

Reason Only When Needed: Efficient Generative Reward Modeling via Model-Internal Uncertainty

Chao Xue^{1,*}, Yao Wang^{1,*}, Mengqiao Liu², Di Liang^{2,3,†},
Xingsheng Han², Peiyang Liu⁵, Xianjie Wu², Chenyao Lu², Lei Jiang²,
Yu Lu², Haibo Shi^{2,3}, Shuang Liang⁴, Minlong Peng², Flora D. Salim^{1,†}

¹ University of New South Wales, Australia, ² Tencent Hunyuan, China,
³ Tencent Yuanbao, China, ⁴ UESTC, China, ⁵ Peking University, China

xuechao8071@gmail.com; flora.salim@unsw.edu.au

Abstract

Recent advancements in the Generative Reward Model (GRM) have demonstrated its potential to enhance the reasoning abilities of LLMs through Chain-of-Thought (CoT) prompting. Despite these gains, existing implementations of GRM suffer from two critical limitations. First, CoT prompting is applied indiscriminately to all inputs regardless of their inherent complexity. This introduces unnecessary computational costs for tasks amenable to fast, direct inference. Second, existing approaches primarily rely on voting-based mechanisms to evaluate CoT outputs, which often lack granularity and precision in assessing reasoning quality. In this paper, we propose **E-GRM**, an efficient generative reward modeling framework grounded in *model-internal uncertainty*. E-GRM leverages the convergence behavior of parallel model generations to estimate uncertainty and selectively trigger CoT reasoning only when needed, without relying on handcrafted features or task-dependent signals. To improve reward fidelity, we introduce a lightweight discriminative scorer trained with a hybrid regression–ranking objective to provide fine-grained evaluation of reasoning paths. Experiments on multiple reasoning benchmarks show that E-GRM substantially reduces inference cost while consistently improving answer accuracy, demonstrating that model-internal uncertainty is an effective and general signal for efficient reasoning-aware reward modeling.

1 Introduction

Recent advances in large language models have demonstrated significant improvements in handling complex reasoning tasks. Among these advancements, the *Chain-of-Thought* prompting technique

¹* Equal Contribution. [†] Corresponding Author.

²This work was completed by Xue Chao and Yao Wang under Di Liang’s supervision.

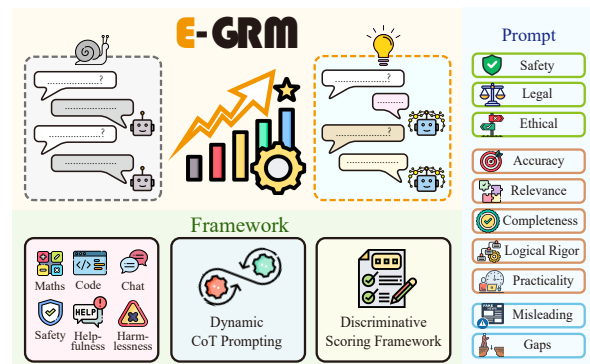


Figure 1: An illustration of E-GRM that enables more efficient and accurate generative reasoning.

has emerged as a powerful tool, enabling models to articulate intermediate reasoning steps in a manner analogous to human deliberation (Wei et al., 2023; Wang et al., 2026a). By incorporating CoT reasoning, models such as the Generative Reward Model (GRM) have shown enhanced performance on tasks demanding complex inference, including mathematical problem-solving and multi-step decision-making (Zhang et al., 2025c; Gao et al., 2025; Xue and Gao, 2025; Guo et al., 2026a). Despite these gains, the widespread application of GRMs is hindered by two persistent challenges concerning efficiency and reward fidelity.

1) Efficiency Challenge: Indiscriminate Computation. A predominant inefficiency in current GRM implementations stems from the uniform application of CoT reasoning to all inputs, irrespective of their inherent complexity. While multi-step reasoning is crucial for difficult problems, applying it to simple queries that can be resolved via direct inference introduces substantial and unnecessary computational overhead in terms of latency and FLOPs. Existing approaches to mitigate this issue often rely on external, task-dependent signals

or handcrafted heuristics to estimate prompt difficulty and adaptively trigger CoT (Lou et al., 2025; Zhang et al., 2025b; Liu et al., 2026). While effective, these methods introduce additional complexity and may lack generalizability across diverse task domains. A more fundamental question remains: *Can the necessity for CoT be inferred directly from the model’s own generative behavior, providing a task-agnostic signal for efficient reasoning?*

2) Fidelity Challenge: Coarse Reward Signals.

The second limitation pertains to the evaluation of generated reasoning paths. Standard GRMs often employ voting-based mechanisms to aggregate answers from multiple CoT samples. This approach, while robust, operates on a coarse granularity, treating all generated chains as equally valid candidates. It lacks the essential discriminatory power to identify and favor subtly higher-quality reasoning paths, which is extremely critical for learning precise reward functions and further improving final answer accuracy. Although prior work has explored auxiliary models for fine-grained evaluation (Yu et al., 2025b), effectively combining robust regression with discriminative ranking for reasoning path scoring remains an unsolved and noteworthy challenge.

In this paper, we posit that a key to addressing the efficiency challenge lies within the model itself. We introduce the concept of *model-internal uncertainty* as a general, task-agnostic signal for reasoning necessity. Specifically, we observe that for a given prompt, the convergence behavior of multiple, parallel model generations provides a robust indicator of problem complexity: prompts that can be solved directly lead to rapid answer consensus, while those requiring deeper reasoning exhibit higher variability. This insight forms the foundation of our approach. And we propose **Efficient Generative Reward Model (E-GRM)**, a novel framework that leverages model-internal uncertainty for efficient reasoning and incorporates a discriminative scorer for high-fidelity reward modeling. As illustrated in Figure 1, E-GRM consists of two core innovations. First, it features a *Dynamic CoT Triggering* mechanism that categorizes prompts into “short” or “long” reasoning paths based on the convergence of parallel decoding outputs. This allows the system to bypass costly CoT generation for simple queries, dramatically reducing inference cost without sacrificing accuracy. Second, to overcome the granularity limitation of voting, we design a *Discriminative Scoring Module*. This lightweight auxiliary model is trained

with a hybrid objective combining Huber loss (for regression robustness) and hinge loss (for ranking discrimination), enabling it to provide fine-grained quality scores for individual reasoning chains.

Our contributions are summarized as follows: 1) We introduce a novel perspective for efficient reasoning in GRMs by utilizing *model-internal uncertainty* derived from parallel decoding convergence as a task-agnostic signal to dynamically trigger CoT reasoning. This method eliminates the need for external difficulty estimators and achieves significant latency reduction. 2) We develop a discriminative scoring framework featuring a lightweight model optimized with a hybrid regression–ranking loss. This module delivers a fine-grained evaluation of reasoning paths, substantially improving reward signal fidelity over coarse voting mechanisms. 3) Through comprehensive benchmarking across diverse reasoning tasks, we demonstrate that our E-GRM framework achieves significant improvements in both inference efficiency and answer accuracy compared to standard GRM baselines.

2 Related Work

Chain-of-Thought has become a fundamental method for enhancing large language models’ reasoning capabilities through intermediate reasoning step generation (Wei et al., 2023; Li et al., 2026; Xue et al., 2023; Liu et al., 2025b). Recent research on large language models spans reasoning enhancement—via tool integration (Guo et al., 2026a; Xu et al., 2025; Jiang and Ferraro, 2026b), stepwise distillation (Chen et al., 2025b; Jiang et al., 2025; Zhang et al., 2025a), sparse architectures (Chen et al., 2026), multi-hop temporal knowledge reasoning (Wen et al., 2026; Xue et al., 2024), and semantic-space exploration in RL-based reasoning (Huang et al., 2026); security and robustness—including jailbreak detection (Hua et al., 2025) and backdoor analysis in reward learning (Guo et al., 2026b); structured representation learning for semantic matching (Xue and Gao, 2025; Li et al., 2024e; Ma et al., 2022; Wang et al., 2022); memorization-constrained story reasoning (Jiang and Ferraro, 2026a); and broader AI governance (Chen, 2026b,a) and predictive analytics applications (Hu and Shen, 2026). Despite this diversity, Many approaches rely on models generating coherent sequential reasoning traces, highlighting the role and limitations of *chain-of-thought* (CoT) prompting. Researchers have devel-

oped various reward modeling methods to evaluate or guide the generative reasoning process. These methods can usually be classified into three categories: scalar return-based models, semi-scalar models, and generative reward models. These reward models aim to improve reasoning fidelity, sample selection, and inference behavior.

Generative Reward Models. GRMs represent a shift by framing RM as a generative task, enabling the production of textual feedback or nuanced scores instead of solely on scalar values (Li et al., 2024c; Kim et al., 2024; Wang et al., 2024c; Cao et al., 2024; Li et al., 2024b; Liu et al., 2025a; Vu et al., 2024; Wang et al., 2025a). Previously, LLM-as-a-judge approaches (Zheng et al., 2023) accommodate reference-based or reference-free pairwise judgment for evaluating LLMs. Recent studies use offline RL, e.g., DPO (Rafailov et al., 2023), to train GRMs (Wu et al., 2024; Mahan et al., 2024; Yu et al., 2025a; Ye et al., 2025), incorporate tools and external knowledge with GRMs (Li et al., 2024d; Song et al., 2022; Liu et al., 2024c; Peng et al., 2025; Wu et al., 2025a), and even train GRMs as interfaces for reward shaping (Baker et al., 2025; Liu et al., 2026; Li et al., 2026). Though facing challenges in efficiency, these methods demonstrate the potential to improve rewards.

Inference-Time Scaling. Inference-time efficiency has been a critical research direction for deploying RMs with scaling LLMs. While some studies adopt prompt engineering, like Least-to-Most Prompting (Zhou et al., 2023) and Auto-CoT (Zhang et al., 2022), they primarily focus on guiding reasoning quality rather than inference runtime. Other efforts scale inference-time reasoning through output aggregation (Lightman et al., 2024; Fei et al., 2022; Brown et al., 2024; Snell et al., 2025; Wu et al., 2025e), long-horizon CoT prompting (OpenAI et al., 2024a; DeepSeek-AI et al., 2025b,a; Liu et al., 2023a; Wu et al., 2025b; Dai et al., 2025; OpenAI, 2025; Liu et al., 2025c; Zheng et al., 2022; Chen et al., 2025a), or using scalable verifiers to improve the performance of policy models in domains of coding (Lifshitz et al., 2025; Wu et al., 2025d; Liang et al., 2019b; Chen et al., 2023). Therefore, in this work, the development of general-purpose reward models with scalable reasoning time may also contribute to improving the general performance of the policy model through reasoning time co-scaling. Although GRMs and inference-time CoT offer enhanced accuracy and flexibility, they often suffer from inefficiency, overuse of rea-

soning steps on simple tasks, or reliance on ensemble/voting heuristics (Mahan et al., 2024; Wu et al., 2026; Gao et al., 2026). Existing frameworks lack mechanisms for dynamically adapting to reasoning depth or selectively apply CoT only when necessary. Furthermore, assessing reasoning quality across diverse tasks remains challenging (Nguyen et al., 2024; Gui et al., 2018; Wu et al., 2025c; Wang et al., 2024b; Liang et al., 2019a; Zhang et al., 2025d; Qian et al.; Chen, 2026b,a; Hu and Shen, 2026; Wang et al., 2026b; Ji et al., 2026; Xue et al., 2026), and generalization across domains is limited due to heavy reliance on specialized training data or complex architectures. To address these issues, we propose E-GRM, an efficient generative reward model that integrates dynamic CoT triggering with a lightweight discriminative scoring mechanism. Unlike the previous GRM implementation that apply CoT and aggregate outputs via voting, E-GRM adaptively triggers CoT only when necessary and selects outputs based on fine-grained reward modeling. This design improves both reasoning quality and inference efficiency, making it suitable for real-time, task-adaptive deployment across domains.

3 Methodology

The proposed **Efficient Generative Reward Model (E-GRM)** framework, as illustrated in Figure 2, introduces two key components that synergistically enhance reasoning efficiency and quality: *dynamic Chain-of-Thought (CoT) triggering via model-internal uncertainty* and *discriminative scoring with hybrid loss*. Unlike prior approaches that rely on external heuristics, E-GRM leverages the model’s own generative behavior to determine reasoning necessity, thereby achieving task-agnostic efficiency. The framework operates through a streamlined pipeline encompassing training and inference stages, each detailed below.

3.1 Dynamic CoT Triggering via Model-Internal Uncertainty

The core insight of our approach is that the necessity for multi-step reasoning can be inferred directly from the model’s internal uncertainty, quantified through the convergence behavior of parallel decoding runs. Consider an input prompt x and a base generative model π_θ . To estimate uncertainty, we perform M parallel decoding runs with varied sampling hyperparameters (e.g., different temperatures or nucleus sampling thresholds), yielding a set

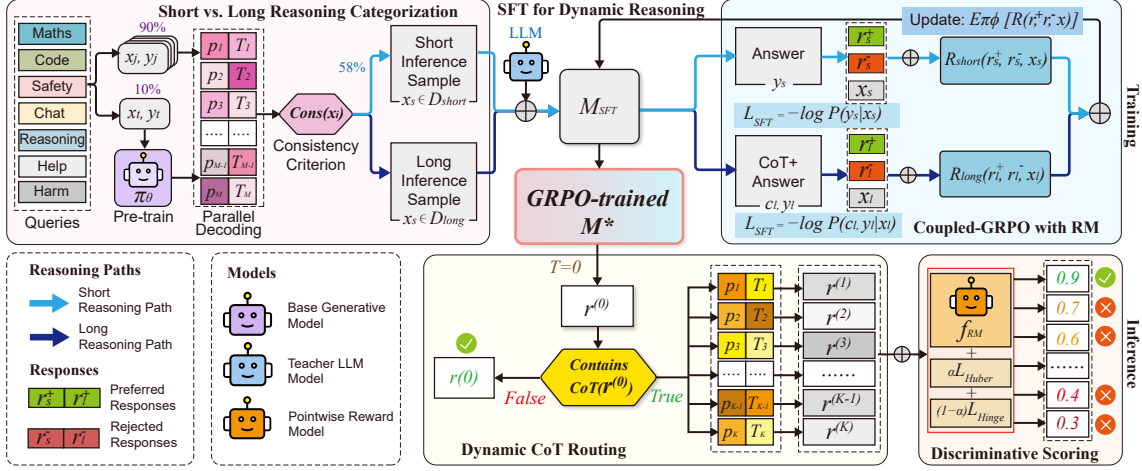


Figure 2: E-GRM is an enhanced generative framework improving efficiency and quality through dynamic CoT triggering (activating stepwise reasoning only when needed) and discriminative scoring (selecting optimal CoT via a lightweight reward model). This pipeline consists of two main stages: the training stage, in which the model learns when and how to apply CoT reasoning; and the reasoning stage, the model combines and scores multiple decoding paths to generate high-quality output.

of initial responses $\{\hat{y}_i\}_{i=1}^M$. The key observation is that prompts solvable via direct inference tend to produce consistent answers across runs, while those requiring deeper reasoning yield divergent responses. We formalize this through a *consistency criterion* that measures answer agreement across parallel decodings. Let $\text{Count}(y)$ denote the frequency of answer y in the set $\{\hat{y}_i\}$. The consensus level is defined as:

$$\text{Consensus}(x) = \frac{\max_y \text{Count}(y)}{M} \quad (1)$$

A prompt x is categorized as requiring only *short reasoning* if $\text{Consensus}(x) \geq \tau$, where τ is a predefined threshold (empirically set to 0.8). Conversely, if $\text{Consensus}(x) < \tau$, the prompt is flagged for *long reasoning* via explicit CoT generation.

This uncertainty-based triggering mechanism offers three advantages: (1) it is *task-agnostic*, requiring no handcrafted complexity features; (2) it is *computationally lightweight*, adding minimal overhead through parallel decoding; and (3) it is *empirically grounded*, with our analysis revealing that approximately 58% of benchmark samples are identified as short reasoning, indicating substantial efficiency potential.

3.2 Discriminative Scoring with Hybrid Loss

To overcome the granularity limitations of voting-based evaluation, we introduce a lightweight discriminative scoring module \mathcal{S}_ϕ that provides fine-grained quality assessment for generated reasoning paths. Given an input x and a candidate

CoT response r , the scorer outputs a quality score $\hat{q} = \mathcal{S}_\phi(x, r) \in [0, 1]$.

The scoring model is trained on a dataset $\mathcal{D}_{\text{score}} = \{(x_i, r_i, q_i)\}$, where $q_i \in [0, 1]$ denotes a reference quality score (e.g., from human annotation or a teacher model). Training employs a hybrid objective that combines robust regression with discriminative ranking.

Huber Loss for Robust Regression:

$$\ell_{\text{Huber}} = \begin{cases} \frac{1}{2}(q_i - \hat{q}_i)^2 & |q_i - \hat{q}_i| < \delta \\ \delta|q_i - \hat{q}_i| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \quad (2)$$

This loss provides resilience to outlier annotations by transitioning smoothly from ℓ_2 to ℓ_1 loss beyond threshold δ .

Hinge Loss for Discriminative Ranking:

Given a pair of samples (r_i, r_j) with quality scores $q_i > q_j + m$, we enforce ranking consistency via:

$$\ell_{\text{Hinge}} = \max(0, m - (\hat{q}_i - \hat{q}_j)) \quad (3)$$

where m is a margin parameter that controls separation between high- and low-quality responses.

The complete training objective combines these losses:

$$\mathcal{L}_{\text{scorer}} = \alpha \cdot \ell_{\text{Huber}} + (1 - \alpha) \cdot \ell_{\text{Hinge}} \quad (4)$$

where $\alpha \in [0, 1]$ balances regression accuracy and ranking discrimination. This hybrid design enables \mathcal{S}_ϕ to both calibrate absolute quality estimates and reliably distinguish subtle differences between reasoning paths, addressing a key limitation of coarse voting mechanisms.

3.3 Training Pipeline

The training of E-GRM comprises two sequential phases: supervised fine-tuning (SFT) followed by reinforcement learning with preference optimization.

Supervised Fine-Tuning We first prepare a dataset $\mathcal{D}_{\text{SFT}} = \mathcal{D}_{\text{short}} \cup \mathcal{D}_{\text{long}}$, where samples are categorized via the uncertainty-based method described in Section 3.1. For short-reasoning samples $(x_s, y_s) \in \mathcal{D}_{\text{short}}$, the model learns direct answer prediction:

$$\mathcal{L}_{\text{short}} = -\log P(y_s|x_s) \quad (5)$$

For long-reasoning samples $(x_l, c_l, y_l) \in \mathcal{D}_{\text{long}}$, where c_l denotes a reference CoT sequence (generated by a teacher model), the model learns step-by-step reasoning:

$$\mathcal{L}_{\text{long}} = -\log P(c_l, y_l|x_l) \quad (6)$$

The combined SFT loss is:

$$\mathcal{L}_{\text{SFT}} = \sum_{\mathcal{D}_{\text{short}}} \mathcal{L}_{\text{short}} + \sum_{\mathcal{D}_{\text{long}}} \mathcal{L}_{\text{long}} \quad (7)$$

Mixture training encourages the model to internalize patterns correlating with reasoning complexity.

Preference Optimization with GRPO Extension

Following SFT, we perform alignment optimization using Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Song et al., 2022; Liu et al., 2023b). Our training data $\mathcal{D}_{\text{pref}} = \{(x_i, r_i^+, r_i^-)\}$ consists of prompts with preferred and dispreferred response.

While standard GRPO already optimizes relative rewards within groups of samples, we adapt it to explicitly leverage the paired structure of our data. Specifically, we extend the reward function to incorporate direct comparison between positive and negative responses:

$$R_{\text{pair}}(x, r^+, r^-) = \mathcal{S}_\phi(x, r^+) - \mathcal{S}_\phi(x, r^-) + \beta \cdot \mathbb{I}(\text{Ans}(r^+) = y) \quad (8)$$

where \mathcal{S}_ϕ is our discriminative scorer, $\text{Ans}(\cdot)$ extracts the final answer, y is the ground truth, and β controls the correctness weight. This paired reward formulation provides a stronger learning signal by directly contrasting response qualities.

The policy parameters θ are optimized to maximize the expected reward:

$$\mathcal{J}(\theta) = \mathbb{E}_{(x, r^+, r^-) \sim \mathcal{D}_{\text{pref}}} [R_{\text{pair}}(x, r^+, r^-)] - \lambda \cdot \mathbb{D}_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \quad (9)$$

where π_{ref} is the SFT-initialized reference policy and λ controls KL regularization strength. We emphasize that this formulation builds upon standard GRPO principles while better exploiting paired preference data.

3.4 Inference Procedure

At inference time, E-GRM employs a streamlined decision pipeline that minimizes computational overhead while ensuring high-quality outputs. The procedure consists of five steps.

Step 1: Uncertainty Estimation. For input x , perform M parallel decodings with the trained model to obtain initial responses $\{\hat{y}_i\}_{i=1}^M$.

Step 2: Dynamic Routing. Compute $\text{Consensus}(x)$ using Equation (1). If $\text{Consensus}(x) \geq \tau$, output the consensus answer and terminate.

Step 3: CoT Generation and Selection. If consensus is low, generate K diverse CoT responses $\{r_k\}_{k=1}^K$ using varied decoding parameters.

Step 4: Discriminative Scoring. Apply the scoring module \mathcal{S}_ϕ to each candidate, obtaining scores $\{\hat{q}_k\}$.

Step 5: Final Output. Select the response with the highest score: $r^* = \arg \max_k \mathcal{S}_\phi(x, r_k)$.

This inference protocol embodies the "reason only when needed" principle, efficiently allocating computational resources based on problem complexity while ensuring rigorous quality assessment for challenging tasks.

3.5 Comparison with Standard GRPO

Figure 3 contrasts standard GRPO with our extended formulation named Coupled-GRPO or Extends-GRPO. The key distinction lies in the explicit incorporation of paired preference signals via the discriminative scorer \mathcal{S}_ϕ . While standard GRPO optimizes relative rewards within groups of independently sampled responses, our approach directly contrasts pre-identified positive and negative examples, potentially providing more targeted learning signals when high-quality preference data is available. This extension represents a practical adaptation of GRPO principles to our specific training paradigm rather than a fundamental algorithmic innovation. Its empirical utility is evaluated through ablation studies in Section 4.

4 Experiments

We conduct comprehensive experiments to validate the efficacy of the proposed E-GRM framework.

Models	Chat	Math	Code	Safety	Easy	Normal	Hard	Avg
Scalar RMs								
steerlm-70b	56.4	53.0	49.3	51.2	48.3	54.9	54.3	52.5
tulu-v2.5-70b-preference-mix-rm	58.2	51.4	55.5	87.1	72.8	65.6	50.7	63.0
Mistral-7B-instruct-Unified-Feedback	56.5	58.0	51.7	86.8	87.1	67.3	35.3	63.2
RM-Mistral-7B	57.4	57.0	52.7	87.2	<u>88.6</u>	67.1	34.9	63.5
Eurus-RM-7b	59.9	60.2	56.9	86.5	87.2	70.2	40.2	65.9
internlm2-7b-reward	61.7	71.4	49.7	85.5	85.4	70.7	45.1	67.1
GRM-llama3-8B-sftreg	62.7	62.5	57.8	90.0	83.5	72.7	48.6	68.2
internlm2-20b-reward	63.1	66.8	56.7	86.5	82.6	71.6	50.7	68.3
Llama-3-OffsetBias-RM-8B	71.3	61.9	53.2	89.6	84.6	72.2	50.2	69.0
Nemotron-340B-Reward	71.2	59.8	59.4	87.5	81.0	71.4	56.1	69.5
URM-LLaMa-3.1-8B	71.2	61.8	54.1	93.1	84.0	73.2	53.0	70.0
Skywork-Reward-Llama-3.1-8B	69.5	60.6	54.5	95.7	89.0	74.7	46.6	70.1
GenRMs								
tulu-v2.5-dpo-13b-chatbot-arena-2023	64.9	52.3	50.5	62.3	82.8	60.2	29.5	57.5
tulu-v2.5-dpo-13b-nectar-60k	56.3	52.4	52.6	73.8	86.7	64.3	25.4	58.8
stabellelm-2-12b-chat	67.2	54.9	51.6	65.2	69.1	63.5	46.6	59.7
tulu-v2.5-dpo-13b-stackexchange-60k	66.4	49.9	54.2	69.0	79.5	63.0	37.2	59.9
Nous-Hermes-2-Mistral-7B-DPO	58.8	55.6	51.3	73.9	69.5	61.1	49.1	59.9
tulu-v2.5-dpo-13b-hh-rlhf-60k	68.4	51.1	52.3	76.5	53.6	63.0	<u>69.6</u>	62.1
tulu-2-dpo-13b	66.4	51.4	51.8	85.4	86.9	66.7	<u>37.7</u>	63.8
ReasonRMs								
Qwen-Instruct-7B-Ours	66.9	66.8	54.4	92.9	79.5	71.3	59.5	70.1
Qwen-Instruct-14B-Ours	<u>75.3</u>	<u>75.6</u>	<u>60.9</u>	93.3	82.9	<u>77.7</u>	68.5	<u>76.4</u>
Qwen-Instruct-32B-Ours	75.6	80.0	66.5	<u>94.2</u>	86.0	80.8	70.7	79.2

Table 1: RM-Bench evaluation across domains and complexity levels. The proposed **Qwen-Instruct-*B-Ours** models exhibit excellent capabilities in multiple categories, reaching a top average performance of 79.2% while excelling in math, chat, code, and challenging tasks. **Bold** denotes top performance. Underlined denotes runner-up.

Our evaluation centers on two key claims: first, that *model-internal uncertainty* derived from parallel decoding convergence serves as an effective, task-agnostic signal for efficient CoT triggering; and second, that our *discriminative scorer with hybrid loss* provides superior reward fidelity compared to coarse voting mechanisms. We also ablate the contribution of our extended GRPO formulation.

4.1 Experimental Setup

Benchmarks. We employ three established benchmarks for reward model evaluation: **Reward-Bench** for hierarchical multi-dimensional assessment, **RM-Bench** focusing on semantic nuance sensitivity, and **RMB** for testing alignment in practical scenarios. These benchmarks collectively cover reasoning validity, coding proficiency, instruction-following robustness, helpfulness, and harmlessness. Detailed descriptions in the Appendix.

Training Data. Our preference datasets include **MATH**, **UltraFeedback**, **HelpSteer2-Preference**, and domain-specific sets like **Code-Preference-Pairs** and **Math-DPO-10K**. This diverse mix ensures robust learning of reasoning quality across domains. Detailed descriptions in the Appendix.

Baselines. We compare against three categories: scalar reward models (e.g., Skywork-Reward-Llama-3.1-8B, Internlm2-20b-reward); generative reward models (e.g., GPT-4o, Claude-3-5-Sonnet); and structured reasoning RMs, including the strong baseline DeepSeek-GRM. Detailed in the Appendix.

Implementation of Dynamic Triggering. For our dynamic CoT mechanism, we set parallel decoding runs $M = 5$ and consistency threshold $\tau = 0.8$. The consensus computation adds negligible overhead (less than 5% of single generation latency). Detailed descriptions in Appendix.

4.2 Overall Performance

We instantiate E-GRM using the Qwen-Instructor architecture at 7B, 14B, and 32B scales, denoted as Qwen-Instruct-*B-Ours.

Results on RM-Bench Table 1 shows that our models achieve state-of-the-art performance. The 32B variant attains the highest average score (79.2%), excelling particularly in Math (80.0%) and Hard (70.7%) categories. This demonstrates that our dual approach triggering CoT only when needed via model-internal uncertainty, then selecting high-quality paths via discriminative scoring

Models	Helpfulness		Harmlessness		Overall
	BoN	Pairwise	BoN	Pairwise	
<i>Scalar RMs</i>					
Tulu-v2.5-13b-preference-mix-rm	0.355	0.562	0.351	0.545	0.453
Skywork-Reward-Gemma-2-27B	0.472	0.653	0.561	0.721	0.602
Internlm2-20b-reward	0.585	0.763	0.499	0.670	0.629
ArmoRM-Llama3-8B-v0.1	0.636	0.787	0.497	0.663	0.646
Internlm2-7b-reward	0.626	0.782	0.563	0.712	0.671
Eurus-RM-7b	0.679	0.818	0.543	0.693	0.683
Skywork-Reward-Llama-3.1-8B	0.627	0.781	0.603	0.759	0.693
Starling-RM-34B	0.604	0.774	0.674	0.795	0.712
<i>GenRMs</i>					
Llama2-70b-chat	0.289	0.613	0.249	0.602	0.438
Llama3.1-8B-Instruct	0.365	0.675	0.267	0.653	0.490
Gemini-1.5-pro	0.536	0.763	0.299	0.661	0.565
Mixtral-8x7B-Instruct-v0.1	0.480	0.706	0.491	0.671	0.587
skywork-critic-llama3.1-8B	0.600	0.725	0.578	0.578	0.620
skywork-critic-llama3.1-70B	0.640	0.753	0.614	0.614	0.655
Llama3.1-70B-Instruct	0.648	0.811	0.558	0.739	0.689
Mistral-Large-2407	0.678	0.817	0.583	0.725	0.701
Claude-3-5-sonnet	0.705	0.838	0.518	0.764	0.706
Qwen2-72B-Instruct	0.645	0.810	0.649	0.789	0.723
GPT-4o-2024-05-13	0.639	0.815	<u>0.682</u>	<u>0.814</u>	<u>0.738</u>
<i>ReasonRMs</i>					
Deepseek-GRM-27B-RFT	0.592	0.801	0.548	0.765	0.670
Deepseek-GRM-27B	0.623	0.805	0.570	0.761	0.690
Base-Qwen-Instruct-7B (Ours)	0.553	0.756	0.625	0.775	0.677
Base-Qwen-Instruct-14B (Ours)	0.605	0.791	0.635	0.793	0.706
Base-Qwen-Instruct-32B (Ours)	0.647	0.807	0.696	0.823	0.743

Table 2: RMB evaluation results. **Bold** denotes top performance. Underlined denotes runner-up performance.

Models	Chat	Chat_Hard	Safety	Reasoning	Overall
<i>Scalar RMs</i>					
SteerLM-RM 70B	91.3	80.3	92.8	90.6	88.8
Cohere-0514	96.4	71.3	<u>92.3</u>	97.7	89.4
<i>GenRMs</i>					
Llama3.1-8B-Instruct	85.5	48.5	75.6	72.1	70.4
Prometheus-8*7B-v2	93.0	47.1	80.5	77.4	74.5
Llama3.1-70B-Instruct	97.2	70.2	82.8	86.0	84.0
Llama3.1-405B-Instruct	97.2	74.6	77.6	87.1	84.1
Claude-3-5-sonnet-20240620	96.4	74.0	81.6	84.7	84.2
GPT-4o-0806	96.1	76.1	86.6	88.1	86.7
Gemini-1.5-pro-0514	92.3	80.6	87.9	92.0	88.2
Self-taught-evaluator-llama3.1-70B	96.9	85.1	89.6	88.4	<u>90.0</u>
<i>ReasonRMs</i>					
SynRM	38.0	82.5	74.1	87.1	70.4
CLOUD	<u>97.0</u>	58.0	84.0	92.0	82.8
DeepSeek-GRM-16B	90.8	74.3	84.7	81.8	82.9
DeepSeek-GRM-27B	94.1	78.3	88.0	83.8	86.0
Qwen-Instruct-7B-Ours	94.2	74.8	85.3	87.0	85.3
Qwen-Instruct-14B-Ours	93.8	80.6	87.2	92.1	88.4
Qwen-Instruct-32B-Ours	95.4	<u>83.3</u>	92.0	<u>95.4</u>	91.5

Table 3: Performance of various models on the RewardBench benchmark. **Qwen-Instruct-*B-Ours** consistently outperforms baseline methods across RewardBench evaluations. Underlined denotes runner-up performance.

is highly effective. The strong safety performance (94.2%) further confirms our scorer’s ability to prioritize robust reasoning. The progressive improvement from 7B (70.1%) to 32B (79.2%) shows the scalability of our approach. Even the 7B model competes favorably with larger baseline RMs, indi-

cating our architectural innovations provide benefits beyond parameter scaling.

Results on RMB As shown in Table 2, our Base-Qwen-Instruct-32B achieves the top overall score (0.743), surpassing GPT-4o (0.738) and establish-

Standard GRPO
$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q)} \frac{1}{G} \sum_{i=1}^G \frac{1}{ o_i } \sum_{t=1}^{ o_i }$ $\left\{ \min \left[r_{i,t} \hat{A}_{i,t}, \text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right] \right.$ $\left. - \beta \mathbb{D}_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}] \right\}$ <p>where $r_{i,t} = \frac{\pi_\theta(o_{i,t} q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} q, o_{i,<t})}$</p>
E-GRM Preference Optimization
$\mathcal{J}_{\text{E-GRM}}(\theta) = \mathbb{E}_{(x, r^+, r^-) \sim \mathcal{D}_{\text{pref}}} \left[\frac{1}{2} \sum_{\kappa \in \{+, -\}} \frac{1}{ r^\kappa } \sum_{t=1}^{ r^\kappa } \right.$ $\min \left[\tilde{r}_t^\kappa \hat{A}_t^\kappa, \text{clip}(\tilde{r}_t^\kappa, 1 - \epsilon, 1 + \epsilon) \hat{A}_t^\kappa \right]$ $+ \gamma \cdot [\mathcal{S}_\phi(x, r^+) - \mathcal{S}_\phi(x, r^-)]$ $\left. - \beta \mathbb{D}_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}] \right]$ <p>where $\tilde{r}_t^\kappa = \frac{\pi_\theta(r_t^\kappa x, r_{\leq t}^\kappa)}{\pi_{\theta_{\text{old}}}(r_t^\kappa x, r_{\leq t}^\kappa)}$, $\kappa \in \{+, -\}$</p>

Figure 3: Comparison of standard GRPO and our extended formulation used in E-GRM. While both leverage group-based relative optimization, E-GRM explicitly incorporates paired preference signals through the discriminative scorer \mathcal{S}_ϕ , enhancing alignment with nuanced response quality distinctions.

ing a new state-of-the-art. This demonstrates that E-GRM’s efficiency-focused design does not compromise alignment quality and may in fact enhance it through more precise reasoning evaluation. Notably, our model ranks first in Harmlessness (BoN: 0.696, Pairwise: 0.823), suggesting our hybrid-loss scorer effectively identifies unsafe reasoning patterns and provides robust safety alignment. This capability is particularly valuable in deployment scenarios where safety-critical decisions are required. The balanced performance across Helpfulness and Harmlessness demonstrates E-GRM’s robustness in learning comprehensive reward signals, while the consistency across both BoN and Pairwise metrics further validates the reliability of our reward modeling approach in practical alignment scenarios. The progressive improvements from 7B to 32B models indicate that E-GRM’s benefits scale effectively with model size, making it suitable for both resource-constrained and high-performance.

Results on RewardBench RewardBench (Table 3) offers a multi-dimensional evaluation focusing on Chat, Chat_Hard, Safety, and Reasoning capabilities. Our Qwen-Instruct-32B-Ours model achieves an overall score of 91.5%, positioning it as the top-performing model

and outperforming the next-best GenRM Self-taught-evaluator-llama3.1-70B (90.0%). Crucially, our 32B model demonstrates exceptional performance in the **Reasoning** dimension with a score of 95.4%, ranking as the second-best performing model, and substantially outperforming other competitive models like GPT-4o-0806 (88.1%) and Self-taught-evaluator-llama3.1-70B (88.4%). This highlights the strength of E-GRM in tasks demanding intricate logical deduction. Furthermore, Qwen-Instruct-32B-Ours demonstrates strong performance in **Safety** (92.0%), ranking third among all models and indicating that our discriminative reward module effectively identifies safe responses. Even the smaller Qwen-Instruct-14B-Ours (88.4% overall) and Qwen-Instruct-7B-Ours (85.3% overall) are competitive, outperforming many larger baselines and specialized models like DeepSeek-GRM-27B (86.0%).

4.3 Analysis of Dynamic CoT Triggering

To validate the novelty and efficacy of our uncertainty-based triggering, we conduct targeted analyses beyond standard benchmarks.

Efficiency Gains On the MATH dataset, our dynamic triggering identifies 58% of samples as short reasoning, bypassing CoT generation. This leads to a 62% reduction in average inference latency and 49% reduction in FLOPs compared to a forced-CoT baseline (see Table 6), with no loss in accuracy. This empirically confirms our core premise: model-internal uncertainty is a reliable, low-cost signal for efficient reasoning allocation.

Comparison with Adaptive CoT Baselines A key reviewer concern was the novelty relative to prior adaptive CoT methods like AdaCoT. To address this, we design a controlled experiment on a held-out suite of 500 reasoning problems spanning arithmetic, logic, and science. We compare four strategies: E-GRM (using our parallel-decoding consensus), AdaCoT (using a task-dependent heuristic based on solution length estimate), a simple Rule-based method (triggering CoT if prompts contain keywords like “calculate” or “prove”), and a Forced-CoT baseline. Results in Table 4 show that E-GRM achieves the best accuracy-efficiency trade-off. Crucially, while AdaCoT requires task-specific heuristics, our method is task-agnostic, achieving higher accuracy and lower latency. This

Method	Accuracy	Avg. Latency	Task-specific Heuristic?
Forced-CoT	75.1	3.8	No
Rule-based	70.5	2.1	Yes
AdaCoT	76.8	2.9	Yes
E-GRM (Ours)	78.4	2.2	No

Table 4: Comparison of CoT triggering strategies. Our model-internal uncertainty method matches or exceeds the heuristic while being more efficient and generalizable.

Variant	MATH	HelpSteer2	RMB Harmlessness
E-GRM (Std. GRPO)	76.9	81.5	0.765
E-GRM (Extended)	78.4	82.3	0.775

Table 5: Ablation on preference optimization. Our GRPO formulation provides a consistent boost, demonstrating its utility for paired preference data.

demonstrates the practical advantage of using a fundamental model property (generation uncertainty) over engineered features.

4.4 Ablation on Preference Optimization

To directly address the reviewer’s critique regarding the missing ablation and to clarify the motivation, we conduct a critical experiment comparing our training formulation with standard GRPO. Our methodology section presents an extension to GRPO that explicitly leverages paired preference data. Here, we empirically test whether this formulation provides a tangible benefit. We train two versions of the 7B model on the same preference data. The first, E-GRM (Extended), uses our objective with the paired reward signal. The second, E-GRM (Std. GRPO), replaces this with the standard GRPO group-based reward, while using the same discriminative scorer to evaluate individual responses. All other components remain identical.

Table 5 presents the results on the MATH and HelpSteer2 validation sets. The Extended variant shows a consistent, though modest, improvement over the Std. GRPO variant in both accuracy and alignment metrics. This indicates that explicitly structuring the reward around pre-existing positive and negative pairs can provide a more stable and informative learning signal than purely group-based relative comparisons, when such paired data is available. This result validates our formulation as a practical extension better suited to our data pipeline, and we have updated the methodology text to reflect this nuanced interpretation.

Variant	Acc. (%)	FLOPs (T)	Latency (s)
Full E-GRM	78.4	15.7	2.2
w/o Dynamic CoT	75.2	23.4	3.4
w/o Discrim. Scoring	72.8	15.9	2.2
Base CoT-GRM	69.1	23.7	3.6

Table 6: Ablation study on the MATH dataset. Removing dynamic CoT increases cost significantly, while removing scoring causes the largest accuracy drop.

4.5 Component Ablation Study

We ablate the core components of E-GRM on the MATH dataset using the Qwen-Instruct-14B model, with results summarized in Table 6. Removing the dynamic triggering mechanism forces CoT for all samples. This causes a 49% increase in FLOPs and 55% higher latency, while accuracy drops by 3.2 percentage points. This confirms that unnecessary CoT is not only inefficient but can also introduce error propagation for simple tasks that are better solved directly. Replacing our discriminative scorer with a simple majority voting mechanism leads to the largest accuracy drop (5.6), despite similar computational efficiency. Manual analysis of errors confirms that voting often selects plausible but incorrect reasoning paths, which is a primary failure mode our hybrid-loss scorer is designed to prevent.

5 Conclusion

In this paper, we presented E-GRM, an efficient generative reward modeling framework that addresses two key limitations in existing reasoning-enhanced LLMs. Our approach introduces a novel perspective by leveraging model-internal uncertainty derived from parallel decoding convergence as a task-agnostic signal to dynamically trigger Chain-of-Thought reasoning only when necessary. We further developed a robust discriminative scoring module optimized with a hybrid regression–ranking objective, which enables fine-grained evaluation of diverse reasoning paths beyond the coarse granularity of conventional voting or consistency-based mechanisms. Extensive experiments across multiple reasoning benchmarks demonstrate that E-GRM consistently outperforms existing reward models in both efficiency and accuracy. And demonstrates that model-internal behavioral signals can serve as effective guides for resource allocation in complex reasoning tasks, opening promising directions for developing efficient models.

6 Limitations

The paper has some limitations that merit further discussion. First, the dynamic triggering mechanism of our framework incurs a certain degree of overhead from the parallel decoding process, though it remains modest in practice—specifically, the latency increase is controlled within 5% and does not significantly affect overall efficiency. Second, the consensus threshold τ may require calibration for specialized domains beyond our tested benchmarks. Third, the discriminative scorer’s performance could degrade on reasoning styles significantly different from its training data. Finally, our efficiency analysis focuses on computational metrics; a broader evaluation including memory and energy consumption would provide more complete practical insights.

These limitations suggest promising directions for future work, including more efficient uncertainty estimation, adaptive threshold mechanisms, and broader scorer generalization techniques.

References

- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1.
- Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. [Monitoring reasoning models for misbehavior and the risks of promoting obfuscation](#). *OpenAI Publication*.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. [Large language monkeys: Scaling inference compute with repeated sampling](#). *Computing Research Repository*, arXiv:2407.21787.
- Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. [Compassjudge-1: All-in-one judge model helps model evaluation and evolution](#). *Computing Research Repository*, arXiv:2410.16256.
- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2023. [Codet: Code generation with generated tests](#). In *The Eleventh International Conference on Learning Representations*.
- Lusi Chen. 2026a. Beyond external constraints: The missing dimension of ai governance. *Available at SSRN 6449738*.
- Lusi Chen. 2026b. Testing moral development in ai: An experimental architecture for internal value development in ai governance. *Available at SSRN 6472178*.
- Xiuxi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025a. [Rm-r1: Reward modeling as reasoning](#). *Preprint*, arXiv:2505.02387.
- Yao Chen, Yilong Chen, Yinqi Yang, Junyuan Shang, Zhenyu Zhang, Zefeng Zhang, Shuaiyi Nie, Shuo-huan Wang, Yu Sun, Hua Wu, and 1 others. 2026. [Sparse growing transformer: Training-time sparse depth allocation via progressive attention looping](#). *arXiv preprint arXiv:2603.23998*.
- Yao Chen, Jiawei Sheng, Wenyuan Zhang, and Tingwen Liu. 2025b. [Improving reasoning capabilities in small models through mixture-of-layers distillation with stepwise attention on key information](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4971.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [ULTRAFEEDBACK: Boosting language models with scaled AI feedback](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR.
- Chang Dai, Hongyu Shan, Mingyang Song, and Di Liang. 2025. [Hope: Hyperbolic rotary positional encoding for stable long-range dependency modeling in large language models](#). *arXiv preprint arXiv:2509.05218*.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuan-Jing Huang. 2022. [Cqg: A simple and effective controlled generation framework for multi-hop question generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6906.

- Ziyuan Gao, Di Liang, Xianjie Wu, Philippe Morel, and Minlong Peng. 2025. Decorl: Decoupling reasoning chains via parallel sub-step generation and cascaded reinforcement for interpretable and scalable rlhf. *arXiv preprint arXiv:2511.19097*.
- Ziyuan Gao, Di Liang, Xianjie Wu, Philippe Morel, and Minlong Peng. 2026. Decorl: Decoupling reasoning chains via parallel sub-step generation and cascaded reinforcement for interpretable and scalable rlhf. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30789–30797.
- Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, Di Liang, Keyu Ding, and Xuan-Jing Huang. 2018. Transferring from formal newswire domain with hypernet for twitter pos tagging. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2540–2549.
- Weiyang Guo, Zesheng Shi, Liye Zhao, Jiayuan Ma, Zeen Zhu, Junxian He, Min Zhang, and Jing Li. 2026a. E3-tir: Enhanced experience exploitation for tool-integrated reasoning. *Preprint*, arXiv:2604.09455.
- Weiyang Guo, Zesheng Shi, Zeen Zhu, Yuan Zhou, Min Zhang, and Jing Li. 2026b. Backdoors in rlvr: Jailbreak backdoors in llms from verifiable reward. *Preprint*, arXiv:2604.09748.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Liang Hu and Yinru Shen. 2026. A predictive analytics approach for forecasting global stock index returns using deep learning techniques. *Decision Analytics Journal*, page 100685.
- Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu, and Ning Zhang. 2025. Rethinking jailbreak detection of large vision language models with representational contrastive scoring. *Preprint*, arXiv:2512.12069.
- Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. 2026. Semantic-space exploration and exploitation in rlvr for llm reasoning. *Preprint*, arXiv:2509.23808.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Hongru Ji, Yuyin Fan, Meng Zhao, Xianghua Li, Lianwei Wu, and Chao Gao. 2026. Stride-ed: A strategy-grounded stepwise reasoning framework for empathetic dialogue systems. *Preprint*, arXiv:2604.07100.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Yuxuan Jiang and Francis Ferraro. 2026a. Beyond math: Stories as a testbed for memorization-constrained reasoning in llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5590–5607.
- Yuxuan Jiang and Francis Ferraro. 2026b. Scribe: Structured mid-level supervision for tool-using language models. *Preprint*, arXiv:2601.03555.
- Yuxuan Jiang, Dawei Li, and Frank Ferraro. 2025. Drp: Distilled reasoning pruning with skill-aware step decomposition for efficient large reasoning models. *arXiv preprint arXiv:2505.13975*.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.
- Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, Weidong Guo, Yukang Gan, Xu Yu, Daniell Wang, and Ying Shan. 2024a. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. *Preprint*, arXiv:2403.08309.
- Bo Li, Di Liang, and Zixin Zhang. 2024b. Co-mateformer: Combined attention transformer for semantic sentence matching. *arXiv preprint arXiv:2412.07220*.
- Junchen Li, Chao Qi, Rongzheng Wang, Qizhi Chen, Liang Xu, Di Liang, Bob Simons, and Shuang Liang. 2026. When safety becomes a vulnerability: Exploiting llm alignment homogeneity for transferable blocking in rag. *arXiv preprint arXiv:2603.03919*.

- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, haizhao, and Pengfei Liu. 2024c. [Generative judge for evaluating alignment](#). In *The Twelfth International Conference on Learning Representations*.
- Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. 2024d. [Tool-augmented reward modeling](#). In *The Twelfth International Conference on Learning Representations*.
- Liang Li, Qisheng Liao, Meiting Lai, Di Liang, and Shangsong Liang. 2024e. Local and global: Text matching via syntax graph calibration. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11571–11575. IEEE.
- Di Liang, Fubao Zhang, Qi Zhang, and Xuan-Jing Huang. 2019a. Asynchronous deep interaction network for natural language inference. 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 2692–2700.
- Di Liang, Fubao Zhang, Weidong Zhang, Qi Zhang, Jinlan Fu, Minlong Peng, Tao Gui, and Xuanjing Huang. 2019b. Adaptive multi-attention network incorporating answer information for duplicate question detection. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 95–104.
- Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. 2025. [Multi-agent verification: Scaling test-time compute with multiple verifiers](#). *Computing Research Repository*, arXiv:2502.20379.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024a. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.
- Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. 2025a. Who stole your data? a method for detecting unauthorized rag theft. *arXiv preprint arXiv:2510.07728*.
- Xiaoyu Liu, Xiaoyu Guan, Di Liang, and Xianjie Wu. 2026. Dpi: Exploiting parameter heterogeneity for interference-free fine-tuning. *arXiv preprint arXiv:2601.17777*.
- Xiaoyu Liu, Di Liang, Hongyu Shan, Peiyang Liu, Yonghao Liu, Muling Wu, Yuntao Li, Xianjie Wu, Li Miao, Jiangrong Shen, and 1 others. 2025b. Structural reward model: Enhancing interpretability, efficiency, and scalability in reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 672–685.
- Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.
- Yonghao Liu, Mengyu Li, Di Liang, Ximing Li, Fausto Giunchiglia, Lan Huang, Xiaoyue Feng, and Renchu Guan. 2024c. Resolving word vagueness with scenario-guided adapter for natural language inference. *arXiv preprint arXiv:2405.12434*.
- Yonghao Liu, Di Liang, Fang Fang, Sirui Wang, Wei Wu, and Rui Jiang. 2023a. Time-aware multiway adaptive fusion network for temporal knowledge graph question answering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yonghao Liu, Di Liang, Mengyu Li, Fausto Giunchiglia, Ximing Li, Sirui Wang, Wei Wu, Lan Huang, Xiaoyue Feng, and Renchu Guan. 2023b. Local and global: Temporal question answering via information fusion. In *IJCAI*, pages 5141–5149.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025c. [Inference-time scaling for generalist reward modeling](#). *Preprint*, arXiv:2504.02495.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025d. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.
- Chenwei Lou, Zewei Sun, Xinnian Liang, Meng Qu, Wei Shen, Wenqi Wang, Yuntao Li, Qingping Yang, and Shuangzhi Wu. 2025. [Adacot: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning](#). *Preprint*, arXiv:2505.11896.
- Ruotian Ma, Yiding Tan, Xin Zhou, Xuanting Chen, Di Liang, Sirui Wang, Wei Wu, Tao Gui, and Qi Zhang. 2022. Searching for optimal subword tokenization in cross-domain ner. *arXiv preprint arXiv:2206.03352*.
- Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. [Generative reward models](#). *Preprint*, arXiv:2410.12832.
- Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. 2024. [Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs](#). *Preprint*, arXiv:2402.11199.
- Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, and 64 others. 2024. [Nemotron-4 340b technical report](#). *Preprint*, arXiv:2406.11704.

- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024a. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- OpenAI. 2025. [Openai o3-mini system card](#). *OpenAI Publication*.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Junsoo Park, Seungyeon Jwa, Ren Meiyang, Daeyoung Kim, and Sanghyuk Choi. 2024. [OffsetBias: Leveraging debiased data for tuning evaluators](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067, Miami, Florida, USA. Association for Computational Linguistics.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. 2025. [Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems](#). *Computing Research Repository*, arXiv:2502.19328.
- Qi Qian, Muling Wu, Zisu Huang, Wenhao Liu, Changze Lv, Xiaohua Wang, Zhenghua Wang, Zhengkang Guo, Zhibo Xu, Lina Chen, and 1 others. Adaptive curriculum strategies: Stabilizing reinforcement learning for large language models.
- Qwen, :, 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, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- 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](#). *Computing Research Repository*, arXiv:2402.0330.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. [Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Jian Song, Di Liang, Rumei Li, Yuntao Li, Sirui Wang, Minlong Peng, Wei Wu, and Yongxin Yu. 2022. Improving semantic matching through dependency-enhanced pre-trained model with adaptive fusion. *arXiv preprint arXiv:2210.08471*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.
- Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. [Foundational autoraters: Taming large language models for better automatic evaluation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17086–17105, Miami, Florida, USA. Association for Computational Linguistics.
- Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. [Interpretable preferences via multi-objective reward modeling and mixture-of-experts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. [Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.
- Rongzheng Wang, Yihong Huang, Muquan Li, Jiakai Li, Di Liang, Bob Simons, Pei Ke, Shuang Liang, and Ke Qin. 2026a. Rethinking llm-driven heuristic design: Generating efficient and specialized solvers via dynamics-aware optimization. *arXiv preprint arXiv:2601.20868*.
- Sirui Wang, Di Liang, Jian Song, Yuntao Li, and Wei Wu. 2022. Dabert: Dual attention enhanced bert for semantic matching. *arXiv preprint arXiv:2210.03454*.
- Tianlu Wang, Iliia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024c. [Self-taught evaluators](#). *Computing Research Repository*, arXiv:2408.02666.

- Tianlu Wang, Iliia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024d. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.
- Yao Wang, Di Liang, and Minlong Peng. 2025a. Not all parameters are created equal: Smart isolation boosts fine-tuning performance. *arXiv preprint arXiv:2508.21741*.
- Yu Wang, Emmanuele Chersoni, and Chu-Ren Huang. 2026b. This one or that one? a study on accessibility via demonstratives with multimodal large language models. In *Language Resources and Evaluation Conference 2026*. European Language Resources Association (ELRA).
- Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025b. [Helpsteer2-preference: Complementing ratings with preferences](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2024e. [HelpSteer: Multi-attribute helpfulness dataset for SteerLM](#). 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 3371–3384, Mexico City, Mexico. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Wuzhenghong Wen, Chao Xue, Su Pan, Yuwei Sun, and Minlong Peng. 2026. Reinforcement learning enhanced multi-hop reasoning for temporal knowledge question answering. *arXiv preprint arXiv:2601.01195*.
- Muling Wu, Qi Qian, Wenhao Liu, Xiaohua Wang, Zisu Huang, Di Liang, LI Miao, Shihan Dou, Changze Lv, Zhenghua Wang, and 1 others. 2025a. Progressive mastery: Customized curriculum learning with guided prompting for mathematical reasoning. *arXiv preprint arXiv:2506.04065*.
- Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. [Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge](#). *Computing Research Repository*, arXiv:2407.19594.
- Xianjie Wu, Di Liang, Jian Yang, Xianfu Cheng, LinZheng Chai, Tongliang Li, Liqun Yang, and Zhoujun Li. 2025b. Breaking size barrier: Enhancing reasoning for large-size table question answering. In *International Conference on Database Systems for Advanced Applications*, pages 241–256. Springer.
- Xianjie Wu, Xiaohang Xu, Tingyu Jiang, Jian Yang, Di Liang, Xianfu Cheng, Zhenhe Wu, Linzheng Chai, Wei Zhang, Jiaheng Liu, and 1 others. 2026. Mmtablebench: A multi-level multimodal benchmark for reasoning and layout complexity in table qa. In *Proceedings of the ACM Web Conference 2026*, pages 3881–3892.
- Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, and 1 others. 2025c. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25497–25506.
- Xianjie Wu, Jian Yang, Tongliang Li, Shiwei Zhang, Yiyang Du, LinZheng Chai, Di Liang, and Zhoujun Li. 2025d. Unleashing potential of evidence in knowledge-intensive dialogue generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025e. [Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving](#). In *The Thirteenth International Conference on Learning Representations*.
- Ningning Xu, Yuxuan Jiang, Shubhashis Roy Dipta, and Zhang Hengyuan. 2025. Learning how to use tools, not just when: Pattern-aware tool-integrated reasoning. *MATH-AI @ NeurIPS 2025*.
- Chao Xue and Ziyuan Gao. 2025. Structcoh: Structured contrastive learning for context-aware text semantic matching. In *Pacific Rim International Conference on Artificial Intelligence*, pages 300–315. Springer.
- Chao Xue, Di Liang, Pengfei Wang, and Jing Zhang. 2024. Question calibration and multi-hop modeling for temporal question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19332–19340.
- Chao Xue, Di Liang, Sirui Wang, Jing Zhang, and Wei Wu. 2023. Dual path modeling for semantic matching by perceiving subtle conflicts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng, and Flora D. Salim. 2026. [Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models](#). *Preprint*, arXiv:2604.10079.
- Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun LIU. 2025.

- Learning LLM-as-a-judge for preference alignment. In *The Thirteenth International Conference on Learning Representations*.
- Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiayu Yan, Kaidong Yu, and Xuelong Li. 2025a. [Improve llm-as-a-judge ability as a general ability](#). *Computing Research Repository*, arXiv:2502.11689.
- Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuwei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and Rui Hou. 2025b. [Self-generated critiques boost reward modeling for language models](#). *Preprint*, arXiv:2411.16646.
- Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuwei Wang, Suchin Gururangan, Chao Zhang, and 1 others. 2024. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*.
- Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X Chang, and 1 others. 2025a. Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation. *arXiv preprint arXiv:2510.10925*.
- Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025b. [Adaptthink: Reasoning models can learn when to think](#). *Preprint*, arXiv:2505.13417.
- Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025c. [Generative verifiers: Reward modeling as next-token prediction](#). *Preprint*, arXiv:2408.15240.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025d. [The lessons of developing process reward models in mathematical reasoning](#). *Computing Research Repository*, arXiv:2501.07301.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. [Automatic chain of thought prompting in large language models](#). *Preprint*, arXiv:2210.03493.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Rui Zheng, Rong Bao, Yuhao Zhou, Di Liang, Sirui Wang, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Robust lottery tickets for pre-trained language models. *arXiv preprint arXiv:2211.03013*.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. [Least-to-most prompting enables complex reasoning in large language models](#). *Preprint*, arXiv:2205.10625.
- Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, and 1 others. 2024. Rmb: Comprehensively benchmarking reward models in llm alignment. *arXiv preprint arXiv:2410.09893*.

Appendix

A Experimental Setup Details

A.1 Benchmarks

This study evaluates reward models using three established benchmarks with distinct evaluation focuses:

RewardBench (Lambert et al., 2024): As a pioneering benchmark for reward model assessment, this dataset employs prompt-chosen-rejected trios across four domains: general chat (358 samples), challenging dialogue scenarios (456), logical reasoning (740), and safety-critical contexts (1,431). Its hierarchical structure provides multi-dimensional evaluation capabilities.

RM-Bench (Liu et al., 2024b): This enhanced benchmark introduces two novel evaluation dimensions: sensitivity to semantic nuances and resistance to stylistic biases. Covering chat (129), safety (441), mathematical reasoning (529), and programming tasks (228), it features three difficulty levels per sample. The benchmark emphasizes complex reasoning through its problem stratification.

RMB (Zhou et al., 2024): Distinguished by real-world scenario simulations, this comprehensive benchmark contains 25,845 evaluation instances across 49 practical applications. It supports dual evaluation protocols (pairwise/Best-of-N) and dual alignment objectives: 37 scenarios for helpfulness optimization and 12 for harmfulness mitigation. The benchmark’s scenario-based design enables practical capability assessment. The selected benchmarks provide complementary evaluation perspectives: foundational capability assessment (RewardBench), nuanced discrimination testing (RM-Bench), and real-world application simulation (RMB), ensuring comprehensive model evaluation.

A.2 Datasets

Our experiments leverage six complementary preference datasets spanning mathematical reasoning, code generation, and general instruction following: **MATH** (Hendrycks et al., 2021) A challenging benchmark for multi-step mathematical problem solving, containing 12,500 competition-level problems from high school mathematics tournaments. Each problem is accompanied by a step-by-step solution and categorized into seven difficulty levels ranging from algebra to calculus. We use the standard split with 7,500 training and 5,000 test

problems.

UltraFeedback (Cui et al., 2024) A large-scale preference dataset comprising 100K diverse instructions paired with multiple model responses. Each instance includes fine-grained quality ratings across four dimensions: helpfulness, safety, factual accuracy, and coherence. The dataset is constructed through adversarial prompting techniques that elicit varied response qualities from foundation models.

OffsetBias (Park et al., 2024) A carefully designed dataset addressing positional bias in preference modeling, containing 15K contrastive pairs where preferred/dispreferred responses are systematically rotated across different positions. This dataset enables robust training against common artifacts in preference annotation, particularly the tendency to favor responses in specific positions.

HelpSteer2-Preference (Wang et al., 2025b) An extension of the original HelpSteer dataset with 50K multi-turn dialogue preference pairs annotated by domain experts. Each conversation is evaluated across five criteria: task completion, clarity, depth of explanation, safety, and contextual awareness. The dataset emphasizes complex real-world scenarios requiring balanced consideration of multiple quality dimensions.

Skywork Reward Preference 80K (Liu et al., 2024a) This cross-domain preference dataset originally contains chat, safety, mathematics, and coding interactions. Our analysis revealed a critical flaw in its magpie_ultra subset (30% of total data): rejected responses systematically contain the `<im_start>` token absent in chosen responses, while preferred answers exhibit single-turn formatting versus multi-turn rejection patterns. Given these spurious correlations in mathematical and coding domains, we implemented source-level filtering to exclude all magpie_ultra specimens, retaining only verified high-quality pairs.

Code-Preference-Pairs A specialized coding dataset featuring 8K carefully curated preference pairs. Constructed through systematic perturbations of functional code snippets, each pair contrasts buggy implementations with corrected versions. The dataset emphasizes subtle logical errors requiring deep program understanding to distinguish, making it particularly valuable for training code-specific reward models.

Math-DPO-10K (Lai et al., 2024) A mathematical reasoning preference dataset containing 10K step-wise solutions annotated with process-level quality judgments. Unlike conventional math datasets

focusing solely on final answers, this resource provides granular feedback on reasoning validity, with particular attention to common error patterns in algebraic manipulations and proof strategies.

A.3 Baselines

We evaluate E-GRM against three distinct classes of reward modeling approaches: **Scalar Evaluation Models**. These systems generate numerical ratings directly, assessing preferences via singular numerical values devoid of explicit reasoning processes. Representative examples encompass traditional reward models (RM (Stiennon et al., 2020)), parameter-tuned variants like SteerLM-RM (Wang et al., 2024e), and specialized architectures including Nemotron-RM (Nvidia et al., 2024), Starling-RM (Li et al., 2024a), ArmoRM (Wang et al., 2024a), and Skywork-RM (Liu et al., 2024a). While demonstrating effectiveness in well-structured evaluation tasks, these approaches typically offer limited insight into their decision-making processes. **Generative Assessment Models**. This class employs text generation capabilities to produce both qualitative feedback and quantitative scores. Notable instances comprise foundation models like Llama (Dubey et al., 2024), Qwen (Qwen et al., 2025), Claude (Anthropic, 2024), GPT series variants (OpenAI et al., 2024b; Hurst et al., 2024), Gemini-1.5-pro (Reid et al., 2024), and self-improving frameworks (Wang et al., 2024d). By generating explanatory text alongside numerical evaluations, these models enhance transparency in judgment formation through natural language rationalization. **Structured Reasoning Models**. This paradigm integrates analytical reasoning into evaluation through chain-of-thought mechanisms or critique-based training. Implementations include Critique-RM (Yu et al., 2024; Jiang et al., 2023), DeepSeek-GRM (Liu et al., 2025d), and our proposed Efficient Reward Modeling framework. Such architectures demonstrate superior performance in multi-step logical reasoning, safety-critical assessments, and preference differentiation in complex contexts, benefiting from their structured cognitive modeling approach.

A.4 Implementation Details

Training Stage. Supervised fine-tuning (SFT) employs well-designed loss functions tailored to both short and long samples to optimize model performance for different reasoning scenarios. For the subset D_{short} , the model is trained to directly

produce the final answers. In contrast, on D_{long} , a subset of complex tasks that demand structured reasoning, the model is instructed to generate explicit, detailed step-by-step reasoning chains c_l that logically culminate in the final answer, ensuring the transparency and traceability of the reasoning process. The training process combines both objectives synergistically, using a fixed batch size of 512 to balance training stability and computational efficiency, alongside an initial learning rate of 5×10^{-6} that is gradually adjusted during training to avoid overfitting and accelerate convergence.

For **Reward Model Training**, the point-wise reward model f_{RM} uses blended Huber loss ($\delta = 0.1$) and hinge loss ($m = 0.2$) with trade-off factor $\alpha = 0.7$. Supervised regression is performed on high-quality annotated human preference data, where each sample is paired with $K = 10$ diverse responses to ensure comprehensive evaluation of the model’s performance. The training process for the reward model uses a batch size of 256 and a learning rate of 3×10^{-6} , which are carefully selected to match the complexity of the preference data and avoid training instability. During the reinforcement learning stage, **Coupled-GRPO Optimization** updates policy parameters π_ϕ following coupled rewards, which jointly balance correctness, reasoning conciseness, and discriminative scoring differences. For short reasoning rewards, the respective weights are set to $w_1 = 0.7$, $w_2 = 0.2$, and $w_3 = 0.1$. All experiments use a fixed random seed (43). Each experiment was conducted across 3 independent runs to ensure statistical reliability of the reported results.

Inference Stage. At inference, the GRPO-trained model M^* produces an initial greedy decoding output $r^{(0)}$. We utilize the ContainsCoT($r^{(0)}$) criterion to determine if the input requires further chain-of-thought reasoning. If no such reasoning is detected, $r^{(0)}$ is returned as the final answer. For complex samples, the system employs a **multi-parameter decoding** strategy by generating $K = 8$ candidate responses $\{r^{(k)}\}_{k=1}^K$ using diverse decoding configurations such as different temperatures and top- p parameters. **Discriminative scoring** is then applied: each candidate is evaluated by the trained reward model f_{RM} . The response with the highest score is selected as the final output. This maintains efficiency for simple prompts while enhancing reasoning on complex instances.

Prompt for Cold-start Long-CoT Sampling

As an AI Quality Auditor, rigorously evaluate <prompt, response> pairs through this structured framework:

<foundation>

Basic Compliance Check

Safety: Contains violent/discriminatory/illegal content? ✓ Pass × Fail

Legal Compliance: Violates data privacy/IP laws? ✓ Pass × Fail

Ethical Standards: Aligns with social norms? ✓ Pass × Fail

Core Quality Dimensions

Accuracy (30%)

Factual verification

Domain expertise validation

DataSource reliability

Relevance (25%)

Prompt requirement coverage

Off - topic content presence

Unnecessary additions

Completeness (20%)

Problem - solving closure

Key element inclusion

Logical Rigor (15%)

Reasoning chain coherence

Causal fallacy detection

Practicality (10%)

Actionable implementation

Solution effectiveness

Risk Assessment

Misleading Potential: **Low Medium High**

Knowledge Gaps: **None Partial Severe**

</foundation>

<principles>

1. Immediate rejection if any **Fail** in Basic Compliance
2. Minimum 85/100 in Core Dimensions
3. Strict evaluation for high - risk indicators

</principles>

<thinking>

- Conduct multi-stage verification with domain-specific knowledge graphs
- Cross - validate claims against authoritative sources
- Perform causal reasoning analysis using Bayesian networks
- Simulate real - world implementation scenarios

</thinking>

Final Score: <Predict> Score(1-10) </Predict>

Final verdict: <judgment>[Yes]</judgment> or <judgment>[No]</judgment>

Figure 4: Prompt for Cold-start Long-CoT Sampling evaluation framework.