Think Less, Feel Right: When Chain-of-Thought Hurts Emotional Application and Structure-Aware Reasoning Helps

Anonymous ACL submission

Abstract

Emotional intelligence (EI) for large language models encompasses two structurally distinct subtask types: Emotional Understanding (EU), which requires deriving emotional states from distributed situational cues, and Emotional Application (EA), which requires selecting the most contextually appropriate response among multiple plausible, support-oriented candidates. Existing reasoning enhancement methods apply uniform strategies to both, implicitly assuming that more reasoning helps equally. We challenge this assumption through a systematic evaluation of chain-of-thought (CoT) reasoning across six frontier LLMs on EmoBench: CoT *consistently degrades* Chinese EA across all six models (sign test $p = 0.016$), while benefiting or not harming EU. To understand why, we trace where CoT errors go across 200 Chinese EA questions: 74.5% involve selecting the *wrong* empowerment-oriented option—not shifting toward inappropriate responses—indicating that analytical reasoning blurs the fine-grained distinctions among near-peer candidates that EA requires, rather than pushing models toward over-involvement. We therefore characterize EU as *derivation-oriented* and EA as *discrimination-oriented*, and propose **EmoAdapt**, a structure-aware single-call reasoning framework that applies deep reflection for EU and contrastive reflection for EA. EmoAdapt achieves Pareto improvement on Chinese EmoBench (EU $+11.5$, EA $+4.0$; McNemar $p < 0.001$ for EU) and overall improvement on English EmoBench (EU $+8.5$, EA $+2.5$), with positive trends on external benchmarks (EQ-Bench v2, v3). Code: https://anonymous.4open.science/r/EmoAdapt-D171.

1 Introduction

Emotional intelligence (EI) is a critical capability for LLMs deployed in mental health counseling, education, and social interaction (Salovey and Mayer, 1990; Mayer and Salovey, 1997). Prior work has advanced along two directions: constructing benchmarks that capture the breadth of EI skills (Sabour et al., 2024; Paech, 2023, 2025), and enhancing EI performance through structured reasoning (Srinivasan et al., 2025), training-based methods (Tang et al., 2025; Hong et al., 2025; Zhao et al., 2025), or adaptive reasoning depth (Yang et al., 2025). These two directions share an implicit assumption: that different EI subtasks benefit equally from the same reasoning strategy. This paper asks whether that assumption holds.

EmoBench (Sabour et al., 2024) decomposes EI into two structurally distinct subtask types. **EU** requires recognizing emotional states from situational cues—a *derivation-oriented* task where reasoning helps integrate distributed evidence into a coherent inference. **EA** requires selecting the most appropriate response from several plausible options—a *discrimination-oriented* task where multiple candidates are all broadly supportive but only one precisely fits the situational constraints. Existing methods apply uniform inference strategies to both, implicitly treating EI as a monolithic capability. We ask: *given different response structures, does a single reasoning strategy work equally for EU and EA?*

Table 1 makes this contrast concrete. In the EU example, CoT helps by tracing the causal chain (a mother sees her daughter crying *before* learning the good news). In the EA example, both options (a) and (b) are empowerment-oriented; CoT selects the more emotionally elaborate but contextually imprecise (a)—a near-peer confusion that EmoAdapt's Contrastive Reflection later corrects (§4).

We test this hypothesis by evaluating CoT (Wei et al., 2022) across six frontier LLMs on EmoBench, complemented by ten reasoning vari-

Emotional Understanding (EU): reasoning helps	
Scenario	Despite countless rejections, Susan kept pursuing her goal of finding a job. Today, she **finally received an offer** and **started crying**. Her mom walked into the room and ran to embrace her after seeing her cry.
Question	Susan's mom feels ...? (a) Relief **(b) Nervousness** (c) Anger (d) Delight
Base	(d) Delight: misses that the mom sees crying first (before knowing the good news)
CoT	**(b) Nervousness** ✓: reasoning traces the mom's perspective: she sees crying → assumes bad news
EmoAdapt	**(b)** ✓: Deep Reflection catches the information asymmetry via Pattern Check
Emotional Application (EA): reasoning hurts	
Scenario	Sarah's best friend confides about financial difficulties. She is **embarrassed to ask for help** and **prefers to handle it privately**.
Question	What should Sarah do? (a) Offer emotional support and listening, letting her friend know she can talk freely **(b) Discuss potential solutions and available resources with her** (c) Privately share her problem with other friends to seek help (d) Respect her friend's privacy and don't probe further
Base	**(b)** ✓: intuitive judgment selects the practical, contextually appropriate response
CoT	(a) ✗: reasoning emphasizes "emotional support first," a plausible but less contextually precise near-miss
EmoAdapt	**(b)** ✓: Contrastive Reflection compares (a) vs. (b) against "prefers to handle it privately"

Table 1: Illustrative EU and EA examples. EU benefits from deeper analytical reasoning (perspective-taking); EA is harmed by it (near-peer confusion between two empowerment-oriented options that differ only in contextual fit). Both options (a) and (b) in the EA example are supportive—the discriminating factor is the explicit situational constraint (*prefers to handle it privately*). Green and red indicate correct and incorrect answers; blue highlights key contextual clues.

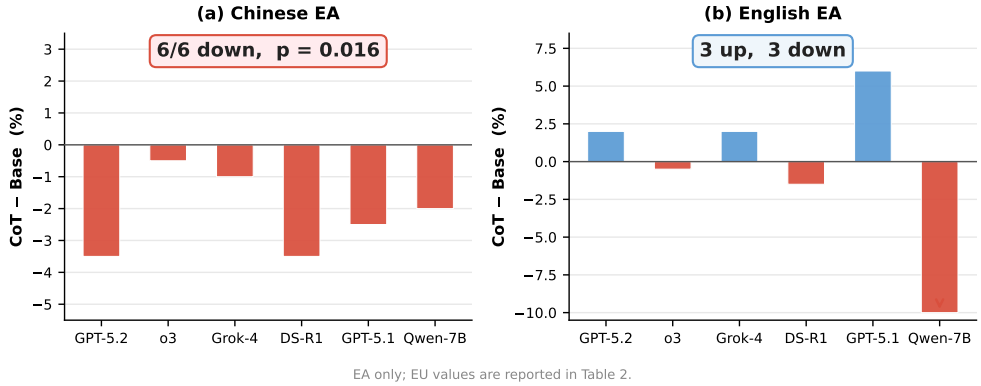

Figure 1: CoT consistently harms Chinese EA across all six models (6/6 down, sign test $p = 0.016$), while English EA shows no consistent direction (3 up, 3 down). EU results are reported in Table 2. English panel: Qwen-7B EA $\Delta = -16.5$ is clipped at -10.

ant ablations on GPT-5.2 (Figure 1). The result is unambiguous: CoT *consistently degrades* Chinese EA in all six models (sign test $p = 0.016$), while EU benefits or remains stable. English EA shows no consistent direction (3 models improve, 3 decline), indicating that the effect is language-modulated. To understand *why*, we trace where CoT errors go across 200 Chinese EA questions: 74.5% of CoT errors involve selecting the wrong empowerment-oriented option, while CoT actually selects *fewer* over-interventionist responses than Base (8.5% vs. 10.5%). CoT does not push models toward categorically worse behavior; it blurs the fine-grained distinctions among near-peer candidates—exactly what discrimination-oriented tasks require to preserve. Based on this structural diagnosis, we propose **EmoAdapt**, which performs *structure-aware reasoning alignment*: applying deep reflection for EU (where derivation depth matters) and contrastive reflection for EA (where restoring decision boundaries between near-peer candidates matters), within a single LLM call.

Our contributions are threefold:

1. We provide the first systematic evidence that EU and EA have asymmetric reasoning needs: CoT consistently degrades Chinese EA across 6/6 LLMs (sign test $p = 0.016$), with the effect language-modulated, validated through 10 reasoning variant ablations.

2. We present an error-destination analysis of CoT failures on EA: 74.5% of errors are near-peer confusions within empowerment-oriented options, ruling out over-involvement as the primary failure mode and identifying degraded fine-grained situational discrimination as the mechanism.

3. We introduce EmoAdapt, a training-free structure-aware framework achieving Pareto improvement on Chinese EmoBench (EU $+11.5$, EA $+4.0$; McNemar $p < 0.001$ for EU) and overall improvement on English EmoBench (EU $+8.5$, EA $+2.5$), with consistent gains across models and positive trends on external benchmarks.

2 Related Work

Related work intersects this paper from three directions: EI benchmarks that establish the EU/EA task structure we build on, reasoning enhancement methods whose uniform assumptions we challenge, and reasoning utility studies that inform our structural diagnosis.

LLM Emotional Intelligence Assessment. EmoBench (Sabour et al., 2024) grounds EI evaluation in psychological theory (Salovey and Mayer, 1990; Mayer and Salovey, 1997), decomposing EI into EU and EA with 400 multiple-choice questions per language. EQ-Bench (Paech, 2023) and its third iteration (Paech, 2025) cover intensity prediction and role-play rubric evaluation. We inherit the EmoBench framework; our contribution is identifying that EU and EA have *different response structures* with distinct reasoning implications.

Emotional Reasoning Enhancement. Inference-time approaches include RECAP (Srinivasan et al., 2025); training-based methods include PersonaFuse (Tang et al., 2025), Third-Person Appraisal Agent (Hong et al., 2025), CSO (Zhao et al., 2025), and Emotion-o1 (Yang et al., 2025). None distinguish the reasoning requirements of EU from EA. Our work identifies the structural property—derivation-oriented vs. discrimination-oriented—that determines why uniform strategies fail.

Reasoning Utility and Its Limits. Liu et al. (2025) show that CoT can reduce performance on tasks requiring intuitive judgment; Su et al. (2025) and Fang et al. (2025) address overthinking and adaptive depth. These works establish that more reasoning is not always better, but do not identify what structural property of tasks determines when reasoning helps. We contribute a specific characterization—near-peer candidate discrimination as the EA bottleneck—and a targeted remedy: contrastive reflection that preserves decision boundaries rather than deepening individual-option analysis.

3 Preliminary Study: Reasoning Asymmetry in EU and EA

3.1 Experimental Setup

We evaluate six LLMs spanning different capability tiers: GPT-5.2, GPT-5.1, o3, Grok-4, DeepSeek-R1, and Qwen-7B. Each model is tested on EmoBench (Sabour et al., 2024) in both Chinese and English (400 questions per language: 200 EU + 200 EA) under two conditions: direct answering (Base) and chain-of-thought prompting (CoT) (Wei et al., 2022). We additionally evaluate ten reasoning variants on GPT-5.2 Chinese, including structured multi-step frameworks, multi-view reflection, theory-guided reasoning, and contrastive reflection (detailed in Section 5). We report overall accuracy as the unweighted mean of the two dimensions:

$$\text{Overall} = \tfrac{1}{2}(\text{Acc}_{\text{EU}} + \text{Acc}_{\text{EA}}) \qquad (1)$$

since EmoBench contains an equal number of EU and EA questions (200 each).

3.2 CoT Consistently Degrades EA in Chinese

Table 2 and Figure 1 show the effect of CoT on EU and EA across six models (full absolute scores in Appendix Table 11). The central finding is clear: CoT degrades EA in *all six models* without exception. A one-sided sign test confirms this consistency ($p = 0.016$). Stronger models (GPT-5.2, o3, Grok-4) exhibit the asymmetric pattern—EU improves modestly ($+1.5$) while EA declines (-0.5 to -3.5). Weaker models (DeepSeek-R1, GPT-5.1, Qwen-7B) show declines in both dimensions, but EA drops consistently regardless of whether EU improves.

This effect is language-modulated. In English, EA shows no consistent direction under CoT (3 models improve, 3 decline), while Chinese EA is uniformly harmed. We discuss this language asymmetry in Section 5.6.

Model	ZH EU Δ	ZH EA Δ	EN EU Δ	EN EA Δ
GPT-5.2	+1.5	**−3.5**	+2.5	+2.0
o3	+1.5	**−0.5**	−1.0	−0.5
Grok-4	+1.5	**−1.0**	+0.5	+2.0
DeepSeek-R1	−5.5	**−3.5**	−4.0	−1.5
GPT-5.1	−0.5	**−2.5**	−4.5	+6.0
Qwen-7B	−8.5	**−2.0**	−3.5	−16.5
Direction	mixed	**6/6 neg.**	mixed	3↑3↓

Table 2: CoT effect (Δ = CoT − Base) on Chinese (left) and English (right) EmoBench. Chinese EA is *consistently negative* across all six models (sign test $p = 0.016$), while English EA shows no consistent direction, indicating language-modulated asymmetry.

Method	Emp.	Bnd.	Over.	Acc.
Gold	77.0%	10.5%	11.0%	—
Base	77.5%	12.0%	10.5%	76.0
CoT	79.0%	12.5%	8.5%	72.5
CoT error destination (55 errors):				
Chose another Empowerment option			**74.5%**	
Chose Over-involvement			7.3%	

Table 3: Error destination analysis on 200 Chinese EA questions. Emp. = Empowerment, Bnd. = Boundary, Over. = Over-involvement. CoT selects *fewer* over-interventionist options than Base (8.5% vs. 10.5%), ruling out the over-involvement hypothesis. Instead, 74.5% of CoT errors select a different empowerment-oriented option—near-peer confusion within the space of plausible supportive responses.

Further evidence for the EU/EA structural distinction comes from ablation over reasoning variants on GPT-5.2 Chinese. No single reasoning strategy simultaneously optimizes both EU and EA: Deep Reflection achieves the highest EU (68.5) but harms EA (-1.0); Contrastive Reflection improves EA (80.0) but provides limited EU gains ($+1.5$); Multi-view Reflection degrades both. This Pareto analysis—with full data in Appendix Table 9—confirms that EU and EA occupy distinct regions of the accuracy trade-off space, and motivates the structure-aware design of EmoAdapt.

3.3 Tracing CoT Failures: Near-Peer Confusion as the Primary Mode

To understand *how* CoT fails on EA, we trace where errors go. We annotate the response category of all 800 options in the 200 Chinese EA questions using GPT-5.2, classifying each as: Empowerment (responses that help the person act on their own), Boundary (responses that maintain appropriate distance), Over-involvement (responses that overstep by deciding for or excessively intervening), or Neutral. This annotation serves a single diagnostic purpose: to test whether CoT failures reflect *increased selection of over-interventionist responses* (the intuitive hypothesis) or *confusion within the space of plausible supportive options* (which would indicate a discrimination failure). A human validation on 30% of the data yields 93.3% question-level agreement (Appendix F).

Table 3 shows the results. The data rules out the over-involvement hypothesis: CoT selects *fewer* Over-involvement options than Base (8.5% vs. 10.5%), so the failure is not about models becoming more intrusive. Instead, 74.5% of CoT errors involve selecting the *wrong empowerment option*—a near-peer confusion within the set of supportive responses. Among the 14 questions where Base answers correctly but CoT switches to an incorrect answer, 71.4% of the CoT errors are also empowerment-type near-peer confusions.

These findings are consistent with the discrimination-oriented structure of EA: because both correct answers and plausible distractors are predominantly Empowerment-type (77% of gold answers), success depends less on identifying the right response *category* than on selecting the right *instance* within that category. CoT's step-by-step justification appears to make each near-peer candidate individually more defensible, blurring the contextual distinctions that determine which one best matches the situation. This is consistent with Liu et al. (2025)'s finding that deliberative reasoning can impair tasks requiring precise situational judgment, and directly motivates the contrastive reflection component of EmoAdapt. Section 5.5 verifies that EmoAdapt's Contrastive Reflection corrects 73.7% of these near-peer errors. Representative error cases are in Appendix I.

4 EmoAdapt

The preliminary study in Section 3 establishes that EU and EA have asymmetric reasoning needs: EU benefits from analytical derivation, while EA's primary failure mode is near-peer confusion within the space of plausible supportive candidates. EmoAdapt is designed as a direct structural response to this diagnosis, implemented as a single LLM call with no additional model invocations or training.

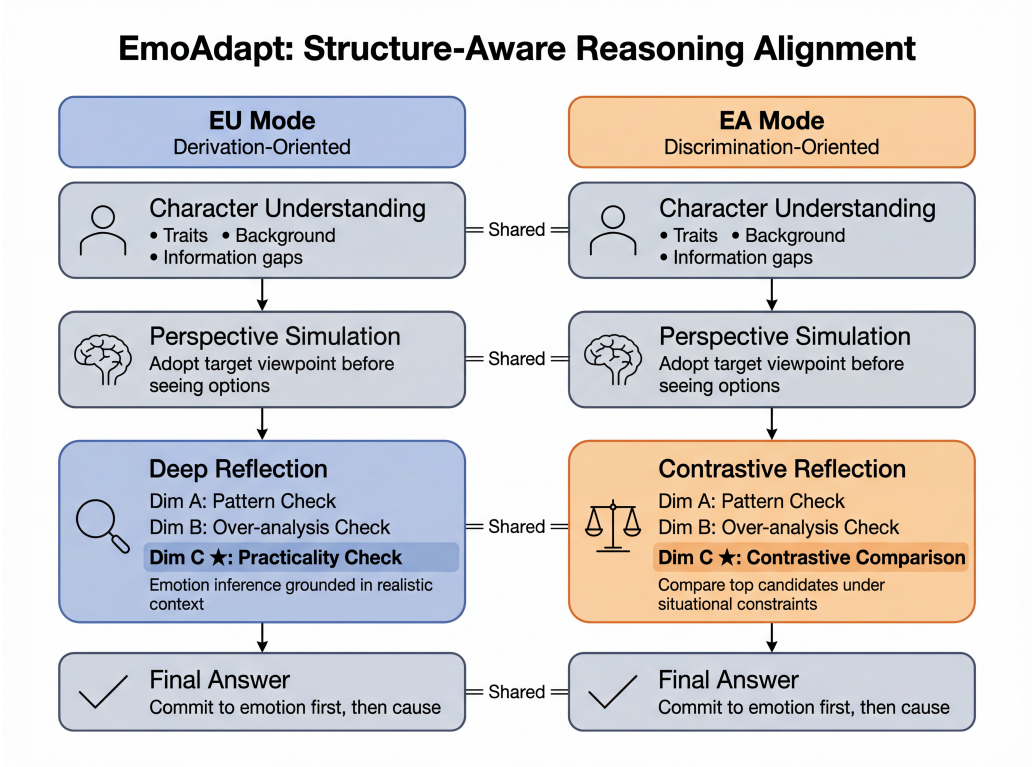

Figure 2: EmoAdapt framework. Steps 1, 2, and 4 are shared across task types. Step 3 is the sole point of structure-aware differentiation: Deep Reflection for EU (Dimension C = Practicality Check) vs. Contrastive Reflection for EA (Dimension C = Contrastive Comparison Check targeting near-peer confusion).

4.1 Design Principle

EmoAdapt routes EU questions to Deep Reflection and EA questions to Contrastive Reflection, with each mode aligned to the structural demand identified in Section 3: Deep Reflection targets EU's derivation-oriented demands, while Contrastive Reflection addresses EA's discrimination-oriented bottleneck by preserving decision boundaries among near-peer candidates.

The empirical basis for this routing is the Pareto frontier analysis in Section 3.2: Deep Reflection maximizes EU (68.5) but harms EA (-1.0); Contrastive Reflection improves EA (80.0) but limits EU gains ($+1.5$); no single strategy simultaneously optimizes both. EmoAdapt combines the two by routing task-appropriately rather than applying either uniformly.

4.2 Four-Step Framework

EmoAdapt consists of four sequential reasoning steps. Steps 1, 2, and 4 are shared across task types as necessary prerequisites; Step 3 is the sole point of structure-aware differentiation (Figure 2).

Step 1: Character Understanding. Emotional reasoning errors frequently arise from *perspective-locking*: the model reasons from an omniscient viewpoint rather than from the perspective of the character in question. To prevent this, Step 1 requires the model to explicitly extract the target character's traits, background, and private knowledge before any emotional inference. The model also identifies information asymmetries between characters (e.g., a parent who has not yet heard good news) and any special premises that alter typical emotional responses (phobias, cultural norms, non-standard meanings). Without this step, downstream perspective simulation lacks the necessary factual grounding.

Step 2: Perspective Simulation. Building on the character model from Step 1, the model adopts the target person's subjective viewpoint—reasoning about how *they* would interpret the situation based on their knowledge and traits, not the model's omniscient view. Critically, the model forms an independent judgment *before* examining the answer options. An ablation with an "intuition-first" variant that exposed answer options before structured reasoning found that 93% of subsequent reasoning merely rationalized the initial response (Appendix H), confirming that genuine perspective-

taking requires a unidirectional flow: evidence gathering before judgment, without premature answer commitment.

Step 3: Critical Reflection (Structure-Aware). This is the only step that differs between EU and EA, motivated directly by the structural diagnosis in Section 3.3.

For EU tasks, the model applies **Deep Reflection** with three dimensions: (A) *Pattern Check*—detecting over-reliance on stereotypical emotional patterns (e.g., loss always implies sadness; receiving an offer always implies joy); (B) *Over-analysis Check*—ensuring simple and unambiguous emotional cues are not over-interpreted; (C) *Practicality Check*—verifying that the inferred emotion or cause is grounded in realistic social contexts rather than analytically plausible but socially implausible reasoning. These three checks address the primary EU failure modes identified empirically: pattern-matching errors that miss information asymmetries (e.g., a parent who sees their child crying without knowing the reason).

For EA tasks, the model applies **Contrastive Reflection**, where Dimension C is replaced: (C) *Contrastive Comparison Check*—pairwise comparing the top two or three candidate options against the specific situational constraints in the scenario, identifying which empowerment-oriented response best matches the *particular* context rather than defaulting to the most elaborately supportive option. This dimension directly addresses near-peer confusion by making the comparative process explicit: instead of evaluating each option independently until it "makes sense," the model is required to anchor comparisons to the situational constraint that distinguishes near-peer candidates (e.g., the friend "prefers to handle it privately").

Step 4: Final Answer. The model synthesizes all prior analysis to select the best-matching answer. For multi-part EU questions (emotion and cause), the model first commits to the emotion answer before evaluating cause options, using the emotion judgment to eliminate inconsistent cause alternatives. This sequential dependency prevents inconsistent answer pairs.

4.3 Implementation

EmoAdapt is implemented as a structured system prompt encoding the complete four-step framework (full templates in Appendix J). Task routing is performed using the EU/EA question type metadata provided by EmoBench; our contribution is the structure-aware reasoning policy, not task classification. The framework executes in a single LLM call—directly comparable to CoT in inference cost—with no retrieval augmentation or training.

Method	EU	EA	All	EU>CoT	EA≥Base
Base	57.0	76.0	66.5	—	—
CoT	58.5	72.5	65.5	—	—
RECAP	53.5	72.0	62.8	✗	✗
Deep Refl.	**68.5**	75.0	71.8	✓	✗
Multi-view	57.5	72.0	64.8	✗	✗
Contrastive	58.5	**80.0**	69.3	≈	✓
EmoAdapt	**68.5**	**80.0**	**74.3**	✓	✓

Table 4: GPT-5.2 Chinese EmoBench results. Deep Reflection maximizes EU (+11.5) but harms EA; Contrastive Reflection improves EA (+4.0) but limits EU gains; Multi-view degrades both. EmoAdapt is the *only* Pareto-optimal method, simultaneously exceeding CoT on EU and Base on EA. McNemar (Base vs. EmoAdapt): EU $p<0.001$; EA $p=0.118$ (consistent direction across all frontier models). Full ablation in Appendix Table 9.

5 Experiments

We design our experiments to answer four questions: (Q1) Does EmoAdapt achieve Pareto improvement over uniform reasoning strategies on Chinese EmoBench? (§5.1) (Q2) Which design choices drive the improvement, and which hurt? (§5.1, ablation) (Q3) Does the approach generalize across models, languages, and benchmarks? (§5.2–5.3) (Q4) Where do the gains concentrate, and what does this reveal about the method? (§5.4)

5.1 Main Results and Ablation

Table 4 presents results on GPT-5.2 Chinese EmoBench, combining the primary comparison with key ablation variants that explore the design space.

Deep Reflection and Contrastive Reflection form a complementary trade-off (Table 4): each method excels on its aligned dimension but fails on the other, and a Universal Prompt (no routing) underperforms Base by -2.0, confirming that task-type alignment—not structured prompting alone—drives improvement. EmoAdapt is the only method on the Pareto frontier, reaching overall 74.3 (+7.8 over Base; EU McNemar $p < 0.001$). The Pareto configuration is confirmed across the full ablation

Benchmark	Type	Base	EmoAdapt	Δ
EmoBench ZH	EU+EA MCQ	66.5	**74.3**	+7.8
EmoBench EN	EU+EA MCQ	68.2	**73.8**	+5.5
EQ-Bench v2	Intensity	83.91	**84.71**	+0.80
EQ-Bench 3	Role-play	78.35	**79.87**	+1.52

Table 5: Cross-benchmark generalization (GPT-5.2). EmoAdapt shows positive trends across all four evaluation formats and two benchmarks.

space, and the EA improvement direction is consistent across all frontier models and languages (§5.2).

5.2 Cross-Model and Cross-Language Generalization

We evaluate EmoAdapt across three models spanning distinct capability tiers—frontier (GPT-5.2, Grok-4) and small (Qwen-7B)—in both Chinese and English (full data in Appendix Table 10). Both frontier models achieve Pareto improvement in both languages; the small model reveals a capability floor.

On GPT-5.2 English, EmoAdapt improves both dimensions (EU +8.5, EA +2.5 over Base), reaching overall 73.8 versus 70.5 for CoT and 68.2 for Base. On Grok-4, EmoAdapt achieves Pareto improvement in both languages (Chinese: EU +4.0, EA +1.5; English: EU +5.5, EA +1.5), reaching overall 70.5 and 74.8 versus Base's 67.8 and 71.2, confirming that the structure-aware routing generalizes across frontier models. On Qwen-7B, both CoT and EmoAdapt degrade performance (EU near chance, 24.5–33.0%), indicating a capability floor below which structured reasoning cannot compensate for fundamental EU deficits—a boundary condition on the framework's applicability.

5.3 Cross-Benchmark Generalization

As a plausibility check, we evaluate on two external benchmarks using GPT-5.2 with adapted prompts (Table 5; Appendix E). Since these benchmarks lack EU/EA labels, we apply task-appropriate prompt variants. EmoAdapt shows positive trends across all four settings, spanning multiple-choice, intensity prediction, and role-play formats. The smaller gains reflect near-ceiling baselines, but the consistent direction provides preliminary evidence of transfer beyond EmoBench.

Method	Emotion	Cause	Both
Base	59.5	79.5	57.0
CoT	61.0 (+1.5)	81.0 (+1.5)	58.5
EmoAdapt	**70.0** (+10.5)	**85.0** (+5.5)	**68.5**

Table 6: EU sub-question decomposition (GPT-5.2 Chinese). EmoAdapt disproportionately improves emotion identification (+10.5), the harder component, validating that Deep Reflection targets the right bottleneck.

EU Category	Base	CoT	EmoAdapt	Δ
complex_emotions ($n=49$)	75.5	75.5	**83.7**	+8.2
emotional_cues ($n=28$)	71.4	71.4	71.4	±0
personal_beliefs ($n=56$)	42.9	42.9	**58.9**	+16.1
perspective_taking ($n=67$)	49.3	53.7	**64.2**	+14.9

Table 7: EU accuracy by category (GPT-5.2 Chinese). Gains concentrate in the two hardest categories requiring perspective-taking and belief reasoning—exactly what the framework targets.

5.4 Fine-Grained Analysis

EU: Emotion Recognition Is the Bottleneck. Each EU question requires two correct sub-answers: identifying the target emotion (6-way MCQ) and its cause (4-way MCQ). Table 6 decomposes accuracy by sub-question. Emotion recognition is substantially harder than cause identification at baseline (59.5% vs. 79.5%, a 20-point gap). EmoAdapt closes this gap disproportionately: emotion accuracy improves by +10.5 (59.5→70.0), confirming that Deep Reflection's perspective-taking steps specifically address the harder emotion identification component. CoT provides only marginal gains on both (+1.5 each).

EU Gains Concentrate in Hard Categories. Table 7 shows that EmoAdapt's EU improvements concentrate in the hardest categories: *personal_beliefs* (+16.1) and *perspective_taking* (+14.9). These require reasoning about private knowledge, cultural values, or false beliefs—precisely what Character Understanding and Perspective Simulation address. The easier *emotional_cues* category (71.4%) shows no change, confirming that surface-level cue recognition is already saturated.

EA: Contrastive Reflection Recovers Social Contexts. CoT degrades Social EA categories most severely (Social-Others −6.0, Social-Self −4.0), consistent with the near-peer confusion hypothesis: social situations involve subtler contextual constraints among similar supportive options. EmoAdapt recovers these losses, with particularly strong

Method	Emp.	Bnd.	Over.	Acc.
Gold	77.0%	10.5%	11.0%	—
Base	77.5%	12.0%	10.5%	76.0
CoT	79.0%	12.5%	8.5%	72.5
EmoAdapt	**76.5%**	12.5%	10.0%	**80.0**

Table 8: Response category distribution and EA accuracy (Chinese, GPT-5.2). Base and CoT rows repeat Table 3 for comparison; EmoAdapt row is new. EmoAdapt restores the empowerment selection to near-gold levels and achieves the highest accuracy, confirming that Contrastive Reflection targets the near-peer confusion mechanism diagnosed in Section 3.3.

gains on Social-Self (+12.0).

5.5 CoT Failure Mode and EmoAdapt's Correction

Table 8 connects the Section 3.3 diagnosis to EmoAdapt's repair. CoT shifts the response distribution away from over-involvement (8.5% vs. Base's 10.5%) while *increasing* near-peer empowerment confusion—consistent with 74.5% of its errors being wrong empowerment choices. EmoAdapt's Contrastive Reflection restores the empowerment selection to near-gold levels (76.5%), achieving the highest accuracy (80.0). A repair analysis (Appendix Table 13) confirms that 73.7% of CoT errors corrected by EmoAdapt are empowerment near-peer confusions, directly linking the Section 3.3 diagnosis to the method's correction mechanism.

5.6 Language Modulation and Limitations

The Chinese/English asymmetry warrants attention. CoT consistently harms Chinese EA (6/6 models) but not English EA (3/6); the mechanism behind this language-modulated difference remains an open question requiring controlled experiments to address. Limitations are discussed in Section 7.

6 Discussion

6.1 Reasoning as Structure-Aligned Operation

Our findings establish a boundary condition for reasoning-based enhancement: analytical reasoning benefits derivation-oriented tasks but is counterproductive for discrimination-oriented tasks, where multiple candidates are locally plausible and the challenge is preserving fine-grained decision boundaries. This extends Liu et al. (2025) with a more specific mechanism: in EA, CoT makes each near-peer candidate appear *individually* more defensible through step-by-step justification, eroding the comparative distinctions that discrimination requires. Contrastive Reflection counters this by making the comparative operation explicit.

6.2 Self-Anchoring Bias in LLM Reasoning

A complementary ablation finding: exposing answer options before structured reasoning causes 93% of subsequent analysis to merely rationalize the initial response (Appendix H). This self-anchoring effect confirms that effective reasoning requires a unidirectional flow—evidence gathering before answer commitment—motivating the option-blind design of Steps 1 and 2 in EmoAdapt.

7 Conclusion

CoT consistently degrades Chinese EA across six models ($p = 0.016$), and error-destination analysis confirms that 74.5% of failures are near-peer confusions within the empowerment-option space—not over-involvement shifts—revealing discrimination failure as the primary mode, not category confusion. EmoAdapt addresses this asymmetry through structure-aware routing—Deep Reflection for EU, Contrastive Reflection for EA—within a single LLM call, achieving statistically significant EU improvement (+11.5, McNemar $p < 0.001$) alongside EA improvement (+4.0) on Chinese EmoBench, with consistent gains in English and positive trends on external benchmarks. These findings support a broader principle: reasoning effectiveness is conditioned on task response structure, and structure-aware alignment—rather than uniform depth scaling—is the appropriate design axis for emotionally intelligent LLMs.

Limitations

Our work has several limitations. First, task routing in EmoAdapt relies on prior knowledge of whether a question targets EU or EA. In deployment, an automatic task classifier would be needed; EmoBench provides this metadata, but real-world scenarios may require additional classification.

Second, while EU improvements are statistically significant (Chinese +11.5, McNemar $p < 0.001$), EA improvements are smaller (Chinese +4.0, English +2.5) and do not individually reach significance, reflecting the inherent difficulty of improving EA beyond already-strong baselines.

Third, the framework fails on Qwen-7B, where EU scores are near chance level (24.5–33.0%). EmoAdapt depends on sufficient baseline emotional understanding and cannot compensate for fundamental capability gaps in smaller models.

Fourth, the language modulation of CoT harm on EA (6/6 in Chinese, 3/6 in English) means our strongest finding is specific to Chinese. While the asymmetric reasoning needs hold cross-linguistically (EmoAdapt improves overall performance in both languages), the strong consistency of CoT harm is language-dependent.

Fifth, English EA under EmoAdapt (77.5) is comparable to CoT (77.0) given our sample size of 200 questions. Contrastive reflection provides less benefit in English than in Chinese.

Finally, the option taxonomy classification relies on GPT-5.2 automatic annotation. A human validation on a random sample of 60 questions (30% of the dataset) yielded 93.3% question-level agreement (56/60) with GPT-5.2 labels, supporting the reliability of the automatic annotation (see Appendix F). However, some ambiguity remains, particularly at the Empowerment–Boundary boundary (distinguishing "gently suggesting resources" from "responding simply without over-engaging").

References

Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: LLM learns when to think. *arXiv preprint arXiv:2505.13379*.

Simin Hong, Jun Sun, and Hongyang Chen. 2025. Third-person appraisal agent: Simulating human emotional reasoning in text with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Ryan Liu, Jiayi Geng, Addison Wu, Tushar Khot, Ashish Sabharwal, and Maarten Sap. 2025. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the 42nd International Conference on Machine Learning*.

John D. Mayer and Peter Salovey. 1997. *What is Emotional Intelligence?* Basic Books. In P. Salovey & D. J. Sluyter (Eds.), Emotional Development and Emotional Intelligence.

Samuel J. Paech. 2023. EQ-Bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Samuel J. Paech. 2025. EQ-Bench 3: Emotional intelligence benchmark via role-play evaluation. https://eqbench.com.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jianyu Zhou, Alvitta S. Suber, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Peter Salovey and John D. Mayer. 1990. Emotional intelligence. *Imagination, Cognition and Personality*, 9(3):185–211.

Adarsh Srinivasan, Jacob Dineen, Muhammad Umar Afzal, Muhammad Uzair Sarfraz, Irbaz B. Riaz, and Ben Zhou. 2025. RECAP: Transparent inference-time emotion alignment for medical dialogue systems. *arXiv preprint arXiv:2509.10746*.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. 2025. Between underthinking and overthinking: An empirical study of reasoning length and correctness in LLMs. *arXiv preprint arXiv:2505.00127*.

Yixuan Tang, Yi Yang, and Ahmed Abbasi. 2025. PersonaFuse: A personality activation-driven framework for enhancing human-LLM interactions. *arXiv preprint arXiv:2509.07370*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Xiang Yang and 1 others. 2025. Emotion-o1: Adaptive long reasoning for emotion understanding in LLMs. *arXiv preprint arXiv:2505.22548*.

Weixiang Zhao and 1 others. 2025. Chain of strategy optimization makes large language models better emotional supporter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

A Full Ablation Results

Table 9 presents the ablation results for all reasoning variants tested on GPT-5.2 Chinese EmoBench.

B Cross-Model and Cross-Language Results

Table 10 presents the full cross-model and cross-language evaluation results.

C Six-Model CoT Effect Data

Table 11 provides the full absolute scores for all six models under Base and CoT conditions in both languages.

D Ablation Variant Summary

Table 12 summarizes the key design differences among ablation variants.

Method	EU	EA	All	Δ_{All}
Baselines				
Base (direct answering)	57.0	76.0	66.5	—
CoT (Wei et al., 2022)	58.5	72.5	65.5	−1.0
RECAP (Srinivasan et al., 2025)	53.5	72.0	62.8	−3.8
EU-oriented reasoning depth				
Deep Reflection	**68.5**	75.0	71.8	+5.3
Multi-view Reflection	57.5	72.0	64.8	−1.8
Dual-process Reflection	59.0	75.5	67.2	+0.8
EA-oriented discrimination				
Universal Prompt	57.0	72.0	64.5	−2.0
Theory-guided (Gross+Cutrona)	58.0	76.5	67.2	+0.8
Contrastive Reflection	58.5	**80.0**	69.3	+2.8
Aggregation / gating				
Confidence Gate	55.0	73.5	64.2	−2.3
3-sample Vote + Gate	60.0	76.5	68.2	+1.8
EmoAdapt (ours)	**68.5**	**80.0**	**74.3**	**+7.8**

Table 9: Ablation on GPT-5.2 Chinese EmoBench. Δ_{All} = Overall accuracy change relative to Base. EU-oriented methods improve EU but often harm EA; EA-oriented methods preserve EA but limit EU gains; EmoAdapt achieves the only Pareto improvement across both dimensions.

Model	Lang	Method	EU	EA	All
GPT-5.2	ZH	Base	57.0	76.0	66.5
	ZH	CoT	58.5	72.5	65.5
	ZH	EmoAdapt	**68.5**	**80.0**	**74.3**
	EN	Base	61.5	75.0	68.2
	EN	CoT	64.0	77.0	70.5
	EN	EmoAdapt	**70.0**	**77.5**	**73.8**
Grok-4	ZH	Base	60.5	75.0	67.8
	ZH	CoT	62.0	74.0	68.0
	ZH	EmoAdapt	**64.5**	**76.5**	**70.5**
	EN	Base	65.0	77.5	71.2
	EN	CoT	65.5	**79.5**	72.5
	EN	EmoAdapt	**70.5**	79.0	**74.8**
Qwen-7B	ZH	Base	**33.0**	**67.0**	**50.0**
	ZH	CoT	24.5	65.0	44.8
	ZH	EmoAdapt	32.5	65.5	49.0
	EN	Base	**30.0**	**68.0**	**49.0**
	EN	CoT	26.5	51.5	39.0
	EN	EmoAdapt	31.0	65.0	48.0

Table 10: Cross-model and cross-language results. EmoAdapt achieves Pareto improvement (EU↑ and EA↑ over Base) on both GPT-5.2 and Grok-4 in both Chinese and English, confirming cross-model and cross-language generalization of the structure-aware routing design. EU gains are consistently larger, reflecting Deep Reflection's targeted effect on the harder perspective-taking component. Qwen-7B represents a capability floor where EU is near chance under all conditions.

E Cross-Benchmark Evaluation Protocol

To evaluate EmoAdapt on external benchmarks beyond EmoBench, we adapt the system prompt while preserving each benchmark's native evaluation pipeline.

EQ-Bench v2 (emotion intensity prediction, 171 questions): Since intensity prediction is an understanding task, we inject the EU-style reasoning prompt as the system message (Character Understanding, Perspective Simulation, and Critical Reflection including Pattern Check, Over-analysis Check, and Bias Check) before the benchmark's standard user prompt. The scoring pipeline (cosine-distance-based intensity matching) is unchanged.

EQ-Bench 3 (role-play rubric, 45 scenarios): Role-play involves both understanding and application, so we inject a blended prompt that combines perspective-taking with EA-oriented response principles (over-intervention awareness, warmth bias check, preference for empowering responses). The multi-turn conversation and rubric scoring by a judge model are unchanged. Both the test model and the judge model use GPT-5.2; only the test model receives the EmoAdapt system prompt.

In both cases, comparison is between Base (no system prompt) and EmoAdapt (with system prompt), using identical model, temperature, and evaluation settings.

F Taxonomy Annotation Validation

To assess the reliability of the GPT-5.2 automatic taxonomy classification (Section 3), a human annotator independently labeled a random sample of 60 EA questions (240 options, 30% of the full dataset) using the same four-category scheme: Empowerment, Boundary, Over-involvement, and Neutral.

At the question level, 56 out of 60 questions had all four option labels matching between GPT-5.2 and human annotation, yielding **93.3%** question-level agreement. The four disagreement cases all involved the Empowerment–Boundary boundary, where the distinction between "gently suggesting resources" (Empowerment) and "responding simply without over-engaging" (Boundary) is inherently ambiguous. No systematic bias toward over- or under-classifying Over-involvement was observed. These results support the validity of the automatic taxonomy used in our diagnostic analysis.

Model	Chinese				English			
	EU_B	EU_C	EA_B	EA_C	EU_B	EU_C	EA_B	EA_C
GPT-5.2	57.0	58.5	76.0	72.5	61.5	64.0	75.0	77.0
o3	64.0	65.5	76.5	76.0	61.0	60.0	75.5	75.0
Grok-4	60.5	62.0	75.0	74.0	65.0	65.5	77.5	79.5
DeepSeek-R1	58.0	52.5	73.5	70.0	56.5	52.5	72.5	71.0
GPT-5.1	58.5	58.0	75.0	72.5	57.0	52.5	70.5	76.5
Qwen-7B	33.0	24.5	67.0	65.0	30.0	26.5	68.0	51.5

Table 11: Full Base (B) and CoT (C) scores for all six models on Chinese and English EmoBench.

Variant	Key Difference from EmoAdapt
Baselines	
CoT	"Let's think step by step" (no structure)
RECAP	Existing structured prompting method
EU-oriented	
Deep Reflection	EmoAdapt's EU branch applied to *all* tasks
Multi-view Refl.	Multiple perspective passes then aggregation
Dual-process	System-1 (fast) then System-2 (slow) reasoning
EA-oriented	
Universal	Single undifferentiated prompt for all tasks
Theory-guided	Uses Gross emotion regulation + Cutrona support theory
Contrastive Refl.	EmoAdapt's EA branch applied to *all* tasks
Aggregation	
Confidence Gate	Adds confidence score; falls back to Base if low
Vote + Gate	3-sample majority vote with confidence gating

Table 12: Summary of ablation variants. Each modifies one design dimension relative to EmoAdapt (task routing + Deep Reflection for EU + Contrastive Reflection for EA).

G EA Repair Analysis

Table 13 examines which CoT errors EmoAdapt corrects and their taxonomy types from Section 3.3.

Metric	Value
CoT errors (Chinese EA)	55 / 200
Corrected by EmoAdapt	19 / 55 (34.5%)
CoT wrong-answer type among corrected items:	
Empowerment (near-miss)	14 / 19 (**73.7%**)
Boundary	4 / 19 (21.1%)
Over-involvement	1 / 19 (5.3%)
Base→CoT regression recovered	11 / 14 (78.6%)

Table 13: EmoAdapt repair analysis on CoT errors. Among the 19 corrected items, 73.7% involved empowerment near-miss confusions, matching the overall CoT error profile (74.5%, Table 3) and confirming that Contrastive Reflection targets the diagnosed failure mode.

H Self-Anchoring Bias Ablation

An "intuition-first" variant (v2) was tested on GPT-5.2 Chinese to examine whether exposing answer options before structured reasoning alters the reasoning process. In this variant, the model was first asked to provide an intuitive gut-reaction answer before proceeding through the four-step framework. Manual inspection of 50 output samples found that 93% of the structured reasoning in subsequent steps merely rationalized the initial answer rather than genuinely reconsidering it—the model consistently found ways to justify its first response regardless of what the structured analysis revealed.

This self-anchoring effect confirms that the unidirectional flow in EmoAdapt—evidence gathering (Step 1) and perspective simulation (Step 2) *before* any answer commitment—is a design requirement, not an arbitrary choice. Once the model commits to a preliminary answer, subsequent "critical reflection" loses its corrective power.

I EA Error Case Studies

We present three representative cases where Base answers correctly but CoT switches to an incorrect answer, illustrating the near-miss confusion pattern identified in our taxonomy analysis (Section 3). In each case, the reasoning process amplifies superficially plausible but contextually inappropriate responses.

Case 1 (QID 11): Near-miss empowerment confusion

Scenario: Sarah's best friend confides about financial difficulties but is too embarrassed to ask for help and **prefers to handle it privately**.
Gold: (B) "Discuss potential solutions and available resources with her"
CoT: (A) "Offer emotional support and listening, letting her friend know she can talk freely"
All nine reasoning methods select (A), demonstrating that reasoning systematically shifts the model toward emotionally elaborate but less practically effective responses.

Case 2 (QID 63): Over-involved reasoning

Scenario: Trevor's colleague **jokes** about heavy workload and lack of sleep, but Trevor notices it affecting their performance and health.
Gold: (B) "Give colleague advice about work-life balance"
CoT: (D) "Offer to take on some of their work so they can rest more"
Reasoning leads to selecting a more interventionist option that takes on the colleague's responsibility rather than empowering them.

Case 3 (QID 144): Emotional over-elaboration

Scenario: Jack's best friend reveals he started dating **Jack's ex-girlfriend** and asks if Jack is okay with it.
Gold: (B) "Wow, that is unexpected, but I hope you are happy together"
CoT: (D) "This makes me sad, but I hope she treats you better than she treated me"
The reasoning process amplifies emotional elaboration at the cost of situational appropriateness.

J EmoAdapt Prompt Templates

Below we provide the complete EmoAdapt prompt structure. EmoAdapt is implemented as a single structured system prompt with JSON-formatted output. Steps 1, 2, and 4 are shared across EU and EA; only Step 3 Dimension C differs. Chinese prompts are direct translations of the English versions. Task routing is performed based on question type metadata provided by EmoBench.

EmoAdapt – EU Version (Deep Reflection)

System: You are an expert in emotional reasoning. Follow these steps strictly, showing your analysis for each:
Step 1: Character Understanding. Read the scenario carefully and extract: Who are the characters involved? What are the target person's traits, experiences, and cultural background? What does each character know or not know? (Note information asymmetry.) Are there special premises or assumptions (phobias, unusual rules, non-standard meanings)?
Step 2: Perspective Simulation. Take the target person's subjective viewpoint: Based on THEIR traits and knowledge (not your omniscient view), how would they interpret this situation? What would they most likely feel or think? Important: Form your own judgment first, do NOT let the options guide you.
Step 3: Critical Reflection.
Dimension A – Pattern Check: Are you applying a common pattern (e.g., acrophobia → must be afraid)? Are there counter-intuitive clues you missed? Do the person's special traits change the typical reaction?
Dimension B – Over-analysis Check: Are you over-complicating a simple situation? If the emotional cues are clear, the most straightforward interpretation is likely correct.
Dimension C – Practicality Check: If the question asks "what to do," is your suggestion natural and appropriate in real social contexts? Are you over-intervening? Is there a simpler, equally effective approach?
If you find issues, correct your previous judgment.
Step 4: Final Answer. Based on all the above analysis, select the best matching answer. If there are multiple sub-questions: first determine the answer to Q1, then review subsequent options for consistency.
Output format: `{"step1_character": "...", "step2_perspective": "...", "step3_reflection": "...", "answer_q1": "A/B/C/D", "answer_q2": "A/B/C/D"}`

EmoAdapt – EA Version (Contrastive Reflection)

Steps 1, 2, 4 and Dimensions A–B are **identical** to the EU version above. Only Dimension C differs:
Dimension C – Contrastive Comparison Check: Compare the top two or three most plausible options side by side. For each pair, identify the subtle contextual constraint (e.g., what the person specifically asked for, their stated preference, the social relationship) that distinguishes the best response from near-miss alternatives. Choose the option that most precisely matches the situational needs rather than the one that sounds most elaborately supportive.
Output format: `{"step1_character": "...", "step2_perspective": "...", "step3_reflection": "...", "answer": "A/B/C/D"}`