Life of Research · Published 9 November 2026 · DOI 10.5281/zenodo.24085149

A Narrower Answer Is Not a Worse Perception: What Think-Versus-Answer Comparisons in Multimodal Emotion Recognition Measure, Why Their Verdicts Split by Arm, Metric and Supervision, and the Matched Design That Would Credit Reasoning

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Emotion-recognition models built on multimodal language models are now routinely trained to reason before they answer, on the expectation that a written chain of cues will make the label both more accurate and easier to inspect. A June 2026 study reports the opposite for accuracy: on the same models, answering directly often scores higher than answering after deliberation, and the authors call this a thinking paradox. Other studies report that explicit, structured reasoning improves speech emotion recognition under matched reinforcement learning by several points. In this analysis the author reviews and synthesises published results to show that these verdicts do not answer one question. A think-versus-answer comparison varies at least four things: how the answer-only arm is produced (an inference-time toggle on a model trained to reason, a separately trained model, a zero-shot prompt, or a different released variant), the scoring rule and therefore the operating point (hit rate, set recall, F1 or single-label accuracy), the supervision that travels with the reasoning (rationales written by stronger models, often with the label in view), and the reliability of the reference labels. Sorted by these, the results stop conflicting. Inference toggles move scores by about a point. Matched training arms split by output space and by whether a warm-up supplied rationales. Released thinking variants lose up to 11 points on foundational emotion recognition. On the anchor study's own tables, its method's gain over its baseline on basic emotion is 10.5 points under the official hit rate and 3.65 points under F1, and reported scores for one prior system fall by up to 12 points when re-implemented on the anchor study's backbone. The paper consolidates the measured values in two tables and a figure, gives an attribution procedure as an algorithm, and specifies the four-arm, multi-metric design that would credit reasoning with what it adds.

Keywords: multimodal emotion recognition; chain-of-thought; reinforcement learning with verifiable rewards; evaluation metrics; open-vocabulary emotion recognition; multimodal large language models; annotation agreement; modality grounding

1. Introduction

Recognising emotion from a face, a voice and a sentence was for a long time a classification problem: an encoder, a label set of six or seven categories, an accuracy score. Multimodal large language models changed the output first and the procedure second. The output became free text, so a model could name "anxious" or "resigned" where the label set offered only "fear" or "sad" (Lian et al., 2024a; Lian et al., 2025a). The procedure then borrowed from mathematical reasoning: after DeepSeek-R1 showed that reinforcement learning on a verifiable reward could induce long written reasoning (DeepSeek-AI, 2025), emotion models were trained the same way, with a think block before the answer and a reward on whether the answer matched the label (Zhao et al., 2025a; Lian et al., 2025b; Wang et al., 2026; Li et al., 2025b). The motivation was twofold. The reasoning would make the decision inspectable, because it names the cues, and it would make the decision better, because integrating a frown, a flat tone and a sarcastic sentence is the kind of multi-step inference that chain-of-thought helps with elsewhere (Wei et al., 2022).

The second half of that motivation is now in dispute. Han et al. (2026a) report that for reasoning-trained emotion models, triggering a direct answer frequently outperforms the answer produced after deliberation on the standard benchmark suite, and they diagnose the gap as a precision-recall trade: deliberation yields fewer, more conservative emotion words, direct answering yields more. Lian et al. (2025b) had found earlier that removing the think block from their reinforcement-learning pipeline raised their open-vocabulary score from 66.49 to 68.39. Benchmarks report the same direction from a different angle: on EmoBench-M, the thinking variant of an 8B vision-language model scores 46.7 on foundational emotion recognition against 57.8 for its non-thinking counterpart (Hu et al., 2025). Against these stand results in which explicit reasoning helps. Li et al. (2025b) train one audio-language model three ways under the same reinforcement learning and find that structured explicit reasoning beats answer-only by 5.7 weighted-accuracy points in-domain and by 7.8 on a held-out dataset. MME-Emotion reports that, across twenty models, the number of reasoning steps a model writes correlates positively with its score (Zhang et al., 2025).

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

The tension is real and matters beyond emotion. If explicit reasoning does not improve perception-bound judgments, the cost of generating it buys interpretability alone, and the interpretability claim then rests entirely on whether the reasoning is faithful to the computation that produced the label - which, for text tasks, it often is not (Turpin et al., 2023; Chen et al., 2025). If it does improve them, the conditions under which it does are what practitioners need. The general literature points both ways: chain-of-thought helps mainly on mathematical and symbolic tasks (Sprague et al., 2024), it can reduce accuracy on perceptual tasks such as matching faces (Liu et al., 2024), reinforcement fine-tuning without a think block matches or beats fine-tuning with one on visual perception tasks (Li et al., 2025a; Yu et al., 2025), and longer reasoning in multimodal models erodes attention to the image (Liu et al., 2025).

This paper argues that the emotion-recognition results conflict less than they appear to, because the studies are not running the same comparison. In this analysis the author reviews and synthesises published results to show that a think-versus-answer comparison has four independent settings, and that the published verdicts sort cleanly once the settings are named. The first is the **arm**: whether the answer-only condition is the reasoning model told to skip its reasoning, a model trained from the start to answer directly, an untrained model prompted with and without "think step by step", or a separately released variant. The second is the **scoring rule**: the field's standard benchmark scores basic emotion recognition by whether the reference label appears anywhere in the predicted set (Lian et al., 2025a), which rewards listing more emotions, so any procedure that shortens the list loses points without having perceived anything less. The third is the **supervision bundled with reasoning**: in nearly every study where reasoning wins, the reasoning arm also received rationales written by a stronger model, frequently with the reference label in view, so the comparison is between more and less supervision as much as between thinking and not thinking. The fourth is the **reference**: emotion labels carry substantial human disagreement (Barrett et al., 2019; Lian et al., 2024a), so a fixed improvement means different things near and far from the agreement ceiling.

The paper does not argue that reasoning is useless for emotion recognition or that it is necessary. It argues that no published comparison yet isolates what written reasoning at inference time contributes to an emotion label, separately from the metric's operating point, from the supervision that came with it and from run-to-run noise, and it specifies the comparison that would. Section 2 sets out the benchmarks, metrics and the reasoning recipe as applied to affect. Section 3 defines the four settings. Section 4 reads the evidence by arm, Section 5 by scoring rule, Section 6 by noise and reference reliability, and Section 7 asks whether the reasoning consults the non-text modalities at all. Section 8 consolidates 22 published comparisons in Table 1 and Figure 1. Section 9 states an attribution procedure as Algorithm 1, Section 10 specifies six experiments, and Sections 11 to 14 take up objections, the unknowns, limitations and the conclusion.

2. Background: how emotion models came to reason

2.1 From fixed labels to free-form emotion words

Benchmark datasets for multimodal emotion recognition were built around small label sets fixed by majority vote. MELD collects about 13,000 utterances from the television series Friends, each labelled with one emotion and one sentiment (Poria et al., 2019); IEMOCAP records dyadic sessions by actors (Busso et al., 2008); the MER challenge series supplies Chinese-language video clips with discrete and dimensional labels (Lian et al., 2023a). The organisers of the 2024 challenge state the limitation directly: fixing the label space and taking the majority vote to raise annotator consistency can produce inaccurate annotations by discarding non-majority and non-candidate labels (Lian et al., 2024b). Open-vocabulary multimodal emotion recognition (OV-MER) was proposed in response: the model may output any number of emotion words, and the reference is itself a set of free-form words (Lian et al., 2024a).

Free-form output needs a free-form metric. OV-MER and its successors map every word, predicted and reference, through emotion wheels - hierarchies that send "joyful", "content" and "cheerful" to a coarser category such as "happy" - and then compute set-level precision, recall and F1 over the mapped categories, averaged over several wheels (Lian et al., 2024a; Han et al., 2026a). AffectGPT, which introduced the MER-UniBench suite now used as the field's standard, applied a different rule to the four basic-emotion datasets in the suite. Because those datasets give one reference label per item while a generative model may output several words, the suite scores a hit when the mapped reference label is among the mapped predictions, and does not penalise the others; the authors explain that extra predicted labels are not necessarily wrong, since they may be fine-grained emotions that the basic label set lacks, and that an appropriate metric for them had not yet been designed (Lian et al., 2025a). The suite's sentiment datasets are scored by weighted F1 after a separate language model maps the predicted emotion words to a polarity (Han et al., 2026a).

2.2 The R1 recipe, applied to affect

R1-Omni was the first application of reinforcement learning with verifiable rewards to an omni-modal emotion model (Zhao et al., 2025a). It warm-starts a 0.5B model on explanatory annotations, then optimises

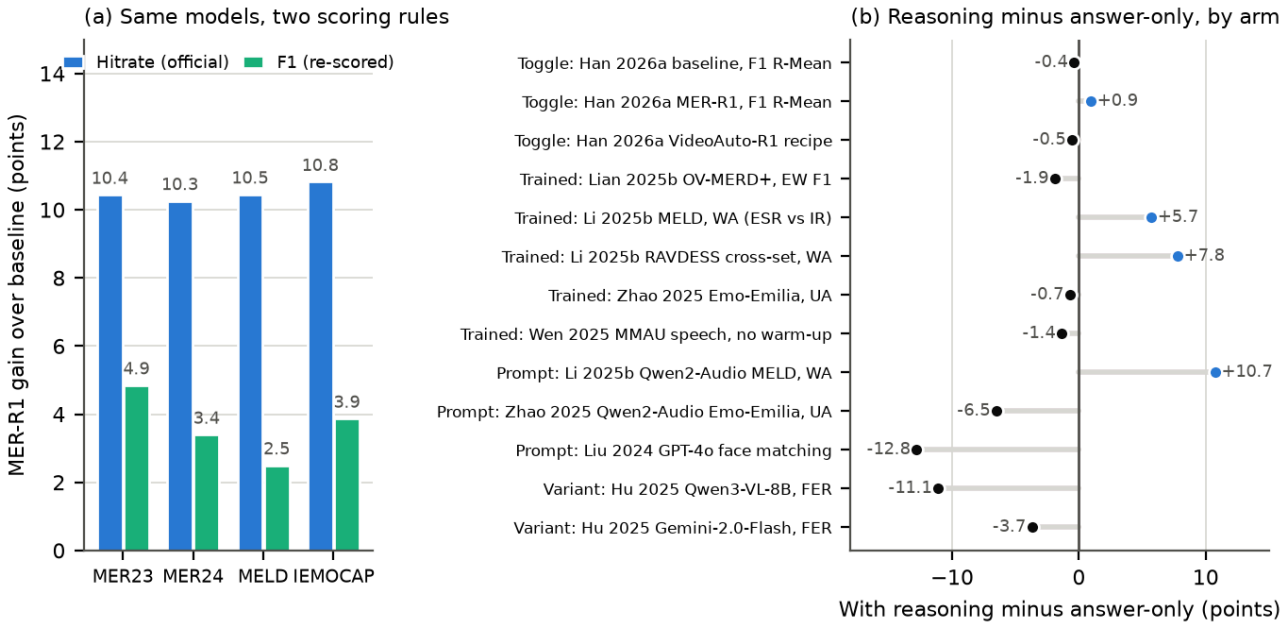


Figure 1. (a) Gain of MER-R1 over its baseline on four basic-emotion datasets, scored by the benchmark's official hit rate and by F1, from Han et al. (2026a), Tables 1 and 3; both use the deliberated answer. (b) Score with written reasoning minus score without, in each source's own metric (percentage points), for thirteen comparisons from Table 1 and Section 4, grouped by how the answer-only arm was produced; values from Han et al. (2026a), Lian et al. (2025b), Li et al. (2025b), Zhao et al. (2025b), Wen et al. (2025), Liu et al. (2024) and Hu et al. (2025). Blue marks a difference favouring reasoning.

with group-relative policy optimisation (GRPO), a policy-gradient method that normalises rewards within a group of sampled answers (Shao et al., 2024), using a reward for a correct label and a reward for the think-then-answer format. Its reported gains over a model fine-tuned directly on the same two datasets are large: on DFEW the pair of recall metrics rises from 60.23 and 44.39 to 65.83 and 56.27, and on the out-of-distribution RAVDESS set from 29.33 and 30.75 to 43.00 and 44.69 (Zhao et al., 2025a). The comparison varies the training algorithm and the presence of reasoning together; no reinforcement-learning arm without a think block is reported. (The source's table header and its prose assign the two recall metrics to the columns in opposite orders; the pairs are reported here without assigning them.)

Successors refined the recipe. AffectGPT-R1 uses the emotion-wheel metric itself as the reward and adds penalties on answer length after observing that the policy learned to list redundant emotion words (Lian et al., 2025b). EmotionThinker builds a 35,000-item corpus of prosody-aware reasoning chains, strengthens the base model's perception of pitch, speed and stress, and adds a learned reward on reasoning quality (Wang et al., 2026). EMO-RL weights the accuracy reward by the similarity between emotions on Plutchik's wheel and compares three output formats under the same optimisation (Li et al., 2025b). HyDRA replaces linear reasoning with a propose-verify-decide protocol over several hypotheses (Liu et al., 2026b). OPPO rewards reasoning that recovers human-annotated visual and acoustic cues and penalises statements about a modality that survive when that modality is masked (Han et al., 2026b). MER-R1, the study that names the thinking paradox, separates recall and precision into two rewards and calibrates the deliberated answer's confidence toward the direct answer's (Han et al., 2026a). Earlier, Emotion-LLaMA had trained an instruction-tuned model to produce emotion explanations alongside labels (Cheng et al., 2024), and surveys of the area now treat reasoning as a defining capability of emotion models (Shou et al., 2025; Guo et al., 2026).

2.3 The general literature on thinking versus answering

Outside affect, the question of when written reasoning helps has a partial answer. A meta-analysis of more than 100 papers found the three task categories that gain most from chain-of-thought to be symbolic reasoning, mathematics and logic, with mean gains of 14.2, 12.3 and 6.9 points; across all other categories the mean score with chain-of-thought was 56.8 against 56.1 without (Sprague et al., 2024). The same authors' own evaluation of 14 models on 20 datasets points the same way. A study that selected tasks on which verbal deliberation is known to impair people found that chain-of-thought lowered GPT-4o's accuracy at picking out a previously seen face from 64.0 to 51.2 percent (Liu et al., 2024), adapting the classic demonstration that describing a face in words impairs recognising it later (Schooler and Engstler-Schooler, 1990); a related human literature reports that analysing the reasons for a preference can lower decision quality (Wilson and Schooler, 1991).

Under reinforcement fine-tuning, two image-domain studies found the think block unnecessary for perception. Li et al. (2025a) report that on a spatial perception benchmark a 2B model scores 70.36 when trained to think first and 76.76 when trained to answer directly, while at 7B the two are within a third of a point (80.36 and 80.67), and on visual mathematics at 7B the thinking model leads, 64.60 to 59.10. Yu et al. (2025) report that adding a thinking process during training and inference lowered referring-expression grounding on RefCOCO from 89.1 to 75.1. For video, VideoAuto-R1 reports that reinforcement learning with thinking and without it yield 66.1 and 66.0 on one general benchmark and 56.4 and 54.4 on a reasoning-heavy one, and that its confidence-gated model chooses to think on 25 percent of items of a perception-oriented benchmark against 51 percent of a reasoning-heavy one (Liu et al., 2026a). The pattern is consistent with a division in which reasoning helps when the bottleneck is combining facts and not when it is extracting them; Section 7 returns to whether emotion recognition is the former or the latter.

3. Four settings of a think-versus-answer comparison

3.1 The arm

A claim that reasoning helps or hurts compares a condition with written reasoning against one without. The condition without can be produced in at least four ways, and they answer different questions.

A1, the inference toggle. A model trained to reason is prompted to answer immediately, or given an empty think block. This asks whether, for a model whose weights were shaped by reasoning, the written trace at test time adds to the answer. It is cheap and is what Han et al. (2026a) and the empty-tag experiment of Li et al. (2025a) do. Its weakness is that the answer-only condition is off the training distribution: a model trained always to think has rarely been rewarded for answering cold, and hybrid models trained to do both often fail to keep the modes apart, with reasoning behaviour leaking into the supposedly direct mode (Wang et al., 2025).

A2, the matched trained arm. Two models are trained from the same starting point on the same data with the same reward, one required to write reasoning and one required to answer directly. This asks whether training with reasoning produces a better labeller. It is what Lian et al. (2025b), Li et al. (2025b), Wen et al. (2025), Li et al. (2025a), Yu et al. (2025) and the reinforcement-learning ablation of Liu et al. (2026a) do. Its weakness is that "the same data" is rarely the same supervision: the reasoning arm is often warm-started on rationales the answer-only arm never sees (Section 3.3).

A3, the zero-shot prompt. An untrained model is asked once for a direct answer and once to reason first. This asks whether the model's existing language-mediated knowledge, when verbalised, helps it read emotion. It is what EMO-RL's and C2SER's zero-shot rows, the financial-sentiment study and the face-recognition study do (Li et al., 2025b; Zhao et al., 2025b; Vamvourellis and Mehta, 2025; Liu et al., 2024). Its weakness is that answer extraction differs between the two formats, and a direct prompt that yields unparseable or off-label text is scored as wrong.

A4, the released variant. A thinking model is compared with a non-thinking sibling released by the same developer. This asks a deployment question - which product to use - and cannot attribute the difference to reasoning, because the two variants differ in post-training data and objectives as well (Hu et al., 2025).

3.2 The scoring rule

Emotion models output sets of words, and the field scores sets in several ways. Hit rate scores one when the reference category is in the predicted set; it is set recall for a single-label reference and is indifferent to how many other categories are predicted (Lian et al., 2025a). Set precision rewards short, correct lists. Set F1 balances the two, and as Han et al. (2026a) compute it, averages precision and recall over items before combining them. MME-Emotion's recognition score for multi-label tasks is the ratio of correctly predicted emotions to reference emotions (Zhang et al., 2025), which is again a recall. Single-label accuracy, weighted or unweighted across classes, applies when the model is constrained to one label (Li et al., 2025b; Zhao et al., 2025b).

These rules disagree about procedures that change the size of the answer set. A reasoning trace that ends in a single committed label and a direct answer that lists three candidate emotions can be equally well perceived and receive very different hit rates. Han et al. (2026a) report exactly this asymmetry for their models - more categories predicted without deliberation, higher recall, lower precision - and note that the benchmark's recall-oriented metric does not reflect deliberation's precision. Whatever a think-versus-answer gap under hit rate measures, it measures it at an operating point the procedure chose.

3.3 The supervision bundled with reasoning

A think block has to come from somewhere. In the emotion literature it usually comes from a warm-up corpus of rationales written by a stronger proprietary or open model: EmotionThinker's chains (Wang et al., 2026), C2SER's explicit chains (Zhao et al., 2025b), Audio-Reasoner's 1.2 million reasoning samples generated with Gemini models, including about 29,000 from MELD (Xie et al., 2025), and SARI's structured and unstructured chains (Wen et al., 2025). Several of these generators see the answer. C2SER's chains are written by an open 9B chat model whose prompt contains discretised pitch, loudness and speaking-rate levels, the transcript and the reference emotion label, and which is asked to describe the speech and conclude with that

emotion (Zhao et al., 2025b). HyDRA's cold-start prompt passes the generator "EVAL-ONLY affect profiles (Ground Truth hints)" with an instruction not to quote them (Liu et al., 2026b); the ambiguity-aware chains of Yu et al. (2026) are synthesised from the input and the reference label distribution, with an instruction that the reasoning be strong enough that a reader would predict the reference labels from it. A rationale written with the label in view is a rationalisation by construction. It can still be useful supervision - it tells the student which cues a stronger model associated with the label - but a gain from training on it is a gain from distillation of cue-label associations, not evidence that writing a rationale at test time improves the label.

The arm-A2 comparisons therefore split in two. Where both arms start from the same checkpoint with no rationale warm-up, the comparison isolates the format. Where only the reasoning arm is warm-started on rationales, it confounds format with extra supervision.

3.4 The reference

Emotion labels are judgments of other people's states from brief clips, and people disagree about them. A review of the evidence on facial movements concluded that how people express anger, disgust, fear, happiness, sadness and surprise varies substantially across cultures, situations and individuals, and that similar facial configurations express more than one category (Barrett et al., 2019). In OV-MER's own annotation, pairwise agreement between annotators' label sets, measured as Jaccard overlap, ranged from 0.46 to 0.57 in the first checking round and from 0.64 to 0.77 in the second, after the labels had been merged and checked (Lian et al., 2024a). The machine-learning literature on learning from disagreement treats such variation as signal about the item rather than noise to be voted away (Uma et al., 2021), and Yu et al. (2026) recast speech emotion recognition as predicting the distribution of annotators' labels.

A reference with this much disagreement changes what a small difference means. A procedure that commits to the majority label and one that hedges across two plausible labels will diverge most on items where annotators diverged, and which one a metric favours depends on the metric, not on which procedure perceived the item better.

4. The evidence, read by arm

4.1 Inference toggles move scores by about a point

Han et al. (2026a) is the clearest A1 study and the anchor of the paradox. Their headline figure compares direct and deliberated answers of a reasoning-trained Qwen2.5-Omni model (Xu et al., 2025) on nine datasets under the suite's official metrics, where direct answering leads. When they re-score the four basic-emotion datasets with F1 instead of hit rate, the gap shrinks and loses its consistent sign. For their baseline model, the deliberated answer scores below the direct one on MER2023 (60.58 against 61.92) and IEMOCAP (51.86 against 53.13), above it on MER2024 (65.06 against 64.37), and level on MELD (46.20 against 46.18) and on the fine-grained set (66.86 against 66.87); averaged over the five recognition sets the deficit is 0.38 points. For a model trained with the VideoAuto-R1 recipe, which supervises both an early and a final answer (Liu et al., 2026a), the deliberated answer trails on all five, by 0.18 to 1.01 points, with a mean deficit of 0.54. For MER-R1, the deliberated answer leads on all five, by 0.02 to 1.81 points, mean 0.94.

These are small numbers, reported without variance or seed counts, and they are the entire empirical content of the paradox once the metric's operating point is removed. They establish that, for reasoning-trained Qwen2.5-Omni models on these benchmarks, writing the reasoning at test time changes the F1 score by roughly a point in either direction depending on the training recipe. They do not establish that reasoning harms emotion perception, and the authors' own method reverses the sign by a comparable margin.

Li et al. (2025a) report a related toggle: when their thinking-trained image model is evaluated with an empty think block appended, part of the advantage of the model trained without thinking is recovered, which they attribute to avoiding overthinking at inference. The general point is that the trained weights, not the presence of the trace at test time, carry most of the behaviour, and A1 measures only the residual.

4.2 Matched trained arms split by output space and warm-up

Six studies train matched reasoning and answer-only arms on emotion or closely related perception tasks, and they disagree.

On open-vocabulary emotion recognition, AffectGPT-R1 removed the think block from both its warm-up and its reinforcement learning and scored 68.39 on OV-MERD+, against 66.49 with it and 62.52 for the supervised starting model (Lian et al., 2025b). The authors offer two explanations: the warm-up reasoning was derived from coarse-grained datasets and may be noisy, and generating reasoning may distract the model from the labelling objective. The extended version reports the same comparison and adopts the answer-only configuration as its default (Lian et al., 2026). The 68.39 configuration, however, is also the one with no length penalty, and the authors report that it produces overlong outputs; with the penalties that suppress them, the score is 67.44 to 68.05 in one version and 67.60 to 68.05 in the other (Lian et al., 2025b; Lian et al., 2026). The answer-only advantage over the think arm therefore survives the penalty - it is about 1.0 to 1.6 points with penalties against 1.9 without - but part of the headline figure is the reward-hacked operating point.

On single-label speech emotion recognition the sign reverses. EMO-RL trains Qwen2-Audio with the same similarity-weighted reward in three formats (Li et al., 2025b). On MELD, weighted accuracy is 63.85 answer-only, 66.17 with unstructured reasoning and 69.56 with structured reasoning; on IEMOCAP 83.90, 84.85 and 87.28; trained on MELD and tested on RAVDESS, SAVEE and IEMOCAP, the structured arm leads the answer-only arm by 7.78, 6.14 and 4.81 points. The structured format requires the model to record the speaker's words and acoustic features before analysing them, and the authors' case study reports that the unstructured arm tends to reason from pitch and rhythm while omitting the semantic content. This is the strongest matched evidence for reasoning in the emotion literature. It is also a single-label task scored by accuracy, where neither arm can gain by listing more labels.

SARI, on multiple-choice audio question answering with sound, music and speech subsets, gives the clearest view of the warm-up confound (Wen et al., 2025). Trained by reinforcement learning from the same Qwen2-Audio checkpoint with no warm-up, the answer-only model averages 59.90 on MMAU test-mini, the structured-reasoning model 59.53 and the unstructured one 62.05; the authors report that without a warm-up the reasoning these models wrote was mostly meaningless. On the speech subset alone, the answer-only model scores 59.76, the two reasoning models 56.16 and 58.41; on MMSU the three score 30.45, 30.71 and 31.52. Only after supervised warm-up on rationales does the reasoning arm pull clearly ahead on the average, to 63.58 to 65.55 - and the speech subset is where the warmed-up structured model is weakest, at 54.88 before and 59.09 after a curriculum. Without a warm-up the reasoning arms were within about two points of answer-only in either direction on the MMAU average, and the authors describe their reasoning as mostly meaningless.

Outside emotion, the matched arms favour answering directly on perception and reasoning on composition. Li et al. (2025a) find no-thinking reinforcement fine-tuning ahead on classification, spatial perception and puzzle tasks across 2B and 7B models, and behind on visual mathematics at 7B; at 7B their thinking models often gave a different answer inside the think block than in the answer tag, and accuracy on those inconsistent responses was below the overall accuracy. Perception-R1's thinking arm loses on all four perception task families it ablates (Yu et al., 2025). VideoAuto-R1's matched pair is level on a general video benchmark and two points apart, in reasoning's favour, on a reasoning-heavy one (Liu et al., 2026a).

4.3 Zero-shot prompts and released variants

The A3 evidence is mixed in sign and large in magnitude, which is what the extraction weakness of Section 3.1 predicts. Prompted directly, Qwen2-Audio scores 39.83 weighted accuracy on MELD and 51.52 on IEMOCAP; prompted to reason, 50.57 and 60.37 (Li et al., 2025b). On a different test set, the same base model's unweighted accuracy falls from 39.07 to 32.57 when prompted to reason, while its macro F1 rises from 31.91 to 38.12 (Zhao et al., 2025b) - the two metrics disagree in sign on the same outputs. For a model that answers in free text, the direct prompt and the reasoning prompt differ in how often the answer can be mapped to a label at all, and a zero-shot gap mixes that with recognition.

Text-only sentiment studies, where extraction is easier, lean toward answering directly. For financial news sentences, GPT-4o without chain-of-thought was the most accurate configuration tested, and a reasoning model was the least (Vamvourellis and Mehta, 2025). The study stratifies by annotator agreement: for GPT-4.1, macro F1 on fully agreed sentences is 0.890 answering directly and 0.799 with long reasoning, and on sentences where only 50 to 65 percent of annotators agreed it is 0.552 and 0.534. The authors read this as reasoning being least harmful where items are ambiguous; for GPT-4o with long reasoning the gap is not monotone in agreement (0.056 at the lowest stratum, 0.029 at the highest), so the pattern is model-dependent. On aspect-level sentiment, chain-of-thought had minimal effect overall, with its improvements confined to the smallest model tested and to one-shot prompts (Zheng et al., 2025).

The released variants on EmoBench-M are the largest effects in the record and the least interpretable (Hu et al., 2025). Qwen3-VL-8B scores 57.8 on foundational emotion recognition and its thinking variant 46.7; on conversational emotion understanding 55.2 and 47.7; on socially complex analysis 67.3 and 67.4. Gemini-2.0-Flash and its thinking variant score 61.4 and 57.7, 53.4 and 54.2, 72.0 and 70.0. The authors conclude that emotion tasks rely more on cross-modal perception than on chain-of-thought. That conclusion fits the numbers, but the variants differ in everything their developers changed between them, and the comparison cannot say which change cost the eleven points.

4.4 Reasoning as training signal versus reasoning as test-time computation

A distinction runs through the A2 results that the think-versus-answer framing hides. Written rationales can matter as supervision during training - extra tokens linking cues to labels - or as computation at test time. C2SER separates the two more cleanly than any other emotion study (Zhao et al., 2025b). Its model is trained first on explicit reasoning chains and then, by a curriculum, on "implicit" targets that contain only the emotion description; at the end it answers without writing reasoning. On the authors' main test set, the explicit-reasoning model scores 68.29 unweighted accuracy, the implicit one 69.00, and a variant trained without the reasoning stages 43.14. The reasoning stages were worth about 26 points as training signal; writing

the reasoning at test time was worth nothing, and slightly less than nothing.

EmotionThinker shows the same structure from the other side (Wang et al., 2026). Supervised fine-tuning on its reasoning corpus raises average speech emotion accuracy from 52.63 to 53.91; GRPO with an outcome reward raises it to 62.91; adding a trained reward on reasoning quality, with the authors' trust weighting and progressive schedule, raises it to 68.89. Every arm writes reasoning, so the study cannot say whether an answer-only arm under the same rewards would match it. What it shows is that rewarding the content of the reasoning moved accuracy by about six points beyond rewarding the label alone - evidence that the reasoning reward is a useful training signal, which is a claim about supervision, not about test-time thinking. HyDRA's ablation is similar: from a baseline of 24.28 on its fine-grained benchmark, R1-style linear reasoning under its rewards reaches 39.51 and the full propose-verify-decide protocol 43.00 (Liu et al., 2026b). Most of the gain is the reinforcement-learning pipeline; the protocol adds 3.5 points; no answer-only arm is reported.

Reinforcement learning itself can account for gains attributed to reasoning. Lian et al. (2026) show that training the supervised AffectGPT model on the same two-stage data used for their reinforcement-learning model moves it from 62.52 to 62.84, against 68.39 for reinforcement learning without reasoning: the gain is the objective, not the data and not the think block. In mathematics, GRPO with random rewards raised one model family's score by 21.4 points against 29.1 with correct rewards, an effect that did not transfer to other families (Shao et al., 2025), and evaluations of reinforcement-learning methods under controlled seeds and settings found smaller improvements than originally claimed (Hochlehnert et al., 2025). An emotion result that compares a reinforcement-learned reasoning model with a supervised direct-answer model, as R1-Omni's does (Zhao et al., 2025a), credits reasoning with whatever reinforcement learning did.

5. The evidence, read by scoring rule

5.1 Hit rate pays for breadth

The basic-emotion rule in MER-UniBench gives credit when the reference category is anywhere in the predicted set (Lian et al., 2025a). The designers' reason was principled - a generative model's extra words may be legitimate fine-grained emotions the basic label set cannot express - but its consequence is that the score cannot fall when a model adds a category, and rises whenever the added category happens to be the reference. Lian et al. (2026) state the same mechanism for the emotion-wheel metrics they used as a reward: synonyms leave the score unchanged, "whereas predicting more words increases the likelihood of matching the ground truth", and their policy duly learned to output long lists until length penalties were added. A reinforcement-learning policy optimised against such a rule will drift toward breadth, and a comparison between two procedures that produce sets of different sizes will favour the broader one.

Han et al. (2026a) report the gap between the official rule and F1 for the same outputs, which allows the size of this effect to be read directly. On the four basic-emotion datasets, their baseline's deliberated answers score 78.16, 82.70, 61.45 and 67.28 by hit rate and 60.58, 65.06, 46.20 and 51.86 by F1; MER-R1's score 88.60, 92.96, 71.90 and 78.10 by hit rate and 65.43, 68.46, 48.70 and 55.73 by F1. MER-R1's improvement over the baseline is therefore 10.44, 10.26, 10.45 and 10.82 points by hit rate and 4.85, 3.40, 2.50 and 3.87 by F1 (Figure 1, panel a). The authors' ablation table gives the same comparison as averages, 72.40 to 82.90 by hit rate and 55.93 to 59.58 by F1. About two thirds of the headline gain on basic emotion is not visible under a rule that charges for wrong categories. The remaining third is a real improvement on a balanced metric and should be credited to the method; the authors' own figures indicate that recall rose while precision held. What the comparison shows is that the official suite score - 83.50 for MER-R1 against 77.87 for its baseline, a 5.63-point gain - mixes a change in recognition with a change in operating point, and that the mixture is not small.

The same structure appears in MME-Emotion, where the recognition score for multi-label tasks is the share of reference emotions the model names (Zhang et al., 2025). Its cross-model observation that models writing more reasoning steps score higher on the combined chain-of-thought score is also partly definitional: that score is an equal-weighted average of the recognition score and a reasoning score computed from the steps a judge model accepts, so half of it depends by construction on what the model writes before its answer. The correlation is between models that differ in size, family and training, and says nothing about whether a given model would recognise emotions better by writing more.

5.2 The thinking paradox is an operating-point result

The diagnosis in Han et al. (2026a) is, read carefully, a statement about operating points: direct answers cover more categories and have higher recall, deliberated answers cover fewer and have higher precision, and the benchmark's metric rewards recall. Under that reading the paradox is not that reasoning damages emotion perception but that deliberation changes how many emotions a model is willing to name, and the field's metric does not price that change. When the same outputs are scored by F1, the baseline's deficit falls to 0.38 points and changes sign on one dataset (Section 4.1). A precision-recall trade-off is not a recognition failure. It becomes one only when the operating point is fixed by the task - for example, when a downstream

system must act on one label - and then the right comparison holds the set size fixed or scores with the rule the task imposes.

This matters for how the paradox is resolved. MER-R1 adds a reward that pushes the deliberated answer's confidence on correct categories toward the direct answer's, and a reward structure that optimises recall and precision separately. Both act on the operating point. That the deliberated answer then beats the direct one by 0.94 F1 points is consistent with the method moving the deliberated answer toward a better trade-off; it is not, on its own, evidence that deliberation now perceives more.

5.3 Single-label tasks remove the breadth channel

Where the output is forced to one label, a procedure cannot gain by breadth, and the comparisons that survive are about recognition. This is the setting of EMO-RL's matched arms (Li et al., 2025b) and of C2SER's explicit-versus-implicit comparison (Zhao et al., 2025b), and it is notable that the two disagree: structured explicit reasoning gains 3.4 to 7.8 points over answer-only in the first, and the implicit model matches or beats the explicit one in the second. Both are single-label; the difference between them lies elsewhere (Section 4.4 and Section 11).

6. The evidence, read by noise and reference

6.1 Reproduced scores, judges and seeds

Han et al. (2026a) re-ran two prior systems on the benchmark instead of copying their published numbers. Table 2 sets the published and re-run values side by side. The published AffectGPT-R1 values exceed the re-run values on every recognition dataset, by 2.63 to 12.24 points, and by 4.38 points on the nine-dataset mean; for AffectGPT the mean gap is 2.81. For AffectGPT-R1 the re-run is a re-implementation rather than a re-evaluation: the authors train on Qwen2.5-Omni, and "The reproduced AffectGPT-R1 [15] follows its original SFT/RL data allocation, while keeping the backbone and optimization framework consistent with our implementation" (Han et al., 2026a, Appendix C). Its gap is therefore the combined effect of a different backbone, a different training implementation and any difference in evaluation, and cannot be read as evaluation noise alone. The source gives no cause for the smaller AffectGPT gap. What Table 2 does show is that a system's published score and the score it is compared at in a later paper can differ by several points, so a reader comparing methods across papers should check which of the two is in the table.

Judges add a second layer. MME-Emotion validated its GPT-4o judge against five human experts, reporting a Spearman correlation of 0.9530 and a Cohen's kappa of 0.8626 on 373 reasoning steps (Zhang et al., 2025). Han et al. (2026a) report MME-Emotion results with a different judge, gemini-3.1-flash-lite-preview, because the original judge was no longer available. The human validation does not transfer to the substitute, and scores produced under the two judges are not on a common scale until it is shown that they are.

Training runs add a third. AffectGPT-RL reports three runs of each length-penalty configuration, with standard deviations of 0.10 to 0.33 points (Lian et al., 2026). That is a useful scale: the answer-only advantage of 1.0 to 1.9 points in the same pipeline exceeds it, while several of the per-dataset toggle differences in Han et al. (2026a) fall within it. On small mathematics benchmarks, seed variation alone has produced standard deviations of 5 to 15 points in pass@1 (Hochlehnert et al., 2025); emotion test sets are larger, so the figure does not transfer, but the lesson that single-run comparisons need a stated variance does.

6.2 The reference has its own ceiling

OV-MER's annotation offers the only directly comparable human figure for the open-vocabulary task: pairwise Jaccard overlap of 0.46 to 0.57 between annotators' label sets in the first checking round and 0.64 to 0.77 in the second (Lian et al., 2024a). For a single item, set F1 and Jaccard overlap are related by $F1 = 2J/(1+J)$, so an item on which two annotators overlap at $J = 0.5$ has an F1 of 0.67 between them. Because that transform is concave, averages do not convert exactly, and because the annotations are fine-grained words while models are scored on wheel-mapped categories, the figure is a scale rather than a ceiling. It is enough to say that current open-vocabulary F1 scores in the high 60s and low 70s sit in the range where human annotators disagree with each other, and that differences of a point near that range are differences in how a model distributes its answers across defensible readings as much as in what it perceived.

Agreement-stratified results show why this matters for the think-versus-answer question in particular. In the financial-sentiment study, every model and method scores higher where annotators agreed more, and the size of the reasoning penalty differs by stratum in a model-dependent way (Vamvourellis and Mehta, 2025). In emotion recognition, HyDRA reports the largest margin over a prior system on the subset with high cross-modal conflict, 11.15 points against 6.69 on the low-conflict subset (Liu et al., 2026b) - but that comparison is between different models, not between reasoning and its absence. No emotion study located reports a think-versus-answer difference stratified by annotator agreement. The prediction is clear in both directions: if deliberation helps by weighing conflicting cues, its benefit should concentrate on low-agreement items; if it hurts by committing early to one reading, its cost should concentrate there too.

Table 2. Published and re-run scores on MER-UniBench for two systems, as reported by Han et al. (2026a), Table 1. OV-MERD+ is scored by emotion-wheel F1, the four basic-emotion sets by hit rate; the mean is over all nine datasets. The published AffectGPT-R1 row is as cited by Han et al. (2026a), taken from version 2 of Lian et al. (2025b); its OV-MERD+ value is the answer-only configuration without length penalty. The current version of Lian et al. (2025b) reports the length-penalised configuration instead (OV-MERD+ 68.05, MER23 84.32, MER24 93.75, MELD 63.12, mean 79.48), against which the MER24 gap is 12.86 points and the mean gap 3.88.

System	Source of value	OV-MERD+	MER23	MER24	MELD	Mean (9)
AffectGPT	published	62.52	78.54	78.80	55.65	74.77
AffectGPT	re-run	59.89	73.69	76.90	52.09	71.96
AffectGPT-R1	published	68.39	84.51	93.13	66.71	79.98
AffectGPT-R1	re-run	64.08	81.88	80.89	57.53	75.60

7. Does the reasoning look and listen?

The case for reasoning in emotion recognition assumes that the trace integrates visual, acoustic and textual evidence. Several measurements say that it often does not.

OPPO's diagnostic benchmark measures two things about emotion reasoning traces (Han et al., 2026b). Utilisation is the share of human-annotated visual and acoustic cues that the trace mentions: 49.14 percent for AffectGPT, 53.00 for Qwen2.5-Omni, 54.85 for AffectGPT-R1 and 58.00 for the authors' reinforcement-learned baseline. Faithfulness is the accuracy of yes-or-no questions about a modality that has been masked, such as asking whether the video shows a frown when there is no video: 35.00, 57.00, 54.20 and 59.20 percent for video and 26.00, 33.80, 30.00 and 33.00 for audio. The authors' method raises these to 70.44, 76.40 and 50.60. Read at face value, existing emotion reasoning traces mention about half the cues annotators found, and the same models answer 41 to 65 percent of masked-video probes and 66 to 74 percent of masked-audio probes incorrectly. The same table lists a fifth model, HumanOmni-V2, at 80.17 percent cue recall; the authors set it apart as an oracle reference because it was cold-started on OV-MERD, and this paper leaves it out of the ranges for that reason. It is no better on the probes, answering 52 percent of masked-video and 62 percent of masked-audio probes incorrectly (48.40 and 37.60 percent accuracy). A trace that asserts a frown inferred from an angry voice is not an inspection of the face. R1-Omni's authors list hallucinated reasoning and under-use of audio cues among their model's limitations (Zhao et al., 2025a).

General multimodal results point the same way. Multimodal reasoning models hallucinate more on perception tasks than their non-reasoning counterparts, and attention to the image declines as the reasoning chain grows (Liu et al., 2025). Audio-language models often score lower with chain-of-thought than with direct answers, a gap its authors hypothesise comes from attention drifting from the audio to the generated text (Bonzi et al., 2026); they fail to integrate speech and audio information in multi-hop reasoning even when they extract each fact correctly (Yang et al., 2025); and chain-of-thought helps them on easy and medium items but confuses them on hard ones (Ma et al., 2025). For an audio-language model fine-tuned with GRPO on question answering, much of the gain was traced to better text-based reasoning, and fine-tuning on text alone, with no audio, raised its audio-benchmark average from 65.9 to 68.2 percent (Rouditchenko et al., 2025). Extending reasoning in language models can also move them from sensible priors toward spurious features (Gema et al., 2025).

The perception bottleneck is visible inside the emotion studies themselves. The off-the-shelf Qwen2.5-Omni-7B that EmotionThinker starts from answers pitch questions with 25.71 percent accuracy, and the authors' prosody-enhanced EmotionThinker-Base with 75.11 (Wang et al., 2026). GPT-4V, evaluated across 21 emotion datasets, could not recognise micro-expressions (Lian et al., 2023b). A reasoning trace cannot describe a pitch contour the encoder does not deliver, and a model that cannot see a micro-expression can write about one only by inference from other cues. Where perception is the bottleneck, test-time reasoning has little to work with, which is consistent with the general finding that thinking helps composition more than extraction (Section 2.3) and with EmoBench-M's conclusion (Hu et al., 2025).

This also bears on the half of the motivation that the paradox leaves standing. Han et al. (2026a) concede accuracy to direct answering but keep interpretability for reasoning. Interpretability requires that the trace reflects the computation behind the label. In text tasks, chain-of-thought explanations omit biasing features that drive the answer and rationalise biased answers (Turpin et al., 2023); larger models condition less on their own chains on most tasks studied (Lanham et al., 2023); and reasoning models mention a hint they used often less than 20 percent of the time (Chen et al., 2025). OPPO's masking probes are the emotion-domain analogue, and they show the same gap. Trained on rationales written with the label in view (Section 3.3), an emotion model learns to produce cue lists that justify labels; whether its cue list names the cues that drove its label is a separate, measurable question that the reviewed studies do not answer for the deliberated answers they evaluate.

8. The evidence, consolidated

Table 1 collects 22 published comparisons between a condition with written reasoning and one without, one row per comparison. The "Arm" column uses the labels of

Table 1. Published think-versus-answer comparisons, by source.

Source	Setting	Arm	Metric	With reasoning	Without
Han et al. 2026a	baseline, 5 recog. sets	A1	F1 mean	58.11	58.49
Han et al. 2026a	MER-R1, 5 recog. sets	A1	F1 mean	61.80	60.86
Han et al. 2026a	VideoAuto-R1 recipe	A1	F1 mean	60.51	61.05
Lian et al. 2025b	OV-MERD+, RL	A2	EW F1	66.49	68.39
Li et al. 2025b	MELD, structured	A2	WA	69.56	63.85
Li et al. 2025b	IEMOCAP, structured	A2	WA	87.28	83.90
Li et al. 2025b	MELD to RAVDESS	A2	WA	73.99	66.21
Zhao et al. 2025b	Emo-Emilia, trained	A2*	UA	68.29	69.00
Wen et al. 2025	MMAU speech, unstructured, no warm-up	A2	acc	58.41	59.76
Li et al. 2025a	CVBench, 2B	A2	acc	70.36	76.76
Li et al. 2025a	CVBench, 7B	A2	acc	80.36	80.67
Li et al. 2025a	MathVista, 7B	A2	acc	64.60	59.10
Yu et al. 2025	RefCOCO grounding	A2	acc	75.1	89.1
Liu et al. 2026a	VideoMME, 7B	A2	acc	66.1	66.0
Liu et al. 2026a	VideoMMMU, 7B	A2	acc	56.4	54.4
Li et al. 2025b	Qwen2-Audio, MELD	A3	WA	50.57	39.83
Zhao et al. 2025b	Qwen2-Audio, Emo-Emilia	A3	UA	32.57	39.07
Liu et al. 2024	GPT-4o, face matching	A3	acc	51.2	64.0
Vamvourellis and Mehta 2025	GPT-4.1, full agreement	A3	macro F1	0.799	0.890
Vamvourellis and Mehta 2025	GPT-4.1, 50-65% agreement	A3	macro F1	0.534	0.552
Hu et al. 2025	Qwen3-VL-8B, FER	A4	mean acc	46.7	57.8
Hu et al. 2025	Gemini-2.0-Flash, FER	A4	mean acc	57.7	61.4

Section 3.1 (A1 inference toggle, A2 matched trained arm, A3 zero-shot prompt, A4 released variant); "Metric" is the source's; the two value columns give the score with reasoning and without, in the source's units. Where a source reports several reasoning formats, the row names the one shown. A2* marks a curriculum variant: C2SER's implicit model is its explicit model trained further on the same speech data, so the two arms do not share a starting point (Section 4.4).

Figure 1 draws the two patterns the table carries. Panel (a) shows how much of one method's improvement depends on the scoring rule. Panel (b) shows that the sign of a think-versus-answer difference tracks how the comparison was built more closely than what task it concerns: inference toggles cluster within a point of zero, matched training arms split, and prompts and released variants produce the largest differences in both directions.

Three regularities hold across the table. First, the largest differences favouring answer-only come from arms in which the reasoning condition had no rationale supervision or differs from the answer-only condition in more than reasoning (A3, A4, and A2 without warm-up). Second, the largest differences favouring reasoning in emotion come from single-label speech tasks with a structured format that requires the model to set down the spoken words and acoustic descriptors before deciding (Li et al., 2025b), or from a zero-shot prompt comparison whose answer-extraction rates are not reported, so that format and recognition cannot be separated (Section 4.3). Third, no row reports a variance for the think-versus-answer difference itself.

9. Crediting a reasoning gain

Algorithm 1 states the procedure that the preceding sections imply. It takes a reported think-versus-answer result and returns what the result can be credited to.

The procedure is deliberately conservative. It returns "test-time reasoning helps" only for a matched comparison in which both arms saw identical supervision and rewards, the metric cannot be gained by breadth or the set sizes are equal, and the difference exceeds a stated run-to-run variance. None of the comparisons in Table 1 meets all three conditions. EMO-RL's matched arms come closest: identical rewards and data, a single-label metric, no rationale warm-up reported for either arm - but with a structured template that changes what the model attends to before answering, no variance reported, and no stratification.

10. Six experiments that would settle the open part

Each experiment below is specified so that its outcome distinguishes the readings this paper has kept apart.

E1. The four-arm factorial. From one checkpoint, train four models on identical data and rewards: rationale warm-up then reasoning at inference; rationale warm-up then answer-only at inference (the C2SER design); no rationales, reasoning required; no rationales, answer-only. Run at least three seeds each and report the variance of each pairwise difference. The rationale main effect measures supervision; the inference main effect measures test-time reasoning; the interaction says whether reasoning at test time pays only after rationale

Algorithm 1. Reading a think-versus-answer result in emotion recognition input: result r comparing condition T (reasoning) with D (direct)

```
arm <- how D was produced: toggle | trained | prompt | variant
if arm == variant:
    return "deployment comparison; not attributable to reasoning"
if arm == prompt and parse failures not reported for both:
    return "format and extraction confounded with reasoning"

if metric pays for set size (hit rate, set recall, judge recall):
    if set sizes of T and D differ or are unreported:
        flag "operating point confounded"
        require precision or F1 on the same outputs

if arm == trained:
    if T saw rationales D did not (warm-up, distillation):
        credit <- "rationale supervision + format"
        if rationales were generated with the label in view:
            note "supervision is label-conditioned rationalisation"
    else:
        credit <- "format under matched supervision"
    if T used RL and D used SFT: credit <- credit + " + RL objective"
if arm == toggle:
    credit <- "test-time trace, given reasoning-trained weights"
    note "D is off the training distribution"

if no variance across seeds or runs reported:
    flag "difference may be within run-to-run noise"
if difference not stratified by annotator agreement:
    note "cannot locate effect relative to reference ceiling"
if no modality-masking or cue-recall check on T's traces:
    note "trace grounding unverified; interpretability claim open"
return credit, flags
```

supervision, as SARI suggests (Wen et al., 2025). This is the experiment the field most lacks.

E2. The metric panel. For every arm, report hit rate, set precision, set recall, F1 and the distribution of predicted set sizes on the same outputs, and add a fixed-size score in which each model's single most confident category is scored by accuracy. If the thinking paradox is an operating-point effect, it vanishes under the fixed-size score; if deliberation harms recognition, it survives.

E3. The rationale-content control. Replace the warm-up rationales with three alternatives of matched length: rationales written without the label in view, rationales written for a randomly permuted label, and neutral descriptions of the clip that do not argue for any emotion. If label-conditioned rationales outperform label-blind ones by a wide margin, the gain from reasoning supervision is distillation of cue-label associations through a stronger model; if neutral descriptions do as well, the gain is from richer perceptual description rather than reasoning.

E4. Modality ablation per arm. Mask video, audio and transcript in turn and report each arm's accuracy drop, alongside OPPO-style cue recall and masked-modality probes on the reasoning arm's traces (Han et al., 2026b). If reasoning gains come with larger drops when a non-text modality is masked, reasoning increased reliance on that modality; if they come with larger drops only when the transcript is masked, reasoning increased reliance on text, as the audio question-answering results suggest it may (Rouditchenko et al., 2025).

E5. Agreement-stratified evaluation. On datasets that release individual annotations, report the think-versus-answer difference separately for high- and low-agreement items, and report a held-out annotator's agreement with the remaining majority as the human reference on the same items (Uma et al., 2021; Yu et al., 2026). This locates any effect relative to the ceiling and tests the two opposite predictions of Section 6.2.

E6. Transcript routing. Give an answer-only model the transcript and acoustic descriptors that EMO-RL's structured format makes its reasoning model write out (Li et al., 2025b), as part of its input. If the answer-only model then matches the structured reasoning model, the structured format's benefit was routing attention to the words and acoustics, which an input change also supplies; if not, the writing itself contributes.

11. Objections

"MER-R1 shows that reasoning can be made to help, so the paradox is resolved." It shows that a method acting on rewards and confidence calibration moves the deliberated answer from 0.38 F1 points behind the direct answer to 0.94 ahead, in single runs on one backbone (Han et al., 2026a). That is a real result about the method. It does not show what written reasoning contributes, because the same method could be applied to an answer-only model, and that comparison was not run.

"Hit rate is the right metric because extra labels can be correct fine-grained emotions." The designers' argument is reasonable (Lian et al., 2025a), and this

paper does not claim that extra labels are wrong. It claims that a metric indifferent to them cannot adjudicate between procedures that produce sets of different sizes. Where fine-grained references exist, as on OV-MERD+, precision against them can be computed; where they do not, the set sizes should be reported so readers can see what the score pays for.

"EMO-RL settles the question in favour of reasoning for speech." It is the best matched evidence, and this paper treats it so. Its structured format requires the model to write down the speaker's words and acoustic features before deciding, and its own case study attributes the unstructured arm's errors to neglecting semantic content (Li et al., 2025b). The gain may come from routing the decision through the transcript rather than from reasoning as such. C2SER, also single-label speech with a two-part description before the label, found that writing the description at test time added nothing once the model had been trained on it (Zhao et al., 2025b). E6 would separate these.

"Interpretability alone justifies the reasoning." Only if the reasoning is faithful. Existing emotion models answer 41 to 74 percent of masked-modality probes incorrectly, affirming cues in a modality they were not given (Han et al., 2026b), and the text literature finds explanations that omit what drove the answer (Turpin et al., 2023; Chen et al., 2025). A trace that is fluent and cue-shaped but not causally tied to the label is a description, not an explanation.

"Released thinking variants are what users actually deploy, so A4 results are the relevant ones." For choosing a product, yes, and the EmoBench-M results are a useful warning that a thinking variant can be worse at emotion recognition by a large margin (Hu et al., 2025). For deciding whether to build reasoning into an emotion model, no: the variants differ in more than reasoning.

12. What is not known

Whether writing reasoning at test time improves emotion recognition, holding supervision, rewards and operating point fixed, is not known; no published study has run that comparison. How large run-to-run variance is for think-versus-answer differences on the standard emotion benchmarks is not known. Whether emotion reasoning traces name the cues that caused the label is not known for any reasoning-trained emotion model; masking probes show that the models often affirm cues in modalities they were not given. Whether reasoning helps more on low-agreement items or less is not known. The gap between published and re-run scores for AffectGPT-R1 in Table 2 is at least partly a change of backbone and training implementation; what accounts for the AffectGPT gap (2.81 points on the mean) is not known. Whether the large released-variant gaps on EmoBench-M come from reasoning, from other differences in post-training, or from answer extraction is not known.

13. Limitations

The evidence base is narrow where it matters most. The emotion-specific think-versus-answer comparisons come from a handful of studies, two of the central ones from one research group (Han et al., 2026a; 2026b), and the benchmark suite, the metric rules and several of the compared systems from another (Lian et al., 2024a; 2025a; 2025b; 2026). All of the emotion studies reviewed are preprints; none of their think-versus-answer differences carries a reported variance. The author worked from the published text and tables of each source, so where a source's prose and table disagree - as for the order of the two recall metrics in R1-Omni (Zhao et al., 2025a) - the paper reports the disagreement rather than resolving it.

Several comparisons in Table 1 come from neighbouring tasks - image perception, video question answering, audio question answering and text sentiment - and are included because they share the structure of the question, not because their results transfer to emotion. The annotator-agreement figure in Section 6.2 comes from one dataset's annotation process, at a granularity different from the scored categories, and is offered as a scale. EMO-RL's weighted accuracies on IEMOCAP in the high 80s are far above the self-supervised baselines it reports; the test protocol behind them was not fully recoverable from the text, and the paper uses those values only for the within-study comparison between arms.

The four settings in Section 3 were chosen with the reviewed results in view. That they sort the results is a consistency check, not a test; Section 10 states tests.

14. Conclusion

The thinking paradox in emotion recognition is, on its source's own numbers, mostly a result about how many emotions a model names and how the benchmark pays for naming them. Under the benchmark's hit rate, direct answers that list more emotions beat deliberated answers that list fewer; under F1 the difference is under a point and changes sign with the training recipe. The opposite verdict - that structured reasoning improves emotion recognition - comes from single-label comparisons in which the reasoning format also forces the model to write down the words and acoustics before deciding, and from training pipelines in which reasoning arrived together with rationales written by a stronger model that had seen the label. The largest differences in either direction come from comparisons that change more than reasoning. None of this says that reasoning is useless for reading emotion. It says that the field has not yet measured what reasoning adds, and that the measurement it needs - the same checkpoint, the same rewards, rationale supervision crossed with test-time reasoning, several seeds, a metric that cannot be gained by breadth, and a check that the trace looks at the face

and listens to the voice - is specified and inexpensive relative to the training runs already published.

References

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest, 20(1), 1-68. doi:10.1177/1529100619832930.

Bonzi, F., Mousavi, P., Subakan, C., & Ravanelli, M. (2026). Enhancing Audio Reasoning via Semantic Summary Prediction. Interspeech 2026. arXiv:2609.20849. doi:10.48550/arXiv.2609.20849.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335-359. doi:10.1007/s10579-008-9076-6.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., et al. (2025). Reasoning Models Don't Always Say What They Think. arXiv:2505.05410. doi:10.48550/arXiv.2505.05410.

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., et al. (2024). Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. NeurIPS 2024. arXiv:2406.11161. doi:10.48550/arXiv.2406.11161.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948. doi:10.48550/arXiv.2501.12948.

Gema, A. P., Hagele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., et al. (2025). Inverse Scaling in Test-Time Compute. Transactions on Machine Learning Research. arXiv:2507.14417. doi:10.48550/arXiv.2507.14417.

Guo, Z., Cui, W., Lin, G.-T., Tan, D., Li, J., Zheng, Q., et al. (2026). A Survey of Audio Reasoning in Multimodal Foundation Models. arXiv:2605.21008. doi:10.48550/arXiv.2605.21008.

Han, Z., Zhu, B., Tong, W., Qin, C., Wang, X., Zhang, J., et al. (2026a). MER-R1: Multimodal Emotion Reasoning via Slow-Fast Thinking Synergy. arXiv:2606.27652. doi:10.48550/arXiv.2606.27652.

Han, Z., Zhu, B., Tong, W., Shao, P., Song, P., Wang, X., et al. (2026b). Omni-Perception Policy Optimization for Multimodal Emotion Reasoning. ICML 2026. arXiv:2606.25325. doi:10.48550/arXiv.2606.25325.

Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., & Bethge, M. (2025). A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility. COLM 2025. arXiv:2504.07086. doi:10.48550/arXiv.2504.07086.

Hu, H., You, L., Xu, H., Wang, Q., Yu, F. R., Ma, F., et al. (2025). EmoBench-M: Benchmarking Emotional Intelligence for Multimodal Large Language Models. arXiv:2502.04424. doi:10.48550/arXiv.2502.04424.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., et al. (2023). Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702. doi:10.48550/arXiv.2307.13702.

Li, M., Zhong, J., Zhao, S., Lai, Y., Zhang, H., Zhu, W. B., et al. (2025a). Think or Not Think: A Study of Explicit Thinking in Rule-Based Visual Reinforcement Fine-Tuning. NeurIPS 2025. arXiv:2503.16188. doi:10.48550/arXiv.2503.16188.

Li, P., Zhao, B., Kang, Z., Peng, J., Qu, X., He, Y., et al. (2025b). EMO-RL: Emotion-Rule-Based Reinforcement Learning Enhanced Audio-Language Model for Generalized Speech Emotion Recognition. Findings of EMNLP 2025. arXiv:2509.15654. doi:10.48550/arXiv.2509.15654.

Lian, Z., Sun, H., Sun, L., Chen, K., Xu, M., Wang, K., et al. (2023a). MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning. arXiv:2304.08981. doi:10.48550/arXiv.2304.08981.

Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z., Gu, H., et al. (2023b). GPT-4V with Emotion: A Zero-shot Benchmark for Generalized Emotion Recognition. arXiv:2312.04293. doi:10.48550/arXiv.2312.04293.

Lian, Z., Sun, H., Sun, L., Chen, H., Chen, L., Gu, H., et al. (2024a). OV-MER: Towards Open-Vocabulary Multimodal Emotion Recognition. arXiv:2410.01495. doi:10.48550/arXiv.2410.01495.

Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., et al. (2024b). MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition. arXiv:2404.17113. doi:10.48550/arXiv.2404.17113.

Lian, Z., Chen, H., Chen, L., Sun, H., Sun, L., Ren, Y., et al. (2025a). AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. arXiv:2501.16566. doi:10.48550/arXiv.2501.16566.

Lian, Z., Zhang, F., Zhang, Y., Tao, J., Liu, R., Chen, H., et al. (2025b). AffectGPT-R1: Leveraging Reinforcement Learning for Open-Vocabulary Multimodal Emotion Recognition. arXiv:2508.01318. doi:10.48550/arXiv.2508.01318.

Lian, Z., Zhang, F., Chen, L., Zhang, Y., Liu, R., Wu, J., et al. (2026). AffectGPT-RL: Revealing Roles of Reinforcement Learning in Open-Vocabulary Emotion Recognition. arXiv:2605.06126. doi:10.48550/arXiv.2605.06126.

Liu, C., Xu, Z., Wei, Q., Wu, J., Zou, J., Wang, X. E., et al. (2025). More Thinking, Less Seeing? Assessing Amplified Hallucination in Multimodal Reasoning Models. arXiv:2505.21523. doi:10.48550/arXiv.2505.21523.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse. arXiv:2410.21333. doi:10.48550/arXiv.2410.21333.

Liu, S., Zhuge, M., Zhao, C., Chen, J., Wu, L., Liu, Z., et al. (2026a). VideoAuto-R1: Video Auto Reasoning via Thinking Once, Answering Twice. CVPR 2026. arXiv:2601.05175. doi:10.48550/arXiv.2601.05175.

Liu, Y., Zhang, L., Li, H., Shi, H., Ding, Y., Qu, L., et al. (2026b). Follow the Clues, Frame the Truth: Hybrid-evidential Deductive Reasoning in Open-Vocabulary Multimodal Emotion Recognition. arXiv:2603.16463. doi:10.48550/arXiv.2603.16463.

Ma, Z., Chen, Z., Wang, Y., Chng, E. S., & Chen, X. (2025). Audio-CoT: Exploring Chain-of-Thought Reasoning in Large Audio Language Model. arXiv:2501.07246. doi:10.48550/arXiv.2501.07246.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. ACL 2019. arXiv:1810.02508. doi:10.48550/arXiv.1810.02508.

Rouditchenko, A., Bhati, S., Araujo, E., Thomas, S., Kuehne, H., Feris, R., et al. (2025). Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio LLM? arXiv:2505.09439. doi:10.48550/arXiv.2505.09439.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 22(1), 36-71. doi:10.1016/0010-0285(90)90003-M.

Shao, R., Li, S. S., Xin, R., Geng, S., Wang, Y., Oh, S., et al. (2025). Spurious Rewards: Rethinking Training Signals in RLVR. arXiv:2506.10947. doi:10.48550/arXiv.2506.10947.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., et al. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300. doi:10.48550/arXiv.2402.03300.

Shou, Y., Meng, T., Ai, W., & Li, K. (2025). Multimodal Large Language Models Meet Multimodal Emotion Recognition and Reasoning: A Survey. arXiv:2509.24322. doi:10.48550/arXiv.2509.24322.

Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., et al. (2024). To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. ICLR 2025. arXiv:2409.12183. doi:10.48550/arXiv.2409.12183.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS 2023. arXiv:2305.04388. doi:10.48550/arXiv.2305.04388.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from Disagreement: A Survey. Journal of Artificial Intelligence Research, 72, 1385-1470. doi:10.1613/jair.1.12752.

Vamvourellis, D., & Mehta, D. (2025). Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis. arXiv:2506.04574. doi:10.48550/arXiv.2506.04574.

Wang, D., Liu, S., Zhang, T., Chen, Y., Li, J., & Meng, H. (2026). EmotionThinker: Prosody-Aware Reinforcement Learning for Explainable Speech Emotion Reasoning. ICLR 2026. arXiv:2601.15668. doi:10.48550/arXiv.2601.15668.

Wang, S., Yang, W., Long, X., Ganguly, D., Wang, Q., Chaudhary, V., et al. (2025). Demystifying Hybrid Thinking: Can LLMs Truly Switch Between Think and No-Think? arXiv:2510.12680. doi:10.48550/arXiv.2510.12680.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903. doi:10.48550/arXiv.2201.11903.

Wen, C., Guo, T., Zhao, S., Zou, W., & Li, X. (2025). SARI: Structured Audio Reasoning via Curriculum-Guided Reinforcement Learning. arXiv:2504.15900. doi:10.48550/arXiv.2504.15900.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. Journal of Personality and Social Psychology, 60(2), 181-192. doi:10.1037/0022-3514.60.2.181.

Xie, Z., Lin, M., Liu, Z., Wu, P., Yan, S., & Miao, C. (2025). Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models. arXiv:2503.02318. doi:10.48550/arXiv.2503.02318.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., et al. (2025). Qwen2.5-Omni Technical Report. arXiv:2503.20215. doi:10.48550/arXiv.2503.20215.

Yang, C.-K., Ho, N., Piao, Y.-T., & Lee, H.-y. (2025). SAKURA: On the Multi-hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information. Interspeech 2025. arXiv:2505.13237. doi:10.48550/arXiv.2505.13237.

Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., Peng, Y., et al. (2025). Perception-R1: Pioneering Perception Policy with Reinforcement Learning. arXiv:2504.07954. doi:10.48550/arXiv.2504.07954.

Yu, X., Dong, J., Honorio, J., Ghosh, A., Jia, H., & Dang, T. (2026). Disentangling Reasoning in Large Audio-Language Models for Ambiguous Emotion Prediction. arXiv:2603.08230. doi:10.48550/arXiv.2603.08230.

Zhang, F., Cheng, Z., Deng, C., Li, H., Lian, Z., Chen, Q., et al. (2025). MME-Emotion: A Holistic Evaluation Benchmark for Emotional Intelligence in Multimodal Large Language Models. arXiv:2508.09210. doi:10.48550/arXiv.2508.09210.

Zhao, J., Wei, X., & Bo, L. (2025a). R1-Omni: Explainable Omni-Multimodal Emotion Recognition with Reinforcement Learning. arXiv:2503.05379. doi:10.48550/arXiv.2503.05379.

Zhao, Z., Zhu, X., Wang, X., Wang, S., Geng, X., Tian, W., et al. (2025b). Steering Language Model to Stable Speech Emotion Recognition via Contextual Perception and Chain of Thought. IEEE Transactions on Audio, Speech and Language Processing. arXiv:2502.18186. doi:10.48550/arXiv.2502.18186.

Zheng, K., Zhao, Q., & Li, L. (2025). Reassessing the Role of Chain-of-Thought in Sentiment Analysis: Insights and Limitations. arXiv:2501.08641. doi:10.48550/arXiv.2501.08641.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24085149