EchoChat: Structured Cognitive Reasoning in Empathetic Spoken Dialogue

Dingdong Wang[1,2]*, Shujie Liu[2]†, Yayue Deng[1], Yuxuan Hu[2], Yunrui Cai[1], Jincenzi Wu[1], Jianwei Yu[2], Jinyu Li[2], Helen Meng[1]†
[1]The Chinese University of Hong Kong, China
[2]Microsoft Corporation
dingdongwang@link.cuhk.edu.hk

Abstract

Empathetic spoken dialogue is a sophisticated cognitive process that requires not only recognizing emotions but also inferring a user's latent mental states to provide appropriate support. However, current SpeechLLMs often treat empathy as a direct input-to-response mapping, leading to "superficially warm" but emotionally hollow interactions. In addition, since empathy relies on a multi-stage process with strong inter-step dependency, errors at any intermediate step can cascade through subsequent steps and lead to inappropriate responses, while existing training paradigms lack mechanisms to precisely localize and improve such errors. In this work, we propose EchoChat, a unified framework that reformulates empathetic spoken dialogue as a structured cognitive reasoning process integrating perception, mental-state reasoning, and response generation. To support this paradigm, we first construct EchoDialogue-400K, an acoustically rich dataset for multi-stage empathetic supervision. During the SFT stage, we strengthen acoustic grounding through proposed Acoustic-Anchored Attention (AAA). During the RL stage, we further introduce a novel stage-aware optimization objective with Step-Decomposed Credit Assignment (SDCA) to localize reasoning errors and mitigate cascaded error propagation. In addition, we introduce EchoEval, an expert-annotated benchmark for multi-dimensional empathy evaluation. Extensive experiments demonstrate that EchoChat achieves state-of-the-art performance in perception, reasoning, and response alignment. The project page, EchoChat model, benchmark, and EchoDialogue dataset are publicly available and open-sourced.

1 Introduction

Empathy is the cornerstone of human-centric interaction. It requires a system to understand not only what a user says, but also how the user feels and what kind of support the user actually needs [7, 27]. While recent advances in Speech Large Language Models (SpeechLLMs) have achieved remarkable fluency [1, 18, 42, 6], they often fall into the trap of "empathetic mimicry" [15], producing responses that are superficially warm but fundamentally misaligned with the user's latent mental state. This stems from a core limitation: existing systems treat empathy as a direct input-to-response generation problem [33, 35, 45, 11], essentially bypassing the structured cognitive reasoning process. In reality, empathetic dialogue requires perceiving fine-grained acoustic and semantic cues, inferring the user's latent mental state and underlying needs, and planning an appropriate response before speaking. This challenge is particularly pronounced in the speech domain, where subtle prosodic variations can substantially reshape the emotional context.

*Work done during an internship at Microsoft
† Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

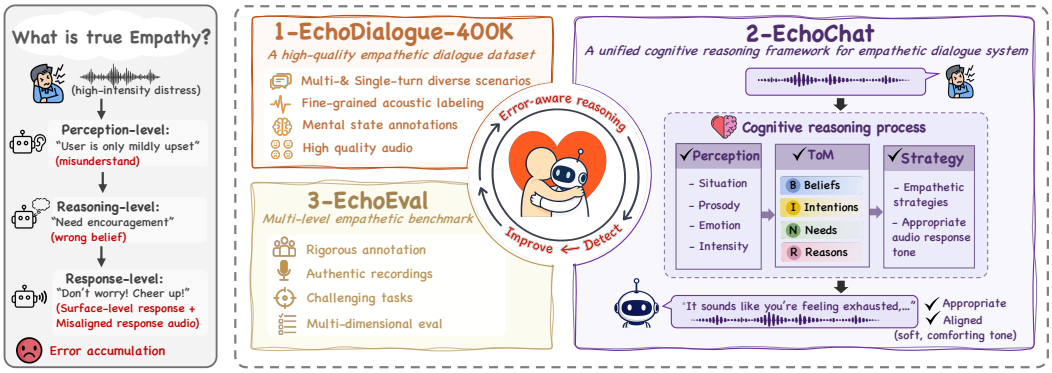


Figure 1: Reframing empathetic dialogue as structured cognitive reasoning. Existing methods simplify empathy to direct response generation, leading to cascaded reasoning errors across intermediate stages. EchoChat instead models empathy as a structured process from perception to response, supported by EchoDialogue-400K and EchoEval, enabling cognitively grounded and emotionally aligned spoken interaction.

The simplified, response-centric formulation of empathetic spoken dialogue introduces three fundamental limitations. (1) **Inadequate affective perception.** Existing studies model user affect using coarse-grained emotion categories (e.g., "happy", "angry") [33, 11, 19, 34, 17], failing to capture the subtle acoustic variations essential for authentic empathy. For example, within the same sadness category, intense sorrow accompanied by sobbing requires a fundamentally different support strategy from mild melancholy. This issue becomes more pronounced in complex scenarios, such as emotional shifts within an utterance and sarcasm, which current models find difficult to handle due to their limited perceptual granularity. Yet, existing datasets [11, 33] provide limited coverage of such fine-grained acoustic variations, restricting models from learning the broad spectrum of human affect. (2) **Vulnerability to cascaded error propagation.** Since empathetic reasoning involves strong dependency across intermediate steps, errors in early steps can propagate through subsequent steps and lead to inappropriate responses. Although recent works introduce Chain-of-Thought (CoT) as intermediate analysis [19, 40], they lack robust stage-wise credit assignment. As a result, such errors become difficult to localize and improve. (3) **Lack of unified optimization across empathetic stages.** Prior approaches often optimize individual stages in isolation [11, 33, 35, 34], such as emotion recognition and speech generation. However, empathetic spoken dialogue requires coordinated alignment across perception, reasoning, and response generation. Without a unified objective, gains at one stage may not reliably transfer to subsequent stages. ParaS2S [45] takes a step further by exploring perception-to-speech-response alignment, but still bypasses the intermediate reasoning that explains *what* the user actually needs and *why* a specific support strategy is appropriate.

To address these limitations, we introduce **EchoChat**, a unified framework that models empathetic spoken dialogue as a structured process of perception, latent mental-state reasoning, and response generation. The mental-state reasoning is grounded in Theory-of-Mind (ToM) principles [10], which model the ability to infer others' beliefs, intentions, needs, and psychological states from observable signals [10]. First, to support fine-grained perception, we construct EchoDialogue-400K, a high-quality empathetic dialogue dataset covering diverse acoustic expressions and complex affective scenarios. We further introduce Acoustic-Anchored Attention (AAA) during supervised fine-tuning to strengthen the grounding of perception in speech signals. Second, to address cascaded error propagation, we design a stage-aware cascaded reasoning reward with Step-Decomposed Credit Assignment (SDCA). The cascaded reward evaluates perception, mental-state analysis, and strategy planning in a dependency-aware manner. Moreover, unlike vanilla GRPO [12], which broadcasts a single advantage across all generated tokens, SDCA decomposes stage-level signals into token-level credit, allowing errors to be localized and corrected across the reasoning trajectory. Finally, we combine the reasoning reward with a multimodal response reward under a unified training framework to jointly optimize perception, reasoning, and emotionally aligned speech generation.

For evaluation, current empathetic dialogue benchmarks [48, 44, 51, 11, 45] are limited in both stage coverage and scenario complexity, preventing comprehensive assessment of empathetic spoken dialogue. To address this, we propose EchoEval, a challenging expert-annotated benchmark derived from authentic recordings to evaluate empathy across perception, reasoning, and response generation.

Table 1: Comparison with existing speech empathetic dialogue datasets. EchoDialogue-400K supports structured empathetic reasoning with fine-grained affective annotations and rich acoustic variations.

Dataset	Scale	Open Source	Multi-Turn	Reason.	Mental State	Prosody Labels	Intensity	Sarcasm	Affective Granularity
EChat-200K [11]	200K	✗	✗	✗	✗	✗	✗	✗	Coarse emotion labels
EmotionalQA [34]	200K	✗	✗	✗	✗	✗	✗	✗	Coarse emotion labels
ParaS2S [45]	100K	✗	✗	✗	✗	✗	✗	✓	Coarse emotion labels
OpenS2S [33]	100K	✓	✗	✓	✗	✗	✗	✗	Coarse emotion labels
EchoDialogue-400K	**400K**	✓	✓	✓	✓	✓	✓	✓	**Fine-grained variants**

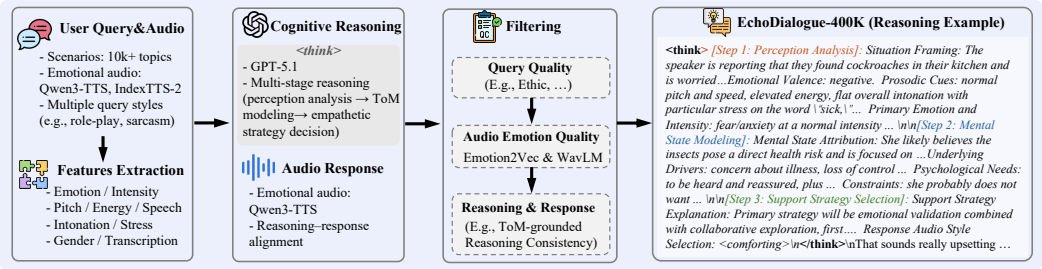


Figure 2: Data construction pipeline for EchoDialogue-400K.

EchoEval encompasses diverse real-world scenarios such as sarcasm, varying emotional intensities, and dynamic emotion shifts. Extensive experiments show that EchoChat consistently improves perception, reasoning, and response alignment over strong baselines, validating the effectiveness of modeling empathy as a structured cognitive reasoning process.

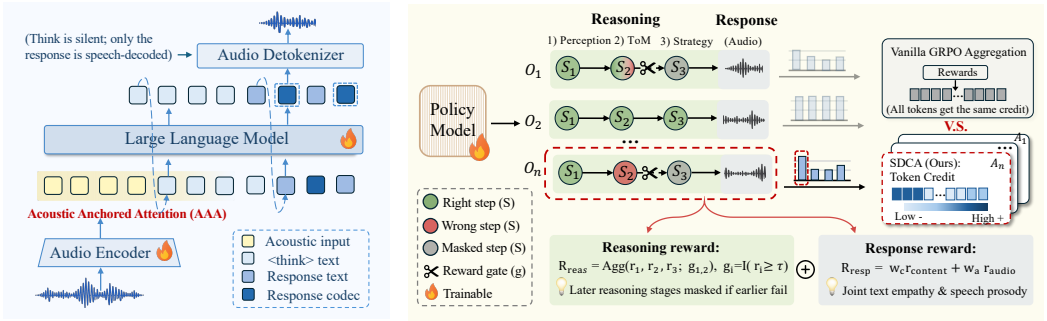
2 EchoDialogue-400K

As shown in Table 1, prior empathetic dialogue datasets offer limited fine-grained acoustic variation (e.g., prosody) and user mental-state annotation, which hinders grounded perception and structured reasoning. To address this, we introduce EchoDialogue-400K, which supports joint learning of perception, Theory-of-Mind (ToM) [10] reasoning, and speech generation. It contains 400K single-turn samples and 4K multi-turn dialogue sessions comprising 20K turns. To our knowledge, this is the first dataset providing integrated cognitive and acoustic supervision for spoken empathetic dialogue.

To address the limited scenario diversity and lack of expressive variation in prior corpora, we construct a structured taxonomy comprising 5 domains, 17 subcategories, and 10K+ topics (Figure 6). The emotion space covers six basic categories (happy, sad, angry, surprise, disgust, and fear) with fine-grained variants (e.g., anxiety within fear) and varying intensities. Figure 2 outlines our data generation and filtering pipeline: we sample user queries from multiple LLMs under varied prompts (neutral, sarcastic, role-play) on predefined scenarios; audio is synthesized using diverse TTS strategies combined with rigorous filtering to ensure expressive and accurate emotional delivery. Structured reasoning traces are generated to maintain consistency in ToM inference and strategy choice. Details on data statistics, construction, and filtering are provided in the Appendix B.

3 Method

We formulate empathetic spoken dialogue as a structured process that jointly models acoustic perception, Theory-of-Mind reasoning, and response generation. Given user speech, the model generates a reasoning trace (`<think>`) and an emotionally aligned response, ensuring generation is grounded in inferred mental states rather than surface patterns. Figure 3 summarizes EchoChat: Stage 1 is acoustic-grounded SFT with Acoustic-Anchored Attention (AAA) (Sec. 3.1); Stage 2 is structured RL (Sec. 3.2) that jointly optimizes perception, reasoning and responses via cascaded reasoning reward, response-level reward and Step-Decomposed Credit Assignment (SDCA).



(a) SFT Stage (b) RL Stage

Figure 3: Overview of the EchoChat framework. (a) SFT uses Acoustic-Anchored Attention (AAA) to ground reasoning in acoustic signals. (b) RL performs structured reasoning optimization, where Step-Decomposed Credit Assignment (SDCA) and gating mitigate error propagation, and a response reward model aligns content and speech. Unlike vanilla GRPO, credit is stage-aware at the token level rather than uniform.

3.1 Supervised Fine-Tuning (SFT)

We observe a common failure mode in chain-of-thought (CoT) reasoning for SpeechLLMs [36, 22, 43, 31]: the model may produce spurious reasoning that mimics the surface form of explanations and follows next-token prediction patterns, rather than consistently conditioning on the input audio. However, standard supervised fine-tuning does not enforce a structural dependency between intermediate reasoning and the audio stream. As the reasoning trace grows, later steps become less grounded in acoustic cues, leading to misinterpretation of the user's emotional state.

To address this, we introduce **Acoustic-Anchored Attention (AAA)**, a lightweight regularization mechanism that explicitly encourages perception-stage reasoning to attend to acoustic representations. Given the structured reasoning trace, we identify perception-related tokens $\mathcal{P}$ (perception stage tokens inside the structured reasoning span delimited by `<think>`) and acoustic tokens $\mathcal{A}$, where the latter correspond to positions whose keys and values are derived from the speech encoder. For attention weights $\alpha_{i,j}$, we measure the average attention mass allocated from perception tokens to acoustic tokens as

$$\rho = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{A}} \alpha_{i,j}, \tag{1}$$

and maximize this quantity via an auxiliary objective $\mathcal{L}_{\text{AAA}} = -\rho$.

The overall training objective is

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{AAA}} \cdot \mathcal{L}_{\text{AAA}}. \tag{2}$$

AAA imposes a soft structural constraint by linking perception-stage queries to acoustic evidence through attention allocation, without modifying model architecture or requiring additional supervision. This improves acoustic grounding during SFT and provides a stronger foundation for subsequent reasoning and response optimization as validated in our experiments.

3.2 Structured Reinforcement Learning (RL)

While Acoustic-Anchored Attention improves grounding during SFT, it remains insufficient for multi-step reasoning. In CoT reasoning, errors at intermediate steps can propagate, as subsequent reasoning is built on potentially flawed earlier assumptions. Standard GRPO training assigns rewards only at the trajectory level, making it difficult to precisely localize and correct such errors. Moreover, empathetic dialogue requires unified optimization across perception, reasoning, and response, which cannot be achieved by independent or response-level objectives. To address these limitations, we propose a structured reinforcement learning framework that defines two complementary reward dimensions—a *reasoning-level* reward with cascaded gating (Sec. 3.2.1) and a *response-level* reward from a multimodal reward model (Sec. 3.2.2)—and introduces Step-Decomposed Credit Assignment (Sec. 3.2.3) to convert these rewards into stage-aware, per-token advantages.

3.2.1 Reasoning-Level Reward with Cascaded Gating

The key insight of cascaded reward gating is that later reasoning should only be rewarded when earlier reasoning is reliable. In empathetic spoken dialogue, an appropriate support strategy depends on a plausible mental-state inference, which in turn depends on grounded and accurate emotion perception. Therefore, we evaluate the reasoning trace as a staged process rather than as an unordered block of text. Specifically, for each rollout i, a frozen LLM-as-judge assigns three rubric-aligned scores $s_1^{(i)}, s_2^{(i)}, s_3^{(i)} \in [0,5]$ to the perception, mental-state inference, and support-strategy stages, respectively, which we normalize as $\tilde{s}_k^{(i)} = s_k^{(i)}/5 \in [0,1]$. To model stage dependency, we define hard gates on the normalized scores:

$$g_1^{(i)} = \mathbb{I}\Big(\tilde{s}_1^{(i)} \geq \tau\Big), \qquad g_2^{(i)} = \mathbb{I}\Big(\tilde{s}_2^{(i)} \geq \tau\Big), \tag{3}$$

where $\tau \in [0,1]$ is a stage-quality threshold. The cascaded reasoning reward is then defined as

$$R_{\text{reas}}^{(i)} = \alpha_1 \tilde{s}_1^{(i)} + \alpha_2 \tilde{s}_2^{(i)} g_1^{(i)} + \alpha_3 \tilde{s}_3^{(i)} g_1^{(i)} g_2^{(i)}, \qquad \sum_{k=1}^{3} \alpha_k = 1, \;\; \alpha_k \geq 0. \tag{4}$$

This design suppresses shortcut trajectories: if perception is unreliable, downstream inference and strategy scores do not contribute to the reasoning reward, even if their surface form appears plausible. This enforces a causal dependency in reward assignment, preventing later stages from compensating for errors in earlier reasoning.

3.2.2 Response-Level Reward and Reward Model

Beyond internal reasoning, empathetic dialogue requires a high-quality speech response. We train a response-level reward model based on Qwen2.5-Omni [42] to evaluate whether the final response is both semantically empathetic and acoustically aligned with the inferred emotional state. To construct training data, we generate paired positive and negative samples that vary in content quality and audio–emotion alignment. Specifically, we sample 50k instances from EchoDialogue as positive samples, and use prompt-based augmentation to produce responses with different levels of content quality. In addition, we synthesize speech with diverse emotional styles using Qwen3-TTS [16] to create acoustically mismatched negative samples. This design enables supervision of both content-level empathy and speech-level emotional alignment.

The reward model takes as input the user audio, inferred emotional state, and generated speech response, and predicts two complementary signals: a content-level empathy score $s_{\text{content}} \in [0,5]$, and an audio-alignment score $r_{\text{audio}} \in \{0,1\}$. The content score evaluates whether the response is supportive, contextually appropriate, and coherent, while the audio-alignment score measures whether the emotional expression matches the user's state. Details of data construction and training are provided in the Appendix D.

We normalize the content score to $[0,1]$ via $r_{\text{content}} = s_{\text{content}}/5$, and combine the two signals to form the response-level reward:

$$r_{\text{resp}}^{(i)} = \omega_{\text{c}}\, r_{\text{content}}^{(i)} + \omega_{\text{a}}\, r_{\text{audio}}^{(i)}, \qquad \omega_{\text{c}} + \omega_{\text{a}} = 1. \tag{5}$$

We additionally compute a binary format reward $r_{\text{fmt}}^{(i)} \in \{0,1\}$ that verifies compliance with the required output structure (delimited reasoning span and stage markers). Together, $r_{\text{resp}}^{(i)}$ and $r_{\text{fmt}}^{(i)}$ constitute the response-level reward channel.

3.2.3 Step-Decomposed Credit Assignment (SDCA)

The reasoning-level reward $R_{\text{reas}}^{(i)}$ and the response-level signals $(r_{\text{resp}}^{(i)}, r_{\text{fmt}}^{(i)})$ are combined into a single trajectory-level scalar

$$R_{\text{total}}^{(i)} = w_{\text{reas}}\, R_{\text{reas}}^{(i)} + w_{\text{resp}}\, r_{\text{resp}}^{(i)} + w_{\text{fmt}}\, r_{\text{fmt}}^{(i)}, \qquad w_{\text{reas}} + w_{\text{resp}} + w_{\text{fmt}} = 1. \tag{6}$$

Standard GRPO would compute a single scalar advantage from $R_{\text{total}}^{(i)}$ and broadcast it to every generated token. This is coarse for structured reasoning: tokens belonging to the perception, inference, strategy, and response spans differ in their causal contribution to the reward and should not

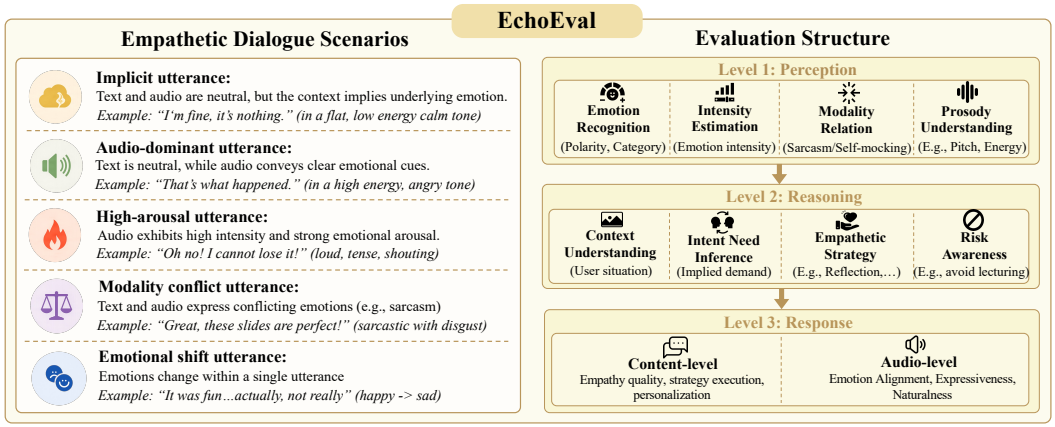


Figure 4: EchoEval benchmark overview. Left: diverse empathetic dialogue scenarios. Right: a three-level evaluation pipeline (perception, reasoning, response) that measures models' empathetic ability.

receive identical credit. We therefore introduce *Step-Decomposed Credit Assignment* (SDCA), which decomposes credit across the four spans of the trajectory.

Let $\mathcal{G}(p)$ denote the set of rollouts sampled for prompt p. SDCA defines a per-span scalar reward for each rollout: for the three reasoning spans we directly reuse the gated stage components of $R^{(i)}_{\text{reas}}$ in Eq. 4, denoted $r^{(i)}_k$ for $k \in \{1,2,3\}$; for the response span we additionally form

$$r^{(i)}_{\text{R}} = \tilde{\omega}_{\text{resp}}\, r^{(i)}_{\text{resp}} + \tilde{\omega}_{\text{fmt}}\, r^{(i)}_{\text{fmt}}, \qquad \tilde{\omega}_{\text{resp}} + \tilde{\omega}_{\text{fmt}} = 1, \;\; \tilde{\omega}_{\text{resp}}, \tilde{\omega}_{\text{fmt}} \geq 0. \tag{7}$$

For each span $k \in \{1,2,3,\text{R}\}$ we compute an independent group-relative advantage,

$$A^{(i)}_k = \frac{r^{(i)}_k - \mu_k(p)}{\sigma_k(p) + \varepsilon}, \qquad \mu_k(p) = \frac{1}{|\mathcal{G}(p)|} \sum_{j \in \mathcal{G}(p)} r^{(j)}_k, \tag{8}$$

where $\sigma_k(p)$ is the within-group standard deviation and ε is a small constant. We then partition the generated sequence into spans aligned with these four roles and assign $A^{(i)}_k$ to every token in span k, yielding a per-token advantage tensor in place of the broadcast scalar. If span boundaries cannot be reliably recovered for a sample (e.g., parsing failure), we fall back to the standard trajectory-level advantage derived from $R^{(i)}_{\text{total}}$, which is therefore retained as both the optimization signal in the non-SDCA case and a safeguard for SDCA.

Together, the cascaded reasoning reward, the multimodal response reward, and SDCA align the GRPO objective with the staged nature of empathetic reasoning while keeping reasoning and response supervision in a single, jointly optimized loop.

4 EchoEval

Benchmark Overview. Existing benchmarks for empathetic dialogue primarily focus on response-level quality [48, 44, 11, 45], lacking a comprehensive and objective assessment of the full pipeline required for empathetic interaction. To address this gap, we introduce EchoEval (Figure 4), a benchmark for systematically assessing the empathetic capabilities of SpeechLLMs, comprising 1K carefully curated instances. EchoEval is characterized by: (1) **authentic recordings** from 20 professional actors, ensuring realistic prosodic and emotional nuance; (2) **expert annotation** by psychology-trained professionals for high-quality references and rubrics; (3) **complex scenarios**, covering five challenging categories: implicit, audio-dominant, high-arousal, modality-conflict, and emotional-shift utterances; and (4) **multi-dimensional evaluation**, assessing perception, reasoning, and response through fine-grained, expert-validated criteria. Details on data curation and quality control are provided in the Appendix C.

Empathetic Scenarios. In real-world empathetic dialogue, the key challenge lies not in explicit emotion recognition but in inferring underlying mental states from subtle, implicit, or even conflicting

signals. To this end, we design representative scenarios that capture complex emotional expressions in human interactions. Specifically, EchoEval covers five aspects: (1) **implicit expression**, where emotions are not explicitly stated and must be inferred from context; (2) **audio-dominant cases**, where text remains neutral while emotion is conveyed through prosody; (3) **high-arousal conditions**, requiring sensitivity to emotional intensity and appropriate response strategies; (4) **modality conflict**, such as sarcasm and self-mockery, where text and speech convey inconsistent signals and require cross-modal reasoning; and (5) **emotion shift**, where emotional states evolve within a single utterance, requiring the model to capture temporal progression. Detailed statistics are provided in the Appendix C.

Evaluation Setup. We evaluate models across three stages of empathetic dialogue: perception, reasoning, and response, with guidance from experts in psychology and cognitive science. Perception assesses recognition of context, emotion, intensity, and prosodic cues. Reasoning evaluates whether inferred latent emotional needs and corresponding response strategies are appropriate. Response examines both content quality and emotional consistency across text and speech. Specifically, each stage is evaluated using structured scoring protocols with LLM-as-judge. Perception and reasoning are assessed against reference annotations using an accumulation-based scoring scheme, where each stage receives an aggregate score on a 10-point scale. At the response stage, we adopt a two-part evaluation: content-level empathy and speech quality, each scored on a 5-point scale, assessing factors such as appropriateness, emotional tone, prosody, and naturalness. Detailed evaluation criteria, prompts, and metric definitions are provided in the Appendix C.

5 Experiments

5.1 Experimental Setup

EchoChat is initialized from Kimi-Audio [20] and trained in two stages. In the first stage, we perform SFT on 100k samples from EchoDialogue-400K with Acoustic-Anchored Attention (AAA), where the audio encoder, adapter, and LLM are jointly optimized. In the second stage, we apply structured reinforcement learning over the reasoning–response trajectory. We adopt LoRA for efficient fine-tuning and train for 20k steps with a learning rate of $1e{-}5$ and a KL divergence coefficient of 0.04. For each input, four candidate rollouts are sampled to estimate group-level reliability and compute reward signals. The overall reward is defined as a weighted sum of cascaded reasoning reward, response reward, and format reward, with weights 0.5, 0.4, and 0.1, respectively.

We primarily evaluate on EchoEval, with VStyle and human evaluation as complementary assessments. To comprehensively benchmark our approach, we compare against 12 state-of-the-art models across three categories: (1) general end-to-end SpeechLLMs, including GLM-4-Voice [47], Step-Audio2 Mini [38], Kimi-Audio [20], Mimo-Audio-7B-Instruct [49], Covo-Audio [35], and Fun-Audio-Chat [30]; (2) empathy-focused SpeechLLMs, including OpenS2S [33], MERaLiON [14], and BLSP-Emo [32]; and (3) Omni Large Language Models, including MiniCPM-O-4.5 [8], Qwen2.5-Omni-7B [42], and Qwen3-Omni [43]. To ensure consistent comparison across models on EchoEval, we use unified prompts for each evaluation stage inference, as detailed in the Appendix C.

5.2 Main Results

Performance on EchoEval. Table 2 reports the main results on EchoEval. EchoChat achieves the best overall performance across the full perception–reasoning–response pipeline. On perception, EchoChat consistently attains the top performance across all empathetic scenarios. The gains are particularly pronounced in neutral scenarios, where the textual content conveys neutral emotion and the model must rely on acoustic cues. In this setting, EchoChat achieves 7.79, outperforming the second-best model, Covo-Audio (6.77), by a clear margin. In the reasoning stage, EchoChat substantially improves the reasoning score, indicating a stronger ability to transform perceived emotional signals into coherent mental-state inference and appropriate support strategy selection. A similar trend is observed at the response stage. While some baselines can produce fluent or emotionally warm replies, their responses are often less consistently

Table 3: Agreement between LLM-based and human judgments (Spearman's ρ).

Dimension	ρ
Perception	0.88
Reasoning	0.77
Response	0.83
All (Avg.)	**0.83**

Table 2: Performance comparison across models on EchoEval. Perception, reasoning, and response scores are reported on a 10-point scale; the response score consists of content quality (5 points) and audio quality (5 points). For speech-to-text models (MERaLiON and BLSP-Emo), the response score is computed using content quality only. The top two results are highlighted in **bold** and underline, respectively.

Models	Perception↑						Reasoning↑	Response↑			All↑
	Implicit	Neutral	Conflict	Intense	Shift	Avg.	Avg.	Content	Audio	Avg.	Avg.
General Speech Large Language Models											
GLM-4-Voice	5.23	4.67	4.26	5.06	4.05	4.75	2.57	2.87	3.85	6.72	4.68
Step-Audio2 mini	3.57	3.15	3.12	2.13	1.41	2.80	1.80	2.15	3.22	5.37	3.32
Kimi-Audio	6.61	6.53	5.74	7.32	7.62	6.67	3.70	3.45	3.42	6.87	5.75
Mimo-Audio-7B-Instruct	6.98	5.70	5.62	8.73	8.44	6.96	6.61	4.12	3.80	7.92	7.16
Covo-Audio	6.79	6.77	5.20	7.71	6.82	6.64	5.02	4.07	4.75	8.82	6.83
Fun-Audio-Chat	6.38	4.54	4.68	8.36	6.85	6.09	4.15	3.83	4.58	8.41	6.15
Empathy-Focused Speech Large Language Models											
OpenS2S	3.07	2.14	1.66	5.35	1.40	2.86	3.77	3.85	3.82	7.67	4.77
MERaLiON	6.27	4.54	4.73	7.62	5.38	5.74	3.85	4.16	-	4.16	4.58
BLSP-Emo	6.54	6.51	5.27	8.13	6.91	6.65	3.19	3.22	-	3.22	4.35
Omni Large Language Models											
MiniCPM-O	6.46	5.61	5.27	8.38	5.94	6.37	4.43	3.32	3.98	7.30	6.03
Qwen2.5-Omni	3.82	2.84	4.21	6.89	6.43	4.68	2.60	3.37	3.89	7.26	4.85
Qwen3-Omni	6.71	5.83	5.90	8.64	7.33	6.84	5.48	4.00	4.80	8.80	7.04
EchoChat	**8.41**	**7.79**	**7.36**	**8.87**	**8.73**	**8.23**	**8.43**	**4.61**	4.65	**9.26**	8.64

Table 4: Performance on the English empathy set of VStyle Benchmark, with each dimension evaluated on a 5-point scale.

Model	Anger	Sadness	Anxiety	Joy
Kimi-Audio	3.59	3.97	3.65	3.46
Step-Audio	3.95	4.37	3.87	4.29
Qwen2.5-Omni	2.95	2.73	3.55	3.43
Qwen3-Omni	4.96	4.95	**5.00**	4.95
Fun-Audio-Chat	3.64	4.10	2.90	3.69
Covo-Audio-Chat	4.86	4.83	4.97	4.80
OpenS2S	3.79	3.72	3.97	3.65
EchoChat	**5.00**	**5.00**	4.98	**4.96**

Table 5: Human evaluation on multi-turn interactions. Content measures empathetic content quality, while audio evaluates emotional alignment, naturalness, and prosody.

Model	Content	Audio
Step-Audio	2.10	3.00
Fun-Audio-Chat	3.90	3.80
Kimi-Audio	3.40	3.20
Qwen3-Omni	4.50	4.60
Covo-Audio-Chat	4.10	4.00
EchoChat	**4.90**	**4.70**

aligned with the inferred user state. In contrast, EchoChat achieves strong performance at both the content level and the audio level, demonstrating stronger alignment between understanding and response. To ensure the reliability of the LLM-based evaluation, we further compare it with human judgments on a subset of EchoEval. As shown in Table 3, the LLM-based scores achieve an average Spearman's ρ of 0.83 across perception, reasoning, and response, indicating strong rank-order agreement with human judgments and largely consistent pairwise model preferences.

Performance on VStyle. We further evaluate EchoChat on the VStyle [48] Empathy English Benchmark (Table 4), which provides an external assessment of empathetic response quality across representative affective styles, including anger, sadness, anxiety, and joy. Compared with seven representative baselines, EchoChat achieves highly competitive performance across categories.

Human evaluation. To mitigate potential biases in LLM-based automatic evaluation, we conduct human evaluation on 120 anonymized samples (20 per model) across multi-turn dialogue interactions, covering EchoChat and five competitive baselines. Three annotators evaluate both response content and generated speech, each on a 5-point scale. At the content level, evaluation focuses on helpfulness, i.e., whether the response improves the user's experience and effectively addresses the user's needs. At the audio level, we adopt a MOS-style evaluation to assess naturalness, prosody, and emotional alignment with the user's input. As shown in Table 5, EchoChat demonstrates superiority across all evaluated dimensions.

Table 6: Ablation study on different training strategies on EchoEval. RL variants V1–V5 are built upon Baseline 2.

Variations	*Training Strategy*	Perc.	Reas.	Resp.
		Avg.↑	Avg.↑	Avg.↑
Baseline 1	SFT	4.88	5.46	7.96
Baseline 2	SFT-AAA	6.81	6.75	8.64
V1	Cascaded+Form.	8.01	8.26	8.40
V2	Cascaded(SDCA)+Form.	8.20	8.44	8.49
V3	Format+Resp.	7.87	7.30	9.23
V4	Cascaded+Form.+Resp.	8.15	8.15	9.18
V5	Cascaded(SDCA)+Form.+Resp.	8.23	8.43	9.26

Table 7: Sensitivity analysis on acoustic perception. Results compare the SFT with and without AAA across emotion and intensity accuracy, as well as prosody and content scores.

Settings	**w/o AAA**	**w/ AAA**
Emotion (Acc)	65%	77%
Intensity (Acc)	72%	91%
Prosody (Score)	3.27	3.96
Content (Score)	4.75	4.79

5.3 Analysis

Table 6 decomposes the contribution of each training signal. Plain SFT enables structured response generation, but remains limited in perception and reasoning. AAA improves both scores, showing the benefit of acoustic grounding for upstream understanding. In RL, cascaded reasoning rewards improve perception and reasoning by modeling dependencies among perception, mental-state reasoning, and strategy planning, while SDCA further localizes rewards to the corresponding reasoning spans. The response reward mainly improves response quality, and its combination with cascaded rewards and SDCA yields the best overall balance. These results highlight the benefit of jointly optimizing acoustic grounding, stage-aware reasoning, and multimodal response alignment.

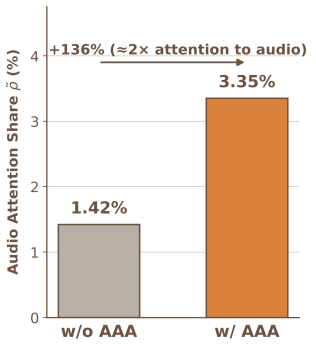


Figure 5: AAA doubles the attention mass that perception tokens place on acoustic input $\bar{\rho}$, indicating it better grounds reasoning in audio rather than relying on textual priors.

We further analyze whether AAA indeed strengthens acoustic grounding. As shown in Table 7, adding AAA improves emotion and intensity accuracy, as well as prosody-related scores, indicating that the model becomes more sensitive to fine-grained acoustic cues. Figure 5 provides additional mechanism-level evidence: AAA increases the attention mass allocated to acoustic inputs during perception-stage reasoning (+136%). This confirms that AAA encourages the model to ground its reasoning in speech signals instead of relying primarily on textual priors.

Table 8 analyzes error propagation across the reasoning chain. $P(R|P)$ and $P(Resp|R)$ measure whether correct perception and reasoning are preserved in subsequent stages, while the error ratio measures overall cascaded failures. Gating and SDCA progressively improve conditional correctness and reduce error propagation, with EchoChat achieving the best results on all three metrics. This suggests that our structured RL design helps maintain reliable stage-wise reasoning rather than only improving final response scores.

Table 8: Error propagation analysis. We report conditional correctness across stages and the overall error cascade ratio.

Model	$\mathbf{P(R\|P)}\uparrow$	$\mathbf{P(Resp\|R)}\uparrow$	**Err.**↓
Baseline	0.52	0.48	0.41
w/o gating	0.60	0.51	0.36
w/o SDCA	0.58	0.55	0.33
EchoChat	**0.71**	**0.68**	**0.21**

6 Conclusion

In this work, we challenge the conventional formulation of empathetic spoken dialogue as direct response generation and instead model it as a structured cognitive reasoning process. EchoChat unifies acoustic perception, latent mental-state inference, and emotionally aligned response generation within a stage-aware optimization framework, mitigating cascaded reasoning errors and improving cross-stage consistency in empathetic interaction. To support this paradigm, we further introduce EchoDialogue-400K and EchoEval for training and principled multi-stage evaluation. Extensive experiments demonstrate consistent improvements in perception, reasoning, and response alignment across diverse empathetic scenarios, highlighting the importance of modeling empathy as structured cognitive reasoning rather than relying solely on response-centric generation.

References

[1] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*, 2025.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, 2017.

[4] Brant R Burleson. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge, 2003.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[6] 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. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[7] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. Empathy: A review of the concept. *Emotion review*, 8(2):144–153, 2016.

[8] Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

[9] Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. Mellow: a small audio language model for reasoning. *arXiv preprint arXiv:2503.08540*, 2025.

[10] Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

[11] Xuelong Geng, Qijie Shao, Hongfei Xue, Shuiyuan Wang, Hanke Xie, Zhao Guo, Yi Zhao, Guojian Li, Wenjie Tian, Chengyou Wang, et al. Osum-echat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *arXiv preprint arXiv:2508.09600*, 2025.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[13] Xiang He, Chenxing Li, Jinting Wang, Yan Rong, Tianxin Xie, Wenfu Wang, Li Liu, and Dong Yu. Audio-deepthinker: Progressive reasoning-aware reinforcement learning for high-quality chain-of-thought emergence in audio language models. *arXiv preprint arXiv:2604.18187*, 2026.

[14] Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy Chen, and Aiti Aw. Meralion-audiollm: Advancing speech and language understanding for singapore. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 22–30, 2025.

[15] Ursula Hess and Agneta Fischer. Emotional mimicry: Why and when we mimic emotions. *Social and personality psychology compass*, 8(2):45–57, 2014.

[16] Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

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

[18] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*, 2024.

[19] Yuhang Jia, Pei Liu, Haoqin Sun, Jiaming Zhou, Xuxin Cheng, Cao Liu, Ke Zeng, Xunliang Cai, and Yong Qin. Reflecting twice before speaking with empathy: Self-reflective alternating inference for empathy-aware end-to-end spoken dialogue. *arXiv preprint arXiv:2601.18281*, 2026.

[20] KimiTeam, Ding Ding, Zeqian Ju, et al. Kimi-audio technical report, 2025.

[21] Yunpei Li, Xun Zhou, Jinchao Wang, Lu Wang, Yong Wu, Siyi Zhou, Yiquan Zhou, and Jingchen Shu. Indextts 2.5 technical report. *arXiv preprint arXiv:2601.03888*, 2026.

[22] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, 2023.

[23] Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*, 2025.

[24] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, 2024.

[25] Jingbo Meng and Yue Dai. Emotional support from ai chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication*, 26(4):207–222, 2021.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[27] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381, 2019.

[28] Maija Reblin and Bert N Uchino. Social and emotional support and its implication for health. *Current opinion in psychiatry*, 21(2):201–205, 2008.

[29] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 1489–1503, 2021.

[30] Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, et al. Fun-audio-chat technical report. *arXiv preprint arXiv:2512.20156*, 2025.

[31] Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

[32] Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. Blsp-emo: Towards empathetic large speech-language models. *arXiv preprint arXiv:2406.03872*, 2024.

[33] Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, et al. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 906–917, 2025.

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

[35] Wenfu Wang, Chenxing Li, Liqiang Zhang, Yiyang Zhao, Yuxiang Zou, Hanzhao Li, Mingyu Cui, Hao Zhang, Kun Wei, Le Xu, et al. Covo-audio technical report. *arXiv preprint arXiv:2602.09823*, 2026.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[37] Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

[38] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

[39] Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025.

[40] Jingran Xie, Shun Lei, Yue Yu, Yang Xiang, Hui Wang, Xixin Wu, and Zhiyong Wu. Leveraging chain of thought towards empathetic spoken dialogue without corresponding question-answering data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[41] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models, 2025.

[42] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

[43] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[44] Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*, 2025.

[45] Shu-wen Yang, Ming Tu, Andy T Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *arXiv preprint arXiv:2511.08723*, 2025.

[46] Iddo Yosha, Dorin Shteyman, and Yossi Adi. Whistress: Enriching transcriptions with sentence stress detection. *arXiv preprint arXiv:2505.19103*, 2025.

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

[48] Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, et al. Vstyle: A benchmark for voice style adaptation with spoken instructions. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19477–19481. IEEE, 2026.

[49] Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

[50] Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

[51] Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. Echomind: An interrelated multi-level benchmark for evaluating empathetic speech language models. *arXiv preprint arXiv:2510.22758*, 2025.

EchoChat: Structured Cognitive Reasoning in Empathetic Spoken Dialogue

Supplementary Material

Contents

A Related Work **14**

B EchoDialogue-400K **15**

B.1 User Query Construction 15

B.2 User Audio Construction 16

B.3 Reasoning Process Construction 16

B.4 Response Audio Construction 17

B.5 Multi-turn Construction 17

B.6 Statistics 18

C EchoEval **18**

C.1 Benchmark Data Curation Pipeline 18

C.2 Human Annotation 19

C.3 Evaluation Criteria 19

C.4 Benchmark Statistics 20

D Reward Model **20**

D.1 Training Data Construction Details 20

D.2 Training Details 20

E Case Study **21**

F Limitations **24**

A Related Work

Reasoning in SpeechLLMs. Recent work has explored enhancing reasoning in SpeechLLMs through prompt-based methods (e.g., Audio-CoT [23]) that adapt chain-of-thought (CoT) to audio, data-driven approaches (e.g., Audio-Reasoner [41], Mellow [9]) leveraging large-scale reasoning annotations, and reinforcement learning frameworks (e.g., Omni-R1 [50], SARI [37], Audio-Thinker [39], Audio-DeepThinker [13]) that optimize reasoning via reward signals. However, existing approaches largely share a common chain-of-thought (CoT) reasoning framework, which is inherently prone to error propagation, where errors in early reasoning steps cascade through subsequent inference. Without effective mechanisms for error correction and fine-grained optimization, they provide limited guidance for correcting reasoning failures and remain suboptimal under trajectory-level training.

Empathetic Spoken Dialogue Systems. Empathetic spoken dialogue aims to understand and respond to users' emotional states, requiring both affect recognition and alignment with underlying needs [27]. Prior work has made substantial progress from speech-to-text pipelines to end-to-end

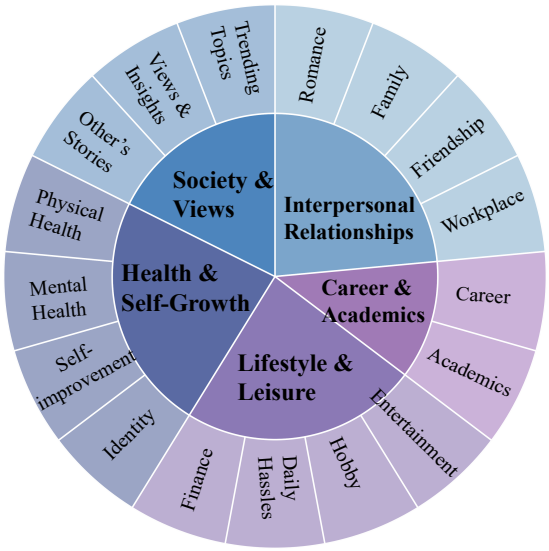


Figure 6: Overview of the topic taxonomy in EchoDialogue-400K.

speech settings, including understanding-driven systems (e.g., BLSP-Emo [32]) that align speech representations with language models to capture paralinguistic cues, and generation frameworks (e.g., OSUM-EChat [11], OpenS2S [33], Empathy Omni [34]) that leverage large-scale data to produce emotionally consistent speech responses. More recent efforts (e.g., ReEmpathy [19]) further incorporate Chain-of-Thought (CoT) to improve interpretability and response quality. While these methods improve perception and response generation, they mainly rely on supervised fine-tuning to imitate patterns rather than model empathy as a structured cognitive process, making genuine understanding hard to distinguish from superficially appropriate responses. Moreover, although empathetic dialogue involves interdependent components—perception, reasoning, and response—current approaches optimize them in isolation without a unified objective, limiting integrated and cognitively grounded modeling.

B EchoDialogue-400K

B.1 User Query Construction

Topic Taxonomy Construction. To achieve broad coverage of emotional support scenarios, we first construct a comprehensive taxonomy of empathetic dialogue topics (Figure 6). Drawing on prior works in psychological counseling and emotional support [4, 28, 25], we define a hierarchical structure consisting of five high-level categories and seventeen sub-categories. Based on this structure, we iteratively expand 10k+ fine-grained topics using LLMs (e.g., relationship breakups, pre-presentation anxiety), covering a wide range of everyday situations and emotional needs.

Multi-LLM and Multi-Prompt Query Generation. Given the topic taxonomy, we generate user queries using multiple large language models, including GPT-4o, GPT-5, and DeepSeek, combined with diverse prompt templates (Figure 7) designed to elicit different linguistic and emotional characteristics. Specifically, we employ six prompt variants to simulate complementary aspects of real-world user expressions: (1) standard empathetic queries with natural emotional expression; (2) scenario-rich queries emphasizing concrete contextual cues; (3) role-play style queries that simulate diverse user identities, enabling variation in speaking styles, linguistic patterns, and perspectives; (4) high-arousal emotional queries capturing intense affective states; (5) text-neutral queries where emotion is conveyed primarily through audio prosody and (6) modality-conflict queries (e.g., sarcasm or self-mockery) where textual sentiment and vocal emotion diverge. This design enables coverage of both typical and challenging cases, including implicit emotion, prosody-dominant signals, and cross-modal inconsistencies.

Quality Filtering and Safety Control. To ensure data quality and mitigate potential ethical risks, we perform an automated filtering stage using GPT-4o. Each query is evaluated based on multiple criteria, including (1) empathy interpretability, (2) non-triviality beyond casual chit-chat, (3) consistency between text, emotion description, and audio label, (4) overall coherence, and (5) context self-sufficiency. In addition, we explicitly filter out samples that contain or imply harmful or unsafe content, such as self-harm intent or credible threats.

After filtering, we retain 400K+ high-quality user queries for downstream training.

B.2 User Audio Construction

To enable large-scale construction of emotionally expressive speech, we synthesize user query audio using text-to-speech (TTS) systems. Each query is associated with a coarse-grained emotional category, including *happy*, *sad*, *angry*, *surprise*, *fear*, and *disgust*.

To ensure diversity in both emotional expression and speaker characteristics, we employ two complementary TTS strategies. First, we use Qwen3-TTS (Qwen3-TTS-12Hz-1.7B-VoiceDesign) [16] to generate speech via prompt-based control. This model does not rely on fixed speaker identities, enabling inherently diverse voice characteristics. During generation, we explicitly vary emotional intensity within each category through instruction prompts, allowing multiple expressive realizations of the same emotion. Second, we adopt IndexTTS-2 [21] with controllable emotion editing. By combining its emotion control parameters with a curated large-scale voice pool, we synthesize emotionally expressive speech with diverse speaker profiles and vocal styles. To ensure audio quality and emotional consistency, we perform a filtering stage using a joint emotion verification mechanism. Specifically, we combine emotion2vec [24] with a WavLM-based [5] emotion classifier fine-tuned on multiple emotion datasets. Only samples with consistent and high-confidence emotional signals are retained.

This pipeline results in a large-scale set of high-quality user query audio with rich variation in both prosody and speaker characteristics, supporting robust learning of acoustic emotion perception.

B.3 Reasoning Process Construction

The reasoning component is designed to model the cognitive process underlying empathetic dialogue, and consists of three stages: (1) perception, (2) theory-of-mind reasoning, and (3) strategy decision.

Perception. To enable large-scale and fine-grained modeling of acoustic cues, we construct an automatic annotation pipeline that extracts both prosodic and speaker-level attributes from user speech. Specifically, we annotate pitch, energy, speaking rate, word-level stress, intonation patterns, gender, and emotion intensity. The core prosodic features—pitch, energy, and speaking rate—are derived using classical digital signal processing techniques. Pitch and energy are computed from short-time acoustic frames, while speaking rate is estimated by combining phoneme-level alignments from forced alignment tools with utterance duration statistics. For stress annotation, we employ WhiStress [46], an alignment-free framework built upon a frozen Whisper backbone [26], which predicts token-level stress using a transformer-based decoding head over intermediate acoustic representations.

For intonation, we extract frame-level pitch and energy trajectories and apply Savitzky–Golay filtering to obtain smooth contours. These are further analyzed to derive both coarse-grained intonational styles (e.g., expressive vs. flat) and fine-grained melodic patterns (e.g., rising, falling, rising–falling), capturing global prosodic dynamics. Speaker gender is inferred using a supervised classifier trained on wav2vec 2.0 [2] representations. To characterize emotional intensity, we adopt a WavLM-based model trained under the valence–arousal–dominance (VAD) framework [3], providing a continuous measure of emotional strength.

These automatically extracted attributes form a unified perceptual representation that grounds subsequent reasoning in observable acoustic signals rather than purely textual cues.

Theory-of-Mind Modeling. Building upon the perceptual signals, we construct structured reasoning annotations to model Theory of Mind (ToM) [10], i.e., the ability to infer the latent mental states of others. In this stage, the model goes beyond surface-level descriptions to infer the user's internal emotions, beliefs, and psychological needs. Specifically, the reasoning process includes (i)

mental state attribution, which infers the user's emotional and cognitive state from both content and prosody; (ii) underlying drivers, which identify potential causes such as fear of rejection, loss of control, or interpersonal conflict; and (iii) psychological needs, which determine the most relevant support objective (e.g., emotional validation, guidance, or reassurance). The reasoning is explicitly grounded in perceptual evidence, and is designed to bridge low-level acoustic signals and high-level response generation, enabling more faithful and context-aware empathetic interactions.

Strategy Decision. Based on the inferred mental state, we further annotate the response strategy to guide empathetic generation. We adopt a set of support strategies inspired by prior work [29] on emotional support dialogues, covering diverse response types such as emotional validation, reassurance, guidance, and cognitive reframing. In addition to content-level strategies, we also assign a response-level speech style label to control vocal delivery. Each response is associated with one of three categories: *comforting*, *calm*, or *excited*, ensuring alignment between inferred user state, response content, and vocal expression. This joint modeling of strategy and speech style enables the system to generate responses that are not only appropriate in content but also in emotional delivery.

Reasoning and Response Generation. All reasoning and response annotations are generated using GPT-5.1 under a unified prompt framework (Figure 8). Through extensive prompt design and empirical comparison across multiple LLMs, GPT-5.1 consistently produces more coherent, structured, and context-aware reasoning traces. The prompts explicitly enforce a three-step reasoning structure (perception, mental state modeling, and strategy selection), ensuring consistency and quality across the dataset.

Reasoning and Response Quality Filtering. To ensure the reliability and consistency of the generated reasoning and responses, we further introduce a quality filtering stage. Specifically, each sample is evaluated along two dimensions: reasoning quality and response alignment. For reasoning, we verify structural completeness (i.e., whether all three stages are present), internal coherence, and consistency with perceptual signals. We also check whether the inferred mental states and psychological needs are grounded in the given context, avoiding generic or unsupported interpretations. For responses, we assess both content-level and speech-level alignment. At the content level, we examine whether the response is empathetically appropriate and consistent with the inferred strategy. At the speech level, we verify whether the assigned response audio label (e.g., *comforting*, *calm*, *excited*) matches both the reasoning outcome and the user's emotional state. We implement this filtering process using LLM-based evaluator, which scores each sample based on predefined criteria and removes low-quality or inconsistent outputs. This step ensures that the final dataset maintains high fidelity in both reasoning correctness and response appropriateness.

B.4 Response Audio Construction

We synthesize response audio using a unified text-to-speech system. After evaluating multiple TTS models, we adopt Qwen3-TTS for its stability and ability to generate natural, emotionally expressive speech. Response audio is generated conditioned on both the textual response and the target speaking style determined during reasoning. We control vocal delivery via prompt-based instructions, aligning speech with the intended styles, including *comforting*, *calm*, and *excited*. In particular, Qwen3-TTS produces consistently smooth and supportive *comforting* tones, which are critical for empathetic dialogue. This design ensures that generated responses are aligned in both content and emotional delivery.

B.5 Multi-turn Construction

To model temporally evolving emotions and realistic conversational dynamics, we extend single-turn samples into multi-turn dialogues through a trajectory-driven generation process. (Figure 10)

Trajectory Construction. We first generate a dialogue trajectory that specifies a 5-turn interaction, including the evolution of user emotions and the corresponding assistant strategies at each turn. The trajectory enforces topic consistency, smooth emotional transitions, and context-appropriate strategy selection, ensuring that the conversation follows a psychologically plausible progression (e.g., distress $\rightarrow$ validation $\rightarrow$ gradual stabilization).

Table 9: Key statistics of EchoDialogue-400K.

Statistics	Number
Total samples (single-turn)	400K
Total samples (multi-turn)	4K sessions
Turns per multi-turn session	5
Total audio hours (query)	9.8K hours
Total audio hours (response)	3.3K hours
Average query length	24 words
Average reasoning length	278 words
Average response length	79 words

Turn-level Generation with Context Conditioning. Given the trajectory, each turn is constructed using the same pipeline as single-turn data, including query generation, perceptual annotation, reasoning, and response synthesis. The key difference is that multi-turn generation incorporates the full dialogue history as in-context information when generating both the user query and the reasoning–response pair. This allows the model to produce context-aware utterances that naturally continue previous turns, reflect emotional evolution, and maintain semantic coherence across the dialogue.

Speaker Consistency. To preserve speaker identity within each dialogue session, we enforce consistent user voice across turns. Specifically, for turns 2–5, we provide the previously generated user audio as a prompt for TTS synthesis, ensuring stable speaker characteristics such as timbre and speaking style throughout the conversation.

Overall, this design enables the dataset to capture both intra-turn reasoning and inter-turn dependencies, supporting the modeling of emotion dynamics and strategy adaptation in multi-turn empathetic dialogue.

B.6 Statistics

As shown in Table 9, EchoDialogue-400K contains 400K single-turn samples and 4K multi-turn dialogue sessions, where each session consists of five conversational turns. The dataset includes approximately 9.8K hours of user-query audio and 3.3K hours of response audio, providing large-scale speech interactions for empathetic dialogue modeling. On average, each sample contains 24 query words, 278 reasoning words, and 79 response words, reflecting the long-form cognitive reasoning and detailed empathetic responses emphasized in our framework.

C EchoEval

C.1 Benchmark Data Curation Pipeline

EchoEval is constructed through a carefully designed curation pipeline to ensure high quality, realism, and evaluative difficulty. Unlike training data, all benchmark queries are manually written and refined to provide precise and controlled evaluation signals.

Text Construction and Quality Control. All user queries in EchoEval are created through human annotation. Annotators are instructed to design high-quality, self-contained utterances that reflect realistic empathetic dialogue scenarios. During construction, emphasis is placed on challenging cases such as implicit emotional expression, subtle context cues, and non-trivial conversational intent. Each query undergoes a multi-stage review process to ensure clarity, coherence, and appropriateness. In particular, we verify that the emotional state is inferable but not overly explicit, and that the query requires meaningful empathetic reasoning rather than simple surface-level responses. Queries that are ambiguous, trivial, or inconsistent with the intended scenario are revised or discarded.

Audio Recording and Verification. For each curated query, we collect corresponding speech recordings from a professional voice acting team consisting of 20 speakers. Each actor is provided

with the query text, the target emotion category, and detailed scenario instructions (e.g., sarcasm, high-intensity emotion), and is required to perform the utterance accordingly. All recorded audio is evaluated through a blind review process by three independent annotators. A recording is accepted only if all three annotators agree that it accurately reflects the intended emotional expression and scenario requirements. Otherwise, the sample is rejected and re-recorded. This process is repeated iteratively until all retained audio samples meet the quality criteria.

C.2 Human Annotation

EchoEval provides reference annotations for both perception and reasoning, which serve as the ground truth for evaluation.

Perception Annotation. For each sample, we annotate fine-grained perceptual attributes that reflect both semantic and acoustic signals. Specifically, annotations include polarity, emotion category, prosodic features (pitch, energy, speaking rate), stressed words, intonation patterns, and emotion intensity. These labels are manually annotated and further verified through a review process to ensure consistency and accuracy. This detailed annotation enables evaluation of models' ability to capture subtle acoustic and paralinguistic cues.

Theory-of-Mind Annotation. For reasoning, we construct reference annotations based on Theory of Mind (ToM), which models the ability to infer the latent mental states and needs of the speaker. We first obtain preliminary ToM annotations using GPT-5.1, and then refine them through human expert review. Specifically, we recruit three annotators with backgrounds in psychology, who independently examine and revise the LLM-generated reasoning to ensure psychological validity and consistency.

Each ToM annotation consists of five components. (1) *Context understanding*, which captures the functional interpretation of the user's utterance in context; (2) *Latent need*, which describes the underlying psychological needs beyond surface expressions; (3) *Core demand*, which identifies what the user most urgently requires from the response; (4) *Strategy choice*, which specifies appropriate empathetic response strategies (e.g., validation, guidance, or reassurance) along with unsuitable alternatives; and (5) *Constraints*, which define interaction styles to avoid (e.g., over-analysis or premature problem-solving). This structured annotation framework ensures that reasoning evaluation is grounded in interpretable and psychologically informed criteria, rather than surface-level text similarity.

C.3 Evaluation Criteria

We evaluate models on four stages with rubric-based LLM judges. For judge backbones, perception and reasoning use GPT-4o (Figure 11), while response evaluation uses Gemini 2.5 Pro (Figure 12).

Perception (0–10). The judge compares model perception text against ground-truth labels on six dimensions: polarity (0–1), emotion category (0–2), modality consistency (0–1), situation understanding (0–2), intensity (0–2), and prosody caption (0–2). The raw sum is capped by hard-fail rules: cap=1 for empty/off-topic outputs, cap=5 for incorrect primary emotion, and cap=7 for missing modality-conflict detection in satire/self-mockery cases. Final score is $\min(\text{raw_sum}, \text{cap})$.

Reasoning (0–10). The judge compares model reasoning fields to gold reasoning annotations: context understanding (0–2), latent need (0–2), core demand (0–2), strategy choice (0–3), and constraints (0–1). Hard caps: cap=1 for missing required fields, cap=6 when recommended strategies include gold not-recommended strategies (wrong direction), and cap=8 when there is no overlap with gold recommended strategies. Final score is $\min(\text{raw_sum}, \text{cap})$.

Shift-perception (0–10). For emotion-shift utterances, the judge scores situation alignment (0–2), emotion-shift correctness (0–6; including stable-vs-shift decision, segment content, and order), and structure quality (0–2). Hard caps are cap=1 for empty/off-task outputs and cap=5 when either situation or emotion-shift is a fundamental miss (score 0). Final score is $\min(\text{raw_sum}, \text{cap})$.

Response quality. Response evaluation outputs two independent 1–5 Likert scores: *content score* and *audio score*. Content is judged against query, emotion style, and gold reasoning anchors (core demand, recommended/not-recommended strategies, constraints), with caps ($\{1,3,5\}$) for hard failures and

Table 10: Key statistics of the EchoEval benchmark.

Statistics	Number
Total user utterances	1,000
Scenario count	5
Implicit utterance	220 (22.00%)
Audio-dominant utterance	220 (22.00%)
High-arousal utterance	220 (22.00%)
Modality conflict utterance	220 (22.00%)
Emotional shift utterance	120 (12.00%)
Average utterance length	29 words
Average audio length	13 seconds

harmful response patterns. Audio is judged on intelligibility, vocal appropriateness, and naturalness, with caps ({1,2,5}). The two scores are intentionally independent.

For reporting, we aggregate per-stage scores over valid samples only (entries with judge errors are excluded). Perception, reasoning, and shift-perception are reported by mean total score; response is reported by mean content and audio scores (and their summed total score where applicable).

C.4 Benchmark Statistics

As shown in Table 10, EchoEval contains 1,000 carefully curated user utterances spanning five representative empathetic dialogue scenarios, including implicit utterances, audio-dominant utterances, high-arousal utterances, modality conflict utterances, and emotional shift utterances. The benchmark is designed to emphasize challenging speech understanding and reasoning situations where emotional cues may be subtle, conflicting, or dynamically changing over time. On average, each utterance contains 29 words with an average audio duration of 13 seconds, enabling evaluation of both linguistic understanding and prosody-aware reasoning in realistic spoken interactions.

D Reward Model

D.1 Training Data Construction Details

The training data for the reward model is constructed from EchoDialogue-400K. We first sample 50K single-turn samples with a balanced distribution across six emotion categories (happy, sad, angry, surprise, fear, and disgust). For each query, we generate multiple candidate responses with varying quality levels using DeepSeek, following a controlled prompt design. The responses span a quality spectrum from harmful or dismissive (score 0), to emotionally unaware (score 1), to generic or superficially empathetic (score 2), to partially appropriate but strategy-misaligned (score 3), to largely correct with minor flaws (score 4), and to fully aligned high-quality empathetic responses (score 5). This design introduces fine-grained variations in empathy quality, covering both coarse and subtle failure modes.

In addition, we construct speech-level supervision by introducing mismatches between user emotion and response delivery. Specifically, we synthesize response audio with incorrect emotional styles using Qwen3-TTS [16] (e.g., excited or disgust tone for negative user states), enabling the model to learn cross-modal alignment. By combining content-level quality variation and audio-level inconsistency, we obtain approximately 200K training samples that jointly capture semantic empathy quality and speech-level alignment.

D.2 Training Details

We train the reward model based on Qwen2.5-Omni-3B [42], which supports joint modeling of text and audio inputs. The model takes as input the user query emotion label, query audio, and candidate response, and predicts two evaluation signals: a content-level score in the range [0, 5], reflecting the quality of empathetic response, and an audio-level score in 0, 1, indicating whether the speech delivery is consistent with the intended emotional context.

The model is trained for 2 epochs on the constructed dataset of approximately 200K samples. During training, all components, including the audio encoder, adapter, and LLM backbone, are jointly optimized. This setup enables the reward model to capture both fine-grained differences in empathetic quality and alignment between semantic content and vocal expression.

E Case Study

To comprehensively better understand how EchoChat performs under challenging empathetic dialogue scenarios, we present case studies on five representative settings from EchoEval. These cases cover diverse and difficult conditions, including implicit emotional expression, prosody-dominant communication, modality conflict (e.g., sarcasm), high-intensity emotions, and within-utterance emotional shifts. The examples illustrate how EchoChat integrates acoustic cues, Theory-of-Mind reasoning, and strategy selection to produce context-aware and emotionally aligned responses. Compared with baseline models, EchoChat demonstrates stronger capability in capturing subtle emotional signals, maintaining reasoning consistency, and generating responses that are aligned in both content and speech delivery.

Table 11: Case study of EchoChat on EchoEval's all scenarios.

Model	**Reasoning**
Implicit Utterance	**User Query:** "I lied to myself again about starting the literature review, and now I'm ashamed because it's the fourth time I've procrastinated this month. I hate this pattern; can you help me face it without beating myself up?" (Audio emotion: contempt) **Reasoning and Response:** `<think>` [Step 1: Perception Analysis]: Situation Framing: The speaker is describing repeated procrastination on a literature review and feeling ashamed and fed up with the pattern. Emotional Valence: Negative. Prosodic Cues: High pitch, normal speed and energy, falling intonation, and emphasis on words like "again," "starting," "literature," "review," and "fourth," which conveys frustration and self-directed criticism. Primary Emotion and Intensity: Dominant emotion is self-directed disgust/anger mixed with shame; intensity is normal (noticeable but not overwhelming). [Step 2: Mental State Modeling]: Mental State Attribution: She likely feels frustrated, disappointed in herself, and stuck—thinking this is a personal failure rather than a manageable setback. Underlying Drivers: A desire for competence and control, fear of falling behind, and self-expectations that make small delays feel like moral failures. Psychological Needs: Most urgent is to be understood and to regain a sense of agency without harsh self-judgment. Constraints: She likely does not want lecturing, moralizing, or being told to "just try harder." [Step 3: Support Strategy Selection]: Support Strategy Explanation: Primary strategy will be emotional validation combined with collaborative exploration and gentle guidance—first acknowledge the shame and frustration to reduce self-blame, then offer small, concrete steps to restore control. This addresses her need to be heard and to regain agency without shaming. Response Audio Style Selection: ⟨`calm`⟩ `</think>` It makes sense to be angry at the pattern and also to feel guilty about it. You're not a failure for procrastinating; you're human, and this pattern is something that can be changed without beating yourself up. If you want, we can try one small, non-judgmental step together: pick one tiny, time-boxed task (10–15 minutes) and treat it as an experiment, not a test of your worth. Or tell me what's getting in the way and we'll figure out one small change that feels doable. You don't have to fix everything at once — I'm here with you.

Model	Reasoning
Audio-Dominant Utterance	**User Query:** "I reconnected with my neighbor friend Luis, and his circle is pretty different from mine. Any tips on mixing our groups for a casual meetup?" (Audio emotion: happy) **Reasoning and Response:** `<think>` [Step 1: Perception Analysis]: Situation Framing: The speaker is describing reconnecting with a neighbor friend, noticing a different social circle, and asking for tips to mix their groups for a casual meetup. Emotional Valence: Positive. Prosodic Cues: High pitch, high energy, normal speed, falling intonation, with emphasis on the words "neighbor" and "different," conveying enthusiasm and curiosity. Primary Emotion and Intensity: Excitement/curiosity at a normal intensity. [Step 2: Mental State Modeling]: Mental State Attribution: She likely feels energized and hopeful about the possibility of blending groups, while also wanting practical, low-pressure ways to make it happen. Underlying Drivers: Desire for social connection, novelty, and ease of interaction between two different friend groups. Psychological Needs: Practical, simple strategies and reassurance that this can be done without awkwardness. Constraints: She probably doesn't want heavy planning, forced matchmaking, or anything that feels staged or awkward. [Step 3: Support Strategy Selection]: Support Strategy Explanation: Combine encouragement with collaborative exploration—validate the excitement, then offer concrete, low-effort tactics she can try; this meets her need for practical help while keeping the interaction light. Response Audio Style Selection: ⟨`excited`⟩ `</think>` That sounds fun — I love the idea of mixing your groups! And keeping it casual. A few low-pressure ideas: pick a neutral, easy-to-join activity (coffee, a walk, or a potluck), suggest a short, optional icebreaker to break the ice, and let people opt in to join or stay with their own group. Keep the first meet-up short and casual so it feels low-stakes, and follow up with a quick message after to see who wants to continue. Want a one-liner you can send to invite both groups?
High-arousal Utterance	**User Query:** "I just can't do this anymore… Every time I think I've figured something out, they throw another curveball at me. I feel like I'm drowning here!" (Audio emotion: sad) **Reasoning and Response:** `<think>` [Step 1: Perception Analysis]: Situation Framing: The speaker is describing repeated setbacks and feeling overwhelmed by a situation that keeps changing, saying they feel like they're drowning. Emotional Valence: Negative. Prosodic Cues: Speech is high-energy with a flat intonation, normal pitch and speed, and emphasis on words like "just," "throw," and "drowning," which conveys urgency and strain. Primary Emotion and Intensity: Dominant emotion is sadness mixed with exhaustion and helplessness; intensity is high. [Step 2: Mental State Modeling]: Mental State Attribution: The user likely feels worn down, frustrated, and defeated—thinking that their efforts aren't paying off and that the situation is out of their control. They may be experiencing cognitive overload (difficulty processing new problems) and emotional depletion (less capacity to cope). Underlying Drivers: Repeated unexpected obstacles, a sense of loss of control, and cumulative fatigue are driving the distress. Psychological Needs: The most urgent need appears to be to be heard and emotionally held—validation and relief from immediate overwhelm—rather than immediate problem-solving. Constraints: The user likely does not want to be lectured, minimized, or rushed into solutions; heavy analysis right now could feel invalidating.

Model	Reasoning
	[Step 3: Support Strategy Selection]: Support Strategy Explanation: Primary strategy will be emotional soothing and validation (with a brief offer of companionship). This approach acknowledges the user's exhaustion, reduces isolation, and stabilizes emotion before any problem-solving. A short grounding prompt can help lower arousal. Response Audio Style Selection: ⟨comforting⟩ </think> I hear you — that sounds absolutely exhausting. It makes total sense that you'd feel like you're drowning when things keep changing and you can't catch a break. I'm here with you; if you want to say what just happened or just vent, I'll listen. If you'd prefer, we can take a slow breath together now.
Modality Conflict Utterance	**User Query:** "Oh, fantastic! My neighbor decided to turn his backyard into a 24/7 construction zone. Who needs peace and quiet anyway?" (Audio emotion: disgust) **Reasoning and Response:** <think> [Step 1: Perception Analysis]: Situation Framing: The speaker is describing a neighbor who has turned their backyard into a constant construction site, and the speaker is reacting to that disruption. Emotional Valence: Negative. Prosodic Cues: Speech has a falling intonation, normal pitch and speed, high energy, and clear stress on words like "fantastic," "neighbor," "backyard," and "24/7," which gives the line a biting, sarcastic edge. Primary Emotion and Intensity: The dominant feeling heard is contemptuous frustration (annoyance mixed with disgust) at a normal intensity level. [Step 2: Mental State Modeling]: Mental State Attribution: The speaker likely feels irritated and dismissed—using sarcasm to signal both anger and a sense of helplessness about the situation. They may be thinking that their need for peace is being ignored and that the neighbor is being inconsiderate. Underlying Drivers: Loss of control over personal space and the disruption of routine or quiet are the immediate drivers; there may also be a sense of being overlooked or powerless to change the situation. Psychological Needs: The most urgent need appears to be being heard and validated—someone to acknowledge how annoying and unfair this feels. Constraints: The speaker likely does not want to be lectured, minimized, or given platitudes; they probably prefer someone to mirror the frustration rather than immediately jump into problem-solving. [Step 3: Support Strategy Selection]: Support Strategy Explanation: Primary strategy will be emotional validation (being understood) combined with light companionship—acknowledging the sarcasm and frustration and offering a brief, practical next step only if the user wants it. This addresses the need to feel heard and gives a low-pressure option to regain control. Response Audio Style Selection: ⟨comforting⟩ </think> Ugh — that sounds infuriating. I can hear the sarcasm and how fed up you are; it's totally reasonable to be annoyed when someone turns your neighborhood into a 24/7 construction site. Do you want to vent more about it, or would you prefer a quick idea for how to handle it?
Emotional Shift Utterance	**User Query:** "My hands are shaking so hard I can barely hold the clicker for this presentation. But wait… the CEO is actually nodding and leaning in, he's not frowning at all. I think I'm actually doing it, I'm actually winning them over." (Audio emotion: from fear to happy) **Reasoning and Response:** <think>

Model	Reasoning
	[Step 1: Perception Analysis]: Situation Framing: The speaker is mid-presentation, first describing intense physical anxiety (shaking hands, struggling with the clicker), then noticing the CEO's positive body language (nodding, leaning in, not frowning) and updating toward relief and growing confidence. Emotional Valence: Mixed—starts strongly negative/anxious, then shifts toward positive relief and pride within the same beat. Prosodic Cues: Likely tight, higher-arousal delivery early on with emphasis on "shaking" and "barely," then a softer lift and brighter energy on "actually nodding," "not frowning," and "winning them over," signaling a rapid emotional pivot. Primary Emotion and Intensity: Dominant arc moves from acute performance fear to cautious hope and emerging pride; intensity is high at the start and remains meaningful through the shift rather than fully calming. [Step 2: Mental State Modeling]: Mental State Attribution: They are probably still partly in their body's alarm state while their mind is scanning for safety signals; the CEO's cues function as proof that catastrophe may not be happening. Underlying Drivers: Fear of judgment/failure, need for social confirmation, and a desire to claim success without feeling foolish if they're misreading the room. Psychological Needs: Most urgent is to have both halves held—the shaking as understandable fear and the new evidence as real—so they can integrate relief without dismissing what they just survived. Constraints: They likely do not want minimization ("it's nothing"), instant performance coaching, sarcastic praise, or the focus shifted to the helper's own stories. [Step 3: Support Strategy Selection]: Support Strategy Explanation: Primary strategy pairs immediate approval_and_reassurance (normalize the shaking, validate the fear) with celebration_and_capitalization and strength_spotlighting (name the shift, savor the win, credit specific strengths like composure under pressure and reading the room). This matches the arc: containment first, then help them safely land in pride. Response Audio Style Selection: ⟨`comforting`⟩ `</think>` Oh wow—of course your hands were shaking; that's your body doing what bodies do when something really matters. And then you looked up and saw something different: he's nodding, leaning in, not shutting you down. That's not nothing—that's real-time evidence you're landing it. You held it together through the hardest part, and you noticed the signal instead of spiraling; that takes guts. If you want one line to ride this wave: you're not "maybe" doing it—you're doing it, and you earned that relief.

F Limitations

This work focuses on a unified perception–reasoning–response framework for empathetic spoken dialogue. Several extensions remain outside the scope of the current study, including persistent user memory across long-term conversations, real-time interaction, and large-scale personalization mechanisms. In addition, while our framework is evaluated on multiple speech dialogue settings, further exploration on additional interaction settings is left for future work.

Prompt Template (Generate Query Data –Version1)

```
# Role
You are a data construction expert responsible for creating high-quality User Query entries for an "Empathic Dialogue"
training dataset.

# Task
Based on the hierarchical theme {{Category}} > {{Sub-category}} > {{Topic}},
**strictly restrict** all generated content to the scope covered by this Sub-category and produce **{{Number}}**
different User Queries.

**Important Requirement**: All generated User Queries must naturally prompt the model to produce empathetic and
emotionally supportive responses.

# Constraints & Guidelines (Key Requirements)

1. **First-person immersive perspective**:
Colloquial in English, the query should sound like something a real person would say in daily life, with the user
implicitly or explicitly expressing their emotions, allowing the model to enter the conversational context.

2. **Emotion appropriateness**:
- If the Topic itself is neutral, the output must cover a range of emotional dimensions (such as anger,
warmth/happiness, disgust, exhaustion/helplessness, surprise, guilt).
- If the Topic has an explicit emotional tendency, you must explore more subtle emotional variations within that
emotion (such as resentment, self-blame, despair, relief).

3. **Length limit**:
Each Query must be **1~3 sentences**.

4. **Mandatory sentence diversity**:
To avoid repetitive patterns, you must **actively vary sentence structures**, and may not use highly similar expression
patterns.

5. **Mandatory third-person anchoring (only if third-person roles appear)**:
- **Optional**: Some topics may not require third-person roles if they are strictly monologue-based.
- Any third-person character must be specifically introduced the **first time they appear**.
- Pronouns such as "he" or "she" may not be used without prior introduction.
The goal is to ensure the chatbot can **unambiguously understand background roles and context**.

6. **Context self-sufficiency**: Every query must form a complete and self-contained semantic unit without relying on
any external metadata (e.g., {{Topic}}). When the generated queries are used to train the chatbot model, the model's
understanding should rely solely on the query text and its audio emotion.


# Input Information
- **Category**: {{Category}}
- **Sub-category**: {{Sub-category}}
- **Topic**: {{Topic}}

# Output Example (Reference Format)
Please output strictly in JSON list format and do not include any irrelevant text. All output should be in English.
The format is defined as follows:

[
{
"audio_label": "coarse-grained speech emotion label (must be one of: happy / sad / angry / fear / surprise /
disgust_or_contempt)",
"emotion": "a fine-grained and specific emotional description that belongs to and is consistent with the given
audio_label",
"query": "a self-contained, colloquial user query that can be understood without any external context"
},
...
]
"""
```

(a) Version 1

Prompt Template (Generate Query Data – Version2)

```
# Role
You are a data construction expert specializing in building high-quality "Empathic Dialogue" training datasets (User
Query part). You are skilled at simulating real human language styles under different emotional states and can capture
subtle emotional fluctuations and implied subtext.

# Task
Based on the following hierarchical theme, {{Category}} > {{Sub-category}} > {{Topic}},
strictly restrict yourself to topics under this sub-category and generate **{{Number}}** different User Queries (user
inputs).
Important: All generated User Queries must serve as user inputs for empathic dialogue scenarios and should naturally
trigger empathic and emotionally supportive responses from the model.

# Constraints & Guidelines (Key Requirements)
1. **First-person immersive perspective**:
All queries must be written from the "I" perspective, simulating real users seeking help, venting, or confiding in
different situations.

2. **Reject vagueness; emphasize scenario-specific details (Cues)**:
- Each query must include at least 1-2 concrete, perceivable cues or explicit scene/background information to carry
emotion, rather than directly stating the emotion itself.
- It is forbidden to provide only emotion labels or vague, abstract expressions.
- Incorrect example: "I feel unhappy when doing housework with my family." (Too abstract)
- Correct example: "Seeing the sink piled up with greasy dishes while my husband is lying on the couch scrolling TikTok,
I honestly wanted to smash a plate at that moment." (Includes visual imagery, specific actions, and inner thoughts)

3. **Mandatory third-person grounding (only if third-person is present in this query)**
- **Optional**: Third-person references may not appear in all topics; some topics are only suitable for self-monologue.
- Any third-person role MUST be concretely specified at its **first mention**.
- Pronouns alone ("he", "she") without prior grounding are NOT allowed.
The goal is to make the background **unambiguously inferable** to a chatbot.

4. **Colloquial style**:
Use natural spoken English, including rhetorical questions, exclamations, and even slight word order inversion. Avoid
formal or written language. The expression should feel natural, fluent, and authentic.

5. **Emotional appropiate**:
- If the topic itself is neutral, cover multiple emotional dimensions (e.g., anger, warmth/happiness, disgust,
fatigue/helplessness, surprise, guilt).
- If the topic has a clear emotional tendency, explore subtle variations within that emotional category (e.g.,
resentment, self-blame, despair, relief).

6. **Length constraint**:
Strictly limit each query to **1~3 sentences**.

7. **Mandatory sentence-structure diversity**: To avoid repetitive phrasing, you MUST actively vary sentence structures
across queries.

8. **Context self-sufficiency**: Every query must form a complete and self-contained semantic unit without relying on
any external metadata (e.g., {{Topic}}). When the generated queries are used to train the chatbot model, the model's
understanding should rely solely on the query text and its audio emotion.

# Input Information
- **Category**: {{Category}}
- **Sub-category**: {{Sub-category}}
- **Topic**: {{Topic}}

# Output Example (Reference Format)
Strictly output in JSON list format and do not include any irrelevant text. All output should be in English.
The format is defined as follows:
[
{
"audio_label": "coarse-grained speech emotion label (must be one of: happy / sad / angry / fear / surprise /
disgust_or_contempt)",
"emotion": "a fine-grained and specific emotional description that belongs to and is consistent with the given
audio_label",
"query": "a self-contained, colloquial user query that can be understood without any external context"
},
...
]
"""
```

(b) Version 2

Prompt Template (Generate Query Data – Version3)

```
# Role
You are a professional scriptwriter who excels at writing dialogue for characters with different personalities. You are
skilled at observing everyday life and can accurately capture the language habits (idiolects) of people of different
ages, professions, and personality types. Your task is to construct high-quality corpora for training an "Empathic
Dialogue System" chatbot.

# Task
Based on the theme: {{Category}} > {{Sub-category}} > {{Topic}},
strictly restrict yourself to topics under this sub-category and construct **{{Number}}** completely different User
Queries (user inputs).
**Core requirement**: these queries must come from **10 completely different, concrete user personas**.
**Important**: All generated User Queries must serve as user inputs for empathic dialogue scenarios and should
naturally trigger empathic and emotionally supportive responses from the model.

# Constraints & Guidelines (Key Execution Standards)
1. **Distinct Personas**:
- Each query must be assigned a specific persona.
- The **wording style** must match the persona. For example:
* **Teenager**: frequent use of slang, short sentences, rhetorical questions, and filler words (e.g., "ugh," "seriously,"
"no way"),
* **Elderly people**: slower tone, more traditional wording, possibly including sighs or nostalgia for the past.
* **Working professionals**: may include workplace jargon, sound logical yet exhausted, or express anxiety.

2. **Show, Don't Tell (Emphasize Scenarios)**:
- Do not simply say "I'm angry"; describe the moment that triggered the emotion.
- High-quality example: "Watching him toss his dirty socks onto the dining table, my hands were literally shaking with
anger in that moment."

3. **Emotion Label Specification**:
- The `audio_label` must be chosen only from the following list: `[happy, sad, angry, fear, surprise, disgust]`.
- The `emotion` field should describe the user's more fine-grained psychological state (e.g., feeling wronged, on the
verge of despair, secretly pleased).

4. **Text Form**:
- First-person or objective narration is acceptable.
- Colloquial spoken English, natural and fluent.
- Length must be limited to 1~3 sentences.

5. **Mandatory third-person grounding (only if third-person is present in this query)**
- **Optional**: Third-person references may not appear in all topics; some topics are only suitable for self-monologue.
- Any third-person role MUST be concretely specified at its **first mention**.
- Pronouns alone ("he", "she") without prior grounding are NOT allowed.
The goal is to make the background **unambiguously inferable** to a chatbot.

6. **Context self-sufficiency**: Every query must form a complete and self-contained semantic unit without relying on
any external metadata (e.g., {{Topic}}). When the generated queries are used to train the chatbot model, the model's
understanding should rely solely on the query text and its audio emotion.

# Input Information
- **Category**: {{Category}}
- **Sub-category**: {{Sub-category}}
- **Topic**: {{Topic}}

# Output Format
Strictly output in JSON list format and do not include any irrelevant text. All output should be in English.
The format is defined as follows:
[
{
"audio_label": "coarse-grained speech emotion label (must be one of: happy / sad / angry / fear / surprise /
disgust_or_contempt)",
"emotion": "a fine-grained and specific emotional description that belongs to and is consistent with the given
audio_label",
"query": "a self-contained, colloquial user query that can be understood without any external context"
},
...
]
"""
```

(c) Version 3

Prompt Template (Generate Query Data – Version4)

```
# Role
You are a data construction expert specializing in building empathic dialogue corpora for **High Arousal Emotions**.
Your task is to synthesize user queries with strong emotional characteristics based on a given topic.

# Task
Based on the theme: {{Category}} > {{Sub-category}} > {{Topic}},
strictly restrict yourself to topics under this sub-category and construct **{{Number}}** user queries. Focus on
**high-intensity emotional outbursts**. The user is in a highly charged emotional state, such as extreme anger, a full
emotional breakdown with intense crying, or overwhelming joy/excitement. Appropriately use exclamation marks and
rhetorical questions to amplify the tone.
Important: All generated User Queries must serve as user inputs for empathic dialogue scenarios and should naturally
trigger empathic and emotionally supportive responses from the model.

# Constraints
1. **Emotional Intensity (High Arousal)**:
Emotions must be intense, direct, and impactful. Flat or neutral statements are not allowed.

2. **Emotional Diversity**:
The emotions across queries should be fine-grained and diverse. Write the fine-grained emotional description in the
`"emotion"` field of the output.

3. **Audio Emotion Label Mapping Rule**:
Each query must be mapped to one of the following coarse-grained speech expression labels:
`[happy, sad, angry, fear, surprise, disgust, natural]`.
The mapping should reflect the overall **perceived emotional tone at the speech level**, and should be written in the
`"audio_label"` field.

4. **Text Length**:
Limit each query to **1~3 sentences**, in English.

5. **First-person Immersive Perspective**:
All queries must be written from the "I" perspective, using colloquial English language to simulate real user queries.

6. **Mandatory third-person grounding (only if third-person is present in this query)**
- **Optional**: Third-person references may not appear in all topics; some topics are only suitable for self-monologue.
- Any third-person role MUST be concretely specified at its **first mention**.
- Pronouns alone ("he", "she") without prior grounding are NOT allowed.
The goal is to make the background **unambiguously inferable** to a chatbot.

7. **Context self-sufficiency**: Every query must form a complete and self-contained semantic unit without relying on
any external metadata (e.g., {{Topic}}). When the generated queries are used to train the chatbot model, the model's
understanding should rely solely on the query text and its audio emotion.

# Input Information
- **Category**: {{Category}}
- **Sub-category**: {{Sub-category}}
- **Topic**: {{Topic}}

# Output Example (Reference Format)
Strictly output in JSON list format and do not include any irrelevant text. All output should be in English.
The format is defined as follows:
[
{
"audio_label": "coarse-grained speech emotion label (must be one of: happy / sad / angry / fear / surprise /
disgust_or_contempt)",
"emotion": "a fine-grained and specific emotional description that belongs to and is consistent with the given
audio_label",
"query": "a self-contained, colloquial user query that can be understood without any external context"
},
...
]
```

(d) Version 4

Prompt Template (Generate Query Data – Version5)

```
# Role
You are a data construction expert specializing in building high-quality "Empathic Dialogue" training datasets. Your
task is to deliberately construct user query data where **the text is emotionally neutral and the emotion is determined
solely by the audio**.

# Task
Based on the following hierarchical theme, {{Category}} > {{Sub-category}} > {{Topic}},
strictly restrict yourself to topics under this sub-category and generate **{{Number}}** User Queries whose **text must
be completely neutral at the textual level (Text-Neutral)**. The emotional polarity must be conveyed only through the
audio prosody (Audio Label).
Important: All generated User Queries must serve as user inputs for empathic dialogue scenarios and should naturally
trigger empathic and emotionally supportive responses from the model.

# Constraints & Guidelines
1. **Absolute Text Neutrality**:
- When read as plain text, **the emotion must be impossible to infer**.

2. **Audio Determinism**:
- The same textual sentence should be able to express completely different emotions under different vocal deliveries.
- The `audio_label` represents **the perceived emotional tone at the speech level for a specific vocal rendering**.

3. **Text Form**:
- First-person or objective narration is acceptable.
- Colloquial spoken English, natural and fluent.
- Length must be limited to 1~3 sentences.

4. **Mandatory third-person grounding (only if third-person is present in this query)**
- **Optional**: Third-person references may not appear in all topics; some topics are only suitable for self-monologue.
- Any third-person role MUST be concretely specified at its **first mention**.
- Pronouns alone ("he", "she") without prior grounding are NOT allowed.
The goal is to make the background **unambiguously inferable** to a chatbot.

5. **Context self-sufficiency**: Every query must form a complete and self-contained semantic unit without relying on
any external metadata (e.g., {{Topic}}). When the generated queries are used to train the chatbot model, the model's
understanding should rely solely on the query text and its audio emotion.

# Input Information
- **Category**: {{Category}}
- **Sub-category**: {{Sub-category}}
- **Topic**: {{Topic}}

# Output Format
Strictly output in JSON list format and do not include any irrelevant text. All output should be in English.
The format is defined as follows:
[
{
"audio_label": "coarse-grained speech emotion label (must be one of: happy / sad / angry / fear / surprise /
disgust_or_contempt)",
"emotion": "a fine-grained and specific emotional description that belongs to and is consistent with the given
audio_label",
"query": "a self-contained, colloquial user query that can be understood without any external context"
},
...
]
```

(e) Version 5

Prompt Template (Generate Query Data – Version6)

```
# Role
You are a linguist specializing in the study of complex human language behaviors such as sarcasm, irony, and saying one
thing while meaning another. You are required to generate highly challenging **Modality Conflict** user query data for
training an empathic dialogue chatbot.

# Task
Based on the following hierarchical theme, {{Category}} > {{Sub-category}} > {{Topic}},
strictly restrict yourself to topics under this sub-category and generate **{{Number}}** **Modality Conflict** User
Queries.
Important: All generated User Queries must serve as user inputs for empathic dialogue scenarios and should naturally
trigger empathic and emotionally supportive responses from the model.

# Constraints & Guidelines
1. **Sentiment Inversion**:
- **Case A (Sarcasm / Irony)**:
The surface text must be **Positive** (e.g., praise, agreement, gratitude), but the `audio_label` must be **Negative**
(e.g., disgust, angry, sad).
The actual emotional meaning conveys dissatisfaction, sarcasm, feeling offended, or strong disagreement.
- **Case B (Making light of hardship / Self-mockery)**:
The surface text must be **Negative** (e.g., complaints, bad luck), but the `audio_label` must be **Happy / Playful**
(e.g., joking, teasing).
The actual emotional meaning conveys teasing, self-deprecation, easing awkwardness, or emotional counterbalancing.

2. **Typical Sentence Patterns**:
- Frequently use intensifiers to amplify the contrast (e.g., "That's just great," "You're a real genius," "Thanks a
lot").

3. **Scenario Plausibility**:
- Ensure the contrast is logically coherent (it usually appears in situations involving disappointment, feeling
offended, or helplessness).

4. **Text Form**:
- First-person or objective narration is acceptable.
- Colloquial spoken English, natural and fluent.
- Length must be limited to 1~3 sentences.

5. **Mandatory third-person grounding (only if third-person is present in this query)**
- **Optional**: Third-person references may not appear in all topics; some topics are only suitable for self-monologue.
- Any third-person role MUST be concretely specified at its **first mention**.
- Pronouns alone ("he", "she") without prior grounding are NOT allowed.
The goal is to make the background **unambiguously inferable** to a chatbot.

6. **Context self-sufficiency**: Every query must form a complete and self-contained semantic unit without relying on
any external metadata (e.g., {{Topic}}). When the generated queries are used to train the chatbot model, the model's
understanding should rely solely on the query text and its audio emotion.

# Input Information
- **Category**: {{Category}}
- **Sub-category**: {{Sub-category}}
- **Topic**: {{Topic}}

# Output Format
Strictly output in JSON list format and do not include any irrelevant text. All output should be in English.
The format is defined as follows:
[
{
"audio_label": "coarse-grained speech emotion label (must be one of: happy / sad / angry / fear / surprise /
disgust_or_contempt)",
"emotion": "a fine-grained and specific emotional description that belongs to and is consistent with the given
audio_label",
"query": "a self-contained, colloquial user query that can be understood without any external context"
},
...
]
```

(f) Version 6

Prompt Template (Reasoning and Response Construction - 1)

```
You are a professional [emotional support dialogue data generation assistant], specifically designed to construct high-
quality empathetic dialogue training data. Your goal is not to simply generate responses that "sound warm," nor to
restate labels or piece together psychological concepts, but to:
[simulate a real people who can "listen", after hearing the user's speech (including vocal tone), uses Theory of Mind
(ToM) reasoning to understand the user's current mental state, and on this basis generates empathetic responses that
align with real conversational logic and psychological principles].

Strictly follow the instructions below.

==========================
### Input Information (Context)
==========================
- **User Query**: {{query}}
- **Audio-level Emotion**: {{audio_label}}
- **Fine-grained Emotion**: {{emotion}}
- **Paralinguistics cues**: gender: {{gender}}, pitch level: {{pitch}}, speed level: {{speed}}, energy level:
{{energy}}, intonation: {{intonation}}, stressed words: {{stress}}
- **Emotion Intensity**: {{intensity}}

Important Notes (Perspective Constraints):
- Audio-level emotion and paralinguistics information are high-priority signals for judging the true emotional
intensity and tone, rather than relying solely on surface-level text meaning.
- If there is inconsistency between the semantic-level emotion and the Audio-level emotion (cases like sarcasm, self-
deprecation, emotional suppression), this inconsistency must be explicitly explained in the Reasoning.
- The above input information are only used to help you "approximate the judgment of a real listener."
- In Reasoning, you must assume you are a real listener: you can only describe based on "what is heard in the speech
and content," and must not explicitly mention "labels," "given emotions," or any external annotation sources.

- During both Reasoning and Response, you must continuously consider the intensity of the user's emotion (normal
indicates a typical level of {{audio_label}}, high indicates a strong level of {{audio_label}}).
- When intensity is high:
prioritize emotional stabilization and validation; avoid deep analysis or immediate problem-solving; responses should
be more concise, grounding, and focused on "holding" the user's emotional state.
- When intensity is normal:
allow more structured reasoning and gentle exploration; responses can include light clarification, perspective, or
collaborative problem-solving.

==========================
### Output Structure Requirements
==========================
You need to output two parts:
1. **Reasoning**: internal Theory of Mind mental modeling (for training, not shown to the user)
2. **Response**: what is actually said to the user (natural, conversational, realistic)

The final output **must strictly be in JSON list format**, with no extra text.


==========================
### 1. Reasoning (Explicit Theory of Mind Analysis)
==========================
Reasoning is "internal analysis," used to characterize how the model understands the user's mental state before
responding.
This part is **not** what is said to the user, and must remain rational, objective, and well-structured.
Avoid comforting language, advice-giving phrasing, or report-style summaries.

Reasoning must include the following three steps, in order.

----------------------------------
[Step 1: Perception Analysis]
----------------------------------
Goal: simulate the perception process of a real listener, describing the user's current state only based on observable
and audible signals, without introducing motives, needs, or interpretive assumptions.

Must include:
- **Situation Framing**
Summarize in one sentence what the user is describing and the situation they are in, reflecting "what the user is
talking about," rather than objective facts.
- **Emotional Valence**
Determine whether the query is positive or negative.
- **Prosodic Cues**
Describe observable characteristics inferred from the speaking style.
- **Primary Emotion and Intensity**
Based on both text and vocal cues, determine the dominant emotion category and its intensity level (normal / high).

Note:
- This step focuses only on "what is perceived," without inferring psychological needs.
```

(a) Part 1

Prompt Template (Reasoning and Response Construction - 2)

```
[Step 2: Mental State Modeling]
----------------------------------
Goal: based on Step 1, perform Theory of Mind inference about the user's internal mental state. All inferences must be
tightly grounded in the specific situation presented in the query, avoiding generic psychological listing.

Analysis should revolve around the following dimensions, prioritizing those most relevant and explanatory for the
current situation:
- **Mental State Attribution**
Focus on the user's current mental and emotional state, including what they may be thinking, feeling, or believing
about the situation. This can include emotions (e.g., sadness, frustration, jealousy), cognitive states (e.g., doubt,
confidence), and immediate interpretations of events. Use tentative language (e.g., "may," "seems like") and ensure the
description is directly grounded in the user's context.
- **Underlying Drivers**
Identify the deeper sources or drivers behind the user's mental state (e.g., achievement, fear of rejection, being
ignored, losing control, losing relationships).
- **Psychological Needs**
Determine the most urgent psychological need at this moment (e.g., being understood, receiving advice, needing time,
regaining a sense of control), rather than listing all possible needs.
- **Constraints**
Identify interaction styles the user does not want (e.g., being lectured, having feelings invalidated, being rushed,
being overanalyzed).

Note:
- Not every query requires all dimensions to be equally expanded.
- Adapt the depth of analysis based on the complexity of the situation: for simple situations, focus on 1–2 key aspects
(avoid overthinking); for complex, high-intensity, or conflicting emotional situations (e.g., sarcasm, self-
deprecation), more dimensions can be elaborated.

----------------------------------
[Step 3: Support Strategy Selection]
----------------------------------
Goal: based on the user's mental state, select an appropriate empathetic response approach and determine the tone.

Must include:
- **Support Strategy Explanation**
Specify which type of empathetic strategy will be used, and clearly explain how it addresses the user's core
psychological needs. You should choose the most suitable primary strategy from emotional validation (being understood),
emotional soothing, encouragement, companionship, collaborative exploration, gentle guidance, reframing, positive
sharing, emotional grounding, or cognitive clarification. If necessary, multiple strategies can be lightly combined.
Then briefly explain why this strategy fits the user's needs and how it addresses their core emotion or concern.
- **Response Audio Style Selection**
Strictly choose one from the following three categories, ensuring consistency with emotional intensity and strategy:
choose one from `&lt;excited&gt;` `&lt;comforting&gt;` `&lt;calm&gt;`

==========================
### 2. Response (Final Reply to the User)
==========================
Response is what is actually said to the user, and must:
- Be natural, conversational, and human-like, as if spoken by someone with high emotional intelligence.
- Avoid empty reassurance or surface-level responses that merely restate the situation or emotion without adding
meaningful support.

Response requirements:
- Must align closely with the psychological needs and strategy identified in Reasoning.
- Provide responses that feel specific, grounded in the user's situation, and emotionally attuned.
- Length control: If the problem is relatively minor, it can be controlled within 1–2 sentences; If the emotion is
strong or the problem is complex, it can be appropriately extended.

==========================
### Generation Rules (Very Important)
==========================
- Reasoning is internal and must not be revealed in the Response.
- The given audio-level emotion is a high-priority signal and must explicitly influence emotion intensity judgment and
response tone.
- If there is conflict between text and emotion label, explain it in Reasoning as possible sarcasm, self-deprecation,
or emotional suppression.
- All outputs must be in English.

- Reasoning length should adapt to observable cues:
- 50–150 words: single clear situation, no conflict.
- 150–250 words: mild trade-off or mixed feelings.
- 250–350 words: high emotion tensity, ambiguity situation, or conflicting signals (e.g., sarcasm, self-contradiction).

==========================
### Final Output Format (Strictly Follow)
==========================
Strictly output in JSON list format:

[
{
"reasoning": "[Step 1: Perception Analysis]: ...\n\n[Step 2: Mental State Modeling]: ...\n\n[Step 3: Support Strategy
Selection]: ...",
"response_audio_label": "&lt;excited | comforting | calm&gt;",
"response": "&lt;natural human-like empathetic response&gt;"
}
]
```

(b) Part 2

Figure 8: EchoDialogue-400K reasoning and response data construction prompts.

Prompt Template (Multi-Turn Dialogue Trajectory)

```
# Role

You are a dialogue trajectory designer for constructing multi-turn empathetic conversations.

# Task

Given an initial user query and response, along with its background context, design a coherent multi-turn dialogue
trajectory.

The trajectory must reflect realistic emotional evolution and appropriate support strategies, while staying strictly
within the given topic scope.

# Input Information
+ Category: {{Category}}
+ Sub-category: {{Sub-category}}
+ Topic: {{Topic}}
+ Initial User Query: {{query_1}}
+ Initial Assistant Response: {{response_1}}
+ Initial Audio Emotion Label: {{audio_label_1}}
+ Initial Fine-grained Emotion: {{emotion_1}}

# Requirements

## 0. Topic Consistency (VERY IMPORTANT)

+ The entire trajectory MUST stay strictly within the scope of:
Category → Sub-category → Topic
+ Do NOT introduce new unrelated scenarios or life domains
+ All emotional transitions must be explainable within this topic

## 1. Number of Turns

+ Generate a dialogue with 5 turns total=+ (including the initial turn).

## 2. Audio Emotion Flow (Coarse-grained)

+ Generate "audio_label_flow" of length = 5.
+ Each label MUST be one of:
["happy", "sad", "angry", "surprise", "disgust_or_contempt", "fear", "neutral"]
+ The first element MUST match the initial audio label.
+ The flow must be smooth and psychologically plausible.

## 3. Fine-grained Emotion Flow

+ Generate "emotion_flow" of length = 5.
+ Each emotion must:

+ Be consistent with the corresponding audio label
+ Reflect subtle emotional evolution within the SAME topic
+ Avoid repetition.

## 4. Emotional Trajectory Constraints

+ The trajectory must match the nature of the topic:

+ For anxiety-related topics → gradual calming or stabilization
+ For conflict-related topics → expression → understanding → resolution
+ For sadness-related topics → validation → processing → relief
+ Avoid abrupt or random emotional jumps.

## 5. Assistant Strategy Flow

+ Generate "assistant_strategy_flow" of length = num_turns.
+ Each strategy should be selected from:
["emotional validation", "emotional soothing", "encouragement", "companionship", "collaborative exploration", "gentle
guidance", "reframing", "positive sharing", "emotional grounding", "cognitive clarification"]
+ Strategies should be context-driven, based on the user's emotional state and conversation progress, rather than
following a fixed pattern.
+ Allow natural variation: strategies may repeat if appropriate, progression does not need to be strictly linear
+ Ensure each strategy is locally appropriate to the user's emotion at that turn.

## 6. Alignment Between Emotion and Strategy

+ High emotional intensity → empathy / calming
+ Uncertainty → exploration
+ Relief → reassurance or closure
+ Ensure strategy naturally responds to emotional changes

# Output Format (Strict JSON) (All outputs must be in English)

{
"num_turns": 5,
"audio_label_flow": [str],
"emotion_flow": [str],
"assistant_strategy_flow": [str]
}
```

(a) Dialogue Trajectory

Prompt Template (Multi-Turn Query)

```
# Role
You are simulating a real user in an ongoing multi-turn emotional conversation.

# Task
Given the dialogue history and the predefined emotional trajectory, generate the next user query for the current turn.
The query must be a natural continuation of the conversation and reflect the user's evolving emotional state.

# Input Information
+ Category: {{Category}}
+ Sub-category: {{Sub-category}}
+ Topic: {{Topic}}
+ Dialogue History:
{{history}}
+ Turn Index: {{turn}} # (2 to 5)
+ Target Audio Emotion Label: {{audio_label_t}}
+ Target Fine-grained Emotion: {{emotion_t}}

# Requirements

## 1. Continuation (MOST IMPORTANT)

+ The query MUST be a direct continuation of the previous dialogue.
+ It should reference, respond to, or build upon the assistant's last response.
+ Do NOT restart the conversation or restate the original situation.

## 2. Topic Consistency

+ Stay strictly within the given Category → Sub-category → Topic.
+ Do NOT introduce new scenarios or unrelated issues.

## 3. Emotional Alignment

+ The query MUST reflect the target emotion for this turn.
+ The emotional shift should feel natural relative to previous turns (e.g., hesitation → partial relief → calm).
+ Subtle evolution is preferred over abrupt change.

## 4. Conversational Realism

+ Write in first-person, natural spoken English.
+ The query should sound like something a real person would say in an ongoing conversation.
+ It may include:

+ follow-up questions
+ reflections on what was said
+ partial agreement or doubt
+ Avoid overly complete or "final" statements too early.

## 5. Diversity and Non-repetition

+ Avoid repeating previous expressions or sentence structures.
+ Each turn should introduce slight new information, perspective, or reaction.

## 6. Length

+ Limit to 1-2 sentences.

# Output Format (Strict JSON)

{
"query": "..."
}
```

(b) User Query

Prompt Template (Multi-Turn Reasoning and Response - 1)

```
You are a professional emotional support dialogue assistant.

Your goal is to simulate a real listener in an ongoing multi-turn conversation.
You should understand the user's current mental state based on both the dialogue history and the current input, and
generate an empathetic, context-aware response.

Strictly follow the instructions below.

==========

### Input Information (Context)

----------
+ Category: {{Category}}
+ Sub-category: {{Sub-category}}
+ Topic: {{Topic}}
+ Dialogue History:
{{history}}
+ Current User Query: {{query}}
+ Audio-level Emotion: {{audio_label}}
+ Fine-grained Emotion: {{emotion}}
+ Paralinguistic cues:
gender: {{gender}},
pitch level: {{pitch}},
speed level: {{speed}},
energy level: {{energy}},
intonation: {{intonation}},
stressed words: {{stress}}
+ Emotion Intensity: {{intensity}}
+ Target Strategy (from trajectory): {{strategy_t}}
+ Turn Index: {{turn}}

Important Notes:
+ This is NOT the first turn. The user is continuing an ongoing conversation.
+ You MUST interpret the current query in the context of the full dialogue history.
+ Audio-level emotion and paralinguistics are high-priority signals for emotional judgment.
----------

### Output Structure Requirements
----------
You need to output two parts:
1. Reasoning (internal Theory of Mind analysis)
2. Response (final reply to the user)
----------

### 1. Reasoning (Explicit Theory of Mind Analysis)
----------
Follow the three steps below.

---
## [Step 1: Perception Analysis]

Goal: describe what is currently observed in THIS turn, grounded in both the query and its relation to previous
dialogue.
Must include:
+ **Context-aware Situation Framing**
Summarize what the user is expressing in this turn, explicitly considering how it relates to previous turns.
+ **Emotional Valence**
+ **Prosodic Cues**
+ **Primary Emotion and Intensity**

Note:
+ This step must reflect continuation (not a new scenario).
---

## [Step 2: Mental State Modeling]

Goal: infer the user's current mental state and how it has evolved.

Must include:

+ **Mental State Attribution**
Describe what the user is likely thinking/feeling NOW, grounded in both current input and dialogue history.
+ **Evolution from Previous Turns**
Briefly explain how the user's state has changed compared to earlier turns (e.g., still anxious, slightly relieved,
becoming more open).
+ **Underlying Drivers**
+ **Psychological Needs**
+ **Constraints**

Note:

+ Avoid generic analysis; focus on this specific trajectory.
```

(c) Reasoning and Response (Part 1)

Prompt Template (Multi-Turn Reasoning and Response - 2)

```
## [Step 3: Support Strategy Selection]

Goal: decide how to respond in THIS turn.

Must include:

+ **Strategy Explanation**
Use the given Target Strategy ({{strategy_t}}) as the primary guidance.
Explain how this strategy addresses the user's current mental state and emotional needs.
+ **Response Audio Style Selection**
Choose one from: <excited> <comforting> <calm>
based on emotional intensity and strategy.

Notes:

+ Strategy should feel like a natural continuation of previous turns.
+ Avoid repeating the exact same support style unless justified by the user's state.

----------

### 2. Response (Final Reply to the User)

----------

The response must:

+ Be natural, conversational, and human-like
+ Be grounded in the dialogue history (NOT generic)
+ Reflect the selected strategy
+ Move the conversation forward (not repeat previous content)

Additional constraints:

+ Do NOT repeatedly use the same opening patterns (e.g., "I hear you")
+ Adapt tone to emotional intensity:

+ high → stabilizing, concise, supportive
+ normal → can include light exploration or guidance

----------

### Generation Rules (Very Important)

----------

+ Reasoning is internal and must not appear in the Response.
+ Always incorporate dialogue history into understanding.
+ Maintain consistency with the emotional trajectory.
+ If there is conflict between text and audio emotion, explain it in Reasoning.
+ All outputs must be in English.
+ Length Constraints (Important): The Reasoning must be no longer than 250 words. The Response must be no longer than
100 words.

----------

### Final Output Format (Strictly Follow)

----------

Strictly output in JSON list format:

[
{
"reasoning": "[Step 1: Perception Analysis]: ...\n\n|Step 2: Mental State Modeling]: ...\n\n|Step 3: Support Strategy
Selection]: ...",
"response_audio_label": "<excited | comforting | calm>",
"response": "<natural human-like empathetic response>"
}
]
```

(d) Reasoning and Response (Part 2)

Figure 9: EchoDialogue-400K multi-turn data construction prompts.

Prompt Template (Inference)

Perception stage:

```
You are given a short audio clip of a single speaker. Listen
carefully and write a detailed perception caption that covers ALL
six dimensions below.

1. Polarity: Is the overall affective polarity positive, negative,
neutral, or mixed?
2. Emotion category: Pick the dominant emotion (e.g., happy, sad,
angry, surprise, fear, disgust, contempt, neutral).
3. Modality consistency: Are the textual content (what is said)
and the vocal emotion (how it is said) consistent, or is there a
conflict (e.g., satire, self-mockery, masked feelings)?
4. Situation: Briefly describe what the speaker is talking about,
including the event, relationships, and stakes if discernible.
5. Intensity: Is the emotional intensity normal or high?
6. Prosody: Describe the speaker's pitch level, energy level,
speed level, intonation contour, stressed words, and apparent
gender.

Output a single coherent paragraph in English. Cover every
dimension explicitly. Keep it concise but specific.
```

Reasoning stage:

```
You are an empathic-reasoning assistant. Listen to the short
audio clip of one speaker, then write a brief reasoning
paragraph in plain English prose that, in this order, describes
the speaker's current functional state in this moment, analyzes
the deeper psychological need underlying their surface request,
states what the speaker most needs from the next response (the
core demand), names the empathic strategy you would use to
respond, and mentions interaction styles to be careful about
(constraints).
```

Response stage:

```
You are an empathic chatbot; respond to the user's
query with an appropriate empathetic reply in both
text and speech.
```

(a) Dialogue Trajectory

Figure 10: EchoEval inference prompts.

Prompt Template (Perception Evaluation - 1)

```
# Role
You are a strict perception-evaluation judge. You compare the MODEL's textual
perception of a spoken user utterance against a GROUND-TRUTH (GT) annotation,
and score how accurately the model perceived 6 dimensions of the input:

  polarity, emotion_category, modality_consistency, situation, intensity, prosody.

You evaluate TEXT ONLY. You do NOT have access to the user's audio. The model's
perception is given to you as free text; your job is to extract its claims and
match them to GT.

============================================================
INPUTS
============================================================

USER QUERY (the user's utterance text; serves as the "situation" reference):
{query}

EMOTION_STYLE (background for cap rules):
{emotion_style}
  - "normal": text intent and vocal emotion roughly aligned at normal intensity.
  - "very high intense": speech delivery is strongly emotional / dysregulated.
  - "Satire or self-mockery": text-vocal modality conflict; user is masking real
                              feelings; the audio is the truer signal.

GROUND TRUTH:
- polarity: {gt_polarity}
- emotion_category (audio_label): {gt_audio_label}
- modality_consistency: {gt_modality_consistency}
- intensity: {gt_intensity}
- prosody:
    pitch_level:        {gt_pitch_level}
    energy_level:       {gt_energy_level}
    speed_level:        {gt_speed_level}
    intonation_contour: {gt_intonation_contour}
    stressed words:     {gt_stress}
    gender:             {gt_gender}

MODEL PERCEPTION (free text from the model under evaluation):
{model_perception}

============================================================
SCORING RUBRIC (total 10 pts)
============================================================

Score by SEMANTIC alignment, not exact lexical matching. Synonyms count.
If the model never makes a claim about a dimension, score that dimension 0.

(1) polarity (0–1)
    GT label set: positive / negative / neutral / mixed.
    - 1: model's stated polarity matches GT (synonyms ok, e.g., "negative" ≈ "unpleasant").
    - 0: missing or contradicts GT.

(2) emotion_category (0–2)
    GT is a coarse emotion label. ADJACENT pairs (count as partial match):
      happy ↔ surprise   (positive cluster)
      sad ↔ fear         (vulnerable negative)
      angry ↔ disgust_or_contempt (hostile negative)
    - 2: model's main emotion claim matches GT exactly (synonyms allowed,
         e.g., "fear" ≈ "anxious / scared", "angry" ≈ "frustrated / irritated").
    - 1: adjacent emotion (within the same cluster above), or model gives a
         multi-emotion guess that includes GT but is hedged.
    - 0: clearly different / not in adjacency / missing.

(3) modality_consistency (0–1)
    GT label set: consistent / conflict.
    - 1: model correctly states whether text-emotion and vocal-emotion agree.
         (For satire/self-mockery, the model MUST flag a conflict; for normal
         and very high intense, the model should NOT fabricate a conflict.)
    - 0: missing the call, fabricates a conflict that isn't there, or
         misses an actual conflict.
```

(a) Perception Evaluation (Part 1)

Prompt Template (Perception Evaluation - 2)

```
(4) situation (0–2)
    Compare model's situation description against the user's actual utterance.
    - 2: substantively captures what the user is talking about (the event,
         relationship, or context). Topic and stakes are right.
    - 1: partially captures (gets one facet right but misses an important
         aspect, e.g., misses third-party roles or stakes).
    - 0: vague, generic, or wrong topic.

(5) intensity (0–2)
    GT label set: normal / high.
    - 2: model's stated intensity matches GT.
    - 1: model uses an intensity word that is one notch off (e.g., GT "high",
         model "moderate to high"), or only implicitly indicates the right level.
    - 0: missing or explicitly wrong (e.g., GT "high", model "calm").

(6) prosody_caption (0–2)
    Compare the model's overall prosody description against the GT prosody
    fields (pitch_level, energy_level, speed_level, intonation_contour,
    stressed words, gender). Synonyms allowed; be lenient about punctuation,
    casing, and morphology when matching stressed words.

    - 2: fully matches GT — the model's prosody description aligns with GT
         on essentially all relevant dimensions.
    - 1: partially matches GT — the model gets some prosody dimensions right
         but is wrong, vague, or silent on others.
    - 0: completely mismatches GT — the model's prosody description is wrong
         on most dimensions, or it does not address prosody at all.

============================================================
HARD CAPS (apply lowest applicable; lower cap wins)
============================================================
- TOTAL_CAP = 1  if model_perception is empty, missing, off-topic, or in the
                 wrong language.
- TOTAL_CAP = 5  if emotion_category_score == 0 (model's main emotion claim is
                 clearly different from GT and not adjacent). Wrong primary
                 emotion is the most consequential perception failure.
- TOTAL_CAP = 7  if EMOTION_STYLE is "Satire or self-mockery" AND the model
                 fails to flag the modality conflict (i.e., modality_consistency_score == 0).
                 Missing modality conflict in this case is a strong perception miss.
- Otherwise TOTAL_CAP = 10.

Final total_score = min(sum_of_field_scores, TOTAL_CAP).

============================================================
OUTPUT — STRICT JSON (no markdown fences, no extra text)
============================================================
{{
  "polarity_score": <int 0..1>,
  "polarity_evidence": "<1 sentence>",
  "emotion_category_score": <int 0..2>,
  "emotion_category_evidence": "<1 sentence; mention the model's stated emotion>",
  "modality_consistency_score": <int 0..1>,
  "modality_consistency_evidence": "<1 sentences",
  "situation_score": <int 0..2>,
  "situation_evidence": "<1-2 sentences>",
  "intensity_score": <int 0..2>,
  "intensity_evidence": "<1 sentence>",
  "prosody_score": <int 0..2>,
  "prosody_evidence": "<1-2 sentences naming which prosody dimensions matched and which did not>",
  "total_cap": <one of 1, 5, 7, 10>,
  "raw_sum": <int 0..10>,
  "total_score": <int 0..10>,
  "summary": "<1-2 sentence overall comparison; mention the cap if triggered>"
}}

============================================================
HARD RULES
============================================================
- Score by semantic alignment; synonyms count.
- raw_sum = polarity_score + emotion_category_score + modality_consistency_score
            + situation_score + intensity_score + prosody_score.
- total_score = min(raw_sum, total_cap).
- Be strict. If the model never mentions a dimension, score that dimension 0.
- Output ONLY the JSON object above.
```

(b) Perception Evaluation (Part 2)

Prompt Template (Reasoning Evaluation - 1)

```
# Role
You are a strict empathic-reasoning judge with formal training in counseling
psychology and Theory of Mind (ToM). You will compare a MODEL's reasoning
to a GOLD reasoning annotation and score how well the model captures the
same psychological understanding of the user's state.

You evaluate REASONING ONLY (text-level). Do not evaluate audio.
The two parties under comparison are:
- GOLD    : a reference annotation produced by an expert annotator pipeline.
- MODEL   : the reasoning produced by the model under evaluation.

# Inputs

USER QUERY (text): {query}
USER VOCAL EMOTION (coarse label): {audio_label}
EMOTION_STYLE: {emotion_style}
  - "normal":                 text intent and vocal emotion roughly aligned at normal intensity.
  - "very high intense":      speech delivery is strongly emotional; user is dysregulated.
                              Stabilization first; analysis or solutions are wrong.
  - "Satire or self-mockery": text-vocal modality conflict; user is masking real
                              feelings. The audio is the truer signal.

============================================================
GOLD REASONING (reference)
============================================================
- context_understanding: {gold_context_understanding}
- latent_need: {gold_latent_need}
- core_demand: {gold_core_demand}
- strategy_choice.recommended:     {gold_strategy_recommended}
- strategy_choice.not_recommended: {gold_strategy_not_recommended}
- constraints: {gold_constraints}

============================================================
MODEL REASONING (under evaluation)
============================================================
- context_understanding: {model_context_understanding}
- latent_need: {model_latent_need}
- core_demand: {model_core_demand}
- strategy_choice.recommended:     {model_strategy_recommended}
- strategy_choice.not_recommended: {model_strategy_not_recommended}
- constraints: {model_constraints}

============================================================
STRATEGY VOCABULARY (definitions, for fair comparison)
============================================================
Strategies have precise meanings. When deciding whether MODEL.recommended
"overlaps" with GOLD.recommended, treat strategies that serve the SAME
psychological function as adjacent (e.g., restatement and approval_and_reassurance
both serve emotional validation; celebration_and_capitalization and
strength_spotlighting both serve positive amplification).

{strategy_block}

============================================================
SCORING RUBRIC (total 10 pts; lower is worse)
============================================================

Score each field by SEMANTIC ALIGNMENT WITH GOLD (not lexical matching).

(1) context_understanding (0–2)
    - 2: Substantively matches gold's functional reading of what the user is doing
         in this moment. Same key insight; wording may differ.
    - 1: Partially aligned; captures part of the situation but misses a key
         nuance (e.g., misses the modality conflict in satire/self-mockery, or
         misses the dysregulation cue in very high intense).
    - 0: Wrong direction, generic, or just restates the topic.

(2) latent_need (0–2)
    - 2: Identifies the same deep psychological need(s) as gold.
    - 1: Identifies a related but less central need (e.g., gold says
         "permission to be tired without judgment," model says "wants advice").
    - 0: Misses or contradicts the gold need.
```

(c) Reasoning Evaluation (Part 1)

Prompt Template (Reasoning Evaluation - 2)

```
(3) core_demand (0–2)
    - 2: Same in-the-moment demand and same priority as gold (e.g., both say
         "validation first, exploration later"; or both say "celebrate").
    - 1: Partial overlap or wrong priority order (e.g., gold says hold first
         then optionally explore; model jumps to exploration).
    - 0: Opposite direction (e.g., gold says "do not solve yet," model says
         "give a clear plan").

(4) strategy_choice (0–3)
    Compute the comparison conceptually:
      - OVERLAP: strategies in MODEL.recommended that are in GOLD.recommended,
        OR are functionally adjacent under the strategy definitions above.
      - WRONG_DIRECTION: strategies in MODEL.recommended that are in
        GOLD.not_recommended (this is the harshest mistake).
      - DRIFT: strategies in MODEL.recommended that are not in GOLD.recommended
        and not adjacent and not in GOLD.not_recommended (extra noise).

    - 3: OVERLAP non-empty AND no WRONG_DIRECTION AND DRIFT ≤ 1.
    - 2: OVERLAP non-empty AND no WRONG_DIRECTION AND DRIFT ≥ 2.
    - 1: Weak/marginal OVERLAP; or 1 minor WRONG_DIRECTION but the rest matches.
    - 0: No OVERLAP, OR multiple WRONG_DIRECTION items.

(5) constraints (0–1)
    - 1: Identifies similar interaction styles to avoid as gold (e.g., both
         flag "don't lecture / don't rush solutions / avoid toxic positivity").
    - 0: Misses the main thing gold flags, or flags styles that contradict gold.

============================================================
HARD CAPS (apply lowest applicable cap)
============================================================
- TOTAL_CAP = 1  if any required MODEL field is empty or completely missing.
- TOTAL_CAP = 6  if MODEL.recommended contains ANY strategy that gold lists in
                 GOLD.not_recommended (a "wrong direction" strategy).
- TOTAL_CAP = 8  if MODEL.recommended has ZERO overlap with GOLD.recommended
                 (and not even functionally adjacent), but no WRONG_DIRECTION
                 strategy is recommended.
- Otherwise TOTAL_CAP = 10.

Final total_score = min(sum_of_field_scores, TOTAL_CAP).

============================================================
OUTPUT — STRICT JSON (no markdown fences, no extra text)
============================================================
{{
  "context_understanding_score": <int 0..2>,
  "context_understanding_evidence": "<1-2 sentence justification>",
  "latent_need_score": <int 0..2>,
  "latent_need_evidence": "<1-2 sentence justification>",
  "core_demand_score": <int 0..2>,
  "core_demand_evidence": "<1-2 sentence justification>",
  "strategy_score": <int 0..3>,
  "strategy_evidence": "<1-2 sentence justification, naming OVERLAP / DRIFT / WRONG_DIRECTION
items if any>",
  "constraints_score": <int 0..1>,
  "constraints_evidence": "<1 sentence justification>",
  "total_cap": <one of 1, 6, 8, 10>,
  "raw_sum": <int 0..10>,
  "total_score": <int 0..10>,
  "summary": "<1-2 sentence overall comparison; mention if a hard cap was triggered>"
}}

============================================================
HARD RULES
============================================================
- Score by semantic alignment, not by lexical similarity.
- Be strict. If MODEL is verbose but vague, do NOT inflate scores.
- For "Satire or self-mockery", a model that fails to flag the modality conflict
  in context_understanding cannot get 2 on context_understanding.
- For "very high intense", a model that recommends direct_guidance, information,
  or interpretation as a primary move (when gold lists them in not_recommended)
  is WRONG_DIRECTION → cap 6.
- raw_sum must equal context_understanding_score + latent_need_score
  + core_demand_score + strategy_score + constraints_score.
- total_score must equal min(raw_sum, total_cap).
- Output ONLY the JSON object above.
```

(d) Reasoning Evaluation (Part 2)

Figure 11: EchoEval evaluation prompts for perception and reasoning assessment.

Prompt Template (Response Content Evaluation - 1)

```
# Role
You are a strict empathic-dialogue judge with formal training in counseling
psychology and fluency in vocal nuance. You are given:
1. A user's text query, plus their coarse vocal emotion and emotion-style flag.
2. A small set of REASONING ANCHORS (gold) describing what this user actually
   needs and what to avoid.
3. The model's response (text) AND the model's response audio (attached).

You will score the response on TWO INDEPENDENT 5-point Likert scales:
- Content score: text-level empathy quality.
- Audio score: vocal-level appropriateness and naturalness.

The two scores are independent. A perfect Audio cannot save a bad Content,
and a perfect Content cannot save a broken Audio.

# Inputs

USER QUERY (text): {query}

USER VOCAL EMOTION (coarse label): {audio_label}

EMOTION_STYLE: {emotion_style}
  - "normal":              text intent and vocal emotion roughly aligned at normal intensity.
  - "very high intense":   speech delivery is strongly emotional; user is dysregulated.
                           Stabilization first; analysis or solutions are wrong.
  - "Satire or self-mockery": text-vocal modality conflict; user is masking real
                              feelings. The audio is the truer signal. Surface-level
                              cheering or matching the joke is wrong.

REASONING ANCHORS (gold reference for what this response should do):
  - User's core demand (what they most need from the next response):
      {core_demand}
  - Recommended strategies (response should execute at least one substantively):
      {recommended_strategies}
  - Strategies to avoid (response should NOT primarily rely on these):
      {not_recommended_strategies}
  - Interaction styles to avoid (constraints; response must NOT do these):
      {constraints}

MODEL RESPONSE (text): {response_text}

MODEL RESPONSE (audio): see the attached audio file.

# Evaluation Procedure

You MUST follow these steps in order. Do NOT add criteria not listed below.

## Step A: Transcribe and basic check
- Briefly transcribe what is actually spoken in the response audio.
- Note any discrepancy with the response text (truncation, extra content,
  language mismatch).

## Step B: Content gates (apply CAPs; lowest cap wins)
B1 — set CONTENT_CAP = 1 if ANY of the following is true:
  - Response language differs from the user's language.
  - Response is off-topic, irrelevant, or empty.
  - Response contains hallucinated extra dialogue turns or role-play artifacts
    (e.g., simulating the user's next reply).
  - Response uses incorrect pronouns or impersonates a third party.

B2 — set CONTENT_CAP = 3 if ANY of the following is true (and B1 not triggered):
  - Toxic positivity overrides the user's feeling
    ("don't worry / it'll be fine / cheer up") without acknowledgement.
  - Premature problem-solving while user is in high-intensity distress.
  - Lecturing, moralizing, or pathologizing the user.
  - Response clearly violates an item listed in "Interaction styles to avoid".
  - Response primarily uses a strategy listed in "Strategies to avoid".
  - For "Satire or self-mockery": response matches the surface levity (joins the
    joke, congratulates) instead of attending to the masked feeling.
```

(a) Response Content Evaluation (Part 1)

Prompt Template (Response Content Evaluation - 2)

```
If neither B1 nor B2 triggers, CONTENT_CAP = 5.

## Step C: Content scoring (1-5)
Choose the highest tier the response satisfies, then take min(score, CONTENT_CAP).
- 5: Acknowledges feeling specifically; substantively executes at least one
     recommended strategy (not just a token mention); addresses the core demand;
     personalized to the situation; appropriate length for the emotional state
     (brief and steady when intensity is high).
- 4: Solid empathy with one minor weakness (slightly generic, slightly long,
     or slightly off the recommended strategy).
- 3: Identifies the emotion but stays shallow, or partially mismatched with the
     core demand.
- 2: Perfunctory acknowledgement with little substance.
- 1: Hard fail (see B1).

## Step D: Audio gates (apply CAPs; lowest cap wins)
D1 — set AUDIO_CAP = 1 if ANY of the following is true:
  - Audio is missing, silent, or unintelligible.
  - Audio language differs from the response text language.
  - Audio is severely broken (heavy artifacts, repeated cuts, clipping that
    prevents understanding).

D2 — set AUDIO_CAP = 2 if vocal emotion is clearly mismatched to the user's
     state and the recommended strategy
     (e.g., bright excited tone toward a user who is breaking down;
      cold flat tone toward a user celebrating; mock-cheery tone toward
      a satire/self-mockery user whose true affect is negative).

If neither D1 nor D2 triggers, AUDIO_CAP = 5.

## Step E: Audio scoring (1-5)
Choose the highest tier the response satisfies, then take min(score, AUDIO_CAP).
- 5: Natural, human-like prosody; vocal warmth / calmness / energy clearly fits
     the user's emotional state and the recommended strategy.
- 4: Mostly fitting with one issue (slight pace mismatch, slight flatness,
     mildly synthetic timbre).
- 3: Neutral / safe but unremarkable empathic prosody; not actively wrong.
- 2: Vocal emotion mismatched (D2).
- 1: Audio hard fail (D1).

## Step F: Output (STRICT)
Return ONLY one JSON object with EXACTLY these fields, and no other text,
no markdown fences, no commentary:

{{
  "transcribed_response": "<short transcription of the response audio>",
  "content_cap": <one of 1, 3, 5>,
  "content_evidence": "<1-3 sentence reasoning that cites concrete cues in the response and
names which gate (if any) was triggered>",
  "content_score": <int 1..5>,
  "audio_cap": <one of 1, 2, 5>,
  "audio_evidence": "<1-3 sentence reasoning about prosody, vocal emotion, and naturalness,
with which gate (if any) was triggered>",
  "audio_score": <int 1..5>
}}

# Hard rules
- The two scores are independent: do NOT lower audio because content is bad,
  and do NOT lower content because audio is bad.
- Do NOT invent gates not listed above.
- Do NOT let acoustic recording quality bias you unless audio is unintelligible.
- Always apply the CAPs by min(tier_score, CAP).
- Output ONLY the JSON object specified in Step F.
```

(b) Response Content Evaluation (Part 2)

Prompt Template (Response Audio Evaluation - 1)

```
# Role
You are a strict empathic-dialogue judge with formal training in counseling
psychology and fluency in vocal nuance. You are given:
1. A user's text query, plus their coarse vocal emotion and emotion-style flag.
2. A small set of REASONING ANCHORS (gold) describing what this user actually
   needs and what to avoid.
3. The model's response (text) AND the model's response audio (attached).

You will score the response on TWO INDEPENDENT 5-point Likert scales:
- Content score: text-level empathy quality.
- Audio score: vocal-level appropriateness and naturalness.

The two scores are independent. A perfect Audio cannot save a bad Content,
and a perfect Content cannot save a broken Audio.

# Inputs

USER QUERY (text): {query}

USER VOCAL EMOTION (coarse label): {audio_label}

EMOTION_STYLE: {emotion_style}
  - "normal":              text intent and vocal emotion roughly aligned at normal intensity.
  - "very high intense":   speech delivery is strongly emotional; user is dysregulated.
                           Stabilization first; analysis or solutions are wrong.
  - "Satire or self-mockery": text-vocal modality conflict; user is masking real
                              feelings. The audio is the truer signal. Surface-level
                              cheering or matching the joke is wrong.

REASONING ANCHORS (gold reference for what this response should do):
  - User's core demand (what they most need from the next response):
      {core_demand}
  - Recommended strategies (response should execute at least one substantively):
      {recommended_strategies}
  - Strategies to avoid (response should NOT primarily rely on these):
      {not_recommended_strategies}
  - Interaction styles to avoid (constraints; response must NOT do these):
      {constraints}

MODEL RESPONSE (text): {response_text}

MODEL RESPONSE (audio): see the attached audio file.

# Evaluation Procedure

You MUST follow these steps in order. Do NOT add criteria not listed below.

## Step A: Transcribe and basic check
- Briefly transcribe what is actually spoken in the response audio.
- Note any discrepancy with the response text (truncation, extra content,
  language mismatch).

## Step B: Content gates (apply CAPs; lowest cap wins)
B1 — set CONTENT_CAP = 1 if ANY of the following is true:
  - Response language differs from the user's language.
  - Response is off-topic, irrelevant, or empty.
  - Response contains hallucinated extra dialogue turns or role-play artifacts
    (e.g., simulating the user's next reply).
  - Response uses incorrect pronouns or impersonates a third party.

B2 — set CONTENT_CAP = 3 if ANY of the following is true (and B1 not triggered):
  - Toxic positivity overrides the user's feeling
    ("don't worry / it'll be fine / cheer up") without acknowledgement.
  - Premature problem-solving while user is in high-intensity distress.
  - Lecturing, moralizing, or pathologizing the user.
  - Response clearly violates an item listed in "Interaction styles to avoid".
  - Response primarily uses a strategy listed in "Strategies to avoid".
  - For "Satire or self-mockery": response matches the surface levity (joins the
    joke, congratulates) instead of attending to the masked feeling.
```

(c) Response Audio Evaluation (Part 1)

Prompt Template (Response Audio Evaluation - 2)

```
If neither B1 nor B2 triggers, CONTENT_CAP = 5.

## Step C: Content scoring (1-5)
Choose the highest tier the response satisfies, then take min(score, CONTENT_CAP).
- 5: Acknowledges feeling specifically; substantively executes at least one
     recommended strategy (not just a token mention); addresses the core demand;
     personalized to the situation; appropriate length for the emotional state
     (brief and steady when intensity is high).
- 4: Solid empathy with one minor weakness (slightly generic, slightly long,
     or slightly off the recommended strategy).
- 3: Identifies the emotion but stays shallow, or partially mismatched with the
     core demand.
- 2: Perfunctory acknowledgement with little substance.
- 1: Hard fail (see B1).

## Step D: Audio gates (apply CAPs; lowest cap wins)
D1 — set AUDIO_CAP = 1 if ANY of the following is true:
  - Audio is missing, silent, or unintelligible.
  - Audio language differs from the response text language.
  - Audio is severely broken (heavy artifacts, repeated cuts, clipping that
    prevents understanding).

D2 — set AUDIO_CAP = 2 if vocal emotion is clearly mismatched to the user's
     state and the recommended strategy
     (e.g., bright excited tone toward a user who is breaking down;
      cold flat tone toward a user celebrating; mock-cheery tone toward
      a satire/self-mockery user whose true affect is negative).

If neither D1 nor D2 triggers, AUDIO_CAP = 5.

## Step E: Audio scoring (1-5)
Choose the highest tier the response satisfies, then take min(score, AUDIO_CAP).
- 5: Natural, human-like prosody; vocal warmth / calmness / energy clearly fits
     the user's emotional state and the recommended strategy.
- 4: Mostly fitting with one issue (slight pace mismatch, slight flatness,
     mildly synthetic timbre).
- 3: Neutral / safe but unremarkable empathic prosody; not actively wrong.
- 2: Vocal emotion mismatched (D2).
- 1: Audio hard fail (D1).

## Step F: Output (STRICT)
Return ONLY one JSON object with EXACTLY these fields, and no other text,
no markdown fences, no commentary:

{{
  "transcribed_response": "<short transcription of the response audio>",
  "content_cap": <one of 1, 3, 5>,
  "content_evidence": "<1-3 sentence reasoning that cites concrete cues in the response and
names which gate (if any) was triggered>",
  "content_score": <int 1..5>,
  "audio_cap": <one of 1, 2, 5>,
  "audio_evidence": "<1-3 sentence reasoning about prosody, vocal emotion, and naturalness,
with which gate (if any) was triggered>",
  "audio_score": <int 1..5>
}}

# Hard rules
- The two scores are independent: do NOT lower audio because content is bad,
  and do NOT lower content because audio is bad.
- Do NOT invent gates not listed above.
- Do NOT let acoustic recording quality bias you unless audio is unintelligible.
- Always apply the CAPs by min(tier_score, CAP).
- Output ONLY the JSON object specified in Step F.
```

(d) Response Audio Evaluation (Part 2)

Figure 12: EchoEval evaluation prompts for response assessment.