

# OPENREWARD: LEARNING TO REWARD LONG-FORM AGENTIC TASKS VIA REINFORCEMENT LEARNING

Ziyou Hu<sup>1\*</sup> Zhengliang Shi<sup>2\*</sup> Minghang Zhu<sup>2</sup> Haitao Li<sup>3</sup> Teng Sun<sup>2</sup>

Pengjie Ren<sup>2</sup> Suzan Verberne<sup>1</sup> Zhaochun Ren<sup>1†</sup>

<sup>1</sup>Leiden University, Leiden, Netherland <sup>2</sup>Shandong University, Qingdao, China

<sup>3</sup>Tsinghua University, Beijing, China

{retrro.hu, zhengliang.shii}@gmail.com

{s.verberne, z.ren}@liacs.leidenuniv.nl

## ABSTRACT

Reward models (RMs) have become essential for aligning large language models (LLMs), serving as scalable proxies for human evaluation in both training and inference. However, existing RMs struggle on knowledge-intensive and long-form tasks, where evaluating correctness requires grounding beyond the model’s internal knowledge. This limitation hinders them from reliably discriminating subtle quality differences, especially when external evidence is necessary. To address this, we introduce OPENRM, a tool-augmented long-form reward model that systematically judges open-ended responses by invoking external tools to gather relevant evidence. We train OPENRM with Group Relative Policy Optimization (GRPO) on over 27K synthesized pairwise examples generated through a controllable data synthesis framework. The training objective jointly supervises intermediate tool usage and final outcome accuracy, incentivizing our reward model to learn effective evidence-based judgment strategies. Extensive experiments on three newly-collected datasets and two widely-used benchmarks demonstrate that OPENRM substantially outperforms existing reward modeling approaches. As a further step, we integrate OPENRM into both inference-time response selection and training-time data selection. This yields consistent gains in downstream LLM alignment tasks, highlighting the potential of tool-augmented reward models for scaling reliable long-form evaluation. Our code is available on  OpenRM.

## 1 INTRODUCTION

Reward models (RMs) have emerged as scalable and effective substitutes for human evaluators, playing a central role in aligning Large Language Models (LLMs) during both training and inference (Guo et al., 2025; Chen et al., 2025b). By learning to predict human preferences over model outputs, RMs provide reliable supervision signals that guide reinforcement learning and support preference optimization at inference time (Ouyang et al., 2022; Song et al., 2024; Zuo et al., 2025). Benefiting from the rich world knowledge embedded in LLMs, RMs exhibit a certain degree of generalization capability, enabling them to perform effectively across diverse scenarios. As a result, building accurate, robust, and well-aligned reward models has become a critical step toward the development of safer and more useful intelligent systems.

Despite the success, existing RMs still faces significant challenges when evaluating the knowledge-intensive long-form outputs generated by Deep Researcher systems (Zheng et al., 2025; Zhu et al., 2025). These texts typically require integrating various external information sources, which goes beyond the model’s internal knowledge. For example, evaluating the novelty of a research paper often depends on cross-referencing external scientific corpora (e.g., arXiv) (Zhu et al., 2025), while comparing travel itineraries drafted by downstream LLM applications demands up-to-date knowledge of specific destinations (Chen et al., 2024). Existing studies either rely on prompting commercial LLMs (e.g., GPT-4) to perform fact-checking via search engines (Wei et al., 2024), or use simple tool

---

\*Equal contributions

†Corresponding author

interactions for short-form factual verification (e.g., date checking via a calendar) (Li et al., 2023). However, these approaches exhibit limited flexibility and scalability, and their effectiveness in long-form reward modeling remains unclear. These limitations highlight a key open question: *How can we design reward models that effectively evaluate long-form outputs in open-domain settings?*

To address the above question, we propose *empowering reward models with the ability to use external tools during evaluation*. Similar to task-solving agents (Jin et al., 2025; Wang et al., 2024a), tool-augmented reward models can retrieve and reason over external information sources in real time, allowing them to make more accurate and context-aware judgments. Building on this intuition, we introduce OPENRM, a reward modeling framework designed to assess complex, knowledge-intensive long-form responses with the assistance of external tools. OPENRM explores how tool-use capabilities can be flexibly integrated into the reward modeling pipeline via reinforcement learning (RL), enabling the model to actively retrieve, verify, and reason over external information when forming evaluation judgments. Given two candidate responses, OPENRM first plans and executes a sequence of tool calls to retrieve supporting evidence, then verifies each response against this evidence, and finally selects the better answer. We train both the stepwise reasoning and the final judgment of the reward model via reinforcement learning using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). The objective combines (i) an intermediate tool-use penalty that discourages irrelevant or excessive calls while encouraging task-appropriate tool selection, and (ii) a terminal outcome-accuracy reward that incentivizes reliably comparing the two responses and choosing the higher-quality one.

A key obstacle in training long-form reward models is the lack of reliable pairwise response data from real-world sources to serve as training data, as demonstrated in Table 1. To address this, we propose a simple yet controllable partial synthesis framework that automatically generates queries and constructs positive–negative pairs. Specifically, we first perform target-aware query generation, prompting a strong LLM (e.g., DeepSeek-V3) to formulate a self-contained query from a sampled document. Then, we obtain preference-ordered responses by prompting the LLM answer the query with and without access to the reference document, thereby yielding a positive and a negative response, respectively. This method enables scalable supervision without costly human annotations.

As a further step, we explore the utility of OPENRM in scaling LLM alignment at both training and inference stages. At inference time, we show that OPENRM can select a better response among candidates, improving long-form answer quality. At training time, we employ OPENRM as a data selector, where we compare task-solving trajectories sampled from LLMs and filtering with OPENRM, thereby obtaining a higher-quality training dataset. Our experiments demonstrate that LLMs trained with OPENRM-selected data achieve significant performance gains compared with those trained on data filtered by existing RMs, highlighting the potential of tool-augmented reward modeling.

Our contributions are summarized as follows: (i) We introduce OPENRM, a reward model that can judge the long-form answer with the assistance of external tools, enabling accurate rewarding of complex agentic tasks. (ii) We propose a simple but controllable pipeline for automatic reward data construction, collecting 30K+ high-quality pair-wise training data with binary label in an unsupervised manner. (iii) Extensive experiments on five benchmarks, showing that OPENRM outperforms existing baselines significantly. (v) We demonstrate that OPENRM not only improves inference-time response selection but also serves as an effective data selector for training, paving the way for future research on LLM alignment.

## 2 RELATED WORK

### 2.1 REWARD MODELING

Reward modeling serves as a crucial bridge connecting human intent with model behavior, and is a key component in the development of artificial intelligence. Existing research mainly explores two paradigms of reward modeling: (1) Scalar Reward Models (Scalar RMs): This approach typically involves training a sequence classifier on top of a frozen large language model (LLM), turning reward modeling into a supervised classification problem that predicts human preferences or rating labels (Wang et al., 2024c; Liu et al., 2024b). (2) Generative Reward Models (a.k.a., LLM-as-Judge): This paradigm retains the generative capabilities of LLMs, leveraging their powerful language understanding and free-form reasoning abilities to produce preference judgments for pairwise

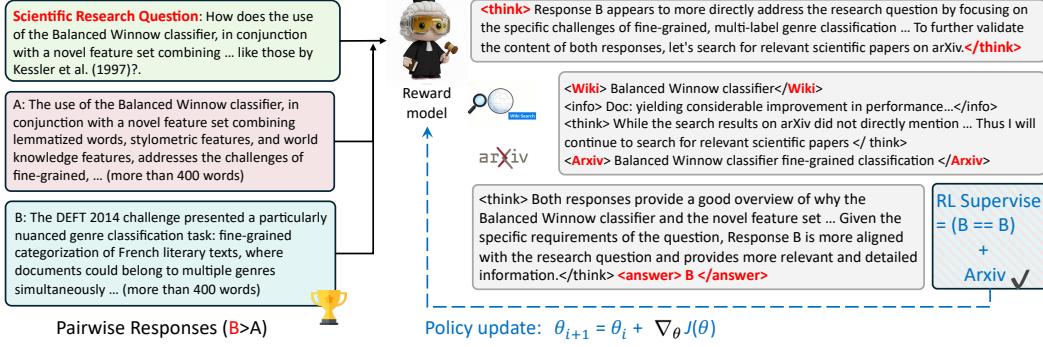


Figure 1: Illustration of the OPENRM framework where the reward model, when receiving the candidate responses, progressively invokes external tools to gather useful evidence, and then make the final judgment.

comparisons or natural language evaluations (Li et al., 2024a; Zheng et al., 2023; Gu et al., 2024). Compared to traditional scalar reward models, generative reward models offer significant advantages in expressiveness and flexibility. They not only provide fine-grained natural language explanations but also perform more complex value judgments in context, thereby better capturing human preferences and alignment objectives. Therefore, in this work, we focus on enhancing the capabilities of generative reward models. Our approach focuses on addressing the performance limitations of generative reward models in long-form text scenarios.

## 2.2 REINFORCEMENT LEARNING WITH VERIFIABLE REWARD

Reinforcement Learning with Verifiable Reward (RLVR) has recently emerged as a foundational paradigm for improving the reasoning capabilities and robustness of large language models. By leveraging programmable or automatically verifiable criteria as reward signals, RLVR reduces dependence on subjective human preferences and helps mitigate issues such as reward hacking. A growing body of research has applied RLVR to the training of LLM-as-a-judge systems, aiming to encourage deeper reasoning and more accurate evaluations through outcome-driven rewards Chen et al. (2025a); Whitehouse et al. (2025); Li et al. (2024b). For instance, JudgeLRM Chen et al. (2025a) combines structural and content-based rewards, yielding significantly better performance than standard supervised fine-tuning (SFT) across a range of evaluation tasks. Similarly, J1 Whitehouse et al. (2025) transforms both verifiable and non-verifiable prompts into judgment tasks with verifiable reward signals, using reinforcement learning to foster thoughtful reasoning while reducing systematic biases. In contrast to these prior approaches, the present work focuses on leveraging reinforcement learning to improve the model’s ability to effectively incorporate external tool.

## 3 METHOD

This section details our proposed OPENRM. We begin by detailing how OPENRM evaluates two candidate responses with the support of external tools. Then, we describe our controllable data synthesis strategy for generating large-scale training pairs. Finally, we introduce the reinforcement learning framework that empowers OPENRM with robust judgment capabilities.

### 3.1 REWARDING AGENTIC TASKS WITH TOOLS

OpenReward is a tool-augmented evaluation framework for knowledge-intensive long-form responses, where the answer quality depends on information beyond the model’s internal parameters. As illustrated in Figure 1, OPENRM autonomously invokes external tools during evaluation when necessary, enabling more accurate and informed judgments. Specifically, given an input query  $q$  and two candidate responses  $x_1$  and  $x_2$ , the model iteratively decides which tool to invoke in order to verify or complement the information contained in the responses. Formally, the tool selection at step  $i$  can be expressed as:

$$t_i = \mathcal{RM}(t_i \mid P, (x_1, x_2, q); \theta), \quad (1)$$

where  $P$  denotes the system prompt provided to the model, specifying the available toolset, and  $t_i$  indicates the tool selected at step  $i$ . A concrete example of the prompt is illustrated in Appendix A.3. The selected tool  $t_i$  is executed to obtain an external result  $e_i$ , which is appended to the model’s context to guide subsequent reasoning and tool selection. This process can be formalized as a sequential decision process: the state encodes the current context including the query, candidate responses, and prior evidence, the action corresponds to selecting a tool or terminating with a final decision; and the transition updates the context with the result  $e_i$  returned by the selected tool. The overall judgment trajectory  $o$  can be summarized as:

$$(q, x_1, x_2) \rightsquigarrow t_1 \xrightarrow{\text{Exec}} e_1 \cdots \rightsquigarrow t_i \xrightarrow{\text{Exec}} e_i \rightsquigarrow \cdots \rightsquigarrow y, \quad (2)$$

Here  $y$  denotes the final judgment, which contains selecting the better response between  $x_1$  and  $x_2$  (e.g., <answer> A </answer>), accompanied by a brief explanation that supports the decision for interpretability purposes. We set the maximal tool calling to  $n$  during the practice, and thus the  $o = \{(t_i, e_i) \mid i \in [n]\} \cup \{y\}$ .

### 3.2 LEARNING TO REWARD VIA RLVR

Enabling the agentic rewarding process in OPENRM is challenging. A straightforward approach is to directly prompt the reward model to select tools and make judgments. However, this strategy may result in limited performance, particularly for smaller models such as Qwen-7B, which tend to struggle with multi-step reasoning in the absence of explicit supervision. Another alternative is to fine-tune the model using supervised objectives, such as Rejection sampling Fine-Tuning (RFT). While this approach is more structured than prompting, it suffers from two key limitations. First, it passively imitates fixed trajectories, which restricts generalization beyond handcrafted templates. Second, it fails to capture the dynamic and iterative nature of tool-augmented reasoning.

To overcome these challenges, we adopt a reinforcement learning framework that enables LLMs to actively explore tool-use strategies. It requires only a final preference label as supervision, while encouraging the discovery of effective multi-step reasoning through trial and error. This improves judgment accuracy and promotes better generalization to unseen queries and tool configurations.

**Training Signal.** Relying solely on final correctness as a reward is too sparse to effectively guide multi-step reasoning in many real-world scenarios. To address this, we design a composite reward function that supervises both the intermediate tool-use behavior and the final prediction outcome. Specifically, the reward consists of two complementary components. First,  $\mathcal{R}_{\text{tool}}$  evaluates the accuracy of tool selection before the final decision. It rewards the model for invoking appropriate tools relevant to the task and penalizes unnecessary or irrelevant tool usage. Second,  $\mathcal{R}_{\text{EM}}$  measures the correctness of the final prediction by checking whether the model’s output  $a_i$  exactly matches the ground-truth label, i.e.,  $\text{EM}(a_i)$ . The final reward is defined as follows:

$$\mathcal{R} = \mathcal{R}_{\text{EM}} + \text{sign}(\mathcal{R}_{\text{EM}}) \cdot \lambda \cdot \mathcal{R}_{\text{tool}}, \quad (3)$$

where  $\text{sign}(x)$  is set to 1 if and only if  $x > 0$  and  $\lambda$  is a weighting factor that balances the two reward components.  $\mathcal{R}_{\text{EM}}$  enforces the correctness of the final outcome, while  $\mathcal{R}_{\text{tool}}$  provides dense feedback during the decision process by assigning +1 to correct tool selections and 0 to incorrect ones. To ensure factual accuracy remains the primary learning objective, the reward function only assigns credit for tool usage when the final prediction is correct. This design strikes a balance between promoting accurate reasoning and encouraging effective tool-use behavior.

**Training Process.** We optimize OPENRM with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Formally, for each query  $(q, x_1, x_2)$ , OPENRM generates a group of  $m$  candidate responses using different sampled trajectories of tool use. Each trajectory  $\mathcal{T}_i$  is evaluated by our composite reward function  $\mathcal{R}$ , which jointly considers both the intermediate tool usage quality and the final answer correctness. GRPO then computes a group-relative advantage by comparing each trajectory’s reward to the mean within its group:

$$\mathcal{A}_i = \frac{\mathcal{R}(\mathcal{T}_i) - \frac{1}{m} \sum_{j=1}^m \mathcal{R}(\mathcal{T}_j)}{\text{std}(\{\mathcal{R}(\mathcal{T}_j) \mid j \in [m]\})}, \quad (4)$$

where  $\text{std}(\cdot)$  is the standard deviation of group rewards. The final objective is a clipped policy gradient with KL regularization:

$$\mathcal{J}(\theta) = \mathbb{E}_{o_i \sim \mathcal{O}} [\min(\rho_i \mathcal{A}_i, \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) \mathcal{A}_i) - \beta \text{KL}(\theta \parallel \theta_{\text{ref}})], \quad (5)$$

Statistic	Ours	RewardBench	TARA	JudgeLM
# Training data	✓	✗	✓	✓
- # of data Scale	27K	-	13K	100K
- # of Answer / Response Length	582.45	-	49.04	117.60
# Evaluation data	✓	✓	✓	✓
- # of data Scale	2K	2K	1.5K	5K
- # of Answer / Response Length	601.03	93.31	52.19	116.09

Table 1: Basic statistics of our benchmark.

where  $\rho_i$  is the important ratio between the updated and old model parameter,  $\epsilon$  controls the clipping range and  $\beta$  weights the penalty for diverging from a reference model  $\theta_{\text{ref}}$ .

### 3.3 CONTROLLABLE DATA SYNTHESIS

A key challenge in training reliable reward models for long-form text evaluation lies in the lack of suitable training data. Compared to short-form tasks, knowledge-intensive long-form responses are more difficult to collect due to their high information density and annotation cost. To address this bottleneck, we propose a straightforward and controllable data synthesis framework that scales the construction of high-quality pairwise training instances for OPENRM. Specifically, the core idea of our framework is to generate preference pairs by prompting the same LLM to produce responses to the same query under different input conditions. By controlling the availability of supporting information, such as including or excluding the reference document, we induce a clear and consistent quality gap between the resulting responses. This strategy eliminates the need for manual annotation while ensuring that the resulting preference pairs are both diverse and reliable, making them well-suited for training reward models in knowledge-intensive long-form evaluation settings.

**Target-aware Query Generation.** We begin by sampling a set of domain-specific documents (e.g., Wikipedia passages, arXiv papers, or travel guides) and use them as context to prompt a strong LLM to generate a *target-aware query* for these documents. This approach ensures that the generated queries are closely grounded in the source content and can be meaningfully answered with access to the corresponding document. As a result, the synthesized queries are inherently knowledge-intensive and require external evidence for accurate answering, providing a solid foundation for constructing high-quality training data for long-form reward modeling. It is worth noting that a single query may be associated with multiple relevant documents, which allows for the generation of queries with varying difficulty levels and promotes data diversity.

**Positive–Negative Pair Synthesis.** Given a query, we generate preference pairs by prompting the LLM under different input conditions. The *positive response* is generated using both the query and its corresponding reference document, guaranteeing that the answer is grounded, informative, and factually accurate. In contrast, the *negative response* is generated using the query alone, without access to the reference document, which typically leads to incomplete, hallucinated, or less reliable content. This contrastive setup introduces a clear quality gap between the two responses, allowing us to construct controllable and scalable pairwise training data tailored for long-form reward modeling.

Finally, we synthesize over 27K high-quality training instances and 2K evaluation instances across three representative scenarios: (i) Wikipedia, where responses address open-domain questions about specific entities; (ii) Scientific research, where responses typically provide comprehensive surveys or technical introductions; and (iii) Medical, where responses focus on answering health-related questions. The detailed statistics are presented in Table 1.

## 4 EXPERIMENT SETUP

**Benchmark and Evaluation Protocol** We evaluate OPENRM on five benchmarks: three newly collected datasets and two widely used ones. Specifically, the newly collected datasets include 500 Wikipedia QA examples, 500 scientific research questions, and 1,000 medical QA examples. These datasets are consistent with the training data source and are used to assess the model’s in-domain performance. Additionally, we validate OPENRM on two popular reward model benchmark

Domain	Wikipedia ( $\uparrow$ )	Scientific ( $\uparrow$ )	Medical ( $\uparrow$ )	Average ( $\uparrow$ )
<i>Direct reasoning without tool utilization.</i>				
Deepseek-V3.1 (Liu et al., 2024a)	75.00	46.00	33.00	51.33
GPT-4o (Hurst et al., 2024)	70.00	48.20	44.00	54.07
Gemini-2.5-Pro (Comanici et al., 2025)	72.20	46.60	36.00	51.60
Claude-Opus-4-1-20250805 (Anthropic Team, 2025)	74.60	49.20	51.10	58.30
Skywork-Reward-Gemma-2-27B (Liu et al., 2024b)	45.20	55.40	47.74	49.45
JudgeLRM-7B (Chen et al., 2025a)	50.80	50.60	48.44	49.94
RRM-7B (Guo et al., 2025)	56.90	52.95	53.10	54.32
RM-R1-Qwen2.5-Instruct-7B (Chen et al., 2025b)	55.40	54.80	52.30	54.17
RM-R1 trained on our data	66.00	73.00	65.00	68.00
<i>Reasoning with tool utilization.</i>				
Deepseek-V3.1 (Liu et al., 2024a)	77.00	48.00	34.00	53.00
GPT-4o (Hurst et al., 2024)	76.40	58.60	53.40	62.80
Gemini-2.5-pro (Comanici et al., 2025)	72.50	54.60	42.40	56.50
Claude-Opus-4-1-20250805 (Anthropic Team, 2025)	75.60	56.10	58.00	63.23
OPENRM-Qwen-2.5-7B	<b>93.00</b>	<b>90.00</b>	<b>91.00</b>	<b>91.33</b>

Table 2: Comparison between OPENRM and baseline models in terms of **Accuracy** ( $\uparrow$ ) across three in-domain tasks (Wikipedia, Scientific, and Medical). We report results both under direct reasoning without external tool usage and under reasoning with tool utilization.

datasets, RewardBench Lambert et al. (2024) and PandaLM Wang et al. (2024b), to assess the out-of-domain performance. For all datasets, we report the average accuracy on the test sets.

**Baselines** We compare OPENRM against two categories of baselines. First, direct reasoning without tool utilization (*a.k.a.*, naive LLM-as-Judge), which typically prompts or trains a LLM to reason over its internal knowledge and select a better response from multiple candidates. We include: *Skywork-Reward* (Liu et al., 2024b), *JudgeLRM* (Chen et al., 2025a), *RRM* (Guo et al., 2025), and *RM-R1* (Chen et al., 2025b), all of which are trained through supervised fine-tuning (SFT) on hand-crafted golden data or RL with accuracy as a training signal. We also include top-ranked LLMs from the Arena Leaderboard, such as DeepSeek and Gemini.

Second, *agentic reward Modeling*, which augments the LLM-as-Judge approach by integrating external tools, enabling LLMs to use tools on demand. In more detail, we prompt the LLM to retrieve relevant information using external tools before making a final judgment. We implement this approach using top-ranked LLMs based on FacTools Chern et al. (2023), a widely used toolkit for tool-augmented fact-checking. More detailed implementation information for all baselines can be found in Appendix A.2.

**Implementation Details** We adopt Qwen-2.5-7B-Instruct as the backbone model for training, consistent with prior approaches Chen et al. (2025b;a). During training, the reward model is allowed to call two external retrieval tools as evidence sources: (i) Wikipedia Search, indexed from the 2018 Wikipedia dump; (ii) arXiv Search, served via LitSearch (Ajith et al., 2024) over publicly available literature. Unless otherwise specified, ColBERT-v2.0 (Santhanam et al., 2021) is employed as the retrieval model, chosen for its openness and transparency. All experiments are conducted over two epochs, with a maximum prompt length of 4096 tokens, a response length capped at 2048 tokens, and a batch size of 512. The group size of our GRPO is set to 5, the clip range is set to 0.5, the KL divergence factor  $\beta$  is set to  $10^{-3}$ , and smooth term is set to  $10^{-6}$ .

## 5 EXPERIMENT RESULTS

### 5.1 OVERALL PERFORMANCE

**In-domain Evaluation.** Table 2 compares OPENRM with baseline models on our three newly collected benchmarks. We derive the following observations from the experiment results. (1) Train-based reward models generally underperform generic LLM-as-judge, likely due to limitations in parameter scale and training method. (2) Agentic reward models outperform directly prompting LLMs as judges, suggesting that tool usage contributes to improved evaluation accuracy. However,

Methods	Backbone	Data (↓)	PandaLM (↑)	RewardBench (↑)	Average (↑)	Average Δ% (↑)
<b><i>Out-of-domain Evaluation</i></b>						
OPENRM	Qwen-2.5-7B-Instruct	<b>27K</b>	<b>79.42</b>	77.66	<b>78.54</b>	-
RM-R1	Qwen-2.5-7B-Instruct	72K	72.71 <sub>±6.71</sub>	68.34 <sub>±9.32</sub>	70.52	↓11.37%
JudgeLRM-7B	Qwen-2.5-7B-Instruct	100K	72.37 <sub>±7.05</sub>	74.45 <sub>±3.21</sub>	73.41	↓6.99%
Prometheus-v2.0	Mistral-7B-Instruct	200K	72.80 <sub>±6.62</sub>	71.55 <sub>±6.11</sub>	72.17	↓8.83%
RRM-7B	Qwen-2.5-7B-Instruct	420K	77.74 <sub>±1.68</sub>	<b>78.54</b> <sub>±0.88</sub>	78.14	↓0.52%

Table 3: Experiment results in terms of accuracy on two out-of-domain benchmarks (PandaLM and RewardBench). We further report the training data scale (Scale) of each reward model to highlight efficiency differences across methods. OPENRM achieves the best average accuracy (78.54) while using substantially less training data.

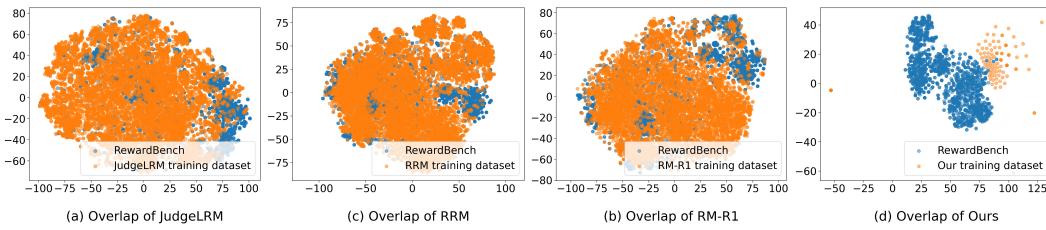


Figure 2: Illustration of the overlap between RewardBench and the training datasets of different reward models. The embeddings are visualized in 2D space. The orange points denote training data and blue points denote RewardBench instances. We observe substantial overlap in JudgeLRM, RRM, and RM-R1, whereas our dataset shows almost no intersection, ensuring a fairer and less biased evaluation .

as most LLMs are not explicitly trained to use tools effectively, prompting them to invoke tools still yields unsatisfactory performance. (3) Fine-tuning RM-R1 on our training data improves its performance from 54.17 to 68.00. However, due to limitations in the training approach, its overall performance remains suboptimal, even when using the same data. (4) By combining carefully curated synthetic data with a tailored training procedure, OPENREWARD can autonomously use tools during evaluation and achieves the best in-domain results, surpassing large models such as GPT-4o and DeepSeek-R1, which demonstrates the effectiveness of our method.

**Out-of-domain Evaluation.** To evaluate generalization ability, we further assess OPENRM on out-of-domain benchmarks, where the test sets lie in a distribution space distinct from the training data. As shown in Table 3, OPENRM, despite being trained on only 27K examples, substantially outperforms strong baselines such as RM-R1 and JudgeLRM-7B. We attribute this improvement to two main factors. First, the reward model is trained to evaluate knowledge-intensive, long-form responses that demand nuanced reasoning in order to produce accurate judgments. This capability contributes to improved generalization across domains. Second, the agentic modeling framework strengthens the model’s ability to reason over retrieved external evidence. These two factors together equip OPENRM with notable robustness and the ability to generalize across diverse domains.

**A Closer look at training data overlap.** We find that OPENRM achieves relatively smaller improvements on RewardBench compared to other benchmarks. To investigate the underlying cause of this discrepancy, we examine the overlap between the training data and the RewardBench test cases. As shown in Figure 2, the training sets of several baselines, such as JudgeLRM and RM-R1, have substantial overlap with the RewardBench test cases. In contrast, the training data used in our method exhibits only minimal overlap. Nevertheless, OPENRM still achieves competitive performance despite relying on significantly less and out-of-distribution data. This indicates that the effectiveness of OPENRM primarily stems from its stronger generalization capability, rather than memorization of seen examples.

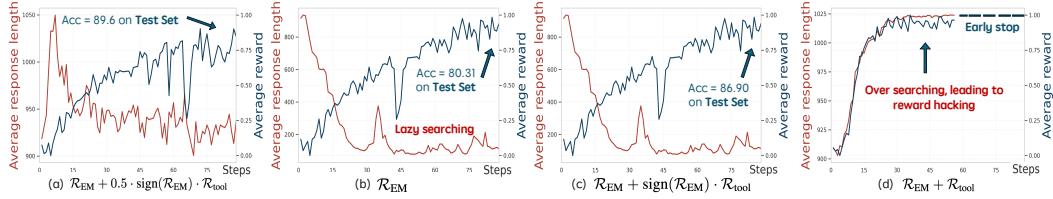


Figure 3: Training process of the models trained with different variants of the vanilla training supervision in Eq. 3.1. We plot average response length (red) and average reward (blue) over training steps, with final test set accuracy annotated. The results highlight distinct failure modes: (b) lazy searching under  $R_{EM}$ , (d) over-searching and reward hacking under  $R_{EM} + R_{tool}$ , while (a) the proposed composite reward achieves stable improvement and the best accuracy.

## 5.2 ABLATION STUDIES

**Setup.** In OPENRM, the reward model is trained with a composite function (Eq. 3.1), which combines two supervision signals: tool selection and judgment accuracy. To better understand the contribution of each component, we design several reward variants for comparison:

- (1) *only  $R_{EM}$* : we use only the judgment accuracy signal  $R_{EM} \in [0, 1]$ .
- (2)  $\mathbf{1}_{R_{EM}} \cdot R_{tool}$ : the weight of  $R_{tool}$  is increased from 0.5 to 1, formulated as  $R_{EM} + \text{sign}(R_{EM}) \cdot R_{tool} \in [0, 2]$ ;
- (3)  $0.5 \cdot R_{tool}$ : we remove the indicator  $\mathbf{1}_{R_{EM}}$  when supervising tool usage, formulated as  $R_{EM} + R_{tool} \in [-0.5, 1.5]$ . This decouples tool-use learning from judgment learning, encouraging more explicit tool usage.

Since these reward functions operate on different scales, we normalize them to a unified  $[0, 1]$  range for fair comparison.

**Results.** We derive three key insights from the ablation studies. **First, removing the tool selection reward leads to lazy searching.** using only  $R_{EM}$  produces shorter responses and a noticeable drop in overall performance. Detailed analysis shows that the model avoids using tools and instead relies on direct prediction, a behavior referred to as lazy searching, which is inadequate for knowledge-intensive tasks. **Second, model performance is relative robust regarding hyper-parameter  $\lambda$ .** The w/  $\mathbf{1}_{R_{EM}} \cdot R_{tool}$  variant achieves results comparable to the vanilla version. For example, the vanilla model achieves 89.60 across our three newly collected datasets, while this variant reaches 86.90, with no statistically significant difference under a two-tailed  $t$ -test. This suggests that the model is not overly sensitive to the weighting of the tool reward. **Finally, decoupling tool supervision from judgment accuracy leads to reward hacking.** In the variant with decoupled tool selection and judgment accuracy supervision, we observe excessive tool usage during both training and evaluation, i.e., *over-searching*. A likely reason is that the model receives relatively strong positive signals for correct tool usage even when failing to make correct judgments, leading to a form of reward hacking.

## 5.3 HUMAN EVALUATION

To further validate the reliability of judgments made by OPENRM, we conduct a human evaluation based on two key criteria: (1) **Self-containment**, i.e., whether the judgment is internally consistent and logically coherent; and (2) **Factuality**, i.e., whether the chosen response is more factually correct. We recruited three well-educated volunteers to evaluate 30 randomly sampled cases from each of the five datasets. Each case is rated on a three-point scale, where a score of 3 indicates the highest quality and 1 the lowest. As presented in Table 4, OPENRM outperforms RM-R1 in both self-consistency and factual accuracy. This improvement stems from its ability to incorporate external evidence, enabling it to generate more trustworthy and accurate judgments. Additionally, we conduct qualitative case studies to better understand the reasoning capabilities of OPENRM. These studies reveal that OPENRM is more effective at orchestrating tool usage to handle complex tasks. A representative example is provided in Appendix A.

Methods	Backbone	Self-contain ( $\uparrow$ )	Factuality ( $\uparrow$ )	Average ( $\uparrow$ )	Average $\Delta\%$ ( $\uparrow$ )
Metrics		Accuracy ( $\uparrow$ )	Accuracy ( $\uparrow$ )	Accuracy ( $\uparrow$ )	Accuracy ( $\uparrow$ )
OPENRM	Qwen-2.5-7B-Instruct	2.73	2.83	2.78	-
LLM-as-Judge	DeepSeek-V3.1	2.67	2.47	2.57	$\downarrow 7.55\%$
RM-R1 Chen et al. (2025b)	Qwen-2.5-7B-Instruct	2.13	1.73	1.93	$\downarrow 30.58\%$

Table 4: Human evaluation results comparing OPENRM with baseline reward models. The self-contain metric measures internal consistency of the generated judgments, while factuality assesses whether the judgments contain counterfactual or factually incorrect statements.

Methods	TrutuFulQA ( $\uparrow$ )	MMLU-Pro( $\uparrow$ )	Triviaqa ( $\uparrow$ )	Average ( $\uparrow$ )	Average $\Delta\%$ ( $\uparrow$ )
<i>Out-of-domain Evaluation</i>					
Qwen-2.5-3B-Instruct	32.19	38.17	42.33	37.56	-
-w/ RM-R1	32.07	38.23	43.69	37.99	$\uparrow 1.14\%$
-w/ OPENRM	<b>34.03</b>	<b>38.60</b>	<b>44.56</b>	<b>39.03</b>	$\uparrow 3.91\%$
Qwen-2.5-7B-Instruct	39.04	48.55	50.19	45.93	-
-w/ RM-R1	41.00	48.81	51.53	47.11	$\uparrow 2.57\%$
-w/ OPENRM	<b>43.33</b>	<b>50.63</b>	<b>50.12</b>	<b>48.03</b>	$\uparrow 4.57\%$

Table 5: Experiment results in terms of accuracy on three out-of-domain benchmarks. We compare Qwen-2.5-3B-Instruct and Qwen-2.5-7B-Instruct trained with either RM-R1 or OPENRM as reward models under direct preference optimization (DPO).

## 6 UTILITY STUDY

In this section, we demonstrate the practical utility of OPENRM in LLM training, extending its use beyond standalone evaluation. We experiment on the widely used UltraFeedback dataset Cui et al. (2023), which consists of diverse prompts paired with multiple candidate responses.

**OPENRM for data selection.** We apply OPENRM to score and filter training data, allowing the model to learn from higher-quality responses and reducing noise in supervision. Table 6 shows the selection accuracy on the UltraFeedback dataset, where the accuracy is measured by how often the model selects the response originally labeled better. OPENRM achieves 75.18%, outperforming RM-R1 Chen et al. (2025b) with 68.87%, indicating stronger alignment with human preferences. Additionally, we conduct qualitative case studies to better understand the reasoning capabilities of OPENRM. These studies reveal that OPENRM is more effective at orchestrating tool usage to handle complex tasks. A representative example is shown in Table 7.

**OPENRM for model alignment.** We further apply OPENRM to Direct Preference Optimization (DPO) Rafailov et al. (2023), a widely adopted reinforcement learning method for aligning LLMs with human feedback (i.e., RLHF). This experiments evaluate whether the high-quality responses identified by OPENRM can enhance downstream alignment. In this setting, the previously identified positive and negative responses are used as preference pairs for DPO training. As shown in Table 5, the 3B model trained on data curated by OPENRM achieves 39.03% and the 7B model trained on data curated by OPENRM achieves 48.03% in terms of accuracy on preference prediction, representing a substantial improvement over the vanilla backbone without post-training. Moreover, we find that it surpasses the counterpart trained on data filtered by RM-R1. This result demonstrates that the higher-quality preference data identified by OPENRM can more effectively guide preference-based optimization, leading to stronger alignment. These findings highlight the broader potential of tool-augmented reward modeling: by producing more accurate and reliable preference signals, OPENRM strengthens the effectiveness of downstream alignment algorithms such as DPO, thereby contributing to more robust and human-aligned LLMs.

Dataset	UltraFeedback
Metric	Accuracy ( $\uparrow$ )
<b>OPENRM</b>	<b>75.18</b>
RM-R1 Chen et al. (2025b)	68.87

Table 6: Selection results accuracy comparing **OPENRM** (75.18) and **RM-R1** Chen et al. (2025b) (68.87) on the UltraFeedback dataset.

---

**Dataset:** UltraFeedback

**Question:** How have foreign influences played a role in shaping Romania's cultural identity?

**Answer A:** Foreign influences have played an important role in shaping Romania's cultural identity. Due to its geographic location, Romania has been exposed to various cultures for centuries, from the neighboring Hungarians and Austrians to further away Ottomans and Russians. As a result, Romanian culture has incorporated aspects from all of these cultures... For example, the traditional Romanian New Year's Eve dinner includes ntura...

**Answer B:** The influence of foreign cultures has played a significant role in shaping Romania's cultural identity throughout history. Here are some examples: 1. Greek Influence: Ancient Greece had a profound impact on Romanian culture... 2. Ottoman Influence... 3. German Influence... 4. French Influence... 5. Italian Influence...

**RM-R1 Judgment:**

Answer B is better due to its systematic, numbered breakdown of five specific cultural influences (Greek, Ottoman, etc.). This provides a more complete and structured inventory than Answer A's general narrative, enhancing clarity. This organization is more useful for a user seeking a comprehensive list and easy-to-digest information. < answer >B< /answer >

**X Wrong answer: the explanation for each response is brief and lacks in-depth analysis, which only focus on the structure while neglecting content factuality.**

**OPENRM Judgment:**

Both answers provide information about the role of foreign influences in shaping Romania's cultural identity. However, Answer A is more detailed, providing concrete examples such as... I will perform a search to see if there are any critical steps missed by either answer, before making a final decision. <search>WIKI(foreign influences on Romanian culture)<search>... The search results returned content related to Slavic influences on Romanian culture, consistent with the description in Answer A. Answer A is more detailed and specific, providing concrete examples such as ... Answer B, while informative, is more structured as a list of examples and lacks the same level of detail and specificity. Therefore, Answer A is more comprehensive and better answers the input question. <answer>A<answer>

**✓ Correct answer: a search tool is invoked to assist judgment. The model eventually identifies that the content of Answer A is more detailed than Answer B.**

Table 7: A showcase comparing the judgments of RM-R1 and OPENRM on a representative case.

## 7 CONCLUSION

In this paper, we have proposed OPENRM, a tool-augmented reward model designed to evaluate knowledge-intensive long-form responses. OPENRM addresses the key limitation of existing reward models, which struggle to reliably assess outputs requiring external grounding and multi-step reasoning. By integrating tool usage with reinforcement learning, OPENRM learns to actively retrieve evidence and make more accurate judgments. Extensive experiments across three newly collected datasets and two widely used benchmarks have demonstrated that OPENRM consistently outperforms strong baselines, achieving superior generalization while requiring fewer training instances. These results highlight the potential of tool-augmented reward modeling to serve as a more faithful and scalable evaluator for knowledge-intensive reasoning tasks.

Despite the promising results, our study still has two limitations. First, the effectiveness of OPENRM relies on the availability and reliability of external tools, which may introduce bias or extra latency in evidence retrieval. Second, our current implementation only focuses on text-based evaluation, and has yet to be extended to multimodal settings involving visual or tabular inputs. Looking forward, we plan to extend our framework to more complex scenarios involving a broader set of external tools, thereby enhancing its applicability to diverse tasks. We also aim to adapt OPENRM to multimodal settings, enabling models to seamlessly integrate textual reasoning with visual and tabular evidence. Such extensions would pave the way for building more general and robust evaluation agents capable of aligning large language models with complex real-world requirements.

## REFERENCES

- Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. Lit-search: A retrieval benchmark for scientific literature search. *arXiv preprint arXiv:2407.18940*, 2024.
- Anthropic Team. Claude Opus 4.1. Anthropic News, <https://www.anthropic.com/news/clause-opus-4-1>, August 2025. Accessed: August 06, 2025.
- Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. Travelagent: An ai assistant for personalized travel planning. *arXiv preprint arXiv:2409.08069*, 2024.
- Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgeirm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025a.
- Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*, 2025b.
- I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. Factool: Factuality detection in generative ai—a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Reward reasoning model. *arXiv preprint arXiv:2505.14674*, 2025.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024a.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024b. URL <https://arxiv.org/abs/2412.05579>.
- Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. Tool-augmented reward modeling. *arXiv preprint arXiv:2310.01045*, 2023.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

- Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024b.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*, 2021.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18990–18998, 2024.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*, 2024a.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. 2024b.
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Eger, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024c.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.
- Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning, 2025. URL <https://arxiv.org/abs/2505.10320>.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.
- Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

## A APPENDIX

### A.1 LLM USAGE

In this work, large language models (LLMs) were not used for research ideation or for generating original scientific content. LLMs were only employed as general-purpose assistive tools for grammar checking and minor wording adjustments during the manuscript preparation process. All conceptual development, experimental design, implementation, analysis, and writing of original content were carried out entirely by the authors.

### A.2 BASELINE IMPLEMENTATION DETAILS

We provide additional details on the baseline reward models evaluated in our experiments.

**Train-based Reward Models.** These baselines are directly trained to predict preference scores or rank candidate responses. **Skywork-Reward** (Liu et al., 2024b) and **JudgeLM** (Chen et al., 2025a) are trained using large-scale human preference datasets such as UltraFeedback and JudgeLM. **RRM** (Guo et al., 2025) adopts a multi-stage preference optimization pipeline with iterative reward model updates. **RM-R1** (Chen et al., 2025b) is trained with reinforcement learning signals derived from rule-based reward functions. For all train-based reward models, we use the 7B versions released and follow their official evaluation setups.

**LLM-as-Judge.** In this category, large language models are prompted to act as judges without external retrieval or verification tools. We select top-ranked models from the *Arena Leaderboard*, including **DeepSeek-V3.1**, **GPT-4o**, **Gemini-2.5-Pro** and **Claude-Opus-4**. Each model receives the same pairwise comparison prompt as OPENRM, and outputs are parsed into binary preferences(A or B). Temperature is set to 0.3 for deterministic judgment generation.

**Agentic Reward Models.** These methods extend the LLM-as-Judge paradigm by incorporating external tools. We adopt **FacTool** (Chern et al., 2023), a popular toolkit for tool-augmented fact-checking and evidence retrieval. Specifically, the model first retrieves relevant documents using external tools like WikiSearch and Arxivsearch, then re-evaluates candidate responses conditioned on the retrieved evidence. We test this approach with top-ranked models mentioned in LLM-as-Judge.

**Evaluation Protocol.** All baselines are evaluated on the same preference alignment datasets as OPENRM, using accuracy as the main metric (i.e. whether the predicted preference matches the ground-truth annotation). Each judgment is computed independently without majority voting, and the models are evaluated under identical sampling configurations.

### A.3 IMPLEMENTATION DETAILS

#### System Prompt Template of Our Reward Model

You are an impartial judge tasked with evaluating two candidate responses and determining which one better answers the input question with higher quality. Your evaluation should focus on factual correctness, clarity, completeness, and helpfulness.

In judging the response, you must start **\*\*concise\*\*** reasoning inside `<think>` and `</think>`. Whenever you want to ground your judgement with external evidence, you are ENCOURAGED to pause the reasoning, and then activtely **\*\*search for reference\*\*** from relevant corpora inside "`<search>`" and "`</search>`". The search results will be enclosed and returned in `<information>` and `</information>`. After obtaining the results, you can continue your reasoning.

Note: For research-oriented questions, search for relevant scientific papers on arXiv with <search>ARXIV("YOUR QUERY")</search>; for everyday open-domain questions, search Wikipedia with <search>WIKI("YOUR QUERY")</search>. Since retrieval is commonly imperfect, you should always try more DIFFERENT ISSUES when failing to find useful information. But if you still fail after 2 attempts, you should terminate the search process and continue to reason the final judgment.

Starting below, you should interleave "<think> </think>" and "<search> </search>" multiple times to progressively gather relevant evidence in a step-by-step manner. After gathering enough evidence, you must conclude your final judgment inside <think> and </think>, followed by your final choice (ONLY 'A' or 'B') inside <answer> and </answer>.

### System Prompt of Our Reward Model

Please help me Judge which of the two candidate responses better answers the question: {question}

Answer A: {answer a}

Answer B: {answer b}

NOTE: Using relevant evidence from the search results to support your judgment, focusing on \*\*factual correctness\*\*, \*\*clarity\*\*, \*\*completeness\*\*, and \*\*helpfulness\*\*. Trivial factors such as the order of the responses, their length, or writing style should not influence your decision. Do not infer polarity from position: neither the first response nor the second is inherently positive or negative.

### Input Question and Two Candidate Prompts

Question: How does the use of the Balanced Winnow classifier, in conjunction with a novel feature set combining lemmatized words, stylometric features, and world knowledge features, address the specific challenges of a fine-grained, multi-label genre classification task for French literary texts, as opposed to more traditional classifiers and broader genre definitions used in foundational studies like those by Kessler et al. (1997)?

Answer A: The provided information does not discuss the Balanced Winnow classifier, genre classification, or French literary texts. Therefore, I will address the question based on my knowledge.

The use of the Balanced Winnow classifier, in conjunction with a novel feature set combining lemmatized words, stylometric features, and world knowledge features, addresses the specific challenges of fine-grained, multi-label genre classification for French literary texts by leveraging its ability to handle high-dimensional feature spaces and its effectiveness in binary and multi-label classification tasks. Unlike traditional classifiers such as Naive Bayes or decision trees, which may struggle with the complexity and nuance required for fine-grained distinctions, balanced Winnow - a multiplicative weight-update algorithm - excels in environments where feature relevance is sparse but

critical, making it well-suited for literary analysis where discriminative features might be subtle and varied. The novel feature set enhances this capability: lemmatized words help normalize morphological variations, capturing core lexical patterns without noise from inflectional changes; stylometric features (e.g., sentence length, lexical richness, syntactic structures) quantify authorial and stylistic nuances that are pivotal in distinguishing literary genres; and world knowledge features (e.g., historical context, thematic elements derived from external databases) incorporate semantic and cultural depth, addressing limitations of purely statistical approaches. This combination allows the model to capture both surface-level patterns and deeper semantic cues, which is essential for multi-label tasks where texts often exhibit hybrid genre characteristics.

In contrast, foundational studies like Kessler et al. (1997) often relied on broader genre definitions (e.g., fiction vs. non-fiction) and simpler feature sets (e.g., raw word frequencies) with classifiers like Naive Bayes, which prioritize probabilistic coherence over nuanced feature interactions. These approaches are less adaptable to fine-grained categories (e.g., subgenres like "symbolist poetry" or "existentialist novel") and multi-label scenarios, as they may oversimplify linguistic and extralinguistic cues. The Balanced Winnow classifier, with its iterative weight adjustment and emphasis on informative features, reduces error propagation in complex label sets and improves precision in identifying overlapping genres, thus offering a more robust framework for the intricacies of French literary texts.

**Answer B:** The DEFT 2014 challenge presented a particularly nuanced genre classification task: fine-grained categorization of French literary texts, where documents could belong to multiple genres simultaneously. This differs significantly from the broader genre definitions used in foundational studies, which often distinguished between fundamentally different text types (e.g., prose vs. poetry, written vs. spoken language). The task here required distinguishing between subtler distinctions within the same literary domain, making it both multi-label and fine-grained. The choice of the Balanced Winnow classifier, combined with a novel feature set, directly addresses these specific challenges. While traditional classifiers might struggle with the high dimensionality and feature sparsity common in text classification, Balanced Winnow is well-suited for such scenarios. It is an online linear classifier related to the Perceptron, but designed to handle a large number of features efficiently—a critical capability when combining diverse feature types. The novel feature set itself is key to capturing the fine-grained nuances of literary genres. Lemmatized words help standardize the vocabulary to capture core thematic content (e.g., recognizing that "crime" and "crimes" convey the same theme relevant to a police novel). Stylometric features are essential for detecting structural differences, such as the varied use of punctuation or sentence length that might distinguish poetry from prose or even one type of prose from another. World knowledge features incorporate a deeper understanding beyond surface-level text, potentially capturing context or connotations that are genre-specific.

By combining these features, the approach captures both content (lemmatized words, world knowledge) and style (stylometric features), which is necessary because literary genres often depend on a blend of these elements. For instance, two genres might share similar themes but differ significantly in stylistic

presentation. The multi-label aspect is addressed through the classifier's ability to manage multiple, potentially overlapping categories by not forcing mutually exclusive decisions.

In contrast, the broader genre definitions in earlier work might have relied on more obvious distinguishing features, making simpler classifiers or single feature types sufficient. The fine-grained, multi-label nature of the DEFT task demands a more sophisticated combination of feature engineering and a robust classifier like Balanced Winnow to achieve high precision without sacrificing computational efficiency.