Experience-Augmented Perspective Transition via Reinforcement Learning for Subjective Reasoning in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities on objective tasks such as mathematical reasoning and code generation through long Chain of Thought (CoT) and Reinforcement Learning (RL). However, their performance on subjective tasks, such as cultural understanding and dark humor recognition remains limited, as these tasks require capturing contextual cues and using diverse perspectives. Directly applying existing RL methods to subjective tasks is challenging because optimization trajectories often miss implicit pragmatic signals, while reward supervision is sparse and provides little guidance for policy improvement. To address this, we propose **Ema** (**E**xperience-Aug**m**ented Perspective Tr**a**nsition), a reinforcement learning framework that combines perspective transition with experience augmentation. During training, **Ema** encourages the model to reason from alternative perspectives, improving exploration over diverse subjective interpretations. It further synthesizes effective experiences from low-scoring trajectories and incorporates them into training prompts as explicit guidance for optimization. These experiences are internalized into model parameters during training and introduce no additional cost at inference time. Experiments on 11 subjective reasoning benchmarks, including 3 in-domain and 8 out-of-domain tasks, show that Ema achieves 75.98% overall accuracy, surpassing vanilla GRPO by +3.32% with particularly notable gains in out-of-domain settings (+5.39%). In addition, **Ema** also yields improvements on objective reasoning benchmarks, including CommonSenseQA, GSM8K, and AIME2024.

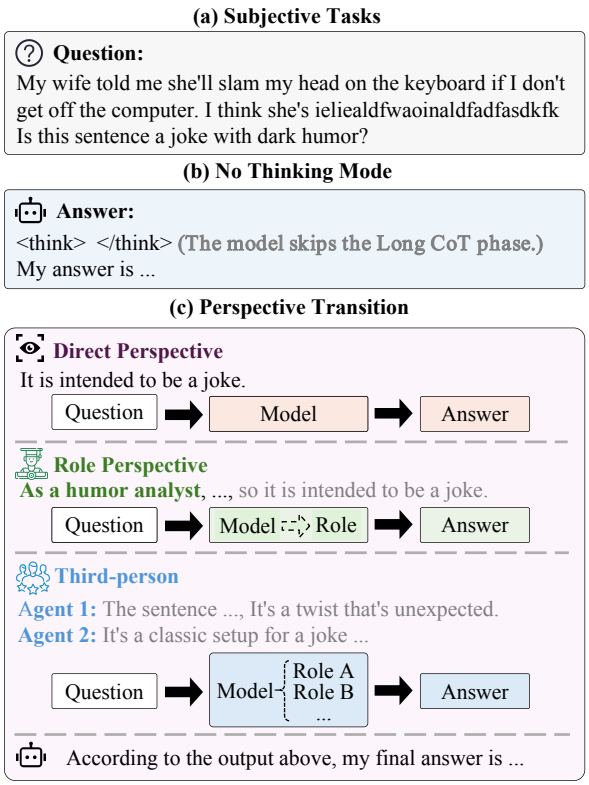

Figure 1: Figure 1(a) presents an example of subjective tasks, where the question constructs a form of dark humor. Figure 1(b) shows an example of "No Thinking" mode. Figure 1(c) provides an example of perspective transition. Leveraging the analysis derived from the aforementioned three perspectives, the LLM produces its final answer.

1 Introduction

Recently, Large Language Models (LLMs) (Brown et al., 2020; Team et al., 2025; Zhao et al., 2026) have advanced rapidly and achieved remarkable success on objective tasks, such as mathematical reasoning (Shao et al., 2024, 2025) and code generation (GLM-5-Team: et al., 2026; OpenAI et al., 2025). However, their capability on subjective tasks still remains limited. Unlike objective tasks, subjective tasks such as cultural understanding and dark humor recognition require LLMs to capture subtle contextual cues, implicit pragmatic meanings, and diverse perspectives (Song et al., 2025). This limitation motivates exploring techniques specifically suited for subjective reasoning.

Motivated by their success on objective reasoning tasks, recent studies have explored reasoning-

oriented techniques, including long Chain-of-Thought (CoT) prompting (Wei et al., 2022) and Reinforcement Learning (RL) (Wang et al., 2026), especially Group Relative Policy Optimization (GRPO) (Shao et al., 2024). While these methods have shown strong effectiveness on many objective reasoning benchmarks, they also present clear limitations. First, they often rely on long reasoning chains, resulting in substantial token consumption and inference latency (Ma et al., 2025). Second, these approaches may not directly benefit subjective tasks, where optimization trajectories often fail to capture implicit pragmatic signals and reward supervision is sparse, providing limited guidance for policy improvement. As Figure 1 (b) shows, although recent open-source models such as Qwen3 (Yang et al., 2024a,b, 2025) offer a "No Thinking" mode to improve efficiency by skipping long CoT, how to improve subjective reasoning under this setting remains underexplored.

A key insight is that subjective reasoning is inherently perspective-dependent (Meaney et al., 2021; Jentzsch and Kersting, 2023; Mao et al., 2024; Wang et al., 2025). As shown in Figure 1 (a)(c), understanding subjective inputs often requires reasoning from multiple perspectives to uncover implicit meanings. However, existing RL pipelines typically operate under a single implicit perspective and rely on scalar rewards for supervision. This raises the question of how RL can more effectively optimize LLMs for subjective tasks without relying on long CoT. Inspired by recent experience-based RL studies (Lu et al., 2025; Zhang et al., 2025; Li et al., 2025; Shi et al., 2026), we conducted a pilot study to examine whether existing experience replay strategies can be transferred directly to this setting. Surprisingly, as Section 2 shows, naively introducing experience replay does not improve subjective-task performance in "No Thinking" mode, and may even underperform vanilla GRPO. This observation suggests that subjective reasoning requires more than directly reusing techniques designed for objective tasks; it necessitates exploration across diverse interpretations and more explicit guidance beyond sparse scalar rewards.

To address these challenges, we propose **Ema** (**E**xperience-Aug**m**ented Perspective Tr**a**nsition), a reinforcement learning framework that integrates perspective transition with experience augmentation. During training, **Ema** encourages the model to reason from alternative perspectives, enhancing exploration across diverse subjective interpretations. It further synthesizes effective experiences from low-scoring trajectories and incorporates them into training prompts as explicit guidance for policy optimization. By internalizing these experiences into model parameters, **Ema** expands the exploration space of subjective reasoning while introducing no additional cost during inference.

Experiments on 11 subjective reasoning benchmarks, including 3 in-domain and 8 out-of-domain (OOD) tasks, show that **Ema** consistently outperforms vanilla GRPO and generalizes well to unseen task distributions. In addition, **Ema** also yields improvements on objective reasoning benchmarks, including CommonSenseQA (Talmor et al., 2019), GSM8K (Cobbe et al., 2021), and AIME2024[1]. These results indicate that enhancing subjective reasoning in the "No Thinking" mode does not compromise objective reasoning ability.

Main contributions are summarized as follows:

- We propose **Ema**, a RL framework that improves LLM performance on subjective tasks by combining perspective transition with experience augmentation.
- We introduce a training mechanism that couples perspective transition with experience augmentation, using synthesized experiences from low-scoring trajectories as guidance during training with no extra inference cost.
- Extensive experiments on 11 subjective reasoning benchmarks show that **Ema** consistently outperforms vanilla GRPO, generalizes to unseen task, and also improves objective reasoning performance.

2 Pilot Analysis

This pilot analysis investigates the efficacy of the experience replay mechanism in the absence of long CoT. Inspired by ERL (Shi et al., 2026), which generates structured self-reflections from low-scoring trajectories on objective tasks, we devise an experience synthesis mechanism, which is detailed in Section 3.2. Briefly, we employ Qwen3-32B to extract lessons from low-scoring trajectories generated by Qwen3-8B during GRPO, with both models operating strictly in the "No Thinking" mode. These extracted experiences are then directly concatenated with the training prompts (see Section 4.1 for full training configurations).

[1] https://huggingface.co/datasets/Maxwell-Jia/AIME_2024

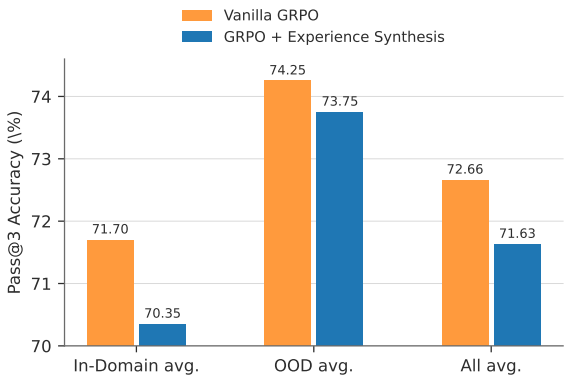

Figure 2: Results of the pilot analysis. Naively augmenting GRPO with synthesized experiences consistently degrades Pass@3 accuracy across in-domain, OOD, and overall averages compared to vanilla GRPO.

As illustrated in Figure 2, **naively applying this experience replay mechanism degrades model performance**. This finding motivates our introduction of a perspective transition mechanism. In subsequent sections, we empirically demonstrate that coupling perspective transition with experience replay yields superior performance compared to applying either approach in isolation.

3 Method

We propose **Ema** (**E**xperience-Aug**m**ented Perspective Tr**a**nsition), a method that unifies perspective transition and experience augmentation within the GRPO training loop to tackle subjective tasks. For each question $q \in \mathcal{D}$, a transformation function $\mathcal{T}(q, \mathcal{P})$ converts q into a multi-perspective prompt q_p, which instructs the policy model to generate a group of i outputs $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_i\}$. Each output $\mathcal{O}_j$ is scored by the environment to obtain a reward r_j. Trajectories that fall below the group mean $\bar{r} = \frac{1}{i}\sum_{j=1}^{i} r_j$ are sampled via an ε-greedy strategy and sent to a frozen summarization model, which synthesizes failure experiences into an experience buffer E. The buffer is concatenated with q_p to form the final prompt $\tilde{q}_p = q_p \oplus E$. Since these experiences are internalized into model parameters through gradient updates, the experience buffer is entirely discarded at inference time, incurring zero additional computational overhead. An overview of the complete pipeline is illustrated in Figure 3.

3.1 Perspective Transition

Building on the perspective taxonomy from RPT (Wang et al., 2025), we adopt the specific combination of perspectives defined in their work.

#	Perspective	XML Tag
1	Direct	`<Direct Perspective>`
2	Role	`<Role Perspective>`
3	Third-person	`<Third-person>`
4	Final Answer	`<Answer>`

Table 1: Output segments and their XML-style tags. Each output $\mathcal{O}_j$ consists of four segments delimited by matching opening/closing tags. Segments 1–3 correspond to the three perspectives in $\mathcal{P}$; Segment 4 synthesizes them into a final answer. The full prompt template is in Appendix D.

Specifically, we define a set of three perspectives $\mathcal{P} = \{p_{third}, p_{role}, p_{direct}\}$, corresponding to the Third-person Perspective, Role Perspective and Direct Perspective respectively. The *Direct Perspective* generates answers directly through prompting strategies. The *Role Perspective* assigns specific identities to elicit specialized reasoning patterns. The *Third-person Perspective* utilizes contextual signals from multi-agent interactions. For each input question $q \in \mathcal{D}$, a prompt transformation function $\mathcal{T}(q, \mathcal{P})$ restructures q into a perspective transition prompt q_p. Concretely, q_p requires the model to organize each output $\mathcal{O}_j$ into four sequential segments delineated by explicit XML-style tags, as shown in Table 1.

verte After generating a group of i outputs for a given q_p, each output $\mathcal{O}_j$ receives a scalar reward $r_j \in [0, 2.0]$: (1) an accuracy reward of $+1.0$ if the answer within `<Answer>` tags matches the ground-truth label, and (2) a format reward of up to $+1.0$, where each correctly closed tag pair contributes $+0.25$. The group-level mean reward $\bar{r} = \frac{1}{i}\sum_{j=1}^{i} r_j$ and the relative advantage $\mathcal{A}_j = \frac{r_j - \bar{r}}{\sigma}$ are then computed, where σ is the standard deviation of $r = \{r_1, r_2, \ldots, r_i\}$.

3.2 Experience Synthesis

Below-average trajectories contain valuable diagnostic information about *why* the model fails on subjective tasks. We introduce an experience synthesis mechanism that extracts actionable experiences from these outputs to guide the policy toward correct reasoning pathways.

Sampling and Summarization. After advantage computation, we identify the set of below-average outputs $\mathcal{O}_{low} = \{\mathcal{O}_j \in \mathcal{O} \mid r_j < \bar{r}\}$. We adopt an ε-greedy strategy: each trajectory in $\mathcal{O}_{low}$ is sampled with probability ε for experience extraction. The sampled $\mathcal{O}_j$, together with q_p and the

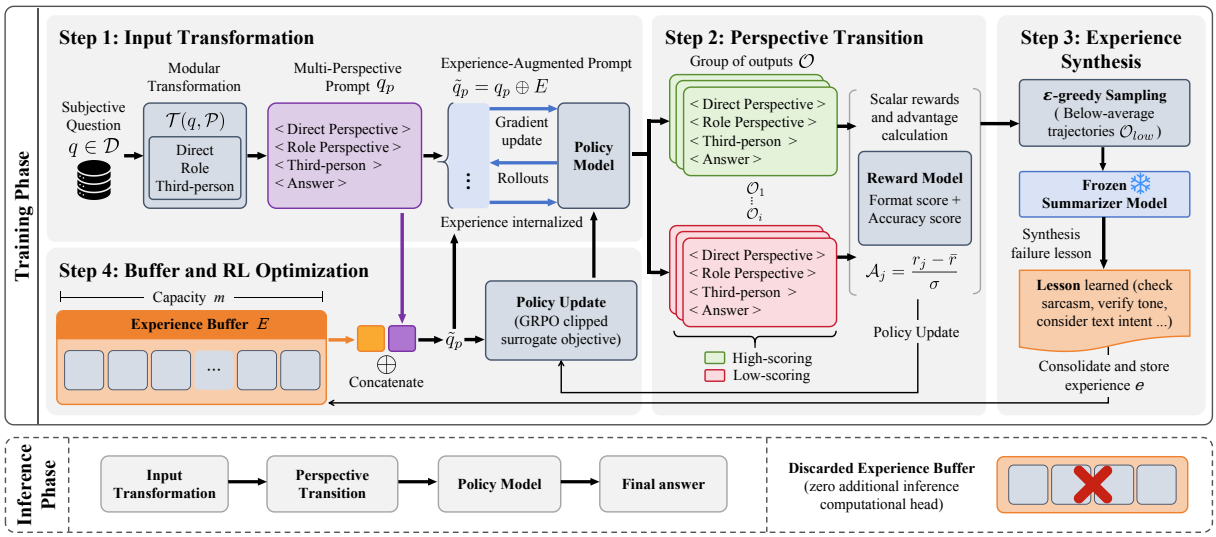

Figure 3: Overview of the proposed **Ema** method. During training, a question q is transformed into a perspective transition prompt q_p to generate a group of outputs. Each output is evaluated for format and accuracy to compute a scalar reward. Trajectories with below-average rewards are fed into a frozen summary model to synthesize diagnostic experiences into an experience buffer E. This buffer is then concatenated with q_p to optimize the policy model via GRPO. The experience buffer is entirely discarded during inference.

ground truth of this question, are fed to a frozen summarization model, which produces a concise experience e that captures the lesson learned. An example is shown in Appendix E.

Experience Buffer Management. The generated experience e is appended to the experience buffer E, which is updated every n training steps with a maximum capacity of m entries. When the total number of accumulated experiences exceeds m, all experiences are summarized into no more than m consolidated entries.

Experience concatenation and Policy Optimization. The updated E is concatenated with the multi-perspective prompt to form the final training prompt $\tilde{q}_p = q_p \oplus E$. The policy is then optimized by maximizing the GRPO clipped surrogate objective (Shao et al., 2024):

$$\mathcal{J}(\theta) = \mathbb{E}_{\{\mathcal{O}_j\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \underbrace{q_p \oplus E}_{\tilde{q}_p})} \left[\sum_{j=1}^{i} \min\Big(\rho_j \, \mathcal{A}_j, \; \text{clip}\big(\rho_j, \, 1 \pm \epsilon_{\text{clip}}\big) \, \mathcal{A}_j \Big) \right]$$

where ρ_j is the importance sampling ratio between π_θ and $\pi_{\theta_{old}}$. The critical distinction from vanilla GRPO is that all rollouts are conditioned on the experience-enriched prompt $\tilde{q}_p$ rather than the original q_p. This means the accumulated experiences E participate in every forward pass, and through the subsequent gradient update, are progressively *internalized* into the model parameters θ. Consequently, the experience buffer E can be entirely discarded at inference time.

4 Experiments

4.1 Settings

Datasets. We evaluate on 11 subjective task datasets, including SocNorm (CH-Wang et al., 2023), CALI (Huang and Yang, 2023), Semeval (Mohammad et al., 2016), and 8 datasets from BigBench[2] (Srivastava et al., 2023). We designate SocNorm, CALI, and Semeval as our training sets, with their respective held-out test sets serving as In-Domain evaluation data. All 8 BigBench datasets are used exclusively as Out-of-Domain (OOD) test data to assess generalization. Detailed descriptions of each dataset are provided in Appendix B.1.

Models. Proprietary baselines include Gemini-3-Flash (Google DeepMind, 2025), Claude-Sonnet-4.5 (Anthropic, 2025), Grok-4-Fast (xAI, 2025), and GPT-5-mini (OpenAI, 2025), evaluated under Directly Answer. Open-source baselines include Qwen3-14B, Qwen3-30B-A3B, Qwen3-32B (Yang et al., 2024a,b, 2025) and GPT-oss-20B (OpenAI, 2025), evaluated under both Directly Answer (Dir.)

[2] https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks

Method	Set.	In-Domain				Out-of-Domain									Avg.
		SocNorm (Acc.)	CALI (Acc.)	Semeval (Acc.)	Avg. (Acc.)	Anach. (Acc.)	Entail. (Acc.)	Humor (Acc.)	Disambi. (Acc.)	IPA (Acc.)	Metaphor (Acc.)	SEQ (Acc.)	SNARKS (Acc.)	Avg. (Acc.)	
Random	-	33.33	33.33	50.00	41.77	50.00	50.00	50.00	33.33	33.33	50.00	25.00	50.00	44.67	42.86
Proprietary Models															
Gemini-3-Flash	Dir.	50.91	19.55	18.83	27.51	32.17	68.57	35.00	55.43	61.90	74.26	22.61	60.77	58.16	39.02
Claude-Sonnet-4.5	Dir.	60.81	46.52	80.63	67.59	73.48	77.14	73.75	66.28	76.19	94.12	98.26	82.32	83.39	73.52
Grok-4-Fast	Dir.	59.38	35.30	82.33	65.51	52.17	82.86	60.00	46.51	72.22	90.74	96.52	74.59	74.71	68.97
GPT-5-mini	Dir.	58.20	56.52	81.72	69.73	89.57	88.57	62.50	72.87	84.92	95.15	100.00	89.50	88.33	76.72
Open-Source Models															
Qwen3-14B	Dir.	61.72	48.13	72.71	64.19	56.96	78.57	63.75	54.52	64.29	90.15	93.33	65.19	74.54	68.08
	Pers.	59.72	47.47	71.65	62.97	71.01	**89.52**	67.50	33.72	62.96	86.91	96.23	79.19	74.21	67.19
Qwen3-30B-A3B	Dir.	62.33	44.70	83.29	68.92	72.32	75.71	71.25	43.54	45.24	91.67	85.51	77.35	75.13	71.26
	Pers.	60.29	46.87	76.17	65.27	72.46	81.90	70.83	34.75	62.96	89.02	95.65	78.82	75.15	68.98
Qwen3-32B	Dir.	66.15	48.64	80.29	69.32	68.41	78.10	65.83	40.83	**76.19**	**92.60**	**97.97**	73.48	77.09	72.23
	Pers.	61.85	46.16	75.63	65.25	74.64	88.57	67.92	34.11	72.75	88.77	97.68	**82.14**	76.57	69.50
GPT-oss-20B	Dir.	32.24	32.58	40.34	36.44	67.39	29.05	21.67	46.64	28.04	24.56	57.10	56.35	39.25	37.50
	Pers.	36.20	37.98	67.92	52.67	49.57	81.90	70.00	52.71	61.90	81.67	87.83	80.29	71.44	59.72
Ours (Qwen3-8B)															
Base	Dir.	59.38	41.82	73.87	62.72	55.65	74.29	67.50	47.67	61.11	88.09	89.57	59.67	71.49	66.01
	Pers.	55.47	43.33	67.74	58.92	70.87	82.86	57.50	34.50	30.95	75.44	93.91	64.64	65.11	61.24
SFT	-	61.94	**55.56**	77.94	68.59	50.00	65.71	50.00	62.14	52.91	82.94	51.30	53.04	65.92	67.59
Ema	-	67.88	51.31	**86.99**	**73.78**	75.36	78.57	68.75	**67.31**	61.38	91.57	95.07	65.93	**79.64**	**75.98**
w/o Exp.	-	**69.01**	50.61	85.06	72.94	**75.65**	72.86	70.00	58.91	60.32	90.15	94.78	62.98	77.30	74.58
w/o Pers.	-	65.63	49.75	82.10	70.35	58.70	74.29	70.00	51.81	69.31	88.43	91.30	62.43	73.75	71.63
w/o Exp. + Pers.	-	65.23	52.12	83.90	71.70	61.30	82.86	**75.00**	53.10	63.49	88.38	90.43	61.33	74.25	72.66

Table 2: Main results of various baselines and our proposed **Ema** in zero-shot settings. **Dir.**, **Pers.**, and **Exp.** denote *Directly Answer*, *Perspective Transition*, and *Experience Synthesis* experimental settings, respectively. *Random* represents the result of random prediction with uniform probability. Note that **w/o Exp. + Pers.** corresponds to the vanilla GRPO baseline. For each dataset, the best result is **bolded** and the second-best result is underlined (compared between open-source models and ours). Due to budget constraints, we do not evaluate the performance of proprietary models under the **Pers.** setting.

and Perspective Transition (Pers.) settings.

We select Qwen3-8B as the primary model and additionally train Ema on Qwen3-4B, Qwen3-1.7B, and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) to assess scalability across scales and architectures. Prompt details are in Appendix D.

Training and Inference. We use VeRL (Zhang et al., 2024; Sheng et al., 2024) as our RL framework. All models are trained with a batch size of 1,024, a learning rate of 1×10^{-6}, and a rollout group size of 6. A frozen Qwen3-32B serves as the summarization model. The experience buffer is updated every $n = 30$ steps with a capacity of $m = 30$ and an exploration rate of $\varepsilon = 0.5$. The reward of the main experiment follows the four-segment format in Section 3.1. The experiment without Perspective Transition (w/o Perspective) uses a simplified reward. Additionally, we leverage LlamaFactory (Zheng et al., 2024) to fine-tune a model as the Supervised Fine-Tuning (SFT) baseline. Appendix B.2 shows more details of our experiments. All inference is performed at a temperature of 0 in "No Thinking" mode. Apart from proprietary models, all results are reported under pass@3 settings. Also, in order to assess experience quality, we manually sampled and examined the experiences of the main experiment. See Appendix C.1 for more details. We further analyze the output token consumption of Ema under "No Thinking" mode in Appendix C.2, confirming that perspective transition incurs substantially fewer tokens than long-CoT methods.

4.2 Main Results

Ema shows competitiveness with GPT-5-mini and surpasses much larger open-source models. As Table 2 shows, Ema achieves the highest overall average accuracy of 75.98%, surpassing the strongest open-source baseline Qwen3-32B (Dir.). Compared to the vanilla GRPO (w/o Exp. + Pers.), Ema yields a consistent improvement of 3.32% overall, with particularly notable gains in OOD (5.39%) settings. Despite being significantly smaller, Ema (Qwen3-8B) matches Claude-Sonnet-4.5 (73.52%) and surpasses Grok-4-Fast (68.97%), while remaining competitive with GPT-5-mini (76.72%). Furthermore, Figure 4 demonstrates that

Model	CSQA	GSM8K	AIME 2024
Zero-shot CoT	80.81	85.41	76.00
SFT	81.00	91.62	78.30
SFT+GRPO	81.19	91.97	79.60
MultiRole-R1	82.10	94.98	**80.10**
Ema	**87.14**	**95.15**	80.00

Table 3: We present the performance on objective tasks across the GSM8K, AIME 2024, and CommonSenseQA (CSQA) datasets. For each dataset, the best result is shown in **bold**, and the second-best is underlined.

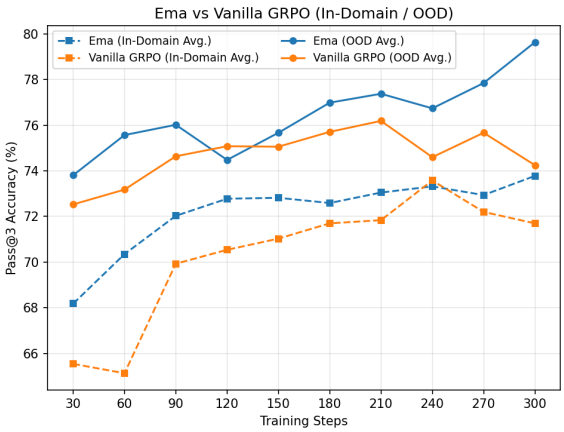

Figure 4: Comparison of Pass@3 accuracy of our proposed method **Ema** and Vanilla GRPO.

Ema exhibits steady improvement throughout training compared to vanilla GRPO. Details of each dataset are shown in Appendix C.3.

Ema exhibits strong OOD generalization without compromising objective reasoning. Ema achieves an OOD average of 79.64%, demonstrating generalization across unseen tasks. Moreover, our method does not degrade objective reasoning capabilities. As shown in Table 3, we evaluate Ema *with long-CoT* (i.e., enabling the model's inherent reasoning steps) on mathematical and logical benchmarks such as GSM8K and AIME 2024. Ema achieves the best performance on CSQA (87.14) and GSM8K (95.15), while remaining highly competitive on AIME 2024 (80.00), closely matching the strongest baseline MultiRole-R1 (80.10). These results indicate that optimizing subjective reasoning through experience-aware reinforcement learning does not impair the model's general reasoning capability, but instead maintains robust performance across both subjective and objective tasks. Notably, except for Ema, all other results are directly reported from the original MultiRole-R1 paper (Wang et al., 2026).

Case study. Figure 5 provides a concrete example from the SNARKS sarcasm detection task. Under Perspective Transition settings, the model treats "It's been a fun semester" as straightforwardly positive and selects the incorrect answer (A). After training with Ema, the model recognizes that "fun" contradicts the negative experiences in the context, correctly identifying the sarcasm.

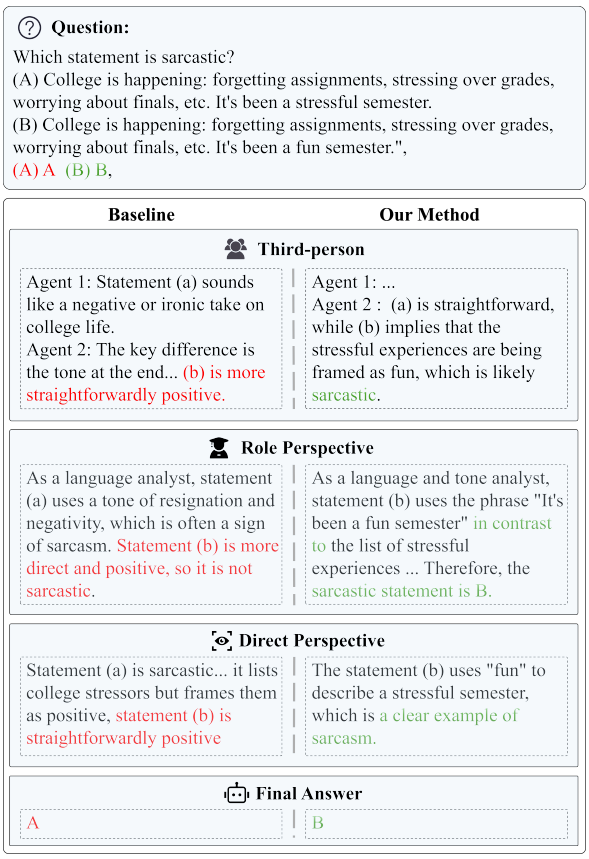

Figure 5: A case study on the SNARKS sarcasm detection task. Using our proposed method, the LLM identifies the contradiction between "fun" and the stressful college experiences as a clear sarcastic device, correctly selecting answer (B).

5 Analysis and Discussion

5.1 Ablation Study

Both components contribute complementary gains. As mentioned in our pilot analysis, naively applying experience replay degrades model performance. However, as Figure 6 shows, we surprisingly find that when successfully coupled, the components yield complementary gains. These results confirm the two mechanisms are mutually reinforcing rather than merely additive.

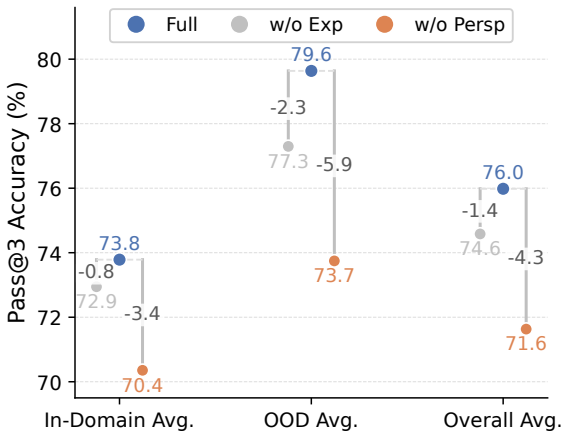

Figure 6: We conduct an ablation study on **Ema**. The Results show that removing any component leads to a sharp drop, particularly on OOD tasks.

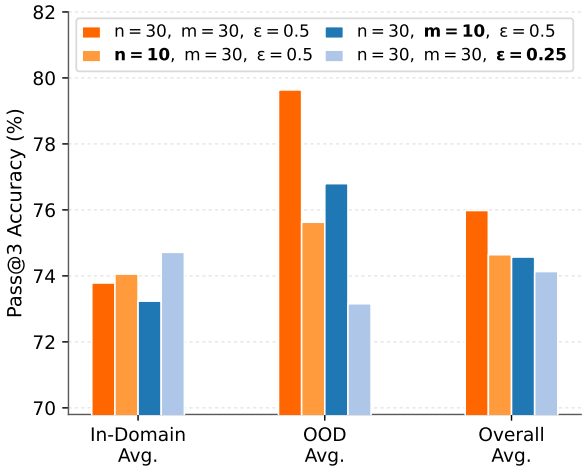

Figure 7: We conduct experiments for our proposed method **Ema** to analyze the impact of varying hyperparameter values on the performance of Qwen3-8B. The parameters n, m, and ϵ denote the *update frequency*, *maximum buffer capacity*, and *exploration rate*, respectively. Note that n=30, m=30 and ε=0.5 are the settings of our main experiment.

Scaling behavior across different models. As Table 4 shows, on Qwen3-4B, Ema achieves an overall average of 74.06%, improving over vanilla GRPO (69.31%) by +4.75%, confirming that a 4B-parameter model possesses sufficient capacity to leverage both perspective transition and experience synthesis effectively. However, when scaled down to Qwen3-1.7B, Ema performs comparably to vanilla GRPO. We hypothesize this is due to capacity limitations. On Llama-3.1-8B-Instruct, vanilla GRPO achieves a marginally higher In-Domain average but suffers severe OOD degradation. Most notably, SEQ accuracy collapses from 82.32% to 43.48%, while Ema maintains stable OOD performance.

5.2 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of Ema to three key experience buffer hyperparameters on Qwen3-8B: update frequency n, maximum buffer capacity m, and sampling probability ε. As Figure 7 shows, a consistent pattern emerges across all three ablations: In-Domain accuracy remains relatively stable, while OOD accuracy absorbs the majority of the performance loss. The full results are reported in Appendix C.4.

Update frequency (n). Reducing n from 30 to 10 (i.e., updating the buffer more frequently) significantly degrades OOD performance (−4.01%). We hypothesize that prematurely refreshing the buffer disrupts the gradient consistency required for knowledge consolidation.

Buffer capacity (m)). Reducing the buffer capacity m from 30 to 10 slightly degrades overall performance by 1.41%. Notably, this restriction causes a sharp performance drop on the Disambiguation QA task. A smaller buffer is less able to maintain diverse experiences across different failure modes, effectively limiting the model's exposure to a broader range of corrective signals.

Exploration rate (ε). Reducing ε from 0.5 to 0.25 leads to the most revealing behavior: we observe that the experience buffer stops updating during training. With a lower sampling probability, too few below-average trajectories are selected for experience extraction. As a result, the model over-exploits a fixed set of early-stage experiences rather than adapting to its evolving policy. While this yields the best In-Domain average (+0.93%), OOD performance drops sharply (−6.48%). IPA collapses from 61.38% to 30.16%, and Disambiguation QA drops to 47.29%.

6 Related Work

Subjective Tasks. Subjective NLP tasks, including metaphor recognition and dark humor detection (Rottger et al., 2022; Kanclerz et al., 2023; Sun et al., 2025; Song et al., 2025), require models to interpret meanings that depend heavily on context, implicit knowledge, and human perception. Existing studies (Wang et al., 2024b; Lv et al., 2024; Cheng et al., 2024; Wang et al., 2025; Aoyagui

Method	Set.	In-Domain				Out-of-Domain									Avg.
		SocNorm (Acc.)	CALI (Acc.)	Semeval (Acc.)	Avg. (Acc.)	Anach. (Acc.)	Entail. (Acc.)	Humor (Acc.)	Disambi. (Acc.)	IPA (Acc.)	Metaphor (Acc.)	SEQ (Acc.)	SNARKS (Acc.)	Avg. (Acc.)	
Qwen3-4B															
Base	Dir.	58.85	41.67	66.10	58.60	63.04	58.57	65.00	53.49	47.62	80.59	82.61	61.88	68.45	62.30
	Pers.	60.42	41.97	67.46	59.67	**76.09**	**80.00**	60.00	48.84	46.03	81.03	**93.04**	61.88	70.86	63.87
Ema		**69.27**	47.73	**86.97**	**73.32**	74.78	67.14	**67.50**	**60.85**	**63.49**	**87.50**	89.57	56.35	**75.29**	**74.06**
w/o Exp. + Pers.		63.54	**48.18**	81.63	69.20	65.65	61.43	66.25	51.03	49.74	81.62	85.22	**63.54**	69.50	69.31
Qwen3-1.7B															
Base	Dir.	50.74	33.38	54.57	48.72	50.14	64.29	56.25	42.76	19.31	**81.03**	64.35	53.96	61.07	53.36
	Pers.	49.13	14.75	51.41	42.44	63.04	**79.52**	66.67	30.36	21.69	78.48	75.65	28.18	59.27	48.76
Ema		**63.54**	45.91	81.81	68.77	**63.48**	64.29	58.75	49.48	**52.65**	76.91	**81.16**	51.75	**65.63**	67.59
w/o Exp. + Pers.		63.02	**46.97**	**82.72**	**69.34**	53.77	64.29	**72.50**	**51.16**	42.06	79.71	67.83	**60.77**	65.61	**67.94**
Llama-3.1-8B-Instruct															
Base	Dir.	56.86	41.06	79.26	64.60	47.83	72.38	**80.42**	56.59	45.24	85.25	82.32	55.25	69.10	66.29
	Pers.	54.86	38.54	80.97	64.36	58.55	68.57	65.83	59.04	**52.91**	83.58	**91.59**	**60.22**	71.09	66.89
Ema		66.67	57.12	**87.38**	74.98	**74.35**	72.86	67.50	59.69	48.77	**88.92**	90.14	47.88	**73.87**	**74.56**
w/o Exp. + Pers.		**68.49**	**57.27**	87.27	**75.44**	50.00	**75.71**	55.00	**61.50**	51.59	87.70	43.48	52.49	67.64	72.51

Table 4: Ablation study of Ema across different model scales and different models. For each dataset, the best result is **bolded** and the second-best result is underlined. **Ema** denotes our full proposed method, while **w/o** rows indicate the impact of removing specific components. Note that **w/o Exp. + Pers.** corresponds to the vanilla GRPO baseline.

et al., 2025) have explored no-training methods to address these challenges. Recent work, such as MultiRole-R1 (Wang et al., 2026), has leveraged RL to enhance performance on subjective NLP tasks. However, these methods typically rely on long CoT, leaving a gap for approaches that can function effectively without long CoT processes.

LLM Perspective. Following the perspective taxonomy of Wang et al. (2025), existing LLM-based approaches can be categorized into three groups. The *direct perspective* generates answers directly through prompting strategies such as standard prompting (Brown et al., 2020), Chain-of-Thought reasoning (Wei et al., 2022), and zero-shot CoT (Kojima et al., 2022). The *role perspective* assigns specific identities (e.g., experts or domain specialists) to elicit specialized reasoning patterns (Xu et al., 2023; Wang et al., 2024c; Wilf et al., 2024). The *third-person perspective* utilizes contextual signals from multi-agent interactions, debates, or simulated dialogues to support reasoning (Wang et al., 2024d,a; Du et al., 2024).

Reinforcement Learning. Reinforcement learning has been widely adopted to align and improve large language models. Recent studies further demonstrate that RL can directly enhance reasoning capabilities, as seen in GRPO for DeepSeek-Math (Shao et al., 2024) and large-scale RL in DeepSeek-R1 (Guo et al., 2025). However, GRPO often provides only scalar rewards, which can be too sparse for effective model training (Lu et al., 2025). To address this, several studies, including RLEP (Zhang et al., 2025), RePO (Li et al., 2025), and ERL (Shi et al., 2026), leverage experience replay mechanisms to select successful trajectories or extract lessons from failures. Training-free GRPO (Cai et al., 2025) attempts to maintain an experience buffer to improve performance without additional training. However, these methods focus exclusively on objective tasks and optimize outputs under fixed reasoning setups. In contrast, **Ema** improves LLM performance on subjective tasks by combining perspective transition with experience augmentation. Furthermore, **Ema** handles subjective tasks efficiently without relying on long CoT during either training or inference.

7 Conclusion

We presented **Ema** (**E**xperience-Aug**m**ented Perspective Tr**a**nsition), a RL framework that improves LLM performance on subjective tasks that unifies perspective transition and experience augmentation. Perspective transition helps the model to infer through multi-view reasoning, while experience augmentation provides additional guidance by extracting lessons from low-reward trajectories. Experiments on 11 subjective reasoning benchmarks illustrate that **Ema** outperforms vanilla GRPO by 3.32%, and achieves an especially remarkable improvement of 5.39% in OOD settings. Our findings also bridge a research gap in subjective reasoning without long CoT phases.

Limitations

We acknowledge three limitations: (1) Ema requires task-specific tuning of buffer hyperparameters (n, m, ϵ), which may limit plug-and-play applicability. (2) Our evaluation focuses on English text-based subjective tasks; extension to multimodal subjective reasoning (e.g., visual humor, sarcasm in images) and cross-lingual settings remains unexplored. (3) our experience synthesis relies on a frozen 32B summarization model, which introduces additional computational cost during training and may bias the extracted experiences toward the capabilities of that specific model.

References

Anthropic. 2025. Introducing claude sonnet 4.5. Online.

Paula Akemi Aoyagui, Kelsey Stemmler, Sharon A Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A matter of perspective(s): Contrasting human and llm argumentation in subjective decision-making on subtle sexism. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–16. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Yong Mao, Ke Li, and Xing Sun. 2025. Training-free group relative policy optimization. *Preprint*, arXiv:2510.08191.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Ruoxi Cheng, Haoxuan Ma, Shuirong Cao, Jiaqi Li, Aihua Pei, Zhiqiang Wang, Pengliang Ji, Haoyu Wang, and Jiaqi Huo. 2024. Reinforcement learning from multi-role debates as feedback for bias mitigation in llms. *Preprint*, arXiv:2404.10160.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

GLM-5-Team:, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, and 167 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Google DeepMind. 2025. Gemini 3 flash - best for frontier intelligence at speed. Online.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao 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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values. *Preprint*, arXiv:2008.02275.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Kamil Kanclerz, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocon, and Przemyslaw Kazienko. 2023. PALS:

Personalized active learning for subjective tasks in NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13326–13341, Singapore. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. 2025. Repo: Replay-enhanced policy optimization. *Preprint*, arXiv:2506.09340.

Fanbin Lu, Zhisheng Zhong, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2025. Arpo:end-to-end policy optimization for gui agents with experience replay. *Preprint*, arXiv:2505.16282.

Fangrui Lv, Kaixiong Gong, Jian Liang, Xinyu Pang, and Changshui Zhang. 2024. Subjective topic meets LLMs: Unleashing comprehensive, reflective and creative thinking through the negation of negation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12341, Miami, Florida, USA. Association for Computational Linguistics.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *Preprint*, arXiv:2504.09858.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

OpenAI, :, Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, and 7 others. 2025. Competitive programming with large reasoning models. *Preprint*, arXiv:2502.06807.

OpenAI. 2025. Gpt-5 system card. Online.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Zhihong Shao, Yuxiang Luo, Chengda Lu, Z. Z. Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *Preprint*, arXiv:2511.22570.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Taiwei Shi, Sihao Chen, Bowen Jiang, Linxin Song, Longqi Yang, and Jieyu Zhao. 2026. Experiential reinforcement learning. *Preprint*, arXiv:2602.13949.

Changhao Song, Yazhou Zhang, Hui Gao, Ben Yao, and Peng Zhang. 2025. Large language models for subjective language understanding: A survey. *Preprint*, arXiv:2508.07959.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024a. DEEM: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541, Torino, Italia. ELRA and ICCL.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024b. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15880–15893, Bangkok, Thailand. Association for Computational Linguistics.

Xiaolong Wang, Yuanchi Zhang, Ziyue Wang, Yuzhuang Xu, Fuwen Luo, Yile Wang, Peng Li, and Yang Liu. 2025. Perspective transition of large language models for solving subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9686–9704, Vienna, Austria. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Yumeng Wang, Zhiyuan Fan, Jiayu Liu, Jen tse Huang, and Yi R. Fung. 2026. Diversity-enhanced reasoning for subjective questions. *Preprint*, arXiv:2507.20187.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024d. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. 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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

xAI. 2025. Grok. Online.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. ExpertPrompting: Instructing large language models to be distinguished experts. *ArXiv preprint*, abs/2305.14688.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Chi Zhang, Guangming Sheng, Siyao Liu, Jiahao Li, Ziyuan Feng, Zherui Liu, Xin Liu, Xiaoying Jia, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. A framework for training large language models for code generation via proximal policy optimization. https://i.cs.hku.hk/~cwu/papers/gmsheng-NL2Code24.pdf.

Hongzhi Zhang, Jia Fu, Jingyuan Zhang, Kai Fu, Qi Wang, Fuzheng Zhang, and Guorui Zhou. 2025. Rlep: Reinforcement learning with experience replay for llm reasoning. *Preprint*, arXiv:2507.07451.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2026. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Use of Large Language Model

We use Large Language Models to aid or polish writing.

B Experimental Details

B.1 Dataset Details

We evaluate Ema on 11 subjective reasoning datasets, which cover a wide spectrum of tasks including Linguistic Rhetoric, Disambiguation QA, Stance Detection, Cultural-Related understanding, and Traditional NLI. The statistics and train/test splits for each dataset are provided in Table 5.

Dataset (names in short)	**Subjective Tasks**	**#Train/Test**
(*Linguistic Rhetoric*)		
Metaphor (Mohler et al., 2016)	Metaphor Understanding	-/680
SNARKS (Khodak et al., 2018)	Sarcasm Detection	-/181
Dark Humor (Meaney et al., 2021)	Dark Humor Detection	-/80
(*Disambiguation QA*)		
Pronoun (Rudinger et al., 2018)	Pronoun Resolution	-/258
Anachronisms (Geva et al., 2021)	Identifying Anachronisms	-/230
(*Stance Detection*)		
SEQ (Hendrycks et al., 2023)	Simple Ethical Questions	-/115
SemEval (Mohammad et al., 2016)	Opinion Analysis	2,194/707
(*Cultural-Related*)		
SocNorm (CH-Wang et al., 2023)	Sociocultural Norm NLI	2,301/768
CALI (Huang and Yang, 2023)	Culturally Aware NLI	1,757/440
(*Traditional NLI*)		
Entailment (Srivastava et al., 2023)	Analytic Entailment	-/70
IPA (Williams et al., 2018)	NLI in the International Phonetic Alphabet	-/126

Table 5: Statistics and resources of datasets.

B.2 Implementation and Baseline Details

Because the experiment without Perspective Transition (w/o Perspective) completely omits the multi-perspective prompt, the model is no longer required to generate the intermediate perspective tags. Consequently, we apply a simplified format reward that assigns $+0.25$ solely for correctly enclosing the final output within the `<Answer>` tags, alongside the standard $+1.0$ reward for correctness.

For the Supervised Fine-Tuning (SFT) baseline, we leverage LlamaFactory to fine-tune the model. The model is fine-tuned for 3 epochs with a learning rate of 2×10^{-5}, following a cosine learning rate scheduler with a warmup ratio of 0.1. We use an effective batch size of 16. The SFT dataset is a mixture of samples from CALI, SemEval, and SocNorm, utilizing a specific "No Thinking" template to align with our inference settings.

C Additional Analysis

C.1 Quality Assessment of Synthesized Experiences

To assess the quality of the synthesized experiences, we manually inspected model outputs sampled at training steps 30, 90, 150, 210, and 270. Our evaluation focuses on two critical dimensions: (1) *intra-step duplication*, which measures content redundancy within a single generation step; and (2) *inter-step overlap*, which quantifies the persistence of experiences across different sampled steps.

As summarized in Table 6, although the experience buffer capacity was set to 30, the model consistently yielded slightly fewer entries in practice. We manually verified each entry as string matching might fail to capture semantic redundancy. The

Step	Actual Entries	Intra-step Duplicates	Inter-step Overlap (vs. Previous)
30	28	1 pair	-
90	28	0	22 (78.6%)
150	28	0	7 (25.0%)
210	27	0	25 (92.6%)
270	28	0	25 (89.3%)

Table 6: Unified quality assessment of synthesized experiences. *Intra-step Duplicates* includes approximate semantic matches. *Inter-step Overlap* denotes the count and percentage of entries carried over from the previously sampled step.

results indicate negligible intra-step duplication. However, inter-step overlap becomes significantly more pronounced in the later stages of training. We hypothesize that this trend reflects policy convergence: as the model stabilizes, it tends to generate homogeneous low-reward trajectories, thereby limiting the discovery of novel failure modes.

C.2 Output Token Consumption Analysis

To verify that Ema's perspective transition in "No Thinking" mode does not simply amount to long CoT, we compare four inference configurations on Qwen3-8B. "Long CoT" means we enable the reasoning steps provided by the model itself.

All settings use greedy decoding. The maximum output length is set to 1024 tokens for "No Thinking" and 16384 for Long CoT to ensure complete generation, as we observed that a short maximum output length may prematurely truncate the output, preventing the final answer from being generated. As shown in Table 7, two findings emerge:

Ema in "No Thinking" mode is far more efficient than Long CoT. Ema generates only **288.71** output tokens per sample on average, which is **55.3%** fewer than its own Long-CoT counterpart (645.80) and **58.6%** fewer than the base model's Long CoT (696.88). Despite the substantial token reduction, Ema's overall accuracy (75.98%) remains comparable to its Long-CoT variant (76.87%) and substantially surpasses the base Long-CoT setting by 4.36%.

Ema training does not degrade Long-CoT performance. This finding is consistent with our observations in Table 3. When switched to Long-CoT inference, Ema achieves 76.87% overall accuracy, outperforming the base Qwen3-8B under the same setting (71.62%) by +5.25%. This confirms that Ema training enhances general reasoning capabilities and does not compromise the model's ability to leverage long CoT when it is available.

Method	Set.	Avg. Tokens			Accuracy (%)		
		ID	OOD	All	ID	OOD	All
Qwen3-8B	Base	**226.3**	**245.3**	**233.4**	58.92	65.11	61.24
	Long CoT	762.4	587.9	696.9	66.00	80.98	71.62
Ema	Base	284.8	295.3	288.7	**73.78**	79.64	75.98
	Long CoT	709.3	540.2	645.8	73.39	**82.64**	**76.87**

Table 7: Output token consumption and accuracy across four configurations on Qwen3-8B. The comparison shows that Ema is far more token-efficient than standard Long CoT. The maintained performance in Long-CoT mode confirms that Ema does not compromise the model's ability to leverage extended reasoning steps.

C.3 Training Results

Figure 8 and Figure 9 illustrate the training dynamics of Ema and Vanilla GRPO on both in-domain and OOD datasets. Overall, Ema demonstrates more stable optimization and consistently stronger subjective reasoning performance throughout training. On in-domain tasks, Ema achieves competitive or superior Pass@3 accuracy with smoother trajectories, indicating more reliable convergence. The advantage becomes more pronounced on OOD datasets, where Ema maintains higher accuracy and better stability across most tasks, while Vanilla GRPO often exhibits larger fluctuations or late-stage degradation. These results suggest that Ema helps the model learn more transferable subjective reasoning patterns rather than overfitting to unstable reward signals, leading to improved robustness and generalization.

C.4 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of Ema to three key experience buffer hyperparameters on the Qwen3-8B model: update frequency (n), maximum buffer capacity (m), and exploration rate (ϵ). As illustrated in Table 8, In-Domain accuracy remains relatively stable across variations, while OOD accuracy is highly sensitive to these parameters. Especially, we find a significant performance drop on the Disambiguation QA task and IPA dataset.

D Prompts for Model

Tables 9 to 13 detail the specific prompts used throughout our experiments.

E Synthesized Experiences Example

Table 14 shows representative examples of synthesized experiences that serve as actionable lessons

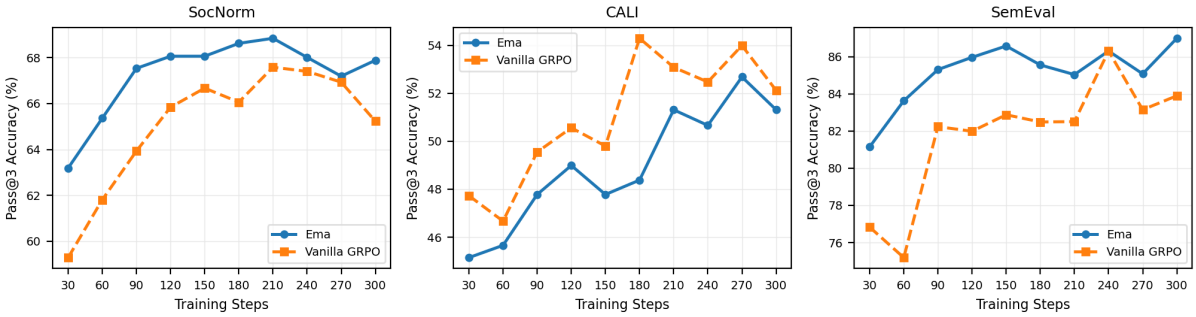

Figure 8: Comparison of Pass@3 accuracy of our proposed method **Ema** and Vanilla GRPO on In-Domain datasets.

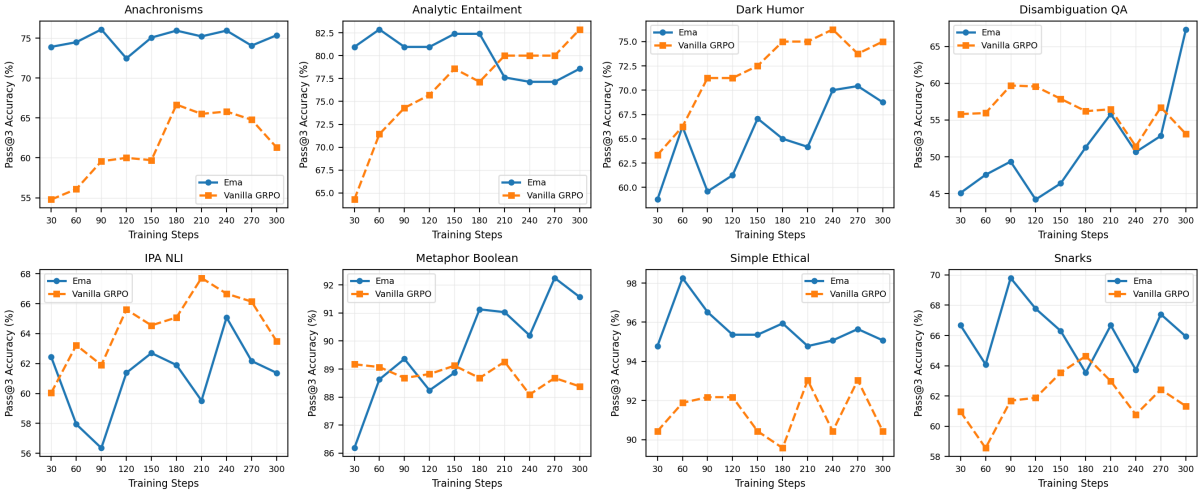

Figure 9: Comparison of Pass@3 accuracy of our proposed method **Ema** and Vanilla GRPO on OOD datasets.

Ema Hyperparameters		In-Domain				Out-of-Domain									Avg.
		SocNorm (Acc.)	CALI (Acc.)	Semeval (Acc.)	Avg. (Acc.)	Anach. (Acc.)	Entail. (Acc.)	Humor (Acc.)	Disambi. (Acc.)	IPA (Acc.)	Metaphor (Acc.)	SEQ (Acc.)	SNARKS (Acc.)	Avg. (Acc.)	
Qwen3-8B															
Base (Qwen3-8B)	Dir.	59.38	41.82	73.87	62.72	55.65	74.29	67.50	47.67	61.11	88.09	89.57	59.67	71.49	66.01
	Pers.	55.47	43.33	67.74	58.92	70.87	82.86	57.50	34.50	30.95	75.44	93.91	64.64	65.11	61.24
Ema															
$n = 30, m = 30, \epsilon = 0.5$ (Main)		67.88	51.31	86.99	73.78	75.36	78.57	68.75	**67.31**	61.38	**91.57**	95.07	**65.93**	**79.64**	**75.98**
$n = 10, m = 30, \epsilon = 0.5$		67.97	51.67	**87.31**	74.05	73.04	**85.71**	67.50	48.06	**61.90**	88.68	**95.65**	65.75	75.63	74.64
$n = 30, m = 10, \epsilon = 0.5$		**68.75**	50.40	85.86	73.23	**76.67**	81.90	**71.25**	50.52	57.67	91.08	93.04	64.27	76.80	74.57
$n = 30, m = 30, \epsilon = 0.25$		67.97	**56.82**	86.29	**74.71**	70.87	80.00	68.75	47.29	30.16	90.29	94.78	64.09	73.16	74.13

Table 8: Analysis of hyperparameters for Experience Buffer Management on Qwen3-8B. The parameters n, m, and ϵ denote the *update frequency*, *maximum buffer capacity*, and *exploration rate*, respectively. For each dataset, the best result is **bolded**, the second-best is underlined.

for improving model performance.

Prompts for LLMs to Answer Directly
Your Task: We are dealing with subjective tasks, you should address them by following the instructions. Directly output your answer in '<Answer>' '</Answer>'. Instructions: Question: {question} Options: {options} An Example: Question: What is the attitude toward "Donald Trump" in the tweet "Get the truth from Trump! @realDonaldTrump #SemST"? Options: (A) FAVOR (B) AGAINST (C) NEUTRAL Your output: <Answer> A </Answer> ATTENTION! **ONLY** provide responses strictly following the format above **WITHOUT any additional explanations**.

Table 9: Prompts for LLMs to Answer Directly.

Prompts for Synthesizing Experiences
You are an assistant for analyzing reinforcement learning rollout errors. [Reward Function Design] Full marks will be awarded if and only if the model outputs in the specified format, the answers are correct, and there is no output of abnormal length. [Question] {prompt} [Model Output] {response} [Ground Truth for the answer] {ground_truth_for_answer} Based on the above, the model made a mistake. Please generate ONE concise, actionable experience lesson that tells the model what to do next time to avoid this mistake. Requirements: 1. Start with an action verb (e.g., "Always...", "When...", "Focus on...", "Check...", "Ensure...") 2. Be specific and actionable, not just a description of the error 3. No more than {max_length} characters 4. Output ONLY the lesson, no explanations Bad example: "Misunderstood the question" (just describes error) Good example: "When answering, first identify the core question before formulating a response" (actionable guidance)

Table 10: Prompts for Synthesizing Experiences.

Prompts for Summarizing Experiences.
You are an assistant for summarizing reinforcement learning experiences. Below is the existing experiences: {existing_experiences} Please merge, deduplicate, and summarize these experiences, keeping the most important {max_count} experiences. Each experience should be on a separate line, **NO NUMBERING**. Output **ONLY** the experiences, **NO OTHER EXPLANATIONS**.

Table 11: Prompts for Summarizing Experiences when the Total Number of Accumulated Experiences Exceeds m.

Prompts for Model Training
We are dealing with subjective tasks, and different issues require selecting appropriate perspectives for resolution. Please analyze and solve the problem. 1. Third-person Perspective (simulate several agents having a discussion, then summarize and give the final answer) 2. Role Perspective (assume the role of an expert or specific character and answer the question) 3. Direct Perspective (answer the question directly) You should output your answer in the following format: 1. Firstly, use <Third-person> [Third-person Perspective simulation] </Third-person> for the Third-person Perspective. 2. Secondly, use <Role Perspective> [Role Perspective simulation] </Role Perspective> for the Role Perspective. 3. Thirdly, use <Direct Perspective> [Direct Perspective simulation] </Direct Perspective> for the Direct Perspective. 4. Finally, combine the above outputs and use <Answer> [Your answer HERE] </Answer> to present your final answer. Answer Only with the option letter. Example: Question: What is the attitude toward "Donald Trump" in the tweet "Get the truth from Trump! @realDonaldTrump #SemST"? Options: (A) FAVOR (B) AGAINST (C) NEUTRAL Your output: <Third-person> Lily: The tweet says "Get the truth from Trump!" with a mention and hashtag supporting him. It sounds positive. Tom: Yes, the phrase "Get the truth from Trump!" indicates trust or endorsement. Tim: There's no negative wording; it's clearly favorable. </Third-person> <Role Perspective> As a sentiment analysis expert, I identify that the expression "Get the truth from Trump!" conveys a positive stance toward him. The mention of his official handle and hashtag use implies endorsement and alignment rather than criticism. Therefore, the sentiment is FAVOR. </Role Perspective> <Direct Perspective> The tweet expresses confidence and support toward Donald Trump; hence, the attitude is FAVOR. </Direct Perspective> <Answer> A </Answer> Instructions for Response: Only provide responses strictly following the format above without any additional explanations. Now, let's apply this structured approach to solve specific question. Question: {question} While answering the question, you should consider the following experiences: {experiences}

Table 12: Prompts for Model Training.

Prompts for Model Inference via Perspective Transition
We are dealing with subjective tasks, and different issues require selecting appropriate perspectives for resolution. Please analyze and solve the problem. 1. Third-person Perspective (simulate several agents having a discussion, then summarize and give the final answer) 2. Role Perspective (assume the role of an expert or specific character and answer the question) 3. Direct Perspective (answer the question directly) You should output your answer in the following format: 1. Firstly, use <Third-person> [Third-person Perspective simulation] </Third-person> for the Third-person Perspective. 2. Secondly, use <Role Perspective> [Role Perspective simulation] </Role Perspective> for the Role Perspective. 3. Thirdly, use <Direct Perspective> [Direct Perspective simulation] </Direct Perspective> for the Direct Perspective. 4. Finally, combine the above outputs and use <Answer> [Your answer HERE] </Answer> to present your final answer. Answer Only with the option letter. Example: Question: What is the attitude toward "Donald Trump" in the tweet "Get the truth from Trump! @realDonaldTrump #SemST"? Options: (A) FAVOR (B) AGAINST (C) NEUTRAL Your output: <Third-person> Lily: The tweet says "Get the truth from Trump!" with a mention and hashtag supporting him. It sounds positive. Tom: Yes, the phrase "Get the truth from Trump!" indicates trust or endorsement. Tim: There's no negative wording; it's clearly favorable. </Third-person> <Role Perspective> As a sentiment analysis expert, I identify that the expression "Get the truth from Trump!" conveys a positive stance toward him. The mention of his official handle and hashtag use implies endorsement and alignment rather than criticism. Therefore, the sentiment is FAVOR. </Role Perspective> <Direct Perspective> The tweet expresses confidence and support toward Donald Trump; hence, the attitude is FAVOR. </Direct Perspective> <Answer> A </Answer> Instructions for Response: Only provide responses strictly following the format above without any additional explanations. Now, let's apply this structured approach to solve specific question. Question: {question}

Table 13: Prompts for Model Inference via Perspective Transition.

Synthesized Experiences Example
Check for direct logical dependency or conflict to determine entailment or contradiction.
Avoid inferring unrelated attitudes; focus on actual content.
Clarify sarcasm, irony, or subtle context that may invert apparent sentiment.
Verify the speaker's tone and align the answer with their expressed attitude.
Check for implicit contradictions in value judgments, norms, or perspectives.
Ensure alignment with cultural or contextual nuances in subjective questions.
Check for overlapping contexts before determining contradiction or entailment.
Distinguish between personal opinion and stated attitude in the question.
Check for specific exceptions that may contradict general statements.
Analyze the quoted statement's context and author to avoid misinterpreting sentiment.
Always verify if the scenario contradicts the stated outcome before determining attitude.
Check for implied criticism versus support in hashtags and context.
Focus on identifying entailment in factual descriptions or normative expectations.
Check for subtle differences in timing, subject, or context between sentences.
Check for direct opposition in advice and behavior claims to identify contradictions.
Check for hidden contradictions in statements about values and outcomes.
Ensure the final answer aligns with the majority perspective in third-person discussions.
Check for implicit emotional connections or assumptions in the speaker's knowledge.
Always clarify key terms like "dislike," "hate," or "not get along" before determining attitude.
Check for logical implications between statements before determining their relationship.
Avoid mislabeling stance by checking the target of sentiment.
Check for cultural metaphors and associations to ensure accurate interpretation.
Always verify the correct mapping between scenario and attitude options before finalizing.
Check for conditional contradictions in statements about actions and attitudes.
Focus on the logical relationship between statements, not just thematic similarity.
Check for explicit mentions or clear implications before inferring attitudes.
Check for subtle contradictions in overlapping roles and responsibilities.
Check for neutral context when multiple roles are mentioned without conflict.

Table 14: This table presents representative examples of synthesized experiences that serve as actionable lessons for improving model performance.