EMOTIONADAPT-R1: CONFIDENCE-CALIBRATED ADAPTIVE REASONING FOR SPEECH EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech emotion recognition inherently involves varying ambiguity, ranging from explicit cues to subtle, context-dependent expressions. While recent SpeechLLMs leverage Chain-of-Thought for explainable emotion reasoning, they apply static strategies allocating equal computation regardless of complexity. This fixed-depth method is often inefficient for clear cases and insufficient for ambiguous ones. To address this, we propose EmotionAdapt-R1, a confidence-aware framework enabling adaptive reasoning through Uncertainty-of-Thought, where uncertainty is integrated into generation to realize a "thinking fast and slow" paradigm. Specifically, we introduce Confidence-Adaptive GRPO (CA-GRPO), a policy optimization algorithm that dynamically aligns accuracy, reasoning depth, and confidence expression. Extensive evaluations show our approach outperforms strong baselines on standard SER benchmarks, demonstrating the value of self-aware confidence for efficient, adaptive multimodal reasoning.

1 INTRODUCTION

Speech emotion recognition (SER) (El Ayadi et al., 2011) is inherently uncertain, as emotional cues are not expressed with uniform clarity. Some utterances convey emotions through strong acoustic signals, while others exhibit subtle prosodic cues and overlapping affective states. This inherent ambiguity often leads to inconsistency even among human annotators, as interpretation relies heavily on context and acoustic details. Despite recent efforts in Speech Large Language Models (SpeechLLMs) (Guo et al., 2026) to elevate emotion classification into explainable reasoning (Wang et al., 2026), existing methods overlook the varying degrees of uncertainty within the same emotion categories. Consequently, current approaches treat all emotional utterances uniformly, typically applying fixed-length Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Lyu et al., 2023) to analyze prosodic and semantic cues regardless of their ambiguity.

However, most existing SpeechLLMs adopt a static reasoning paradigm (Xu et al., 2025a; Xie et al., 2025; Tian et al., 2025; Ghosh et al., 2024), leading to two key limitations. (1) Fixed reasoning depth: regardless of emotional ambiguity, models apply the same CoT process across all cases. This "one-size-fits-all" strategy may cause models to overthink clear cases while under-analyzing ambiguous ones, resulting in reasoning inefficiency. (2) Decision-level overconfidence: current models often produce definitive predictions despite input ambiguity, reflecting unwarranted certainty. Although prior works have explored confidence calibration, confidence is typically treated as a post-hoc metric or auxiliary output (Fu et al., 2025; Yang et al., 2025b;a; Hashimoto et al., 2025), rather than a dynamic variable that effectively guides the reasoning and decision-making process.

Inspired by the human "thinking fast and slow" cognitive paradigm (Kahneman, 2011), emotion reasoning can be adaptive to uncertainty: high-confidence cases require minimal deliberation, while uncertain ones demand deeper analysis. Motivated by this, we treat uncertainty not as a passive post-hoc output but as an active control variable that governs reasoning behavior, and introduce the Uncertainty-of-Thought (UoT) framework. Unlike traditional CoT, which applies static reasoning depth to all inputs, UoT reframes confidence from a diagnostic metric into a computation-allocation signal, enabling adaptive reasoning grounded in epistemic awareness.

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

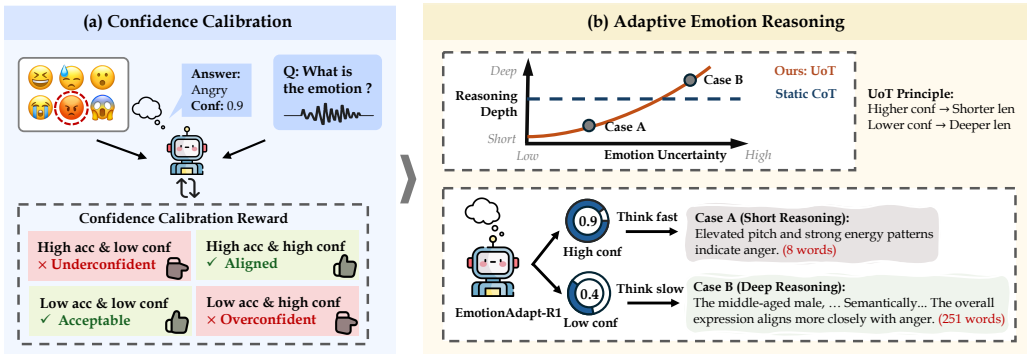


Figure 1: Overview of EmotionAdapt-R1, integrating confidence calibration with adaptive reasoning for uncertainty-aware emotion understanding. (a) Confidence calibration reward aligns predicted confidence with correctness, rewarding calibrated predictions and penalizing overconfident errors. (b) Under UoT, uncertainty regulates reasoning depth, unlike static CoT with fixed length. UoT yields concise reasoning for confident cases and deeper analysis for uncertain ones, improving efficiency and interpretability.

Under the UoT framework, we propose EmotionAdapt-R1 (Figure 1), an adaptive emotion reasoning model that aligns reasoning depth with internal confidence through a three-stage pipeline. First, we use speech processing tools and Valence–Arousal–Dominance (VAD) (Mehrabian, 1996) to obtain fine-grained paralinguistic labels (e.g., pitch, energy, and intonation) and stratify emotion intensity. These features serve as meta-information for prompting LLMs to generate diverse adaptive reasoning paths, forming a high-quality training set. Second, we perform Adaptive Supervised Fine-Tuning (Adaptive SFT) on Qwen2.5-Omni (Xu et al., 2025a) with prosody-perception data to enhance acoustic sensitivity and warm up the model to generate both concise and detailed reasoning chains.

Finally, and most critically, we introduce Confidence-Adaptive GRPO (CA-GRPO), a reinforcement learning framework aligning prediction correctness, confidence estimation, and reasoning depth. Specifically, the confidence calibration reward incentivizes correct high-confidence predictions while penalizing misaligned ones, with milder penalties for low-confidence errors. Together with a confidence–length alignment reward, this optimization enforces an inverse correlation between confidence and reasoning depth. Extensive experiments demonstrate that EmotionAdapt-R1 improves recognition accuracy and reasoning quality while validating the effectiveness of confidence-accuracy calibration and adaptive reasoning.

2 Related Work

2.1 Reasoning in SpeechLLMs

Recent SpeechLLMs have shifted from perception-focused modeling toward complex reasoning (Xu et al., 2025a; Tian et al., 2025; Xu et al., 2025b; Ghosh et al., 2024; Wang et al., 2025a; Sakshi et al., 2024; Tang et al., 2023). Audio-CoT introduces Chain-of-Thought prompting for audio language models (Ma et al., 2025), while subsequent works explore structured reasoning architectures (Xie et al., 2025), reinforcement learning (Li et al., 2025a; Zhong et al., 2025; Wen et al., 2025), and lightweight reasoning models (Deshmukh et al., 2025). For emotion recognition, EmotionThinker introduces prosody-aware reinforcement learning for explainable SER (Wang et al., 2026). However, existing approaches generally apply fixed-depth reasoning across samples, overlooking variations in emotional ambiguity and potentially over-analyzing clear cases while under-analyzing subtle ones.

2.2 Adaptive Reasoning

Adaptive reasoning aims to reduce unnecessary computation and overthinking (Wu et al., 2025a; Wang et al., 2025b; Wu et al., 2025b; Fernandez et al., 2025; Raab, 2007). Prior work mainly considers adaptive reasoning length (Su & Cardie, 2025; Yue et al., 2025; Fernandez et al., 2025) or mechanisms that decide whether explicit reasoning is needed (Zhang et al., 2025; Wu et al.,

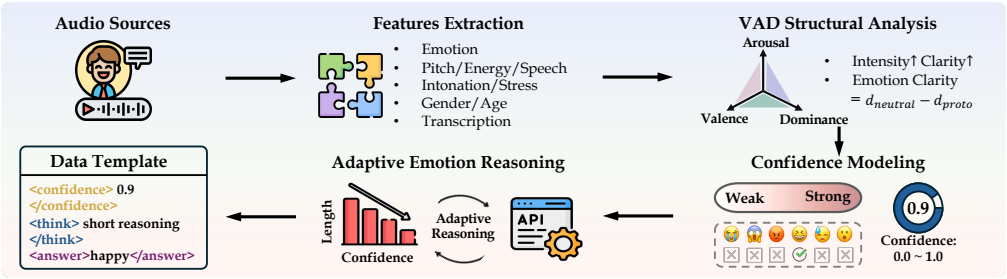


Figure 2: Data construction pipeline of EmotionAdapt-R1 for uncertainty-aware emotion reasoning.

2025a; Zhan et al., 2025). These methods often rely on externally defined difficulty signals, such as human annotations or heuristic rules, to allocate more computation to hard samples. In emotion understanding, however, predefined difficulty can be subjective and affected by annotation bias. Our approach instead grounds adaptivity in the model's own confidence during reinforcement learning, enabling reasoning depth to be guided by internal uncertainty.

2.3 Confidence Calibration for LLMs

Confidence calibration aligns predicted confidence with empirical correctness and has become increasingly important for reliable LLM deployment (Fu et al., 2025; Kang et al., 2025; Li et al., 2025b). Confidence can be estimated implicitly from internal scores such as logits or hidden states (Guo et al., 2017; Li et al., 2025b; Zhang et al., 2020; Mukhoti et al., 2020), or explicitly through verbalized numerical confidence (Li et al., 2025c; Leng et al., 2024; Tao et al., 2024). Existing methods are either training-free, using prompting or repeated sampling (Fu et al., 2025; Lyu et al., 2025; Hashimoto et al., 2025), or training-based, relying on internal representations or auxiliary estimators (Guo et al., 2017; Li et al., 2025c;b). In contrast, we treat verbal confidence as part of the policy and optimize it directly through reinforcement learning, jointly calibrating confidence with correctness and integrating uncertainty into adaptive reasoning.

3 Method

3.1 Problem Formulation

Given a speech input x, the model generates an emotion label y, a reasoning trace r, and a confidence score $c \in [0, 1]$ (discretized in 0.1 increments), forming the structured output (r, y, c). The label y belongs to seven categories: *angry, sad, happy, surprise, disgust, neutral,* and *fear*. We define Uncertainty-of-Thought (UoT) as an adaptive reasoning framework in which model uncertainty controls the reasoning budget. Let $B(r)$ denote the computational budget of reasoning trace r, instantiated in this work by its token length. UoT requires $B(r)$ to decrease with confidence c: uncertain inputs receive deeper reasoning, whereas confident inputs receive shorter reasoning. Unlike static CoT, UoT treats confidence as an endogenous control signal rather than a post-hoc output. Accordingly, we learn a policy $\pi_\theta(r, y, c \mid x)$ that jointly produces an accurate label, calibrated confidence, and confidence-aligned reasoning depth. The policy is optimized as

$$\max_{\theta} \mathbb{E}_{x\sim\mathcal{D},\,(r,y,c)\sim\pi_\theta(\cdot|x)} \left[R(x, r, y, c)\right], \tag{1}$$

where R encourages prediction correctness, confidence calibration against group-level empirical accuracy, and alignment between $B(r)$ and model uncertainty.

3.2 Training Data Construction

As illustrated in Figure 2, we construct an uncertainty-aware training set from IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), Expresso (Nguyen et al., 2023), EARS (Richter et al., 2024), and MSP-Podcast (Busso et al., 2025), comprising over 45k utterances. Following EmotionThinker (Wang et al., 2026), we extract fine-grained paralinguistic and speaker attributes, including pitch, energy, speaking rate, intonation, stressed words, gender, and age group, as structured evidence for interpretable emotion reasoning.

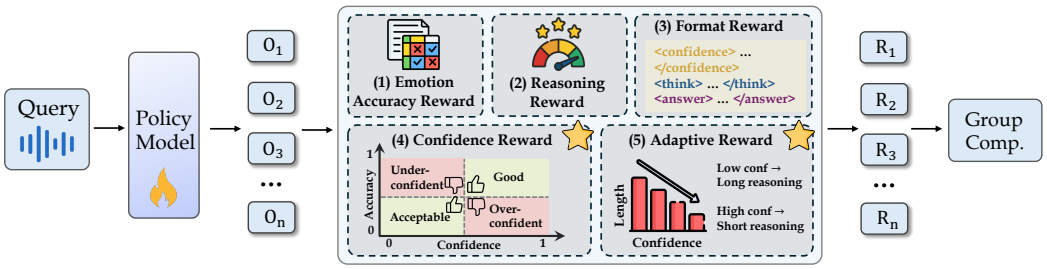


Figure 3: Architecture of EmotionAdapt-R1 with the proposed Confidence-Adaptive GRPO (CA-GRPO). CA-GRPO integrates emotion accuracy, reasoning quality, format compliance, confidence calibration, and confidence–length alignment rewards to jointly enforce accuracy, calibration consistency, and uncertainty-aware adaptive reasoning depth.

To model emotional uncertainty across heterogeneous datasets, we represent each utterance in the continuous Valence–Arousal–Dominance (VAD) space (Mehrabian, 1996). A pre-trained multi-attribute SER model (Goncalves et al., 2024) produces a VAD embedding $v_x \in \mathbb{R}^3$. For each emotion class e, we compute a class prototype $\mu_e = \mathbb{E}[v \mid e]$ on the training split, while v_{neutral} denotes the VAD-space centroid. Let $d_{\text{neutral}} = \|v_x - v_{\text{neutral}}\|$ and $d_{\text{proto}} = \|v_x - \mu_e\|$. We define emotion clarity as

$$\text{Emotion Clarity} = d_{\text{neutral}} - d_{\text{proto}}, \tag{2}$$

where higher clarity indicates expressions that are both distant from neutrality and close to their class prototype, whereas ambiguous samples tend to be weak or atypical.

The clarity score is normalized to $[0, 1]$ and discretized into confidence levels $c \in \{0.0, 0.1, \ldots, 1.0\}$ for SFT supervision. It serves only as a proxy for the relative ordering of emotional ambiguity, rather than ground-truth confidence, and warm-starts the model to produce confidence-aware reasoning formats. CA-GRPO subsequently recalibrates confidence using empirical prediction accuracy across rollouts, separating format acquisition during SFT from calibration refinement during RL. Finally, we prompt DeepSeek-R1 and GPT-4o with the extracted paralinguistic and semantic information to generate emotion reasoning traces under confidence-dependent length constraints. The resulting data follow the structured format `<confidence>`, `<think>`, and `<answer>`. Detailed data construction process is provided in the appendix.

3.3 Adaptive Supervised Fine-Tuning

We employ a two-stage SFT process. First, we fine-tune Qwen2.5-Omni (Xu et al., 2025a) for one epoch on 50k CommonVoice utterances (Ardila et al., 2020). Using the pipeline in data construction section, we extract pitch, energy, speaking rate, intonation, stress, and speaker attributes as multi-task classification targets, aligning low-level acoustic cues with reasoning representations. Second, we perform cold-start SFT on the dataset from the data construction section, training the model to generate `<confidence> c </confidence><think> reasoning </think><answer> emotion label </answer>`. Confidence is linked to reasoning depth, with concise explanations for confident cases and deeper reasoning for ambiguous ones. This initializes joint emotion prediction, confidence estimation, and uncertainty-dependent reasoning for subsequent policy optimization.

3.4 Confidence-Adaptive GRPO

Figure 3 depicts the architecture of our proposed Confidence-Adaptive Group Relative Policy Optimization (CA-GRPO) framework. While the SFT stage initializes structured output generation, it lacks feedback to penalize overconfidence or inefficient reasoning allocation. CA-GRPO addresses this through a composite reward function that jointly optimizes accuracy, calibration, and computational efficiency conditioned on its estimated confidence.

3.4.1 ACCURACY REWARD

We first encourage correct emotion prediction using an outcome-based reward:

$$r_{\text{acc}} = \begin{cases} +1, & \text{if } y = y^*, \\ -1, & \text{otherwise,} \end{cases} \tag{3}$$

where y^* denotes the ground-truth emotion label. This term anchors policy optimization to task performance, preventing degenerate behaviors in which calibration is achieved by uniformly lowering confidence or collapsing predictions.

3.4.2 REASONING PROCESS REWARD

Accuracy alone does not prevent hallucinated explanations. A model may predict the correct emotion while generating reasoning that contradicts observable acoustic evidence. To enforce evidence-grounded reasoning, we introduce a reasoning process reward. For each utterance, we extract verifiable meta-information from annotated audio signals, including speaker attributes (e.g., gender, age group) and paralinguistic statistics (e.g., pitch level, energy, speech rate). The generated reasoning is evaluated by an external multimodal reward model (Qwen2.5-Omni-3B (Xu et al., 2025a)), which assigns an evaluation score $s_{\text{reason}} \in \{0, \ldots, 5\}$ indicating the degree of consistency between the explanation and the acoustic evidence. The reasoning process reward is calculated by normalizing it to [0,1]:

$$r_{\text{reason}} = \frac{s_{\text{reason}}}{5}, \tag{4}$$

which encourages explanations that are grounded in measurable evidence and discourages unsupported or hallucinatory claims.

3.4.3 FORMAT REWARD

We add a format reward to enforce the structured output protocol `<confidence>`, `<think>`, `<answer>`. The reward is defined as:

$$r_{\text{format}} = \begin{cases} 1, & \text{if the output strictly follows the required format,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This constraint stabilizes policy optimization and ensures that confidence and reasoning are explicitly expressed for downstream calibration and evaluation.

3.4.4 CONFIDENCE CALIBRATION REWARD.

For confidence to be epistemically meaningful, it should align with empirical reliability rather than serve as an unconstrained auxiliary output. A well-calibrated model is expected to report higher confidence when its predictions are consistently accurate and to reduce confidence when reliability deteriorates.

To operationalize this alignment, we introduce a confidence calibration reward that regulates predicted confidence using the group-wise accuracy computed over each rollout group. Let $\text{Acc}_{\text{group}}$ denote the fraction of correct predictions within a group of rollouts. Based on this empirical accuracy $\text{Acc}_{\text{group}}$, we define a target confidence interval $\mathcal{I}(\text{Acc}_{\text{group}})$ that specifies the range of confidence values considered consistent with the observed group accuracy:

$$\mathcal{I}(\text{Acc}_{\text{group}}) = \begin{cases} [0.8, 1.0], & \text{Acc}_{\text{group}} \geq 0.75, \\ [0.5, 0.7], & 0.5 \leq \text{Acc}_{\text{group}} < 0.75, \\ [0.0, 0.4], & \text{Acc}_{\text{group}} < 0.5. \end{cases} \tag{6}$$

Intuitively, when most rollouts in a group are correct, the model should express high confidence, and when the group exhibits low accuracy, high confidence would constitute overconfidence and should be discouraged.

Table 1: Performance comparison of speech emotion recognition accuracy (%) and reasoning quality score between EmotionAdapt-R1 and others. Reasoning scores are on a 5-point scale. "G1" and "G2" denote GPT-5 and DeepSeek-R1 evaluation, respectively.

Models	Speech Emotion Recognition ↑					Reasoning Score ↑		
	IEMOCAP	MELD	SAVEE	RAVDESS	Avg.	G1	G2	Avg.
General Multimodal-LLMs								
Qwen-Audio-Chat	38.80	55.70	70.11	71.53	54.86	2.42	2.31	2.37
Qwen2-Audio-Instruct	37.71	51.23	64.98	65.13	51.14	2.51	2.75	2.63
Kimi-Audio	57.72	59.13	61.07	55.21	58.83	3.01	2.94	2.97
MERaLiON2	51.05	51.10	37.02	25.43	46.09	3.11	2.98	3.05
MiniCPM-O	35.54	52.78	40.93	35.47	43.60	2.97	3.06	3.02
Phi-4-Multimodal	36.62	39.81	17.63	15.55	32.15	2.76	2.55	2.66
Qwen2.5-Omni-7B	45.70	54.64	64.77	52.49	50.83	3.04	2.86	2.95
Emotion Recognition-Focused SpeechLLMs								
OSUM-EChat	41.49	53.38	37.34	24.27	44.04	2.04	1.45	1.75
SECap	36.29	34.20	28.34	21.93	32.64	1.55	1.36	1.46
BLSP-Emo	76.00	57.30	72.00	63.73	65.41	2.80	2.79	2.80
Reasoning-Focused SpeechLLMs								
Step-Audio-R1	54.58	46.41	36.27	25.10	44.87	2.93	2.77	2.85
EmotionThinker	**76.68**	59.71	71.56	73.96	68.89	3.91	4.05	3.98
EmotionAdapt-R1	73.45	**61.50**	**74.92**	**75.38**	**70.97**	**4.20**	**4.14**	**4.17**

Given a predicted confidence c and the target interval $\mathcal{I}(\mathrm{Acc}_{\mathrm{group}}) = [l, u]$, with l and u being the lower and upper bounds defined in Equation 6, the calibration reward is defined as:

$$r_{\text{cal}} = \begin{cases} 1, & c \in [l, u], \\ -\mathrm{dist}(c, [l, u]), & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathrm{dist}(c, [l, u]) = \min(|c-l|, |c-u|)$ denotes the distance from the predicted confidence c to the nearest boundary of the target interval. Predictions whose confidence falls within the target interval receive full reward, while deviations are penalized in proportion to their distance from the interval boundary. This mechanism actively discourages both overconfidence in low-accuracy regimes and underconfidence in high-accuracy regimes, thereby grounding the model's self-reported certainty in its empirical performance.

3.4.5 Confidence-Length Alignment Reward.

Calibration alone does not induce adaptive computation. To achieve adaptive reasoning, we posit that computational effort should be inversely proportional to model certainty: inputs with high epistemic uncertainty warrant deeper deliberation (longer reasoning), while unambiguous cases should be resolved concisely (shorter reasoning).

To achieve this, we first map confidence into three discrete regimes: low $[0.0, 0.4]$, medium $[0.5, 0.7]$, and high $[0.8, 1.0]$, which are associated with the long $[L_{\min}^{\text{long}}, L_{\max}^{\text{long}}]$, medium $[L_{\min}^{\text{medium}}, L_{\max}^{\text{medium}}]$, and short $[L_{\min}^{\text{short}}, L_{\max}^{\text{short}}]$ reasoning-length bands, respectively.

Given the predicted confidence c, we determine its corresponding confidence regime, as well as the target reasoning-length band ($[L_{\min}^{c}, L_{\max}^{c}]$). Given the token length L of the reasoning sequence in `<think>`, the reward is defined as:

$$r_{\text{len}} = \begin{cases} 1, & L_{\min}^{c} \le L \le L_{\max}^{c}, \\ -\alpha \cdot \frac{|L - L_{\text{bound}}^{c}|}{L_{\max}^{\text{long}}}, & \text{otherwise}, \end{cases} \tag{8}$$

where $L_{\text{bound}}^{c} \in (L_{\min}^{c}, L_{\max}^{c})$ denotes the nearest boundary of the assigned interval $[L_{\min}^{c}, L_{\max}^{c}]$. Within each regime, a small monotonic adjustment encourages finer alignment, such that higher confidence slightly favors shorter reasoning and lower confidence favors longer reasoning.

This band-based formulation ensures coarse-grained adaptive reasoning while maintaining stability during policy optimization. Confidence thus serves as an operational control signal that regulates reasoning depth.

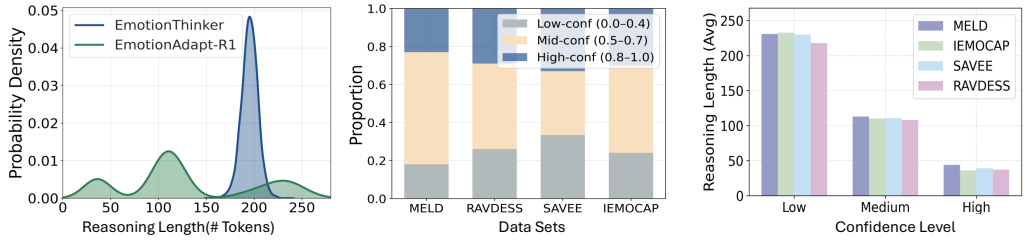
3.4.6 Overall Reward

Collectively, we unify the rewards for correctness, calibration, and adaptive reasoning within a single optimization framework, with the final reward as a weighted combination of these components:

$$\begin{aligned} R_{\text{total}} =& \lambda_{\text{acc}} r_{\text{acc}} + \lambda_{\text{len}} r_{\text{len}} + \lambda_{\text{cal}} r_{\text{cal}} \\ &+ \lambda_{\text{reason}} r_{\text{reason}} + \lambda_{\text{format}} r_{\text{format}}, \end{aligned} \tag{9}$$

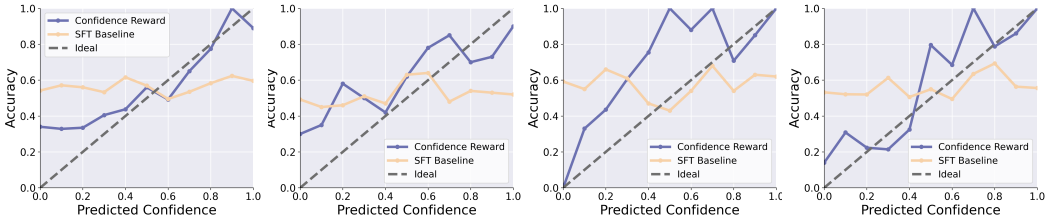
where λ_{acc}, λ_{len}, λ_{cal}, λ_{reason}, and λ_{format} are the corresponding weights for the individual components.

4 Experiment



(a) Reasoning length distribution (b) Confidence distribution (c) Confidence vs. reasoning length

Figure 4: Analysis of adaptive reasoning behavior and confidence distribution in EmotionAdapt-R1. (a) Reasoning length distribution compared with EmotionThinker, showing diversified and confidence-aware length allocation. (b) Distribution of confidence levels across datasets, illustrating calibrated uncertainty modeling. (c) Average reasoning length under different confidence levels, demonstrating inverse correlation between confidence and reasoning depth.



(a) IEMOCAP (b) MELD (c) SAVEE (d) RAVDESS

Figure 5: Confidence–accuracy alignment across datasets. Compared with SFT baseline, the confidence calibrated version shows stronger monotonic alignment between confidence and accuracy, approaching ideal trend.

4.1 Experimental Setup

We use Qwen2.5-Omni-7B (Xu et al., 2025a) as the backbone and apply LoRA adapters (Hu et al., 2022) to the audio encoder, multimodal adapter, and language model. Adaptive SFT is performed for one epoch with a learning rate of 5×10^{-5}, followed by CA-GRPO for 20k steps with a learning rate of 2×10^{-5} and eight rollouts per input. We set the reward weights for accuracy, confidence–length alignment, confidence calibration, reasoning consistency, and format compliance to 1.0, 0.3, 0.3, 0.3, and 0.05, respectively. For evaluation, we conduct experiments on four speech emotion recognition benchmarks: MELD Poria et al. (2019), IEMOCAP Busso et al. (2008), SAVEE Jackson & Haq (2014), and RAVDESS Livingstone & Russo (2018). All datasets are mapped to a unified 7-class emotion taxonomy (angry, sad, happy, surprise, disgust, neutral, and fear) to ensure consistent comparison across corpora. To comprehensively benchmark our approach, we compare against

Table 2: Reward ablation on SER performance, confidence calibration, and token efficiency. Format, Acc, Reas, Conf, and Len denote the format, accuracy, reasoning quality, confidence calibration, and confidence–length alignment rewards, respectively. Acc@HighConf is accuracy restricted to confidence $\in [0.8, 1.0]$. Avg Tokens and HighConf Tokens denote average reasoning length in <think> overall and under high-confidence predictions, respectively.

Method	Reward Components					Performance & Calibration		Token Efficiency	
	Format	Acc	Reas	Conf	Len	Avg Acc ↑	Acc@HighConf ↑	Avg Tokens ↓	HighConf Tokens ↓
SFT Baselines									
SFT (Cold Start)	–	–	–	–	–	49.47	48.75	198	55
SFT + Prosody	–	–	–	–	–	56.41	56.85	193	69
GRPO Baseline									
GRPO	✓	✓	–	–	–	63.57	63.30	176	142
Reward Ablation									
+ r_{cal}	✓	✓	✓	✓	–	67.54	89.17	174	91
+ r_{len}	✓	✓	✓	–	✓	62.55	61.02	122	43
Full Model	✓	✓	✓	✓	✓	70.97	92.61	128	45

12 state-of-the-art models across three categories: (1) general multimodal LLMs, including Qwen-Audio-Chat Chu et al. (2023), Qwen2-Audio-Instruct Chu et al. (2024), Kimi-Audio KimiTeam et al. (2025), MERaLiON2 He et al. (2025), MiniCPM-O MiniCPM-o Team (2024), Phi-4-Multimodal Abouelenin et al. (2025), and Qwen2.5-Omni-7B Xu et al. (2025a); (2) emotion recognition-focused SpeechLLMs such as OSUM-EChat Geng et al. (2025), SECap Xu et al. (2024), and BLSP-Emo Wang et al. (2024); and (3) reasoning-oriented SpeechLLMs, including Step-Audio-R1 Tian et al. (2025) and EmotionThinker Wang et al. (2026). We report SER accuracy and reasoning quality. GPT-5 and DeepSeek-R1 evaluate acoustic-semantic grounding, clarity, and explainability on a five-point scale. Judges receive ground-truth transcripts and extracted paralinguistic attributes for factual verification. Non-reasoning baselines are prompted to explain their predictions. For length-comparable evaluation, only the medium- and long-mode outputs of EmotionAdapt-R1 are included in the reasoning-quality comparison.

4.2 Main Results

As shown in Table 1, EmotionAdapt-R1 achieves the strongest overall trade-off between recognition accuracy and explanation faithfulness, attaining the highest average SER accuracy (70.97%) across the four datasets while also achieving the top reasoning score (4.27). In contrast, general multimodal LLMs (e.g., Kimi-Audio, Qwen2.5-Omni) demonstrate moderate recognition capabilities but lack reasoning faithfulness. While emotion-focused SpeechLLMs (e.g., BLSP-Emo) improve task-specific accuracy, their generated explanations remain inconsistent. Reasoning-oriented models enhance explanation quality, although their recognition performance varies substantially across datasets.

4.3 Analysis and Discussion

Beyond performance, we conduct a series of targeted analyses to dissect the internal mechanisms of EmotionAdapt-R1. Specifically, our investigation is guided by four core questions: (1) **Adaptive Dynamics:** Does the model genuinely adapt its reasoning depth to input uncertainty, and how does confidence modulate computation allocation? (2) **Calibration:** Are the predicted confidence scores well calibrated with respect to empirical correctness? (3) **Component Contribution:** What is the impact of individual reward components on the final performance? (4) **Qualitative Behavior:** How do the generated reasoning traces differ qualitatively across different confidence regimes?

Adaptive Reasoning Analysis Figure 4(a) shows that EmotionThinker produces consistently long explanations, whereas EmotionAdapt-R1 exhibits distinct short, medium, and long reasoning modes aligned with the target ranges of $[0, 60]$, $[70, 140]$, and $[150, 250]$ tokens. This reduces the average reasoning length from 197 to 128 tokens, corresponding to a 35% reduction in generated reasoning tokens while maintaining competitive recognition performance. Figure 4(b) shows that conversational datasets such as MELD and IEMOCAP contain more low- and medium-confidence predictions than the more controlled SAVEE and RAVDESS datasets, indicating sensitivity to dataset-level ambiguity. Figure 4(c) further demonstrates a consistent inverse relationship between confidence and reasoning length, confirming that confidence regulates the allocated reasoning budget.

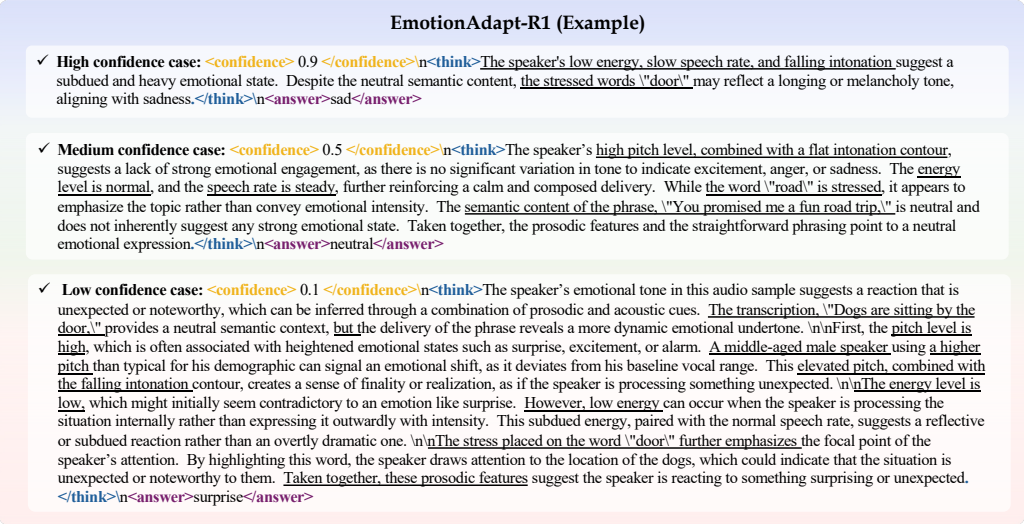


Figure 6: Adaptive reasoning examples across confidence levels in EmotionAdapt-R1. The model maintains accurate prosodic descriptions across cases while progressively increasing reasoning depth and compositional analysis as predicted confidence decreases.

Confidence Calibration Analysis Figure 5 shows that CA-GRPO produces a stronger monotonic relationship between predicted confidence and empirical accuracy than the SFT baseline, particularly in the mid-to-high confidence range, indicating fewer overconfident errors. This improvement does not result from uniformly suppressing confidence: the full model also increases average accuracy from 63.57% under the GRPO baseline to 70.97%. These results show that the calibration reward improves confidence–correctness alignment while preserving predictive performance.

Ablation Study Table 2 shows that SFT improves acoustic initialization, while GRPO substantially increases recognition accuracy but leaves high-confidence accuracy close to overall accuracy, indicating weak calibration. Adding confidence calibration reward r_{cal} raises Acc@HighConf from 63.30% to 89.17%, whereas the confidence-length alignment reward r_{len} reduces the average reasoning length from 176 to 122 tokens. Combining both rewards achieves the best overall trade-off, demonstrating their complementary roles in confidence reliability and reasoning-budget allocation. Additional reward-weight sensitivity analyses are provided in the appendix.

Case Study Figure 6 presents representative high-, medium-, and low-confidence predictions. Across all cases, EmotionAdapt-R1 grounds its decisions in prosodic cues such as pitch and intonation. As confidence decreases, the reasoning becomes progressively more detailed: high-confidence cases focus on dominant cues, medium-confidence cases incorporate semantic context, and low-confidence cases compare multiple signals and alternative interpretations. These examples show that the model adapts both reasoning length and analytical depth to its estimated uncertainty.

5 Conclusion

In this work, we presented EmotionAdapt-R1, an uncertainty-aware adaptive reasoning framework for speech emotion recognition. By introducing the Uncertainty-of-Thought (UoT) formulation, we reconceptualize confidence as an active control variable that regulates reasoning depth rather than a passive post-hoc estimate. Building upon this perspective, we proposed Confidence-Adaptive GRPO (CA-GRPO), which jointly aligns prediction accuracy, confidence reliability, and computation allocation within a unified reinforcement learning objective. Experiments across four benchmarks demonstrate consistent improvements in recognition accuracy, confidence reliability, and reasoning efficiency compared with 12 representative baselines. These findings suggest that uncertainty-aware reward design provides a principled direction for enhancing both trustworthiness and efficiency in speech emotion understanding with large language models.

REFERENCES

Abdelrahman Abouelenin et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Rosana Ardila et al. Common voice: A massively-multilingual speech corpus, 2020.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

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

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, et al. The msp-podcast corpus. *arXiv preprint arXiv:2509.09791*, 2025.

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

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.

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

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

Nigel Fernandez, Branislav Kveton, Ryan A Rossi, Andrew S Lan, and Zichao Wang. Radar: Reasoning-ability and difficulty-aware routing for reasoning llms. *arXiv preprint arXiv:2509.25426*, 2025.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Xuelong Geng et al. Osum-echat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *arXiv preprint arXiv:2508.09600*, 2025.

Sreyan Ghosh et al. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, 2024.

Lucas Goncalves, Ali N. Salman, Abinay Reddy Naini, Laureano Moro-Velázquez, Thomas Thebaud, Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pp. 247–254, 2024. doi: 10.21437/odyssey.2024-35.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Zhihan Guo, Wenqian Cui, Guan-Ting Lin, Daxin Tan, Jingyao Li, Qiyong Zheng, Dingdong Wang, Jing Xiong, Han Shi, Jiaya Jia, et al. A survey of audio reasoning in multimodal foundation models. *arXiv preprint arXiv:2605.21008*, 2026.

Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. Decoding uncertainty: The impact of decoding strategies for uncertainty estimation in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14601–14613, 2025.

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)*, pp. 22–30, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Philip Jackson and S. J. U. Haq. Surrey audiovisual expressed emotion (savee) database. `http://kahlan.eps.surrey.ac.uk/savee/`, 2014.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025.

KimiTeam, Ding Ding, Zeqian Ju, et al. Kimi-audio technical report, 2025. URL `https://arxiv.org/abs/2504.18425`.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025a.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. Confidence is all you need: Few-shot rl fine-tuning of language models. *arXiv preprint arXiv:2506.06395*, 2025b.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *arXiv preprint arXiv:2508.18847*, 2025c.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Qing Lyu et al. 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)*, pp. 305–329, 2023.

Qing Lyu et al. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19260–19268, 2025.

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.

Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current psychology*, 14(4):261–292, 1996.

OpenBMB MiniCPM-o Team. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2024.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*, 2023.

Soujanya Poria et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019. Association for Computational Linguistics.

Markus Raab. Think smart, not hard-a review of teaching decision making in sport from an ecological rationality perspective. *Physical education and sport pedagogy*, 12(1):1–22, 2007.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech*, 2024.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Jinyan Su and Claire Cardie. Thinking fast and right: Balancing accuracy and reasoning length with adaptive rewards. *arXiv preprint arXiv:2505.18298*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. When to trust llms: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5984–5996, 2024.

Fei Tian et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

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.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025a.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning, 2026. URL https://arxiv.org/abs/2601.15668.

Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, Yujun Zhou, and Xiangliang Zhang. Adareasoner: Adaptive reasoning enables more flexible thinking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

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.

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.

Shu Wu, Chenxing Li, et al. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025a.

Siye Wu, Jian Xie, Yikai Zhang, Aili Chen, Kai Zhang, Yu Su, and Yanghua Xiao. Arm: Adaptive reasoning model. *arXiv preprint arXiv:2505.20258*, 2025b.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models, 2025. URL https://arxiv.org/abs/2503.02318.

Jin Xu, Zhifang Guo, et al. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yaoxun Xu et al. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. Uncle: Benchmarking uncertainty expressions in long-form generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30328–30344, 2025a.

Ruihan Yang et al. Logu: Long-form generation with uncertainty expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18947–18968, 2025b.

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

Linan Yue, Yichao Du, Yizhi Wang, Weibo Gao, Fangzhou Yao, Li Wang, Ye Liu, Ziyu Xu, Qi Liu, Shimin Di, et al. Don't overthink it: A survey of efficient r1-style large reasoning models. *arXiv preprint arXiv:2508.02120*, 2025.

Zizheng Zhan et al. Kat-v1: Kwai-autothink technical report. *arXiv preprint arXiv:2507.08297*, 2025.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. Adaptthink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730, 2025.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pp. 11117–11128. PMLR, 2020.

Hao Zhong et al. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

EmotionAdapt-R1: Confidence-Calibrated Adaptive Reasoning for Speech Emotion Recognition

Supplementary Material

CONTENTS

1 Introduction **1**

2 Related Work **2**

2.1 Reasoning in SpeechLLMs 2

2.2 Adaptive Reasoning 2

2.3 Confidence Calibration for LLMs 3

3 Method **3**

3.1 Problem Formulation 3

3.2 Training Data Construction 3

3.3 Adaptive Supervised Fine-Tuning 4

3.4 Confidence-Adaptive GRPO 4

3.4.1 Accuracy Reward 5

3.4.2 Reasoning Process Reward 5

3.4.3 Format Reward 5

3.4.4 Confidence Calibration Reward. 5

3.4.5 Confidence-Length Alignment Reward. 6

3.4.6 Overall Reward 7

4 Experiment **7**

4.1 Experimental Setup 7

4.2 Main Results 8

4.3 Analysis and Discussion 8

5 Conclusion **9**

A Data Construction Details **15**

A.1 Prosody Features Extraction 15

A.2 Emotion Clarity Calculation 15

A.3 Reasoning Traces Generation 17

B Implementation Details **18**

B.1 Unified Inference Protocol. 18

B.2 LLM-as-Judge Evaluation. 19

B.3 Human Evaluation. 19
B.4 Sensitivity Analysis of Reward Weights 20

C AI use statement **20**

A Data Construction Details

A.1 Prosody Features Extraction

Following the prosody extraction procedure adopted in EmotionThinker (Wang et al., 2026), we construct an automatic annotation pipeline to obtain fine-grained paralinguistic and speaker-related attributes for each utterance. Specifically, pitch and energy are computed over short-time acoustic frames to obtain frame-level trajectories describing the temporal evolution of fundamental frequency and vocal intensity. These measurements are subsequently summarized at the utterance level and represented using three categorical levels: *low*, *medium*, and *high*. Speaking rate is estimated by combining phoneme-level timestamps obtained from forced alignment with the total utterance duration, and is categorized as *slow*, *medium*, or *fast*.

For intonation annotation, we smooth the frame-level pitch trajectory using a Savitzky–Golay filter to suppress local fluctuations and extraction noise while preserving its global shape. The smoothed contour is analyzed at two complementary levels. At the coarse-grained level, the utterance is characterized as having either a *flat* or *expressive* intonation style according to the degree of global pitch variation. At the fine-grained level, the pitch contour is categorized as *rising*, *falling*, *rising–falling*, or *falling–rising*, according to its overall direction and turning pattern.

Word-level prominence is identified using WhiStress (Yosha et al., 2025), an alignment-free stress detection model built upon a frozen Whisper encoder. WhiStress applies cross-attention to intermediate Whisper representations and predicts token-level stress labels through a Transformer decoder followed by a feed-forward classification head. The predicted stressed tokens are then aligned with the transcription to obtain the stressed words for each utterance.

We further infer speaker attributes using supervised classifiers built on top of wav2vec 2.0 representations (Baevski et al., 2020). Speaker gender is classified as male or female, while speaker age is mapped into five coarse-grained groups: *child*, *teenager*, *young adult*, *middle-aged adult*, and *older adult*. The resulting annotations, including pitch level, energy level, speaking rate, intonation style, intonation contour, stressed words, gender, and age group, are serialized as structured metadata and provided to the reasoning LLM as perceptually grounded evidence for generating interpretable emotion reasoning traces.

A.2 Emotion Clarity Calculation

To model emotional uncertainty across heterogeneous speech corpora, we observe that discrete emotion categories do not form sharply separated classes. The same categorical label may correspond to substantially different acoustic realizations across speakers, recording conditions, datasets, and application domains. Moreover, utterances near class boundaries may contain mixed or weak affective cues, even when they are assigned a single ground-truth label. Therefore, directly treating the discrete emotion label as an indicator of emotional certainty may overlook the continuous and inherently ambiguous nature of affective expression.

We instead characterize each utterance in the continuous Valence–Arousal–Dominance (VAD) space (Mehrabian, 1996). VAD provides a compact dimensional representation of affective states along three complementary axes. *Valence* describes the pleasantness of an emotional state, ranging from negative or unpleasant affect to positive or pleasant affect. *Arousal* represents the degree of physiological and expressive activation, ranging from calm, passive, or low-energy states to excited, intense, or highly activated states. *Dominance* characterizes the perceived degree of control, power, or agency conveyed by the speaker, ranging from submissive or powerless expressions to dominant or self-assured expressions. Compared with dataset-dependent categorical labels, these continuous dimensions provide a more stable coordinate system for describing gradual affective variation and comparing emotion expressions across heterogeneous sources.

For each utterance x, we use a pre-trained multi-attribute speech emotion recognition model (Goncalves et al., 2024) to obtain a continuous VAD representation:

$$v_x = \left[v_x^{\mathrm{val}}, v_x^{\mathrm{aro}}, v_x^{\mathrm{dom}}\right] \in \mathbb{R}^3, \tag{10}$$

where v_x^{val}, v_x^{aro}, and v_x^{dom} denote the predicted valence, arousal, and dominance values, respectively. The VAD embedding allows emotional ambiguity to be analyzed geometrically rather than solely through discrete class membership.

We define emotion clarity by jointly considering two properties of an utterance: emotional intensity and class typicality. For each emotion category e, we first compute a class-level prototype from the training split:

$$\mu_e = \mathbb{E}\left[v \mid e\right], \tag{11}$$

where v denotes the VAD embedding of an utterance labeled as emotion e. The prototype μ_e summarizes the central affective tendency of the corresponding emotion category in the continuous VAD space. We additionally use v_{neutral} as the affectively neutral reference point in the VAD space.

Given an utterance embedding v_x with ground-truth emotion label e, we compute its distance to the neutral reference as

$$d_{\mathrm{neutral}} = \left\|v_x - v_{\mathrm{neutral}}\right\|_2, \tag{12}$$

and its distance to the corresponding class prototype as

$$d_{\mathrm{proto}} = \left\|v_x - \mu_e\right\|_2. \tag{13}$$

The first distance measures how strongly the utterance departs from an affectively neutral state. A larger d_{neutral} generally indicates a more pronounced emotional expression. The second distance measures how typical the utterance is with respect to its annotated emotion category. A smaller d_{proto} indicates that the utterance lies close to the central VAD pattern of its class, whereas a larger value suggests an atypical, mixed, or boundary-level expression.

We combine these two factors to define the emotion clarity score:

$$\text{Emotion Clarity}(x) = d_{\mathrm{neutral}} - d_{\mathrm{proto}}. \tag{14}$$

This formulation assigns a higher clarity score to an utterance when it is both far from neutrality and close to the prototype of its ground-truth emotion. Such utterances typically contain strong and prototypical emotional evidence. In contrast, an utterance receives a lower score when its affective representation remains close to the neutral region, deviates substantially from its class prototype, or exhibits both properties. The score therefore captures two common sources of emotion ambiguity: weak emotional intensity and atypical within-class expression. An emotionally intense but atypical utterance may still obtain moderate clarity because its large distance from neutrality is offset by its large distance from the class prototype.

To examine whether this geometry-based score reflects human perception, we additionally perform a manual listening inspection of utterances at different regions of the clarity distribution. We find that the resulting ranking is closely aligned with perceptual judgments of emotional explicitness. Utterances assigned high clarity scores generally exhibit salient and internally consistent acoustic cues, such as pronounced pitch movement, strong energy variation, and a clearly recognizable affective tone. In contrast, low-clarity utterances are more frequently perceived as emotionally weak, mixed, context-dependent, or difficult to distinguish from neighboring emotion categories. This perceptual agreement supports the use of the VAD-derived score as a proxy for the relative ordering of emotional ambiguity. Nevertheless, we do not treat it as a direct measurement of human confidence or as an absolute ground-truth uncertainty label.

For SFT supervision, the raw clarity scores are normalized to the interval $[0, 1]$ based on the statistics of the training split:

$$\hat{c}_x = \frac{s_x - s_{\min}}{s_{\max} - s_{\min}}, \tag{15}$$

where s_x denotes the raw emotion clarity score and $s_{\min}$ and $s_{\max}$ are the minimum and maximum scores estimated from the training data. The normalized value is then quantized into eleven confidence levels:

$$c_x \in \{0.0, 0.1, 0.2, \ldots, 1.0\}. \tag{16}$$

These discrete confidence levels are used as reference supervision when constructing confidence-conditioned reasoning traces. Higher-confidence samples are paired with shorter and more direct explanations, whereas lower-confidence samples are paired with more detailed analyses that compare multiple semantic and paralinguistic cues.

Importantly, the VAD-derived clarity score is used only as a warm-start proxy during supervised fine-tuning. Its purpose is to expose the model to diverse confidence expressions and initialize the association between emotional ambiguity and reasoning depth. It is not assumed to be perfectly calibrated with the model's eventual prediction correctness. During the subsequent CA-GRPO stage, confidence is recalibrated according to empirical prediction accuracy across multiple policy rollouts. This allows reinforcement learning to correct deviations introduced by the initial geometry-based proxy. Consequently, the two-stage design separates confidence-format acquisition during SFT from reliability calibration during RL, ensuring that the final reported confidence is shaped by actual predictive behavior rather than solely by the heuristic VAD structure.

A.3 Reasoning Traces Generation

We employ GPT-4o and DeepSeek-R1 as reasoning-trace generators to construct diverse explanations for confidence-aware emotion reasoning. For each utterance x, the generator is provided with its transcription, ground-truth emotion label, normalized emotion clarity score, and the complete set of automatically extracted paralinguistic and speaker attributes. These attributes include speaker gender and age group, pitch level, energy level, speaking rate, intonation contour, and stressed words. Formally, the conditioning information can be represented as

$$m_x = \{y_x^*, t_x, c_x, g_x, a_x, p_x, e_x, s_x, i_x, w_x\}\,, \tag{17}$$

where y_x^* is the annotated emotion, t_x is the transcription, c_x is the clarity-derived confidence target, and the remaining variables represent the speaker and prosodic attributes.

Although the ground-truth emotion is included to anchor the generated trace to the correct training target, the generator is explicitly instructed to reason as if the label were unknown. In particular, it must not reveal the target emotion at the beginning of the explanation or use the label itself as evidence. Instead, the emotional interpretation should emerge gradually from the analysis of the utterance. The prompt emphasizes acoustic and prosodic evidence as the primary basis for inference, while the semantic content of the transcription is treated as complementary evidence. The generator is required to explicitly reference the transcription, naturally incorporate relevant speaker information, and selectively discuss the prosodic cues that provide meaningful evidence for the final decision.

We discourage mechanically enumerating all available attributes. The generator is instructed to focus on the cues that are most discriminative for the current utterance, such as an unusually elevated pitch, reduced vocal energy, a rapid speaking rate, or a salient pitch movement. Stressed words are included only when their prominence contributes to the emotional interpretation. Similarly, intonation annotations marked as uncertain or derived from an utterance that is too short are ignored. When the lexical content is emotionally informative, the trace explains whether it supports the acoustic interpretation. When semantic and acoustic evidence disagree, the generator is asked to explicitly describe this mismatch and explain why the vocal delivery may convey an emotion that is not directly expressed by the words.

A central component of the generation procedure is the use of emotion clarity to control the reasoning budget. The normalized clarity score $c_x \in [0, 1]$ is inversely associated with the target reasoning length: highly prototypical and perceptually explicit utterances receive concise explanations, whereas weak, atypical, or conflicting expressions receive more extensive analysis. Based on the three reasoning modes used in our experiments, we assign the following approximate generation-length intervals:

$$\mathcal{B}(c_x) = \begin{cases} \text{10–60 words} & 0.8 \le c_x \le 1.0, \\ \text{60–150 words} & 0.5 \le c_x \le 0.7, \\ \text{150–250 words} & 0.0 \le c_x \le 0.4, \end{cases} \tag{18}$$

The three length regimes differ not only in verbosity but also in reasoning strategy. For high-clarity utterances, the generator is prompted to produce a concise and intuition-oriented explanation. Such traces identify one or two dominant cues, directly connect them to the perceived affect, and reach the

conclusion without unnecessary consideration of alternative emotions. Short reasoning is therefore designed to be decisive and evidence-focused rather than an arbitrarily truncated version of a longer explanation.

For medium-clarity utterances, the generator integrates several complementary sources of evidence. It may jointly analyze pitch, energy, speaking rate, intonation, and lexical meaning, while briefly explaining how these cues support or weaken the most likely interpretation. Medium-length traces may acknowledge a minor inconsistency or an alternative reading, but they remain centered on the strongest overall explanation.

For low-clarity utterances, the generator adopts a more exploratory reasoning style. It provides fine-grained descriptions of the acoustic and prosodic characteristics, examines potentially conflicting signals, compares multiple plausible emotional interpretations, and considers how speaker characteristics or semantic context affect the interpretation. For example, a low-energy utterance with high pitch may require distinguishing subdued surprise from fear, while emotionally neutral words delivered with a marked pitch shift may require the model to rely more heavily on vocal expression. The final emotion is selected only after these alternatives have been examined. To avoid producing artificially long traces through repetition, the generator is instructed to increase reasoning depth through additional analytical operations—including cue description, cross-cue comparison, semantic–prosodic consistency analysis, and alternative-hypothesis elimination—rather than through paraphrasing or redundant restatement.

B Implementation Details

B.1 Unified Inference Protocol.

We evaluate EmotionAdapt-R1 against 12 representative models, including general-purpose multimodal LLMs, emotion-recognition-focused SpeechLLMs, and reasoning-oriented SpeechLLMs. To ensure a fair comparison, all models are evaluated using the same task instructions and the same unified seven-class emotion taxonomy: *neutral*, *happy*, *sad*, *angry*, *disgust*, *surprise*, and *fear*. For speech emotion recognition accuracy, we use the following prompt for every model:

```
What is the emotion expressed in this audio clip? Please
choose one from the following options: neutral, happy, sad,
angry, disgust, surprise, fear.
```

The model output is mapped to one of the seven candidate labels through exact matching after lowercasing and removing surrounding punctuation. Outputs containing synonymous or morphologically related forms are normalized to the corresponding canonical label whenever the intended category is unambiguous. Samples that do not contain a valid emotion prediction are treated as incorrect. We use deterministic decoding whenever supported by the evaluated model to reduce generation variance.

For reasoning-quality evaluation, each model is explicitly instructed to produce an explanation supporting its emotion prediction. The instruction asks the model to analyze both the semantic content that can be perceived from the audio and the relevant paralinguistic evidence, including pitch, energy, speaking rate, stressed words, and intonation. The model is also encouraged to consider speaker-related information, such as gender and age group, when these attributes provide useful context for interpreting the vocal expression. We use the following unified reasoning prompt:

```
Based on the semantic content you can hear from the audio,
together with its prosodic characteristics, including pitch,
energy, speaking rate, stressed words, and intonation, and
relevant speaker information such as gender and age, provide
a step-by-step reasoning analysis and conclude with the most
likely emotion expressed by the speaker.
```

This prompt is applied to both reasoning-oriented and non-reasoning baselines so that every evaluated model produces an explicit explanation. For EmotionAdapt-R1, we evaluate the generated reasoning contained in the `<think>` field. Following the evaluation setting in the main paper, only its medium- and long-reasoning outputs are included in the reasoning-quality comparison, avoiding an unfair

disadvantage caused by comparing deliberately concise high-confidence outputs with substantially longer explanations from other models.

B.2 LLM-AS-JUDGE EVALUATION.

We employ GPT-5 and DeepSeek-R1 as independent LLM judges to evaluate the quality of the generated emotion reasoning traces. Each judge receives the original model-generated explanation, the predicted emotion, the ground-truth transcription, and the automatically extracted paralinguistic and speaker attributes. The provided reference information includes pitch level, energy level, speaking rate, intonation pattern, stressed words, speaker gender, and age group. These references are used only for factual verification and are not shown to the evaluated model during standard inference unless they are directly available from the audio.

The judges assign each reasoning trace an overall score on a five-point scale according to three complementary criteria:

1. **Acoustic and semantic grounding.** The explanation should be supported by observable information in the audio. Statements about pitch, energy, speaking rate, intonation, stress, speaker characteristics, or semantic content should be consistent with the provided reference annotations and transcription. Unsupported or contradictory acoustic claims are penalized.
2. **Reasoning coherence and clarity.** The explanation should present a logically connected analysis showing how the identified cues lead to the predicted emotion. It should distinguish primary evidence from secondary evidence and avoid merely listing attributes without explaining their emotional implications.
3. **Interpretability and explanatory value.** The reasoning should make the prediction understandable to a human reader. High-quality explanations identify the most discriminative cues, discuss semantic–prosodic consistency when relevant, and provide sufficient justification for choosing the final emotion over plausible alternatives.

The five-point scale is interpreted as follows:

- **5 – Excellent:** The reasoning is well grounded, factually consistent, coherent, and provides a convincing explanation for the predicted emotion.
- **4 – Good:** The reasoning is largely correct and informative, with only minor omissions or imprecise descriptions that do not affect the main conclusion.
- **3 – Acceptable:** The explanation provides partially relevant evidence but contains limited analysis, weak cue integration, or several vague statements.
- **2 – Poor:** The reasoning is weakly grounded, contains substantial inconsistencies, or fails to establish a clear connection between the acoustic evidence and the emotion prediction.
- **1 – Incorrect:** The explanation is largely unsupported, contradictory to the audio evidence, or provides no meaningful reasoning for the prediction.

The outputs are presented to the two judges independently, and the final automatic reasoning score is obtained by averaging their ratings. Model identities are removed from the evaluation input to reduce preference toward particular model families or response styles.

B.3 HUMAN EVALUATION.

Because LLM-based evaluators may exhibit systematic preferences for particular writing styles, response lengths, or model families, we additionally conduct a blinded human evaluation. We recruit three human evaluators with experience in speech or audio analysis. For each evaluated model, we randomly sample 10 test cases, resulting in the same evaluation subset being rated by all three annotators. The evaluators are not informed of the identity of the model that generated each response.

For every case, the evaluators listen to the corresponding audio clip and inspect the generated emotion reasoning trace. They rate the explanation on the same five-point scale used for the LLM-as-judge evaluation, considering whether the reasoning is consistent with the audible semantic and prosodic evidence, whether the intermediate analysis logically supports the predicted emotion, and whether the

explanation is sufficiently clear and interpretable. The final human evaluation score for each model is computed by averaging the ratings across the three evaluators and all sampled cases.

B.4 Sensitivity Analysis of Reward Weights

We analyze sensitivity to key reward component weights in Table 3. Varying λ_{cal} and λ_{len} within a reasonable range does not destabilize performance: all configurations maintain accuracy above 67.7% and Acc@HighConf above 67.3%. We select the moderate setting ($\lambda_{\text{cal}} = \lambda_{\text{len}} = 0.3$) as our final configuration, as it achieves the best trade-off, attaining the highest accuracy (68.9%) alongside strong calibration (92.6%) and competitive token efficiency (128 tokens).

C AI use statement

During the preparation of this manuscript, the authors used generative AI tools for the purpose of language editing and manuscript polishing to improve readability, training data generation and automatic evaluation process. These tools were not used to generate any core scientific ideas. All authors have thoroughly reviewed and approved the final version of the manuscript, and assume full responsibility for the integrity and entirety of its content.

Table 3: Sensitivity analysis *w.r.t.* the weights of the confidence calibration reward λ_{cal} and the confidence-length alignment reward λ_{len}. Higher Acc and Acc@HighConf indicate better accuracy and lower Avg Tokens indicates better token efficiency.

λ_{cal}	λ_{len}	Acc (%) ↑	Acc@HighConf ↑	Avg Tokens ↓
0.1	0.1	67.7	67.3	132
0.3	0.3	68.9	92.6	128
0.5	0.5	68.4	93.3	120