THE CROWD ALREADY VOTES: LEARNING CREATIVE-WRITING REWARDS FROM HUMAN INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Creative writing aims to generate original and engaging literary texts while maintaining long-range coherence, characterization, and expressive language. Despite their strong text-generation capabilities, large language models (LLMs) still struggle to produce writing that receives high evaluations from human audiences. A central obstacle is that writing quality is subjective and real-world judgments emerge from collective rather than individual preferences. We introduce REALFEEDBACK, a data-construction paradigm that treats online communities as continuously operating sources of preference supervision. It converts naturally occurring human interactions into reward signals without new large-scale annotation. We instantiate REALFEEDBACK in two domains: inline critiques from Critique Circle provide passage-level process labels for story generation, while nested discussions from Baidu Tieba provide outcome-level preference pairs for forum responses. We use these data to train REALFEEDBACK-PRM and REALFEEDBACK-ORM. Across established benchmarks and REALFEEDBACK-BENCH, REALFEEDBACK-REWARD achieves state-of-the-art performance and outperforms much larger models. We also apply REALFEEDBACK-REWARD to Best-of-N test-time scaling and find that its selections receive higher human preference scores. These results position online communities as sustainable infrastructure for training reward models that better reflect collective human judgment.

1 INTRODUCTION

The task of creative writing aims to generate original and engaging literary texts, requiring models to maintain long-range narrative coherence, consistent characterization, compelling plot development, and expressive language (Kreminski & Martens, 2022). Unlike domains such as mathematics and coding, where outputs can often be verified against reference solutions or executable specifications (Cobbe et al., 2021; Chen et al., 2021; Su et al., 2025), creative writing does not have a single correct realization. The quality of a story or poem depends jointly on factors such as plot, pacing, characterization, writing style, and the subjective experience and taste of individual readers (Xia et al., 2026; Paech, 2024). Consequently, evaluating and optimizing creative writing remains particularly challenging, as open-ended narratives lack objective ground truth. For example, existing Large Language Models (LLMs) and Reward Models (RMs) often struggle to fully capture human preferences when it comes to high quality creative writing (Xia et al., 2026; Fein et al., 2025).

To address this gap, a common paradigm for improving LLM capabilities is reinforcement learning from human feedback (RLHF; Ouyang et al., 2022), where a reward model (RM) is first learned from human preference annotations and subsequently used to optimize the language model. The success of RLHF in domains such as mathematical reasoning and code generation is closely related to the availability of reliable feedback signals, where correctness can often be objectively verified. However, when it comes to creative writing, its existing reward modeling pipelines face a fundamental dilemma. LLM-as-a-Judge (LAAJ) (Zheng et al., 2023) approaches provide inexpensive and scalable supervision, but they may capture specific biases rather than genuine human preferences (Dubois et al., 2025). In contrast, human annotations provide direct access to human judgments but are expensive, time-consuming, and difficult to scale. Ideally, the most faithful reward signal for creative writing would come from large-scale collective judgments of real readers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, incorporating such a human-in-the-loop evaluation process directly into model training is impractical.

Our key observation is that this supervision already exists in the wild. Online communities continuously collect human responses to human-authored texts: writers comment on passages, readers reply to stories, and discussions naturally emerge around creative works. Rather than treating these interactions as merely social data, we view online communities as continuously operating infrastructures for large-scale preference supervision. By accurately quantifying these naturally occurring interaction behaviors, we can transform them into high-quality reward signals for improving creative writing models.

Specifically, we introduce REALFEEDBACK, a framework that learns reward models from human reactions arising from ordinary human–human interactions. The central question is therefore not how to discover another source of feedback, but how to systematically transform heterogeneous social interactions into reward supervision.

We instantiate REALFEEDBACK at two resolutions and in two communities. For story generation, our training pool contains 12,281 stories comprising 368,230 segments from Critique Circle (CC), an English-language writing community; a separate set of 1,500 stories is reserved for the story track of REALFEEDBACK-BENCH. We quantify peer writers' inline critiques of individual passages as sentiment-based process rewards to train a *process reward model* (PRM) (Lightman et al., 2023) REALFEEDBACK-PRM. For open-ended forum discussion, we collect discussions from Baidu Tieba, a large Chinese online forum, and quantify how strongly real replies engage with each parent floor. The complete Tieba pool contains 209,223 candidate pairs; confidence filtering retains 143,402 pairs for training the *outcome reward model* (ORM) REALFEEDBACK-ORM. The resulting training supervision consists of 368,230 segment-level process annotations and 143,402 outcome-level pairs, each grounded in authentic human-to-human feedback. To validate this interpretation, we conduct an additional human audit on 500 Tieba candidate pairs. The community-derived preference agrees with the majority human judgment in 80.65% of the pairs, providing direct evidence that the interaction quantifier tracks human preference.

To evaluate REALFEEDBACK-REWARD, we construct REALFEEDBACK-BENCH, which has two parts: a *story* part of 1,500 Critique Circle stories held out from training, and a *forum* part of 1,500 Tieba threads held out from training, each built with the same procedure as the corresponding training data. We evaluate REALFEEDBACK-PRM on StoryAlign (Xia et al., 2026), LitBench (Fein et al., 2025), and the story part of REALFEEDBACK-BENCH (RF-Bench). We evaluate REALFEEDBACK-ORM on a Reddit dataset and the forum part of REALFEEDBACK-BENCH, the former drawn from one of the largest English-language forum communities. Our models outperform all existing reward models on most metrics. We further apply REALFEEDBACK-REWARD in test-time scaling, performing Best-of-N (BoN) experiments in which reward models select the better story; stories selected by REALFEEDBACK substantially outperform those from other reward models, further validating its effectiveness in real-world applications. Finally, our analysis reveals that simply having behavioral data is not sufficient: raw reply volume is no better than chance, whereas a simple interaction-aware quantification is strongly predictive. The benefit therefore comes from a reasonable quantification of real feedback, not from treating every observable interaction as a preference.

Our contributions are threefold:

(1) We formulate online communities as continuously operating infrastructures for creative writing preference supervision and introduce a general paradigm for extracting reward supervision from naturally occurring human–human interaction.

(2) We build a large-scale, high-quality training dataset that captures real-world human preferences for both creative writing and open-ended discussion.

(3) We propose REALFEEDBACK-REWARD, an advanced reward model for creative writing preference that achieves state-of-the-art (SoTA) performance and validate its effectiveness in downstream test-time scaling.

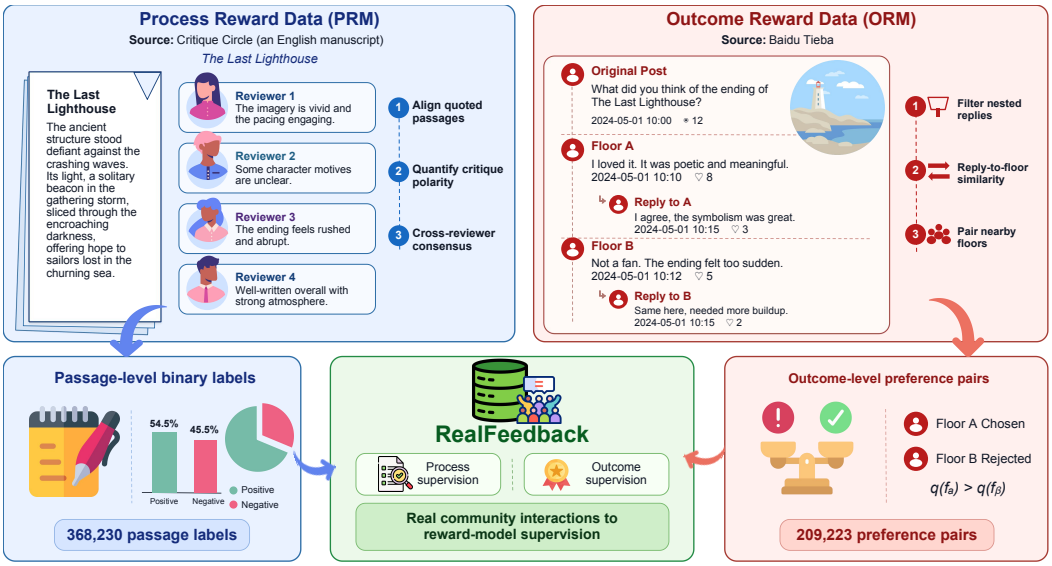

Figure 1: Overview of REALFEEDBACK construction. Critique Circle critiques are aligned and aggregated into passage-level process labels, while Baidu Tieba nested replies are quantified and paired into outcome-level preferences. The two supervision pools form REALFEEDBACK.

Pool	Lang.	Unit	Reward	Scale
CC	EN	Seg.	PRM	368,230
Tieba	ZH	Pair	ORM	209,223

Table 1: Overview of the complete RealFeedback supervision pools. The corresponding quantifiers are peer-critique polarity for Critique Circle and reply-to-floor similarity for Tieba.

2 DATA CONSTRUCTION

This section introduce how we convert interpersonal feedback that arises naturally in online communities into supervision directly usable by reward models. Although such feedback is originally produced for human–human communication, it also records how people respond to a piece of text. Given a text unit x, we collect its naturally occurring feedback $F(x)$ and apply a lightweight quantifier Q matched to the interaction structure of the source community. This abstraction preserves a common source of supervision, real interpersonal feedback, while allowing different communities to produce rewards at different granularities and in different forms.

We instantiate this construction in two communities. Inline critiques in Critique Circle are naturally aligned with story passages, yielding dense passage-level binary process labels. Nested replies in Baidu Tieba surround complete floor posts, so we first aggregate them into floor scores and then form outcome-level preference pairs from temporally proximate floors in the same thread. Figure 1 illustrates the two construction pipelines, and Table 1 summarizes the two complete supervision pools. The training section describes the filtering and data splits used for model optimization.

The quantifiers are used only to construct supervision labels. During both training and inference, the reward models receive only the text to be scored; critique text, nested replies, and interaction metadata are excluded.

2.1 CRITIQUE CIRCLE: PASSAGE-LEVEL PROCESS FEEDBACK

Critique Circle is an English writing community in which authors submit drafts and other writers provide inline critiques and closing comments. Each inline critique contains both a quotation from the draft and a remark on that quotation, giving a local alignment between text and feedback without additional annotation. We collect 14,140 raw story records. We first sample and lock exactly 1,500

unique stories with random seed 42 for the story track of REALFEEDBACK-BENCH. These stories are removed from subsequent deduplication and label filtering and are excluded from model training and training-time evaluation. We then deduplicate the remaining records by story identifier and remove records with no valid remark, leaving 12,281 stories in the complete training supervision pool.

Different reviewers may quote the same passage with slightly different boundaries. To recover a consistent segment sequence, we use the critique with the largest number of inline quotations as the ordered backbone of each story. We lowercase and normalize every quotation, then align a quotation from another reviewer to a backbone segment when their token-set Jaccard similarity (Jaccard, 1912) is at least 0.5. This preserves passage order while grouping feedback from multiple reviewers on the same content.

For segment i and reviewer j, we average the VADER compound sentiment scores (Hutto & Gilbert, 2014) of all aligned remarks from that reviewer to obtain v_{ij}. To identify constructive criticism with a positive lead-in, we use a set of critical cues, including *but*, *however*, *suggest*, *should*, *confusing*, and *needs work*. Let $I_{ij} = 1$ if any aligned remark contains a cue and $I_{ij} = 0$ otherwise. The adjusted feedback score is

$$a_{ij} = v_{ij} - 0.5 I_{ij}. \tag{1}$$

We binarize each reviewer's adjusted score at zero (nonnegative is positive) and assign the segment label by majority vote across reviewers. This produces 368,230 passage labels, comprising 200,592 positive and 167,638 negative instances. We describe the story-level training split in training section.

2.2 TIEBA: RESPONSE-LEVEL OUTCOME FEEDBACK

Baidu Tieba is one of the largest forum community around the world, each discussion of which contains an original post, floor-level contributions, and nested replies beneath each floor. For outcome supervision, the context is the thread title and original post content, and the two candidate responses are floors from the same thread. Whether a floor elicits discussion grounded in its content is an outcome observed in the real community. Instead of using reply volume alone, we quantify this outcome from the reply text itself.

Before scoring, we remove empty or invalid nested replies and those flagged by a toxicity detector. Let R_f be the remaining replies to floor f, and let E denote a Chinese BGE v1.5 text encoder. We define the floor's discussion-grounding score as

$$q(f) = \frac{1}{|R_f|} \sum_{r \in R_f} \cos(E(r), E(f)) \,. \tag{2}$$

Within each thread, we form candidate pairs (f_A, f_B) from floors posted close in time. We mark the floor with the larger q as chosen and the other as rejected, and define $\Delta = |q(f_A) - q(f_B)|$. The complete supervision pool contains 209,223 candidate pairs.

3 REWARD MODEL TRAINING

The last section defines the complete supervision pools derived from community interactions. This section describes the confidence filtering, data splits, and input formats used to train REALFEEDBACK-REWARD.

Process-reward data. Segments from Critique Circle share story-level context, so randomly splitting individual segments would leak content across sets. We reserve the 1,500 stories in the story track of REALFEEDBACK-BENCH before constructing the training pool. We then split the remaining 12,281 stories with random seed 42, assigning 11,053 stories to training and 1,228 to training evaluation. These sets contain 331,219 and 37,011 passage labels, respectively. No story identifier appears in more than one of the training, training-evaluation, and REALFEEDBACK-BENCH story pools.

Each example contains the story title, genre, and m passages in their original order. Its ground truth is an aligned binary sequence $(y_1, \ldots, y_m)$, where each y_i indicates whether passage i received positive or negative feedback. The model therefore predicts one label for each passage.

Outcome-reward data. Using the score margin Δ defined in section of data construction, we stratify the complete Tieba pool as follows: S0 contains $\Delta < 0.050$; S1, $0.050 \leq \Delta < 0.110$; S2, $0.110 \leq \Delta < 0.216$; and S3, $\Delta \geq 0.216$. The four strata contain 17,905, 47,916, 95,514, and 47,888 pairs, respectively. Strata measure only the separation between the two community-derived scores and do not alter the preference direction; larger Δ indicates a more clearly separated pair.

For training, we retain S2 and S3 pairs with $\Delta \geq 0.110$, excluding the near tie S0 and S1. It retains 143,402 pairs, or 68.5% of the complete pool. The selected data span 140 Tieba forums and 22,305 threads.

Each ORM example is represented as a "context,chosen,rejected" triple. The context c concatenates the thread title and original-post content, using the title alone when the post is empty. The candidates r^+ and r^- are the higher- and lower-q floors from the same thread. The model receives the two context–response inputs (c, r^+) and (c, r^-), with the ground-truth preference $r^+ \succ r^-$. We discard empty floors, identical candidates, and invalid preference directions during conversion.

We then reserve 2,000 for optimization validation, and use the remaining 141,402 for training. Interaction scores, forum names, thread identifiers, and floor identifiers are retained only as construction metadata and are excluded from the model's text input.

3.1 Implementation Details

We initialize REALFEEDBACK-PRM and REALFEEDBACK-ORM from Llama-3.1-8B-Instruct (Meta, 2024b) and Qwen3-8B (Yang et al., 2025), respectively. We train REALFEEDBACK-PRM on story-passage sequences with aligned binary labels, using a maximum input length of 5,140 tokens, and train REALFEEDBACK-ORM on Tieba "context,chosen,rejected" preference pairs. Both models use a batch size of 64 and are trained for one epoch. We set the learning rate to 1×10^{-6} for REALFEEDBACK-PRM and 5×10^{-6} for REALFEEDBACK-ORM.

4 Experiments

4.1 Experimental Setup

We evaluate REALFEEDBACK-REWARD through REALFEEDBACK-ORM (outcome-level preferences) and REALFEEDBACK-PRM (passage- and story-level preferences) on REALFEEDBACK-BENCH and three external benchmarks, reporting accuracy throughout.

REALFEEDBACK-BENCH tests whether models recover preferences expressed through real community interactions. We construct both tracks before model training and keep them disjoint from the corresponding training data.

The *story track* consists of exactly the 1,500 unique Critique Circle stories sampled and locked before training-pool deduplication and label filtering. They are not subsequently screened and are disjoint from the 12,281-story training pool, which is split into training and training-evaluation sets. The *forum track* contains 1,500 pairwise Tieba instances drawn from forums outside the 140 forums used for ORM training. No source record in either track is used for model optimization.

Human validation of interaction labels. To test whether the interaction quantifier tracks human preference, we conduct a separate audit on 500 Tieba candidate pairs. Human annotators select the better floor given the thread context. The community-derived preference agrees with the majority human judgment on 80.65% of the pairs. This result supports the use of reply-to-floor similarity as a scalable proxy for human preference in forum discussions; the audit pairs are not used to train REALFEEDBACK-ORM.

Metrics. On the story track of REALFEEDBACK-BENCH, REALFEEDBACK-PRM predicts a binary label for each passage, and we report classification accuracy over all passages. For LitBench and StoryRMB, we split each candidate story at paragraph boundaries; a single-paragraph story forms one passage. For complete-story ranking, we use the length-normalized mean of passage-

Table 2: Pairwise accuracy (%) on Reddit and forum track of REALFEEDBACK-BENCH. **Bold**: best; underline: second best.

Model	Reddit↑	RF-Bench↑
Reward models		
InternLM2-20B-Reward	63.2	58.7
Skywork-Reward-Gemma-2-27B-v0.2	66.8	62.4
ArmoRM-Llama3-8B-v0.1	67.5	63.1
QRM-Llama3.1-8B-v2	61.0	57.2
GRM-Llama3.1-8B-rewardmodel-ft	58.4	55.6
WQRM	64.7	60.3
Closed-source LLM judges		
GPT-5.5	**75.4**	67.7
Claude-Sonnet-4	71.2	66.8
Gemini-2.5-Pro	70.5	65.9
REALFEEDBACK-ORM	74.6	**71.9**

level positive probabilities. Given a story $s = (s_1, \ldots, s_m)$ under context c, its score is

$$R_{\text{PRM}}(s \mid c) = \frac{1}{m} \sum_{i=1}^{m} p_\theta(y_i = + \mid c, s_{1:m}). \tag{3}$$

The mean removes the direct dependence of the score on the number of passages. REALFEEDBACK-ORM and the general reward models directly score each context–candidate input.

For pairwise tasks, the higher-scored candidate is correct when it matches the preference label; on StoryRMB, the highest-scored story is selected from four.

Baselines. We compare six general reward models—InternLM2-Reward (Cai et al., 2024), Skywork-Reward (Liu et al., 2024), ArmoRM (Wang et al., 2024), QRM (Dorka, 2024), GRM (Yang et al., 2024), and WQRM (Chakrabarty et al., 2025)—and three closed-source judges: GPT-5.5 (OpenAI, 2026), Claude-Sonnet-4 (Claude, 2025), and Gemini-2.5-Pro (Comanici, 2025). Reward models score candidates independently; judges compare them directly. Prompts are in Appendix D.

ORM evaluation. We evaluate REALFEEDBACK-ORM on the forum track of REALFEEDBACK-BENCH and the public `reddit-dataset` (Qiu, 2016), a corpus of Reddit threads, comments, and community votes. Each evaluation instance contains a thread post, two comments, and a preference label. For each comment r, we compute $v(r) = \text{ups}(r) - \text{downs}(r)$. We randomly sample 1,500 threads containing at least two comments with different net votes and sample one such comment pair per thread; the higher-v comment is chosen. Reddit measures cross-platform transfer, while the forum track of REALFEEDBACK-BENCH measures transfer to Tieba communities excluded from training.

PRM evaluation. We evaluate REALFEEDBACK-PRM on the story track of REALFEEDBACK-BENCH and two external story benchmarks. LitBench (Fein et al., 2025) is a creative-writing preference benchmark built from Reddit's *WritingPrompts* community. Because its test instances are not included in the public release, we sample 1,500 preference pairs from the released training dataset. StoryRMB (Xia et al., 2026) is a human-verified benchmark for ranking complete stories, which contains 1,133 instances, each consisting of a writing prompt, one chosen story, and three rejected stories.

4.2 RESULTS

Tables 2 and 3 summarize all experimental results. We make three observations.

Table 3: Accuracy (%) on story track of REALFEEDBACK-BENCH, LitBench pairwise preference, and StoryRMB four-way selection (25% random baseline). **Bold**: best; underline: second best.

Model	RF-Bench↑	LitBench↑	StoryRMB↑
Reward models			
InternLM2-20B-Reward	60.5	62.1	21.3
Skywork-Reward-Gemma-2-27B-v0.2	64.2	65.7	32.7
ArmoRM-Llama3-8B-v0.1	65.8	67.3	49.7
QRM-Llama3.1-8B-v2	59.4	61.0	14.9
GRM-Llama3.1-8B-rewardmodel-ft	57.1	58.6	31.1
WQRM	62.9	64.5	24.0
Closed-source LLM judges			
GPT-5.5	68.4	72.1	70.2
Claude-Sonnet-4	69.2	73.4	69.0
Gemini-2.5-Pro	68.9	72.8	62.8
REALFEEDBACK-PRM	**72.4**	**75.8**	**71.6**

(1) General reward models remain limited: their best scores are 67.5% and 63.1% on Reddit and REALFEEDBACK-BENCH, and 65.8%, 67.3%, and 49.7% on the three story settings. Closed-source judges improve over these baselines but still leave substantial room for improvement.

(2) Our models outperform every reward-model baseline in all five settings and achieve the best overall result in four. The only exception is Reddit, where GPT-5.5 exceeds REALFEEDBACK-ORM by 0.8 points, possibly because of greater familiarity with Reddit-style content. These gains support supervision derived from community interactions.

(3) REALFEEDBACK-ORM is the strongest reward model on Reddit and REALFEEDBACK-BENCH, while REALFEEDBACK-PRM leads on REALFEEDBACK-BENCH, LitBench, and StoryRMB. Thus outcome supervision transfers across communities and passage-level supervision supports complete story ranking, providing a scalable path for expanding preference supervision.

5 ANALYSIS

5.1 TEST-TIME SCALING WITH BEST-OF-N SAMPLING

Best-of-N (BoN) sampling generates multiple candidates and selects one with a reward model (Hao et al., 2025; Peng et al., 2025). For forum and story tasks, we retain 150 held-out prompts, generate 16 candidates per prompt with Llama-3.1-70B-Instruct (Meta, 2024a) at temperature 1, and evaluate $N \in \{1, 2, 4, 8, 16\}$ using nested candidate pools. We compare ArmoRM, Skywork, and RealFeedback. Annotators upvote candidates they prefer; each candidate receives at least 20 independent votes, and we report the mean upvotes of each model's selections.

Figure 2 shows that REALFEEDBACK-REWARD consistently selects candidates receiving more human upvotes than the other reward models. The 8B models therefore provide an effective and relatively efficient signal for test-time selection.

5.2 ERROR ANALYSIS

We group the mistakes found during evaluation into common error types. Appendix G gives examples for the ORM and PRM.

REALFEEDBACK-ORM. REALFEEDBACK-ORM's errors occur when the interaction score does not match response quality. Similarity captures whether replies stay on topic, but jokes, repetition, and side topics can also receive high scores. Conversely, corrections, additional information, and empathy may elicit differently worded follow-ups. Errors are more common when the original post is very short or contains only a title.

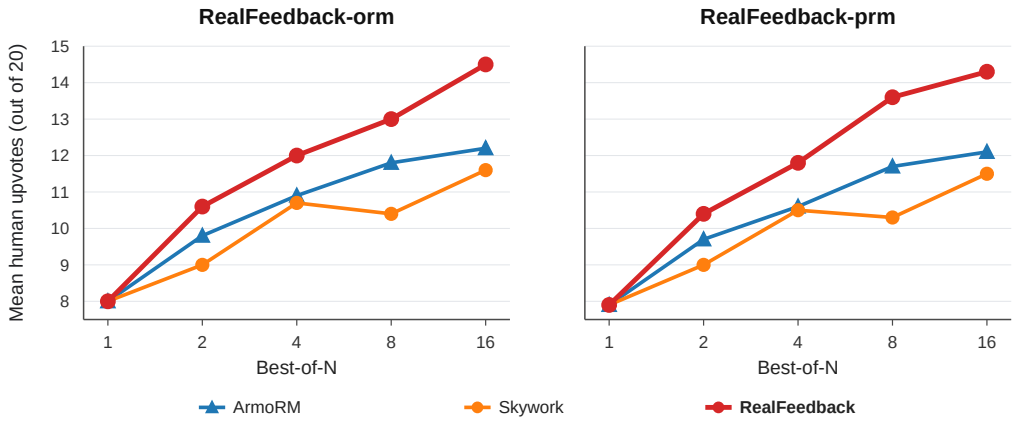

Figure 2: Best-of-N generation results for the ORM and PRM. Each point shows the mean number of upvotes received by the selected candidate, computed from 20 valid human ratings. Higher is better.

Table 4: ORM ablation on real-feedback quantification. Each row trains an ORM with preference labels derived from the listed signal and reports accuracy (%) on the same benchmarks used in the section of training. **Bold** indicates the best result.

Training Signal	Reddit↑	RF-Bench↑
Reply count	54.2	50.9
Aggregate reply sentiment	64.1	62.4
REALFEEDBACK-ORM	**74.6**	**71.9**

REALFEEDBACK-PRM REALFEEDBACK-PRM's errors are concentrated in passages on which reviewers disagree. Some reviewers focus on plot, characters, or emotional impact, while others focus on prose and writing problems, so the same passage can receive both positive and negative comments. Majority voting keeps only one binary label and cannot represent these different judgments.

5.3 ABLATION STUDY

Quantification of Real Feedback. As for REALFEEDBACK-ORM, we construct preference pairs using reply count, aggregate reply sentiment, and reply-to-floor semantic similarity. We evaluate every model on the Reddit and forum track of REALFEEDBACK-BENCH. As shown in Table 4, REALFEEDBACK-ORM outperforming the models trained with reply count or reply sentiment on both tests. The results show that the labels extracted from community interactions directly affect model performance.

Cross-Reviewer Consensus (PRM). We keep the story split, model architecture, and training settings fixed and compare PRMs trained with single reviewer labels and cross reviewer majority vote labels. We evaluate both models on the story track of REALFEEDBACK-BENCH, LitBench, and StoryRMB. As shown in Table 5, REALFEEDBACK-PRM also outperforms reward model trained with single reviewer data.

6 RELATED WORK

6.1 CREATIVE WRITING EVALUATION

Evaluating stories is difficult because surface quality, discourse coherence, originality, and reader appeal need not move together. Early work therefore focused on diagnosing the limitations of au-

Table 5: PRM ablation on cross-reviewer consensus. We report accuracy (%) on the same benchmarks used in Section 4.1. **Bold** indicates the best result.

Label Source	RF-Bench↑	LitBench↑	StoryRMB↑
Single reviewer	66.2	70.1	46.3
REALFEEDBACK-PRM	**72.4**	**75.8**	**71.6**

tomatic metrics. OpenMEVA shows that commonly used story generation metrics correlate weakly with human judgments and often fail tests of discourse coherence, causal reasoning, and robustness (Guan et al., 2021). More recent studies use prompted LLMs as evaluators. They improve over conventional metrics at the system level, but their explanations and reliability remain limited (Chhun et al., 2024). The same gap appears in feedback generation: LLMs can produce plausible comments on a story while still missing its most important flaw or choosing the wrong balance between praise and criticism (Rashkin et al., 2025).

A small number of studies extract a particular signal from a particular creative community. LitBench, for example, converts relative upvotes on complete stories from Reddit's r/WritingPrompts into pairwise preferences (Fein et al., 2025). This is a platform-specific instantiation of community online supervision: one scalar reaction and one outcome level reward. RealFeedback advances a more general view of online communities as preference supervision infrastructure. The problem shifts from discovering a platform to systematically handling heterogeneous human-human interactions. We extract both paragraph level process rewards and response level outcome rewards from the content and structure of textual replies, and test the resulting rewards in downstream Best-of-N generation.

6.2 HUMAN AND AI FEEDBACK FOR REWARD MODELING

Modern preference learning typically collects comparisons specifically for model development, fits a reward model to those comparisons, and optimizes or ranks model outputs with the learned reward. This pipeline has been established for summarization (Stiennon et al., 2022) and instruction following (Ouyang et al., 2022). Human comparisons provide direct preference supervision, but repeated annotation is expensive. Reinforcement learning from AI feedback replaces some of this labor with generated critiques or preferences (Bai et al., 2022); LLM-as-a-judge is similarly attractive because it is inexpensive and scalable (Zheng et al., 2023). However, these approaches trade annotation cost for dependence on another model and its biases. RealFeedback instead reuses human feedback produced during the original community interaction, so no generative model or LLM judge supplies the training preference and no new labeling campaign is required.

Reward supervision also differs in granularity. Outcome reward models assign a value to a completed response, whereas process reward models supervise intermediate steps. Process supervision has been studied most extensively in mathematical reasoning, where correctness can be assessed step by step and has been shown to outperform outcome-only supervision (Lightman et al., 2023).

7 CONCLUSION

In this paper, we presented REALFEEDBACK, a new paradigm for constructing reward-model supervision from naturally occurring human interactions in online communities. We instantiated it in two domains: Critique Circle's critiques were converted into passage-level process labels for story generation, while Baidu Tieba discussions were converted into outcome-level preference pairs for forum responses. We used these data to train REALFEEDBACK-PRM and REALFEEDBACK-ORM and evaluated them on established benchmarks and REALFEEDBACK-BENCH. REALFEEDBACK-REWARD achieves state-of-the-art performance across all evaluations, and Best-of-N experiments further show that it selects outputs receiving higher human preference scores. These results demonstrate that online communities can serve as sustainable sources of preference supervision. We hope this perspective supports the development of reward models that better reflect human preferences and advance human-like creative writing.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language editing, to help prepare and revise scientific figures, and to generate the synthetic training data described in Sections 2 and 3. We did not delegate experimental design, data analysis, or scientific conclusions to generative AI systems. The authors reviewed all AI-assisted material, verified the technical content and experimental results, and take responsibility for the final text, claims, data, and artifacts.

ETHICS STATEMENT

(1) Intellectual Property. Our study uses stories collected from Critique Circle and Baidu Tieba. For Critique Circle, data collection complied with the platform's copyright policy and terms of use. For Baidu Tieba, we collected only publicly accessible content in accordance with the platform's terms of service. All collected data were anonymized before use and do not contain personally identifiable information about either the original authors or annotators. As an additional safeguard, we used GPT-5.5 as an auxiliary screening tool to identify potentially sensitive content (e.g., violent or sexually explicit material), followed by manual verification of randomly sampled stories. We did not identify any sensitive content in the collected dataset. All proprietary LLMs evaluated in this work (e.g., GPT-5.5 and Gemini 2.5 Pro) were accessed through their official paid APIs.

(2) Intended Use. This work investigates reward modeling for creative writing and introduces reward models trained from real-world human feedback. We hope that these resources will facilitate future research on long-form text generation, reward modeling, and creative writing.

(3) Misuse Risks. Although the proposed reward models and benchmark are intended for research purposes, they could potentially be misused to improve the generation of inappropriate or harmful content. We do not endorse such use and encourage researchers and practitioners to deploy these models responsibly and in accordance with applicable laws, regulations, and platform policies.

(4) AI Assistance. The specific uses of generative AI are disclosed in the AI Use Statement above. All technical content, experimental design, analyses, and conclusions were produced and verified by the authors.

(5) Worker Treatment. Details regarding annotator recruitment, compensation, privacy protection, and labor conditions are provided in Appendix F.

REPRODUCIBILITY STATEMENT

The paper documents the main supervision-construction procedures, dataset splits, model input formats, training hyperparameters, benchmark definitions, and scoring rules in Sections 2, 3, and 4. The submitted artifact includes the LaTeX source, tables, and figure files used to prepare the manuscript. Raw community text is not redistributed because of platform and copyright restrictions. The current submission does not include the complete preprocessing, training, and evaluation scripts; consequently, end-to-end reproduction requires reimplementing the documented procedures from the specifications provided in the paper.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024. URL https://arxiv.org/abs/2403.17297.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Ai-slop to ai-polish? aligning language models through edit-based writing rewards and test-time computation. *arXiv preprint arXiv:2504.07532*, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? prompting large language models for automatic story evaluation, 2024. URL https://arxiv.org/abs/2405.13769.

Claude. The claude 3 model family: Opus, sonnet, haiku, 2025. URL https://www.anthropic.com/news/claude-4.

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

Gheorghe Comanici. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Nicolai Dorka. Quantile regression for distributional reward models in rlhf, 2024. URL https://arxiv.org/abs/2409.10164.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL https://arxiv.org/abs/2404.04475.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing, 2025. URL https://arxiv.org/abs/2507.00769.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6394–6407, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.500. URL https://aclanthology.org/2021.acl-long.500/.

Zixu Hao, Jianyu Wei, Tuowei Wang, Minxing Huang, Huiqiang Jiang, Shiqi Jiang, Ting Cao, and Ju Ren. Scaling llm test-time compute with mobile npu on smartphones, 2025. URL https://arxiv.org/abs/2509.23324.

Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 2014. URL https://api.semanticscholar.org/CorpusID:12233345.

P. Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11:37–50, 1912. URL https://api.semanticscholar.org/CorpusID:85574559.

Max Kreminski and Chris Martens. Unmet creativity support needs in computationally supported creative writing. In Ting-Hao 'Kenneth' Huang, Vipul Raheja, Dongyeop Kang, John Joon Young Chung, Daniel Gissin, Mina Lee, and Katy Ilonka Gero (eds.), *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pp. 74–82, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.in2writing-1.11. URL https://aclanthology.org/2022.in2writing-1.11/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

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, 2024. URL https://arxiv.org/abs/2410.18451.

Meta. Llama-3.1-70b-instruct. https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct, 2024a. Accessed: 2025-09-20.

Meta. Llama-3.1-8b-instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024b. Released: July 23, 2024; Accessed: 2025-09-20.

OpenAI. Hello gpt-5.5, 2026. URL https://openai.com/zh-Hans-CN/index/introducing-gpt-5-5/. Accessed: 2025-02-04.

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.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2024. URL https://arxiv.org/abs/2312.06281.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15934–15949, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.775. URL https://aclanthology.org/2025.acl-long.775/.

Linan Qiu. Reddit comment and thread datas. https://github.com/linanqiu/reddit-dataset, 2016. GitHub repository, accessed July 2026.

Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. Help me write a story: Evaluating llms' ability to generate writing feedback, 2025. URL https://arxiv.org/abs/2507.16007.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains, 2025. URL https://arxiv.org/abs/2503.23829.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts, 2024. URL https://arxiv.org/abs/2406.12845.

Haotian Xia, Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. Storyalign: Evaluating and training reward models for story generation, 2026. URL https://arxiv.org/abs/2605.04831.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. In *Advances in Neural Information Processing Systems*, 2024.

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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Table 6: A Critique Circle passage traced from the source text to the majority-vote process label.

Field	Extracted content	Derived signal
Story	*The General* (Literary)	—
Passage	"Oh, dear Thabo," she said, using his first name. "I love you deeply. You are handsome, and you give me a sense of security. But I cannot marry a man who is a gang leader."	Backbone passage
Reviewer 1	"Nice."	Positive
Reviewer 2	"Wow, what a brave young girl! Kudos for her."	Positive
Reviewer 3	"But I cannot marry a man who is a gang leader. This is awkward as dialogue."	Negative
Label	Two positive reactions versus one negative reaction	**Positive**

Table 7: A Tieba outcome pair traced from a thread to the quantifier and preference label.

Field	Extracted content
Thread context	"If Shi and Ye exchanged worlds, could Shi Hao suppress the Dark Turmoil?"
Floor A	A detailed comparison of cultivation environments, abilities, resources, and narrative settings. Nested replies continue the comparison but use several different terms from the floor.
Floor B	A humorous list of imaginary chapter titles ending with the protagonist defeating every opponent. Nested replies repeat and extend the joke around the chapter titles.
Quantification	$q(f_A) = 0.432$, $q(f_B) = 0.547$, and $\Delta = \lvert q(f_B) - q(f_A) \rvert = 0.115$.
Preference	Floor B is chosen and Floor A is rejected because $q(f_B) > q(f_A)$.

APPENDICES

A DATA CONSTRUCTION EXAMPLES

This section gives one Critique Circle example and one Baidu Tieba example to show how raw community interactions become training labels.

A.1 CRITIQUE CIRCLE EXAMPLE

The example below is from the story *The General*. It traces one source passage, three aligned inline critiques, their polarities, and the resulting majority label. The source text and critiques are abridged for space while preserving their original meaning.

The example illustrates the three construction steps: align reviewers' quotes to one passage, quantify each remark's adjusted polarity, and apply a cross-reviewer majority vote. Reviewers may assess the same passage from different perspectives, such as plot and prose.

A.2 TIEBA EXAMPLE

The following example is translated and abridged from the Tieba data. The thread asks whether Shi could end the Dark Turmoil after exchanging worlds with Ye. The two floors were posted close in time and therefore form a candidate pair. We show the floor text, nested discussion, reply-to-floor similarity, and the resulting preference.

Here $q(f)$ measures whether nested replies remain grounded in their parent floor, rather than counting replies or comparing a reply with the original thread post. A concentrated discussion around a joke can therefore obtain a high score even when another floor provides a more complete answer, which is also a source of ORM errors observed during evaluation.

B DATASET STATISTICS

Tables 8 and 9 summarize the pool sizes, label distributions, and training splits. All values match the construction and training descriptions in the main paper.

Table 8: Critique Circle dataset statistics.

Statistic	Count
Raw story records	14,140
Final stories	12,281
Passage labels	368,230
Positive labels	200,592
Negative labels	167,638
Training stories	11,053
Training-time evaluation stories	1,228
Training passage labels	331,219
Training-time evaluation labels	37,011
Final CC test examples	1,500

Table 9: Baidu Tieba dataset statistics.

Statistic	Count
Complete preference pairs	209,223
S_0: $\Delta < 0.050$	17,905
S_1: $0.050 \leq \Delta < 0.110$	47,916
S_2: $0.110 \leq \Delta < 0.216$	95,514
S_3: $\Delta \geq 0.216$	47,888
Selected $S_2 + S_3$ pairs	143,402
Optimization validation pairs	2,000
Training pairs	141,402
Tieba forums	140
Threads	22,305
Final unseen-Tieba test pairs	1,500

C Human Evaluation

For both forum-response and story-generation tasks, we randomly sample 300 held-out inputs and manually retain 150 prompts. Screening removes incomplete contexts, visibly truncated or extremely short texts, prompts that cannot be judged independently, and clearly unsuitable material. We generate 16 candidates for every retained prompt.

Annotators first read the original prompt and then inspect each candidate independently. They upvote a candidate when it matches their overall reading preference; there is no fixed number of upvotes and no requirement to select a single candidate from a group.

A candidate's human score is its number of upvotes; the BoN analysis reports the mean score of the candidate selected by each method.

D Closed-Source LLM Judge Prompts

For pairwise benchmarks, we use the following prompt template. The placeholders `[CONTEXT]`, `[RESPONSE A]`, and `[RESPONSE B]` are replaced with the instance contents, and the judge is required to return only "A" or "B".

Listing 1: Pairwise judge prompt.

```
You are an impartial evaluator of human preferences.

Context:
[CONTEXT]

Response A:
[RESPONSE A]

Response B:
```

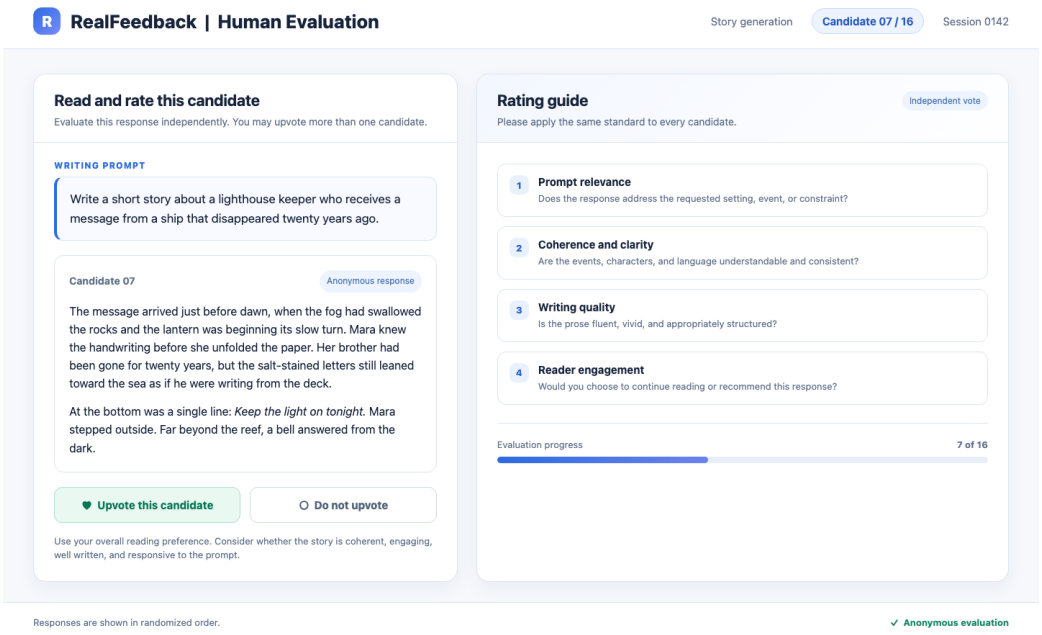

Figure 3: Illustration of the human-evaluation interface for the Best-of-N study. Annotators read the fixed prompt and one candidate at a time, then submit an independent binary upvote decision.

```
[RESPONSE B]

Choose the response that a typical reader would prefer overall. Consider
relevance, coherence, clarity, usefulness, and engagement. Do not explain your
decision. Output exactly one character: A or B.
```

For StoryRMB, we use the four-way template below. The judge returns one candidate index, and the only valid outputs are “A”, “B”, “C”, and “D”.

Listing 2: Four-way story judge prompt.

```
You are an impartial evaluator of creative-writing preferences.

Writing prompt:
[WRITING PROMPT]

Story A:
[STORY A]

Story B:
[STORY B]

Story C:
[STORY C]

Story D:
[STORY D]

Select the story a typical reader would prefer overall. Consider coherence,
prose quality, narrative engagement, and adherence to the prompt. Do not
explain your decision. Output exactly one character: A, B, C, or D.
```

Table 10: Examples of error types identified during evaluation.

Model	Input example	Observed error	Likely source
ORM	Tieba context asks whether Shi Hao could suppress the Dark Turmoil after exchanging worlds with Ye Fan. Floor A gives a detailed comparison; Floor B lists humorous chapter titles.	The quantifier scores A/B as $0.432/0.547$ and prefers B, while human inspection prefers A.	A joke can trigger lexically concentrated replies and a high reply-to-floor similarity even when another floor answers the question more fully.
PRM	In *The General*, Iminathi says she loves Thabo but cannot marry a gang leader. Two reviewers write "Nice" and "what a brave young girl"; another calls the dialogue awkward.	The majority label is positive although one reviewer identifies a real prose problem.	Reviewers emphasize different dimensions, and a binary majority label cannot preserve that disagreement.

For every judge call, candidate order is shuffled randomly and the temperature is set to 0. Accuracy is computed from the returned candidate index.

E From Passage Scores to Story Scores

For complete-story benchmarks, we first split each candidate at paragraph boundaries; a story without paragraph breaks is treated as one passage. For each passage, the PRM reads the writing context and the story prefix ending at that passage, and outputs the probability of the positive class.

Let $p_i = p_\theta(y_i = + \mid c, s_{1:i})$ denote the positive-class probability for passage i in a story with m passages. We compute

$$R_{\mathrm{PRM}}(s \mid c) = \prod_{i=1}^{m} p_i, \tag{4}$$

and select the candidate with the highest story score. The product rewards stories whose passages consistently receive positive process feedback.

F Annotation Coordination

We recruited undergraduate students majoring in English as annotators. The annotator cohort consisted of 58% female and 42% male participants, and all participants had formal training in English reading and writing.

Annotators were compensated according to a predetermined workload-based payment scheme. Formal employment agreements were signed with all annotators, and the process followed applicable local labor regulations. We protected annotator privacy and collected no personally identifiable information beyond what was required for employment.

The total cost of story annotation, outcome-evaluation annotation, and the development and maintenance of the annotation platform was approximately $10,000 USD.

G Error Analysis Examples

Table 10 gives two representative error types found during evaluation. The ORM example shows a similarity signal favoring a joke discussion, while the PRM example shows uncertainty caused by reviewer disagreement.