Instruction Following

1. **Cognitive empathy** v.s. **instruction following**

Instruction following presumes an explicit directive from the user, e.g., "Provide…," "Tell me…," "Answer…," "Give me…," with a clearly defined task-oriented goal.

By contrast, cognitive empathy involves the ability to:

- infer the user's unstated goal, concern, or perspective,
- interpret ambiguous or underspecified context, and
- respond in a way that aligns with the user's internal state, even when it is not explicitly articulated.

2. **Why this distinction applies to AEQ-Bench**

In our setting, the context is a compressed summary of prior dialogue, but it contains no explicit instructions. The model must therefore infer the user's intention from clues in the context + utterance. In example (Fig. 2), the model pays attention to different parts of the same utterance according to different contexts.

The model should read between the lines to infer these concerns when the user does not instruct "Please check accessibility" or "Please provide investment details".

Thus, the task is not a direct instruction-following exercise but a test of whether the model can: i) understand the user's perspective, ii) detect latent concerns, and iii) craft a contextually supportive reply.

3. **Relation to our broader goal**

Most real human conversations are not phrased as tasks or explicit instructions.

People frequently imply their concerns, leave intentions unstated, or expect the listener to infer what they need. For example, when someone says, "That café is up a really steep hill" they rarely add request or instruction, "Please check accessibility for me." A supportive interlocutor is expected to read between the lines and recognize the underlying concern. This ability—inferring perspective, intention, and concern from context without an explicit directive—is precisely what we aim to evaluate through the GigaSpeech subset.

AEQ-Bench is designed to evaluate EQ-oriented companion AI, which requires sensitivity to user perspectives, goals, and concerns/worries, even when emotions are implicit. This complements the prevailing focus on Instruction-following and IQ-oriented logic and reasoning evaluations.

F Construction Steps

Audio filtering. We only retain audios with clear delivery to prevent adding unnecessary difficulty to the speech recognition stage. We exclude all instances of the sleepiness tone present in the EMOV-DB dataset, as these are typically slow, slurred, and often contain non-speech acoustic cues (e.g., yawns).

On the other hand, data instances that require high-level reasoning or factual complexity were excluded. Specifically, :

(i) Asking the other party to perform physical actions. (e.g., *pick up my friends.*)

(ii) Mathematical reasoning or complex task completion. (e.g., *how to arrange the chores to finish them efficiently*; *booking a flight.*)

(iii) Target at real persons, such as intimate confessions, family disputes. (e.g., *You are the most charming woman in the world.*)

MELD and EMOV-DB are filtered manually, and GIGASPEECH is filtered by GPT in early rounds.

Quality Validation Annotators. Three in-house annotators, expertising in linguistics, conducted the verification. They were compensated at 16 USD/hour, the same rate as response evaluation annotators. Approximately 70% of the automatically generated contexts were judged coherent. The remaining contexts were manually revised by the annotators to ensure quality before being used for evaluation.

G Details of Evaluation Metrics

Empathy is inherently multi-dimensional. Therefore, it naturally requires both categorical decisions and graded judgments:

Categorical decisions reflect discrete states (i.e., which modality the model is relying on). Specifically, Modality Reliance: Did the model rely on the audio or the text? We need to know which modality OLMs utilize to give a response in conversation, since previous studies only include instructions in the other modality, leaving it unclear whether context information is involved.

Graded qualities capture nuance in empathic responding that cannot be captured by simple classifications (e.g., how supportive, how natural, how context-adaptive). Ordinal scales allow for subtle differentiation in degree (e.g., "good," "fair," "poor"; or numeric gradations) that reflect human judgments of how well a response makes them feel, not just whether it does or doesn't. (e.g., Barrett-Lennard Relationship Inventory (Chen et al., 2023), Mean Opinion Score (MOS) (Viswanathan and Viswanathan, 2005).

Most rating metrics in AEQ-Bench (i.e., Coherence, Supportiveness, Delivery) use three levels (Good, Fair, Poor) because these qualities tend to vary along coarse but meaningful interpersonal distinctions. This mirrors standard practice in social psychology and counseling assessment tools, where empathy-related constructs are commonly evaluated using 3-point or 5-point Likert scales to capture the essential gradation without overburdening annotators.

For example, Naturalness reflects a more nuanced, continuous progression in conversational behavior, from pure analytical, to analytical+Chatting, to AI-like Chatting, to human-like Chatting. Annotators consistently reported that these degrees of difference are clear and distinguishable, warranting a finer-grained ordinal scale.

Why these metrics are chosen?

Modality Reliance. Recognizing and using both vocal and textual cues ensures the system is sensitive to how emotion is expressed.

Coherence. essential for empathetic accuracy (Ickes et al., 2000). If a response is incoherent to the context, it likely fails to reflect empathic understanding.

Discrimination. Empathic accuracy is not just about detecting emotion, but differentiating subtle differences depending on context (Ickes et al., 2000). A model must adapt to context shifts (or changes in user tone or emotion) rather than produce generic responses.

Naturalness. More human-like language or voice creates a stronger sense of connection and perceived empathy (Kühne et al., 2020).

Supportiveness. Counseling psychology defines empathy in part as validation + perspective-taking + nonjudgmental support (Rogers, 1957).

Delivery and Emotion : Paralinguistic features like tone, pause, and pitch significantly influence perceived empathy in human speech (Burleson and Kunkel, 2009). Emotional expressiveness, rather than neutral tone, plays a crucial role in how people perceive empathy (Loveys et al., 2021).

H Evaluation Metric Scores

This appendix provides a detailed breakdown of the evaluation metrics used in our human and automated assessments, including the specific questions posed to annotators, the corresponding scale or options, and the precise criteria for each score. The metrics are divided into linguistic features (focusing on the content of the response) and paralinguistic features (focusing on the acoustic delivery).

H.1 Linguistic Features

Q1. Modality Reliance (M.R.) This metric evaluates which input source the model primarily utilizes to generate its textual response, verifying if it relies on the provided conversation context, the user's immediate utterance (ASR transcription), or a combination of both.

Evaluation Question: Which input modality (Context or Utterance) does the model primarily rely on, or does it utilize both, to generate the response?
Scale and Criteria:

1. **Context**: Response is based only on the provided CONTEXT.
2. **Utterance**: Response is based only on the current UTTERANCE (ASR).
3. **Both**: Uses CONTEXT and UTTERANCE together, showing no obvious conflict.
4. **Failed**: Response is irrelevant, off-topic, a repetition/translation of inputs, or includes disclaimers.

Q2. Naturalness (Nat.) Naturalness assesses how human-like and spontaneous the generated response is, simulating a natural conversation between peers. Higher scores indicate a more human-like exchange.
Evaluation Question: How natural and human-like is the response? The more human-like the conversation, the higher the score.
Scale and Criteria: NA, 1 to 4.

- **NA (Not Applicable)**: The response is irrelevant or a repetition/translation of the inputs.

1. **Score 1 (Analysis only)**: The response is pure comment or analysis, with no direct conversational reply.
2. **Score 2 (Analysis + Chatting)**: A mix of third-person analysis and first-person reply.
3. **Score 3 (AI-like Chatting)**: Direct reply with clear, formulaic Artificial Intelligence (AI) characteristics.
4. **Score 4 (Human-like Chatting)**: High naturalness, closely resembling a fluent human conversation.

Q3. Coherence (Coh.) Coherence judges the logical and emotional consistency of the response's content relative to the entire conversation and the user's current utterance.
Evaluation Question: Is the response logically and emotionally consistent with the overall CONTEXT and UTTERANCE?
Scale and Criteria: NA, Good, Fair, Poor.

- **NA (Not Applicable)**: The response is irrelevant or a repetition/translation of the inputs.

1. **Good**: Clearly relevant, logically consistent, and appropriate to the situation.
2. **Fair**: Mostly relevant but with minor inconsistencies or vague links to the context.
3. **Poor**: Off-topic, contradictory, or highly inappropriate given the inputs.

Q4. Supportiveness (Sup.) This metric evaluates the degree of supportive intent conveyed by the response's text, considering factors such as emotional acknowledgment, perspective-taking, supportive intent (comfort/encouragement/help), and non-judgmental language.
Evaluation Question: How supportive does the response feel? Considering acknowledgment of feelings, perspective-taking, supportive intent, and non-judgmental language.
Scale and Criteria: NA, Good, Fair, Poor.

- **NA (Not Applicable)**: The response is irrelevant or a repetition/translation of the inputs.

1. **Good**: Clear acknowledgment, perspective-taking, strong supportive intent, and non-judgmental tone.
2. **Fair**: Some supportive elements, but incomplete, formulaic, or mixed with minor judgment.
3. **Poor**: Lacks support, is dismissive, minimizing, or overtly judgmental.

Q5. Discrimination (Disc.) Discrimination measures the model's ability to vary its textual content based on differing contexts or emotional cues in the user's input. A high score indicates the response is highly tailored.
Evaluation Question: Across varying contexts/tones, how much does the model's response content vary? The more tailored for context/tone, the higher the score.
Scale and Criteria: NA, 1 to 6.

- **NA (Not Applicable)**: Response is irrelevant/repetition.

1. **Score 1 (No difference)**: Content remains identical across inputs.

2. **Score 2 (Minor wording)**: Same core content, only small lexical changes.

3. **Score 3 (Some variations)**: Mostly same content, but with isolated tailored variations.

4. **Score 4 (Mostly different)**: Responses show significant differences, adapting to context.

5. **Score 5 (Template-like)**: Clear differences, but the wording is highly standardized or template-based.

6. **Score 6 (Contextually adapted)**: Clear, diverse differences with contextually and emotionally appropriate, varied wording.

H.2 Paralinguistic Features (Audio Assessment)

Q6. Delivery Delivery assesses the acoustic quality and appropriateness of the speech synthesis, focusing on how paralinguistic cues (tone, pacing, prosody, and timing) contribute to the overall supportive or emotional intent of the response.
Evaluation Question: How supportive does the response *sound*? Does the tone of voice, pacing, pitch, and pauses effectively convey empathy or the intended emotion?
Scale and Criteria: NA, Good, Fair, Poor.

- **NA (Not Applicable)**: The response is irrelevant or a repetition/translation of the inputs.

1. **Good**: Warm, calm tone; patient pacing; appropriate prosody; natural pauses and timing.

2. **Fair**: Generally acceptable but inconsistent (e.g., slightly rushed/flat tone or occasional awkward pauses).

3. **Poor**: Cold/harsh tone, stilted or mechanical pacing, mismatched prosody, or unnatural timing.

I Human Annotation for Responses

Specifically, the five OLMs are grouped in pairs, resulting in 10 model pairs. We sample two groups of data for each mdoel pair. Each grouped data consists of one utterance associated with three contexts for MELD, two for GIGASPEECH, and four tonal variations plus one context for EMOV-DB. A total of 180 instances and responses.

Based on the provided context, utterance, response audio, and evaluation instructions, the annotators assess the responses across all metrics. All annotators are fluent English speakers recruited from university student helpers; a total of 12 annotators participated.

Cohen's Kappa (for categorical metrics): modality_dependency=0.72.

ICC (for ordinal scores): naturalness = 0.84, coherence=0.89, supportiveness=0.87, delivery=0.86, discrimination=0.80.

OLM judges are instructed to evaluate the responses with the same set of instructions as human annotators'.

J Fine-grained Paralinguistic Evaluation

This experiment examine whether emotional and paralinguistic qualities of synthesized speech can be faithfully evaluated through textual intermediates, rather than direct auditory perception.

Specifically, we compare: (i) human listening as the gold standard; (ii) direct audio judging by an OLM; (iii) free-form paralinguistic descriptions generated by an OLM; and (iv) constrained, objective textual descriptions of acoustic properties. For each setting, we compute 5-point paralinguistic quality scores across MELD, GigaSpeech, and EmoV-DB subsets. 5-point scale in Table 14.

Score		Evaluation
5	Very Good	Paralinguistic cues are highly expressive, perfectly conveying emotion and intent, resulting in smooth and natural communication (human-like).
4	Good	Paralinguistic cues are mostly clear. Slightly more emotional than typical machine speech, but still not fully natural.
3	Fair	Paralinguistic cues are basically conveyed and do not hinder understanding (typical machine-like speech).
2	Poor	Paralinguistic cues are vague and frequently interfere with communication, making understanding difficult.
1	Very Poor	Paralinguistic cues are almost unrecognizable and severely impair communication effectiveness.

Table 14: Five-point scale for paralinguistic evaluation.