SkillEmo: Skill-Enhanced Reinforcement Learning with an Evolving Emotion Skill Bank

Anonymous ACL submission

Abstract

Emotion reasoning is not a uniform post-training target: on EmoBench, chain-of-thought prompting drops overall accuracy from 45.0% to 43.5% because EU falls sharply from 29.5% to 23.0%, even though EA rises from 60.5% to 64.0%. This split indicates that EU and EA expose different failure patterns, yet standard reinforcement learning still applies a single binary reward and fixed skill priors throughout training. We present SkillEmo, a skill-enhanced RL framework with three components: a psychology-inspired Emotion Skill Bank, a decomposed reward that scores the two EU sub-questions separately, and a rollout-based online skill evolution mechanism that mines failure patterns from training rollouts without additional validation rollouts. Training Qwen2.5-7B-Instruct with GRPO on EmoBench, SkillEmo raises accuracy on the held-out English question-ID split from 44.6% to 50.0% (+5.4%), with balanced EU and EA gains. On the out-of-domain benchmarks SocialIQA and ToMBench, SkillEmo with skill injection reaches 77.1% (+2.9) and 70.4% (+2.1), indicating positive transfer beyond EmoBench. Code is available at https://anonymous.4open.science/r/SkillEmo-2EB6.

1 Introduction

Emotion reasoning, the ability to infer, attribute, and respond to affective states in context, is a foundational capability for socially intelligent AI. Applications from empathetic dialogue (Rashkin et al., 2019) to social question answering (Sap et al., 2019) depend on interpreting subtle social cues, disentangling semantically proximate emotions, and selecting contextually appropriate responses. Yet structured emotion benchmarks reveal that this capability is not monolithic. On EmoBench (Sabour et al., 2024), chain-of-thought prompting reduces overall accuracy for Qwen2.5-7B from 45.0% to 43.5% because EU drops sharply from 29.5% to 23.0%, while EA simultaneously rises from 60.5% to 64.0%. This asymmetric behavior suggests that a single reasoning or reward strategy does not fit both sub-tasks.

Context: Maria finally confronted her colleague after years of unspoken tension. To her surprise, the colleague responded with gratitude rather than defensiveness.
Q1: What does Maria most likely feel? (A) Relief and surprise (B) Guilt (C) Indifference (D) Regret Model selects (B) Guilt Gold: (A) Relief Binary reward: 0 Decomposed: 0.5

Table 1: A near-miss EU case illustrating why binary reward is overly coarse. The model answers Q2 (cause) correctly but misidentifies the emotion (Q1). Binary reward assigns zero, discarding partial progress; decomposed reward preserves this signal as 0.5.

Standard RL post-training therefore faces a structural blind spot on emotion reasoning. Mathematical and code reasoning have deterministic verifiers where correctness is unambiguous; emotion reasoning errors, by contrast, differ systematically by task structure. Error analysis of Qwen2.5-7B on EmoBench zero-shot shows that EU (emotion understanding) errors concentrate on perspective-taking failures and fine-grained emotion confusion such as guilt versus regret, while EA (emotion application) errors concentrate on social-norm misjudgment and oversimplified response strategies. These error modes call for different corrective signals, yet standard GRPO binary reward collapses them into the same outcome signal, offering little guidance toward targeted improvement (Table 1).

Psychological research suggests that proficient emotion reasoning draws on a compact repertoire of reusable strategies: perspective reconstruction, fine-grained emotion distinction, and empathy-depth calibration (Gross, 1998). This motivates externalizing such strategies as retrievable struc-

tured records, injecting them as curriculum signals during training, and dynamically updating them according to the current policy weakness frontier. A further mechanism question arises: the timing and source of evolution matter. Prior approaches rely on separate validation inference to trigger evolution, introducing computational overhead and information lag; we instead collect failure samples directly from training rollouts, decoupling evolution from validation entirely (Section 3.6).

Based on this diagnosis, we present SkillEmo, a unified training framework integrating skill retrieval, task-aware reward design, and automatic skill evolution. Our main contributions are as follows. First, we identify task-structure-dependent error differentiation in emotion RL training and design a decomposed reward that scores the two EU sub-questions separately, enabling gradient signals to distinguish different reasoning failure types (Section 3.5). Second, we propose rollout-based online skill evolution that mines systematic failure patterns directly from training rollouts and generates new skills without additional validation inference, allowing the skill bank to co-evolve with the policy weakness frontier (Section 3.6). Third, SkillEmo achieves +5.4% on the EmoBench held-out English question-ID split—outperforming CoT prompting, which prior work has shown to degrade EU (Sabour et al., 2024)—and, with skill-augmented inference, improves out-of-domain social cognition benchmarks to 77.1% on SocialIQA (+2.9) and 70.4% on ToMBench (+2.1), indicating transfer beyond EmoBench (Section 5, 6).

2 Related Work

Emotion reasoning benchmarks. GoEmotions (Demszky et al., 2020) and EmpatheticDialogues (Rashkin et al., 2019) established foundational corpora for emotion classification and empathetic response generation. SocialIQA (Sap et al., 2019) extended coverage to social commonsense reasoning, and ToMBench (Chen et al., 2024) evaluates theory-of-mind inference over beliefs and intentions. For evaluating structured reasoning rather than surface classification, EmoBench (Sabour et al., 2024) presents multi-step multiple-choice tasks distinguishing EU from EA subtypes. These benchmarks expose systematic weaknesses in current LLMs but do not provide training methods to address them; SkillEmo uses the EU/EA task structure to design type-aware training mechanisms.

Reinforcement learning for LLM reasoning. RLHF (Ouyang et al., 2022) and PPO (Schulman et al., 2017) are standard for preference alignment, while GRPO (Shao et al., 2024) eliminates the critic model and achieves strong results on mathematical reasoning. DeepSeek-R1 (Guo et al., 2025) demonstrates that RL post-training can elicit extended chain-of-thought reasoning at scale. These methods assume deterministic, verifiable reward signals such as correct proofs or passing test cases; emotion reasoning lacks such clean supervision, and binary correctness reward is an impoverished signal for semantically dense tasks. SkillEmo builds on GRPO and introduces task-aware decomposed reward together with domain-knowledge priors to address this gap.

Skill-augmented and knowledge-augmented learning. Retrieval-augmented generation (Lewis et al., 2020) injects relevant passages at inference time to improve factuality but provides no structured guidance during training. The options framework in hierarchical RL (Sutton et al., 1999) and skill injection for text-game agents show that structured decompositions improve generalization, though those skills are extracted mechanically from task structure rather than grounded in domain theory. Recent work such as SkillRL (Xia et al., 2026) also studies recursive skill-augmented reinforcement learning, but focuses on evolving agents in broader agentic settings rather than structured emotion reasoning with task-aware rewards. EmotionPrompt (Li et al., 2023) and chain-of-thought prompting (Wei et al., 2022) improve emotion-task performance through inference-time augmentation without modifying internal representations. SkillEmo differs from these approaches in three respects: skills are updated online from rollout failures rather than fixed offline or via prompt engineering; they serve primarily as training-time curriculum signals that shape parameter updates, not as inference-time prompt scaffolds; and for in-domain use on the primary benchmark, the resulting model achieves its best performance without inference-time augmentation.

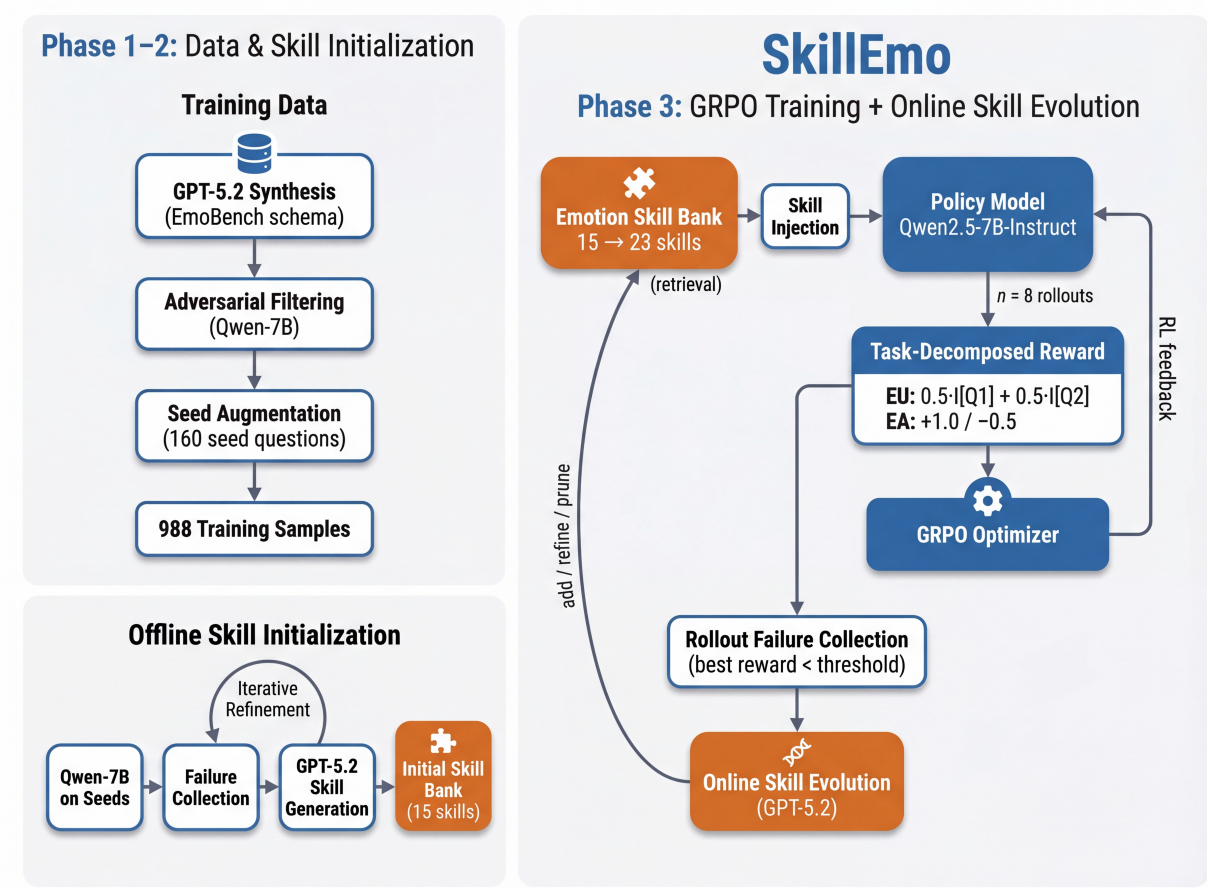

Figure 1: Overview of SkillEmo. Phase 1 generates adversarially filtered training data. Phase 2 initializes the Emotion Skill Bank via iterative failure analysis on seed examples. Phase 3 trains the policy with GRPO using decomposed reward, while the skill bank evolves online by mining failure patterns from training rollouts every T_{evo} steps. Skills are injected as imperative instructions retrieved by a weighted combination of semantic similarity and empirical hit rate.

3 Method

3.1 Task Formulation

We focus on emotion understanding (EU) and emotion application (EA) multiple-choice tasks. Each example provides a natural-language context x and a set of candidate options $\mathcal{Y} = \{y_1, \ldots, y_m\}$ with $m = 4$. The policy π_θ must select the correct option $y^\star \in \mathcal{Y}$ and optionally generate a brief rationale. EU tasks require the model to identify the emotional state of a named agent (Q1) and to attribute the cause of that emotion (Q2); both sub-questions must be answered correctly for the sample to count as correct. EA tasks require selecting an appropriate emotional response or coping strategy in a social scenario and involve a single question. This structural difference between EU and EA directly motivates the reward design in Section 3.5.

We train π_θ with GRPO (Shao et al., 2024): for each example we sample n rollouts $\{\hat{y}_i\}_{i=1}^n$, compute rewards $\{r_i\}$, and normalize advantages within the group:

$$A_i = \frac{r_i - \bar{r}}{\sigma_r + \epsilon}, \qquad \bar{r} = \frac{1}{n}\sum_{j=1}^{n} r_j. \tag{1}$$

A KL penalty with coefficient $\beta = 0.005$ prevents catastrophic capability drift from the reference policy.

3.2 Training Data Construction

The training corpus is generated using GPT-5.2 following the EmoBench schema (Sabour et al., 2024), covering 11 EU fine-grained categories (perspective-taking, emotion transition, mixed emotions, among others) and four EA categories (Personal/Social × Others/Self). EU scenarios produce two sub-questions (emotion identification Q1 and cause attribution Q2), while EA scenarios produce one four-choice question.

Data construction proceeds in three stages. In the synthesis stage, questions are generated with semantically confusable distractors to maximize training signal density. In the adversarial filtering stage, Qwen-7B performs initial inference on all synthetic data; only samples where the model errs on 1/8 to 7/8 of rollouts are retained, filtering out trivially easy and overly ambiguous items. For EA samples with excessively high model accuracy, GPT-5.2 rewrites distractors to increase difficulty. In the seed-data augmentation stage, 160 original EmoBench questions (80 EU, 80 EA) are appended as seed data, ensuring direct training signal on the benchmark task distribution. The final training set contains 988 samples and is disjoint from the 240 held-out English EmoBench question IDs used in the split analysis by construction, corresponding to 480 held-out evaluation instances under the official EN+ZH protocol.

3.3 Emotion Skill Bank

SkillEmo maintains an Emotion Skill Bank $\mathcal{S} = \{s_i\}_{i=1}^{N}$ organized into two tiers: general skills applicable to any EU/EA context and task-specific skills specialized for EU or EA sub-types. Each skill is a structured record:

$$s_i = \{\texttt{skill_id},\ \texttt{category},\ \texttt{principle},\ \texttt{when_to_apply}\}. \quad (2)$$

The `principle` field contains a concise, psychology-inspired instruction such as "Before judging an emotion, reconstruct the agent's subjective situational framing, not the objective facts." The `when_to_apply` field specifies context triggers such as "the agent's expressed emotion conflicts with the likely situational reaction." Skill categories include perspective reconstruction, emotion attribution, intent inference, fine-grained emotion distinction, cultural context sensitivity, over-speculation suppression, and social dynamics awareness. The initial bank is informed by psychological literature on emotional intelligence (Gross, 1998); subsequent versions are maintained by the evolution loop (Section 3.6).

3.4 Skill Injection

Given an input context x, we retrieve the top-k skills (k=2) and prepend them to the model prompt as short imperative instructions. Retrieval uses a weighted combination of embedding similarity and empirical hit rate:

$$\text{score}(s, x) = \alpha \cdot \cos(e_s, e_x) + (1-\alpha) \cdot \text{hit}(s), \quad (3)$$

where e_s and e_x are sentence embeddings (Reimers and Gurevych, 2019), $\text{hit}(s)$ is the empirical rate at which skill s co-occurred with a correct answer in prior training, and α=0.8. The hit-rate term is intended to shift retrieval toward empirically useful skills over time; we set α=0.8 based on preliminary experiments and note that sensitivity to this choice is an area for future investigation. At inference time, skill injection is optional; our experiments show that it improves in-domain accuracy but degrades held-out generalization when retrieval precision is insufficient (Section 5).

3.5 Task-Aware Decomposed Reward

Binary correctness reward is a poor signal for EU tasks: a model that correctly identifies the emotion (Q1) but misattributes the cause (Q2) receives zero reward, discarding the useful gradient information from Q1. We design a decomposed reward that reflects the EU/EA structural difference:

$$r_{\text{EU}} = 0.5 \cdot \mathbb{I}[\text{Q1 correct}] + 0.5 \cdot \mathbb{I}[\text{Q2 correct}], \quad (4)$$

$$r_{\text{EA}} = \begin{cases} +1.0 & \text{if correct} \\ -0.5 & \text{if incorrect.} \end{cases} \quad (5)$$

For EU, the two sub-questions receive separate, equal-weight gradient signals, intended to allow the model to learn emotion identification and causal attribution as separable capabilities. For EA, the asymmetric design provides stronger reinforcement for correct responses than punishment for errors, preventing overly conservative policies. The design motivation comes directly from our error diagnosis: EU sub-questions reflect distinct but related cognitive abilities (identification versus attribution), and scoring them separately reflects the model capability profile more precisely than continuous semantic similarity over the option embedding space (Appendix A.4).

3.6 Rollout-Based Skill Evolution

Prior approaches evaluate on a held-out validation set every T_{evo} steps to collect failure samples and trigger evolution. This introduces additional forward-pass overhead and relies on a fixed validation distribution that may not cover failure modes emerging during training.

We decouple skill evolution entirely from validation inference. Each training step of GRPO naturally produces n candidate responses and their reward scores. We collect failure samples directly from these rollouts: for each prompt, we select the

rollout with the highest reward as the representative and use these representative rollouts for periodic failure analysis. In practice, only representative responses that remain incorrect under the reward signal are retained in the failure pool.

Every T_{evo}=50 steps (chosen to balance evolution frequency against sample accumulation; see Appendix A.7 for details), the evolution pipeline activates. First, failure samples are grouped by EU/EA type and by which distractor was selected, identifying systematic error patterns. Second, representative failure clusters are sent to an external LLM (GPT-5.2), which proposes up to $\Delta_{\max}$=3 updates: adding a new skill targeting the failure mode, refining an existing skill to narrow its scope, or pruning a skill that causes confusion. Third, candidate skills are deduplicated against the existing bank by cosine similarity (threshold 0.9) and checked for length and coherence constraints before being written to the bank.

The updated bank takes effect immediately without a checkpoint reload. Skill injection in the training data is hot-swapped so that subsequent rollouts use the latest skills. The entire pipeline runs as a hook inside the training loop; the vLLM rollout engine operates uninterrupted across evolution boundaries. Skill snapshots are saved automatically before and after each evolution event, providing a complete auditable trajectory.

4 Experimental Setup

We organize the evaluation around four research questions: **Q1** Does SkillEmo improve emotion reasoning on both EU and EA, and do gains transfer to out-of-domain social cognition benchmarks? (Section 5) **Q2** Are gains genuine generalization or memorization of seed data? (Section 5.2) **Q3** What does each component contribute independently? (Section 5.4) **Q4** Why does inference-time skill injection hurt held-out accuracy? (Section 6.1)

4.1 Model and Training Configuration

We use Qwen2.5-7B-Instruct (Qwen Team, 2025) as the base policy. Training uses GRPO (Shao et al., 2024) with vLLM-based rollouts (Kwon et al., 2023) for efficient generation. Core hyperparameters: learning rate 1×10^{-6}, train batch size 64 (PPO mini-batch 8, micro-batch per device 2), rollout count n=8, KL coefficient β=0.005, entropy coefficient 0. Maximum prompt length is 1,024 tokens; maximum response length is 512 tokens. We train with bfloat16 precision and gradient checkpointing on 8 GPUs with tensor parallelism degree 2. Skill evolution uses T_{evo}=50 steps and $\Delta_{\max}$=3; skill embeddings are computed with Sentence-BERT (Reimers and Gurevych, 2019).

4.2 Evaluation Benchmarks

EmoBench (Sabour et al., 2024) is the primary benchmark, containing 800 EU/EA multiple-choice questions (EN + ZH). Our main results use the full official protocol over all 800 EN+ZH instances. For split-level analysis, we follow the English question-ID split used during data construction: 160 seed question IDs (80 EU, 80 EA) participate in training data generation and 240 English question IDs (120 EU, 120 EA) are held out. Under the official EN+ZH evaluation protocol, these correspond to 320 seed instances and 480 held-out instances. SocialIQA (Sap et al., 2019) dev set (1,954 items) and ToMBench (Chen et al., 2024) serve as out-of-domain benchmarks measuring whether emotion reasoning gains transfer to broader social cognition. Neither SocialIQA nor ToMBench data appears in training.

4.3 Compared Systems

We compare four configurations to isolate the contribution of each component: (1) Qwen2.5-7B zero-shot, with no post-training; (2) GRPO-only, using standard GRPO with binary reward and no skill injection, serving as the pure RL baseline; (3) GRPO + Static Skill, using GRPO with the Phase 2 skill bank fixed throughout training, isolating the contribution of online evolution; (4) SkillEmo (full system), using GRPO with decomposed reward and online skill evolution. Each post-trained model is evaluated with and without inference-time skill injection to separate training-time from inference-time effects.

4.4 Evaluation Protocol

All evaluations use greedy decoding (temperature 0). Answers are extracted by matching the first uppercase letter (A through D) following a designated answer marker in the model response. EU questions require both Q1 and Q2 to be correct for the sample to count as correct. Unless otherwise noted, reported benchmark scores come from a single evaluation pass of each checkpoint under this decoding setup.

5 Results

5.1 Main Results (Q1)

Table 2 reports the full evaluation using the official EmoBench protocol. Two patterns emerge from Table 2. CoT harms EmoBench by sharply degrading EU while only slightly improving EA, reinforcing the need for structure-aware training rather than a uniform prompting strategy. In contrast, SkillEmo improves in-domain performance without inference-time augmentation, while skill injection mainly benefits the out-of-domain benchmarks. Section 6.1 provides a detailed analysis of the divergent injection effects across benchmarks.

5.2 Held-Out Analysis (Q2)

To verify that gains reflect genuine generalization rather than memorization of seed questions, we report results on the English question-ID split used during data construction: Seed (80 EU + 80 EA English question IDs that participated in data generation) and Test (120 EU + 120 EA held-out English question IDs). Under the official EN+ZH evaluation protocol, these correspond to 320 seed instances and 480 held-out instances.

Table 3 shows that the held-out split improvement (+5.4 pp, from 44.6 to 50.0) exceeds the seed-split improvement (+4.9 pp), suggesting that gains are not driven solely by memorization of seed questions, although we cannot fully disentangle memorization effects from distributional differences between the seed and held-out subsets. EU-Test improves from 27.5 to 33.3 (+5.8 pp) and EA-Test from 61.7 to 66.7 (+5.0 pp), with gains in both task types.

Skill injection at inference time inflates seed accuracy to 52.8 (+2.3 pp over SkillEmo no-skill) but reduces test accuracy to 47.9 (-2.1 pp). The resulting 4.9 pp seed-test gap suggests that skills function as effective retrieval augmentation on familiar questions but may introduce noise on unseen ones.

5.3 Training Dynamics

Figure 2 compares mean reward trajectories for the three training configurations. All three methods start from the same reward level ($\sim$0.49) and improve rapidly in the first 50 steps. After this initial phase, the curves diverge: SkillEmo (full) reaches a smoothed reward of approximately 0.74 by step 200, while GRPO + Static Skill plateaus near 0.70 and GRPO-only near 0.65. The gap between SkillEmo and GRPO + Static Skill widens after the second skill evolution event (step 100, bank size 15→17), consistent with newly generated skills targeting failure modes that static skills do not cover, though we note this is a descriptive observation from a single run. Step 200 gives the strongest EmoBench accuracy (50.0%) and is used as the reported checkpoint.

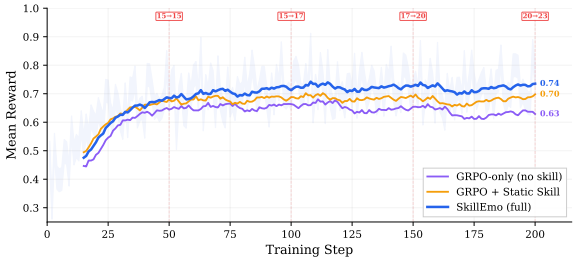

Figure 2: Mean reward over training steps for three configurations (smoothed with a 15-step moving average; single run). Dashed lines mark skill evolution events; labels show skill-bank size changes. SkillEmo (full) achieves the highest reward in this run, with GRPO + Static Skill and GRPO-only reaching progressively lower levels. Step 200 is selected as the reported checkpoint from fixed-interval evaluations in this run.

5.4 Ablation Study (Q3)

Table 4 reports benchmark scores under the same evaluation protocol. The ablation shows a clear ordering: GRPO training provides the base gain, static skills add a smaller EU-focused improvement, and online evolution contributes the largest additional benefit. Removing decomposed reward mainly hurts EU (35.0 to 33.3, -1.7 pp) while leaving EA nearly unchanged, which matches its intended role as a targeted signal for emotion understanding.

CoT prompting is included as a reference baseline: as reported by Sabour et al. (2024), it reduces EmoBench accuracy by 1.5 pp despite a moderate EA gain (+3.5 pp), because EU drops from 29.5% to 23.0% (-6.5 pp). This known limitation underscores the need for training-time methods rather than inference-time prompting strategies.

6 Analysis

6.1 When Does Skill Injection Help? (Q4)

Inference-time skill injection exhibits divergent effects across benchmarks (Table 2, Table 3). On EmoBench, injection slightly degrades the trained model (50.0 to 49.5), and on the held-out English

Model	Setting	EmoBench	EU	EA	SocialIQA	ToMBench
Qwen2.5-7B	zero-shot	45.0	29.5	60.5	74.2	68.3
Qwen2.5-7B	+ CoT	43.5	23.0	64.0	—	—
Qwen2.5-7B	+ skill	46.2	32.8	59.7	75.5	70.1
GRPO (binary)	zero-shot	46.8	31.0	62.5	74.5	69.0
GRPO + Static Skill	zero-shot	47.5	32.2	62.8	75.0	69.5
SkillEmo	zero-shot	**50.0**	**35.0**	**65.0**	74.5	69.8
SkillEmo	+ skill	49.5	33.5	65.5	**77.1**	**70.4**

Table 2: Evaluation across five benchmarks using the official EmoBench protocol. CoT prompting degrades overall accuracy (-1.5), consistent with prior findings (Sabour et al., 2024). SkillEmo improves EmoBench by +5.0 over the base model. With skill injection, SkillEmo reaches 77.1% on SocialIQA (+2.9) and 70.4% on ToMBench (+2.1), indicating positive transfer beyond EmoBench.

Condition	Seed	Test	EU-T	EA-T
Baseline	45.6	44.6	27.5	61.7
+ CoT	41.9	44.6	23.3	65.8
SkillEmo	50.5	**50.0**	**33.3**	**66.7**
+ skill	52.8	47.9	31.6	64.2

Table 3: Seed/test split accuracy on EmoBench using the English question-ID split (reported under the official EN+ZH evaluation protocol). SkillEmo improves the held-out test split by +5.4 pp, exceeding the seed-split gain (+4.9 pp), suggesting that improvement is not driven solely by seed memorization.

Configuration	EmoBench	EU	EA
SkillEmo (full)	**50.0**	**35.0**	**65.0**
w/o Online Evolution	47.5	32.2	62.8
w/o Skill Bank	46.8	31.0	62.5
w/o Decomposed Reward	49.4	33.3	65.4
Baseline (no training)	45.0	29.5	60.5
+ CoT	43.5	23.0	64.0

Table 4: Component ablation on EmoBench. Each row removes one component from the full system. GRPO training contributes +1.8, static skills add +0.7, online evolution adds +2.5, and decomposed reward adds +0.6. CoT is included as a reference baseline following Sabour et al. (2024).

split the drop is sharper (50.0 to 47.9). On the out-of-domain benchmarks, however, injection provides substantial gains: SocialIQA improves from 74.5 to 77.1 (+2.6 pp) and ToMBench from 69.8 to 70.4 (+0.6 pp).

This divergence can be attributed to the interaction between skill specificity and task distribution overlap. Skills evolve based on EmoBench training-data error patterns and are highly tuned to that task distribution. On EmoBench held-out questions, the narrowly specialized skills sometimes retrieve incorrectly, producing retrieval mismatch: for instance, an over-speculation suppression skill can cause errors on questions designed to require deep situational inference. On out-of-domain tasks such as SocialIQA, the same skills provide broadly useful social reasoning guidance without over-specialization, because the task distribution is sufficiently different that the general principles encoded in skills remain helpful while the narrow failure-mode targeting is less harmful.

For in-domain evaluation, additional injection may introduce retrieval noise when the model has already learned the relevant strategies during training.

Step	#Skills	New	Rew.	Direction (example)
0	15	init	.494	Phase 2 init
50	15	0	.678	No new failures
100	17	+2	.683	EU perspective (guilt vs. regret)
150	20	+3	.708	EA social dynamics (caregiver fatigue)
200	23	+3	.765	Mixed-emotion (ambivalent cues)

Table 5: Skill evolution trajectory in this training run. The bank grows adaptively: no additions at step 50 when existing skills cover active failures, then targeted additions as the improving policy exposes finer-grained failure modes.

6.2 Skill Evolution Trajectory

Table 5 traces the co-evolution of the skill bank and policy performance. Different training stages expose different dominant failure modes. At step 50 the existing bank already covers the active failure patterns and no new skills are generated. From step 100 onward in this run, the evolution procedure tends to introduce skills targeting more specific observed failure modes: EU perspective-taking refinement for expectation-violation cases, EA social-dynamics calibration for interpersonal scenarios,

Metric	Base	Ours	Δ
Q1 (emotion ident.)	56.0	63.5	+7.5
Q2 (cause attrib.)	45.5	52.0	+6.5
Both (Q1∧Q2)	29.5	35.0	+5.5

Table 6: EU sub-question decomposition. Both sub-capabilities improve, with Q1 (emotion identification) gaining slightly more than Q2 (cause attribution), consistent with identification being more responsive to gradient signal than multi-step causal reasoning.

and mixed-emotion distinction under ambivalent contexts. The bank grows during early rounds as additions outpace pruning, then stabilizes as later evolution rounds primarily refine trigger conditions rather than add net entries. This sequential specialization pattern mirrors observations in the psychology of expertise, where proficiency often proceeds from broad heuristics to context-sensitive strategies, though we do not claim a formal analogy.

6.3 Decomposed Reward and EU Sub-Question Analysis

Table 6 breaks down EU accuracy into the two constituent sub-questions. Both sub-capabilities improve under SkillEmo, with Q1 (emotion identification) gaining 7.5 pp and Q2 (cause attribution) gaining 6.5 pp. This is consistent with the design rationale: emotion identification, as a relatively shallow judgment, responds more directly to gradient signal, while causal attribution involves more complex reasoning chains and improves more gradually. The ablation removing decomposed reward (Table 4) shows EU dropping by 1.7 pp (35.0 to 33.3) while EA remains stable, consistent with decomposed scoring specifically targeting EU improvement. Under binary reward, samples where Q1 is correct but Q2 is wrong receive zero reward, wasting the useful Q1 signal. Decomposed reward converts this to a 0.5 positive gradient, allowing the model to consolidate Q1 progress while working on Q2. The random-chance accuracy for EU (both correct on two independent four-choice questions) is approximately 1/16; the improvement from 29.5% to 35.0% represents a 19% relative gain over baseline.

7 Discussion

In-domain versus out-of-domain injection. Our results suggest a nuanced interaction between skill-augmented training and inference. On EmoBench, additional injection may introduce retrieval noise after training has already shaped the model. On SocialIQA and ToMBench, the same skills are more helpful, indicating that the evolved bank can still provide useful guidance when the target distribution differs from the training benchmark.

Rollout-based evolution. Collecting failure samples from training rollouts eliminates the overhead of separate validation inference and ensures that failure analysis reflects the current policy state. We note that this design choice has not been compared against validation-based evolution under controlled conditions; such a comparison would strengthen the claim and is left for future work.

8 Conclusion

We presented SkillEmo, a skill-enhanced reinforcement learning framework that improves structured emotion reasoning through task-aware decomposed reward and rollout-based online skill evolution. Three observations guided the framework design. EU and EA exhibit structurally different error patterns, and decomposed reward provides separate gradient signals for each failure type. Mining failure patterns from training rollouts allows the skill bank to co-evolve with the policy without additional validation rollouts. Under skill-augmented inference, the evolved skill bank yields positive transfer on SocialIQA (+2.9) and ToMBench (+2.1).

On the EmoBench held-out English question-ID split, SkillEmo improves accuracy from 44.6% to 50.0% (+5.4%), with balanced gains across EU and EA. Ablation experiments suggest that GRPO training (+1.8%), static skills (+0.7%), and online evolution (+2.5%) each provide complementary contributions, though we do not report multi-seed training statistics and these estimates should be interpreted with that caveat in mind.

SkillEmo provides a practical framework for integrating structured expert knowledge into RL post-training. We will release code, evaluation scripts, and skill-bank versions to support reproducibility.

Limitations

Several limitations should inform how our results are interpreted. Main benchmark tables report single-pass evaluation results for each checkpoint under the same decoding setup. However, we do not report multi-seed training statistics; RL training can exhibit variance across random seeds, and

the training-dynamics and skill-evolution analyses reflect one representative run. Skill evolution currently depends on an external LLM (GPT-5.2) to generate skill candidates from error clusters, and the same model is used for training data synthesis and distractor rewriting. Our goal is not to eliminate external teacher knowledge but to study whether such knowledge can be converted into a compact, evolving skill artifact that improves RL training without additional validation rollouts on the primary benchmark. Disentangling the contribution of the teacher model from the contribution of the SkillEmo framework itself would require ablations with weaker generators, which we leave to future work. All experiments use a 7B-parameter policy; larger models may require different injection templates, evolution frequencies, or retrieval strategies, and the scaling behavior of skill priors is an open question. Emotion reasoning benchmarks contain inherent subjectivity and annotation noise; decomposed reward mitigates signal sparsity but does not resolve disagreements in the ground-truth label distribution, so accuracy estimates carry this underlying uncertainty. The skill bank is initialized from Western psychological frameworks; cultural adaptation of both the initial bank and the evolution procedure is necessary for broader deployment.

Ethics Statement

SkillEmo targets structured emotion reasoning benchmarks in NLP research and does not perform real-world affect inference about individual users. Training data is synthetically generated following public benchmark schemata. Emotional reasoning systems deployed in sensitive contexts such as mental health support, hiring, or law enforcement carry material risks of over-confident or culturally inappropriate inferences; we recommend calibration checks, human oversight, and domain-specific evaluation before any deployment beyond research settings. The skill bank represents a collection of generalizable reasoning heuristics, not clinical or therapeutic guidelines. We will release all components under an open research license with explicit use-case restrictions.

References

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

James J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.

Daya Guo, Dejian Yang, Haowei Zhang, Junxian Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401*.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing, pages 3982–3992. Association for Computational Linguistics.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Liu, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxian Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and 1 others. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. 2026. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *Preprint*, arXiv:2602.08234.

A Appendix

A.1 Training Data Statistics

Adversarial filtering retains approximately 62% of generated questions. Filtering removes items where Qwen-7B zero-shot answers correctly on all 8 rollouts (trivially easy, approximately 25% of generated items) and items answered incorrectly on all 8 rollouts (overly ambiguous, approximately 13%). For EA items with accuracy exceeding 7/8, GPT-5.2 rewrites distractors; approximately 40% of these rewrites pass a second round of filtering.

Category	Synthetic	Seed	Total
EU (English)	312	40	352
EU (Chinese)	296	40	336
EA (English)	158	40	198
EA (Chinese)	62	40	102
Total	828	160	988

Table 7: Training data composition. Seed questions are drawn from the original EmoBench benchmark (80 EU, 80 EA). Synthetic questions are generated by GPT-5.2 following the EmoBench schema and adversarially filtered.

A.2 Checkpoint Selection

We save checkpoints at fixed intervals (every 50 steps) during training. Table 8 reports EmoBench accuracy for checkpoints evaluated during the training run. Step 200 gives the strongest performance among evaluated checkpoints and is used for all reported results. Because this represents a single training run, we note that the ordering may vary across random seeds.

Checkpoint	EmoBench	EU	EA
Baseline (step 0)	45.0	29.5	60.5
Step 50	46.2	30.2	62.1
Step 100	47.2	31.8	62.6
Step 150	48.7	33.0	64.5
Step 200	**50.0**	**35.0**	**65.0**
Step 250	48.7	33.0	64.5

Table 8: EmoBench accuracy across checkpoints saved during training (SkillEmo, single run). Step 200 is selected as the reported checkpoint based on overall accuracy.

A.3 EU/EA Prompt and Answer Format

We standardize prompts and enforce strict output formats to avoid evaluation artifacts.

EU instances require answering two sub-questions. We enforce the following two-line format:

```
Answer Q1: [letter]
Answer Q2: [letter]
```

Any deviation, including missing fields, extra options, or malformed letters, is treated as invalid. The system prompt instructs the model to complete reasoning before producing the answer lines, and rationale length is capped at 256 tokens to reduce format drift.

EA instances require a single choice, formatted as:

```
Answer: [letter]
```

We embed the format constraint in both the system prompt and a closing instruction appended to the task context.

A.4 Decomposed Reward Details

For EU tasks, the decomposed reward assigns separate scores to each sub-question:

$$r_{\text{EU}} = 0.5 \cdot \mathbb{I}[\text{Q1 correct}] + 0.5 \cdot \mathbb{I}[\text{Q2 correct}].$$

This means a rollout that correctly identifies the emotion (Q1) but misattributes the cause (Q2) receives a reward of 0.5 rather than 0. Under binary reward, both this case and a rollout that answers both sub-questions incorrectly receive identical zero reward, discarding the information that Q1 was solved.

For EA tasks:

$$r_{\text{EA}} = \begin{cases} +1.0 & \text{if correct} \\ -0.5 & \text{if incorrect.} \end{cases}$$

The asymmetric design provides stronger reinforcement for correct responses than punishment for errors. This prevents the policy from converging to overly conservative outputs that avoid committing to any specific emotional response.

A.5 Skill Evolution Case Study

Table 9 illustrates two concrete examples of how skill evolution targets specific failure patterns observed during training. Each row shows the training stage at which the failure was identified, the systematic error pattern, the skill generated or refined in response, and the effect on model behavior.

These examples illustrate the key property of online evolution: as the policy improves and easy failure modes are resolved, the evolution mechanism identifies progressively more specific error patterns (e.g., self-attribution subtleties, caregiver-specific social dynamics) that static skills initialized from general psychology literature do not cover.

A.6 Emotion Skill Bank Schema and Examples

Each skill is stored as a structured record with four fields:

```
{skill_id,     category,     principle,
when_to_apply}
```

Representative skills from the initial bank:
Perspective reconstruction (EU): "Before judging an emotion, reconstruct the agent's subjective situational framing, not the objective facts." Trigger: "Agent's expressed emotion conflicts with the likely situational reaction."
Fine-grained distinction (EU): "When two options share the same valence, identify the causal attribution that separates them: self-caused emotions (guilt, pride) differ from externally caused ones (jealousy, envy)." Trigger: "Two or more options are near-synonymous emotions."
Empathy depth (EA): "Acknowledge the root emotional cause before proposing any action; premature advice signals that the speaker's distress has not been understood." Trigger: "Responder must select or generate an emotionally supportive reply."

Skills introduced by evolution target narrower failure modes discovered during training, such as distinguishing caregiver fatigue from interpersonal conflict, or correcting over-reliance on surface social norms in cross-cultural scenarios (see Table 9 for concrete examples).

A.7 Skill Evolution Algorithm

Skill evolution activates every T_{evo}=50 training steps. Failure samples are collected from the training rollouts of the preceding T_{evo} steps: for each prompt, the rollout with the highest reward is selected as the representative. Representative responses that remain incorrect under the reward signal are retained in the failure pool. The pool is grouped by EU/EA type and by selected distractor. Representative failure clusters are sent to GPT-5.2, which proposes up to $\Delta_{\max}$=3 updates (add, refine, or prune). Candidates with cosine similarity above 0.9 to existing skills are rejected as duplicates. Approved updates are written to the bank and take effect immediately for subsequent rollouts.

Step	Failure Pattern	Generated/Refined Skill	Before	After
100	EU: Model confuses guilt with regret when agent caused harm unintentionally	*Fine-grained self-attribution*: "Distinguish guilt (I did something wrong) from regret (I wish it hadn't happened). Check whether the agent views the outcome as caused by their deliberate action or by circumstances."	Predicts *guilt* (wrong)	Predicts *regret* (correct)
150	EA: Model suggests direct advice instead of empathetic acknowledgment in caregiver-fatigue scenarios	*Caregiver emotional validation*: "When a caregiver expresses exhaustion or frustration, prioritize emotional validation over practical suggestions. Premature problem-solving signals the speaker's distress has not been heard."	Suggests time management tips (wrong)	Acknowledges emotional burden first (correct)

Table 9: Skill evolution case study. Each row traces the chain from observed failure cluster to generated skill to behavioral change. Skills target specific, recurring error patterns rather than general heuristics.