Verifying the Subjective: Structured Multilingual Rewards for Low-Resource Alignment

Jiu Sha[1], **Mengxiao Zhu**[2,†]
[1] School of Information Engineering and the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China
[2] School of Artifical Intelligence and Computer Science, North China University of Technology

Abstract

Aligning LLMs in low-resource multilingual settings faces a fundamental reward bottleneck: scalar rewards lack cultural generalization, while unstructured critiques remain noisy and unverifiable. To bridge this gap, we introduce a Structured Multilingual Reward Modeling Framework that extends Reinforcement Learning with Verifiable Rewards (RLVR) to subjective and open-ended tasks. The framework unifies three core components to transform abstract quality into concrete supervision: (1) a Structured Checklist Schema decomposing evaluation into granular universal reasoning steps and task-specific criteria; (2) Structured Generative Critique Modeling, which produces rubric-aligned critiques with grounded justifications; and (3) Adaptive Multilingual Reward Optimization, integrating reasoning quality and language consistency into a verifiable objective. We integrate this framework into a bootstrapped Group Relative Policy Optimization pipeline, augmented by length-aware normalization and variance stabilization to ensure stability. Extensive experiments on a newly constructed suite covering 7 subjective task categories across 50 low-resource languages demonstrate that this checklist-driven approach yields substantial improvements in reasoning capability and response quality, particularly in settings where traditional reward models exhibit significant degradation. We publicly release our models and the corresponding evaluation benchmark to facilitate further research. Our code is available at https://github.com/Shajiu/SGCM.

1 Introduction

Reinforcement learning from human or AI feedback has emerged as the dominant paradigm for aligning Large Language Models (LLMs) (Luong et al., 2024; Lambert et al., 2024). However, its effectiveness is strictly constrained by reward model reliability, a limitation pronounced in low-resource multilingual settings (Zhou et al., 2025). Specifically, scalar reward models (Liu et al., 2024) collapse complex criteria into opaque numerical scores that generalize poorly across cultures, while unstructured generative critiques (Zhang et al., 2024) are often noisy, inconsistent, and difficult to verify. Furthermore, while rule-based verification excels in objective domains like math, it fails to generalize to subjective tasks, and existing LLM-as-a-judge approaches (Desmond et al., 2025) frequently suffer from hallucinations and cultural biases in low-resource contexts (Prabhudesai et al., 2025; Zhang et al., 2025a).

To address these limitations, we introduce a Structured Multilingual Reward Modeling Framework that synergistically combines the verifiability of rule-based systems with the flexibility of model-based judges. Our approach extends Reinforcement Learning with Verifiable Rewards (RLVR) to subjective and open-ended tasks, enabling reliable reward modeling in low-resource multilingual settings. The framework unifies three core components to transform abstract quality into concrete supervision: (1) A Structured Checklist Schema that decomposes subjective evaluation into granular, universal reasoning steps and task-specific answer criteria. (2) Structured Generative Critique Modeling (SGCM), which generates rubric-aligned critiques with grounded justifications and per-dimension scores, reducing the hallucination risks of free-form judges. (3) Adaptive Multilingual Reward Optimization (AMRO), a rubric-grounded objective function that integrates reasoning quality, answer accuracy, and language consistency.

Existing RLVR assumes that reward verification is deterministic, task-grounded, and largely language-independent. Subjective multilingual generation violates all three assumptions simultaneously: correctness is no longer executable, eval-

[†]Corresponding authors. **Correspondence:** zhumx@ncut.edu.cn

Findings of the Association for Computational Linguistics: ACL 2026, pages 23442–23474
July 2-7, 2026

uation criteria become culturally mediated, and reward semantics drift across languages. Therefore, the challenge is not merely to score outputs better, but to transform subjective evaluation into structured signals that remain verifiable enough for stable RL.

We validate this framework by constructing a comprehensive training and evaluation suite covering 7 subjective task categories across 50 low-resource languages. This dataset provides the necessary supervision for structured reward learning, enabling us to integrate the framework into a bootstrapped Group Relative Policy Optimization (GRPO) pipeline. To address the practical instability inherent in low-resource Reinforcement Learning(RL), we incorporate three targeted optimizations: length-aware loss normalization, non-zero reward variance maintenance, and reward variance stabilization. Ultimately, this design allows the model to serve simultaneously as both policy and critic. Extensive experiments demonstrate that this checklist-driven approach yields substantial improvements in reasoning capability and response quality, particularly in extremely low-resource settings where traditional reward models exhibit significant degradation. Our contributions are threefold:

(1) We propose a Structured Multilingual Reward Modeling Framework that extends RLVR to subjective and open-ended tasks. By unifying a checklist schema, generative critique modeling, and adaptive optimization, we bridge the gap between structured, auditable reward signals and flexible judges.

(2) We construct and release a comprehensive critique dataset covering 50 low-resource languages and 7 subjective task categories. This resource provides granular, checklist-aligned supervision, addressing the scarcity of structured evaluation signals in multilingual settings.

(3) We integrate our framework into a bootstrapped GRPO pipeline with targeted stability optimizations. Experiments demonstrate substantial improvements in reasoning and generation quality, particularly in extremely low-resource languages where traditional reward models exhibit significant degradation.

2 Related Work

2.1 Reward Verification Gap

Scalar (Ma et al., 2025) and generative (Zhang et al., 2024) paradigms, including hybrids (Zhang et al., 2025b), face a reward bottleneck: verification is limited to objective domains, while judges hallucinate in non-English settings due to missing structural constraints. We introduce SGCM to bridge this gap, employing a checklist schema to provide structured and auditable reward signals for subjective tasks.

2.2 Low-Resource Multilingual Alignment

Low-resource alignment suffers from cross-lingual reward drift (Hwang et al., 2025), while existing benchmarks lack unified rubrics suitable for stable RL (Que et al., 2024; Zhang et al., 2025c). To address this lack of verifiable signals, we construct a 50-language unified suite enabling AMRO to enforce consistent cross-lingual quality.

2.3 Structured Verifiable Rewards

Process supervision enables RLVR for objective tasks (Wei et al., 2022; Pyatkin et al., 2025) but struggles in subjective scenarios. While Magistral (Rastogi et al., 2025) shows RL-induced reasoning, it lacks explicit structural constraints. We advance RLVR via a Structured Checklist Schema that transforms abstract assessments into granular checkpoints, providing the ground truth to stabilize subjective, multilingual RL.

While these lines of work have each addressed part of the alignment problem, none simultaneously provides (i) verifiable supervision for subjective tasks, (ii) robustness in low-resource multilingual settings, and (iii) reward signals suitable for stable RL optimization. To clarify this distinction, Table 1 summarizes the differences between standard RLHF-style scalar rewards, free-form LLM-as-a-judge approaches, objective RLVR, and our structured subjective RLVR framework.

3 Structured Multilingual Reward Modeling Framework

As illustrated in Fig 1, our Structured Multilingual Reward Modeling Framework supports more reliable alignment through three unified components:

(1) A Structured Checklist Schema providing language-agnostic evaluation dimensions; (2) SGCM, which generates evidence-grounded critiques and scores; and (3) AMRO, which synthe-

Table 1: Comparison of major reward paradigms for alignment. Our method differs from prior work by supporting subjective evaluation, multilingual robustness, process-level grounding, and stable RL through structured critique.

Paradigm	Verifiable?	Subjective Tasks?	Multilingual Robustness?	Process-Level Grounding?	Usable for Stable RL?
Scalar reward model (RLHF/RLAIF style)	Low	Partial	Low	No	Partial
Free-form generative judge (LLM-as-a-judge)	Low	Yes	Low	Partial	Unstable
Objective RLVR	High	No	Partial	Yes	Yes
Our structured subjective RLVR	**High**	**Yes**	**High**	**Yes**	**Yes**

sizes these signals into stable reward targets. Our goal is not to make subjective evaluation fully objective, but to make it sufficiently structured, grounded, and reliability-aware so that it can serve as a stable reward source for RL.

3.1 Structured Checklist Schema

Our framework unifies reasoning and answer evaluation through a structured checklist schema. Formally, a checklist is defined as $K = \{(d_k, w_k)\}_{k=1}^{K}$, where d_k denotes an evaluation dimension and w_k its associated weight. Rather than serving as a prompting device, the checklist functions as an evaluation interface that maps abstract notions of quality into decomposable, weighted dimensions. This checklist-based evaluation interface is inspired by prior checklist-style HelloEval in HelloBench (Que et al., 2024) frameworks, but is redesigned for multilingual reward modeling rather than post-hoc long-form generation assessment. This design separates universal reasoning dimensions from task-instantiated answer dimensions, allowing the same evaluation structure to be reused across languages while remaining adaptable to different task requirements. We instantiate this structure for two distinct levels: task-specific answer checklists instantiate K with dimensions targeting semantic validity and task fulfillment, while the universal reasoning checklist instantiates the same structure with dimensions capturing coherence, correctness, error detection, and conciseness in the chain of thought. Detailed designs are provided in Appendix D. By standardizing these criteria, the schema serves as a language-agnostic foundation for SGCM and AMRO, enabling structured and consistent evaluation across all tasks.

3.2 Structured Generative Critique Modeling

SGCM operationalizes evaluation using the structured checklist schema introduced earlier. Given an input query q, the model generates a reasoning trace t and a final answer y, and SGCM evaluates the tuple (q, t, y) by producing a structured critique aligned with the instantiated checklist (either task-specific or reasoning-level). For a checklist $\mathcal{K}$, SGCM generates a critique:

$$c = \{(d_k, e_k, s_k)\}_{k=1}^{K},$$

where e_k is a concise justification grounded in observable evidence from (q, t, y), and $s_k \in [0, 1]$ is the score assigned to dimension d_k.

Each entry follows a controlled template, ensuring that explanations remain traceable to specific elements of the reasoning trace or final answer. This structure significantly mitigates hallucinated or vague feedback typical of free-form judges.

SGCM is trained on human structured critiques (see Appendix C) to ground evaluations in observable evidence for factuality, task completion, and reasoning. By coupling scores with justifications, it improves interpretability and auditability. Unlike prompt-engineered judges, SGCM produces traceable, decomposable signals aligned with a fixed schema, providing reliable and reusable supervision for AMRO rather than one-off judgments.

3.3 Adaptive Multilingual Reward Optimization

AMRO converts the soft, per-dimension scores by SGCM, with multilingual and task-level signals, into stable reward targets for RL. AMRO aggregates four components: language consistency, task recognition, reasoning-level structure, and answer-level structure.

3.3.1 Language Consistency Reward

Multilingual settings often suffer from cross-language drift, especially in low-resource languages where the model tends to revert to English. To encourage linguistic alignment, AMRO applies a language-consistency reward

$$r_{\text{lang}} = \mathbb{I}[f(y) = \ell],$$

where ℓ is the expected output language inferred from the query and $f(\cdot)$ is a language identification function. In practice, we run three language identification tools, Google Translate,* langid (Lui and Baldwin, 2012), and FastText (Joulin et al.,

*https://translate.google.com

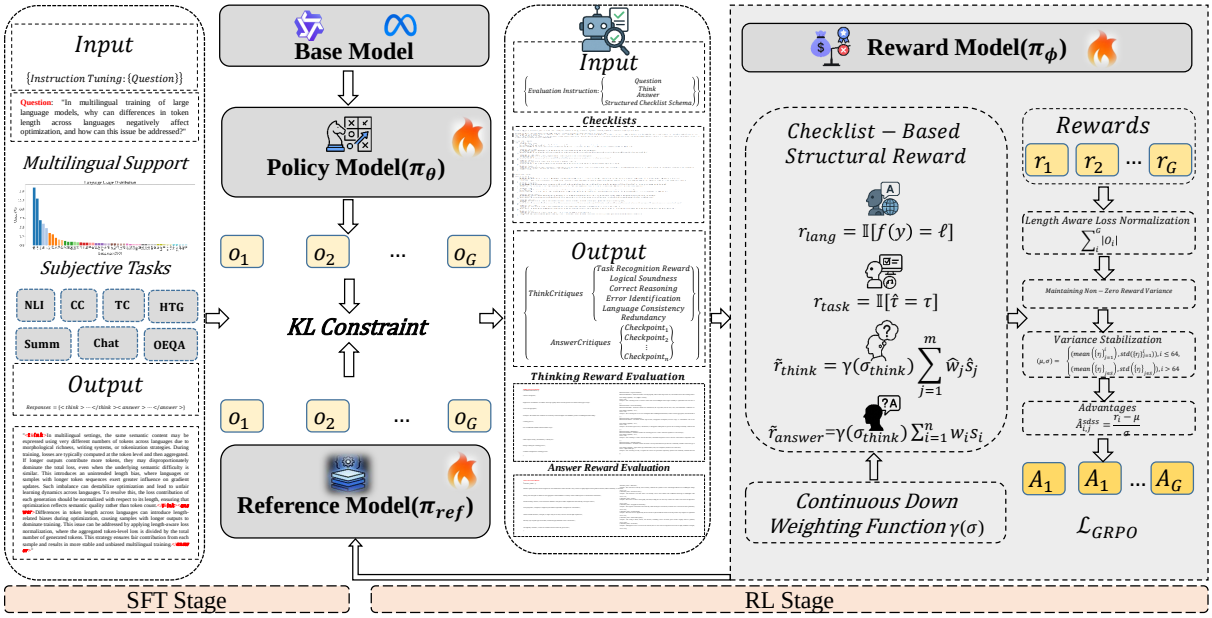


Figure 1: Schematic illustration of the Structured Multilingual Reward Modeling Framework. Highlights how structured critique bridges subjective evaluation and verifiable RL signals.

2016), on both the reasoning trace and the final answer, compare their confidence scores, and use the most confident prediction as the final language label. This ensemble improves robustness for low-resource languages.

3.3.2 Task Recognition Reward

Since SGCM relies on task-dependent checklists, the policy must identify the correct task category. AMRO introduces a task recognition reward

$$r_{\text{task}} = \mathbb{I}[\hat{\tau} = \tau],$$

where τ and $\hat{\tau}$ are the ground-truth and predicted task labels. Using $\hat{\tau}$, the policy selects a schema checklist to drive structured evaluation. We integrate task identification into reasoning to condition the chain-of-thought on the task type (see Appendix D).

3.3.3 Checklist-Based Structural Reward.

Checklist-Based Structural Reward provides a unified, interpretable signal for both the reasoning process and the final answer. AMRO applies a shared variance-aware smoothing strategy to both components.

Reasoning Process Reward. Given SGCM-generated reasoning scores $\hat{s}_j \in [0, 1]$ with weights $\hat{w}_j$, the base reasoning reward is

$$r_{\text{reason}} = \sum_{j=1}^{m} \hat{w}_j \hat{s}_j.$$

To reduce the influence of noisy or inconsistent subscores, we apply variance-based smoothing. Let σ_{reason} denote the standard deviation of the reasoning subscores. The smoothed reward is

$$\tilde{r}_{\text{reason}} = \gamma(\sigma_{\text{reason}})\, r_{\text{reason}},$$

where $\gamma(\sigma)$ is a continuous down-weighting function that suppresses high-variance signals without resorting to discrete filtering.

Final Answer Reward. For the final answer, SGCM provides per-dimension scores $s_i \in [0, 1]$ with weights w_i, yielding

$$r_{\text{answer}} = \sum_{i=1}^{n} w_i s_i.$$

Similarly, with σ_{answer} denoting the standard deviation of the answer-level subscores, we obtain a smoothed answer reward

$$\tilde{r}_{\text{answer}} = \gamma(\sigma_{\text{answer}})\, r_{\text{answer}}.$$

3.3.4 Final Reward Integration.

The overall multilingual reward is defined as:

$$R = \lambda_4 r_{\text{lang}} + \lambda_1 r_{\text{task}} + \lambda_2 \tilde{r}_{\text{reason}} + \lambda_3 \tilde{r}_{\text{answer}} - p_{\text{invalid}},$$

where λ_1, λ_2, λ_3, and λ_4 are tunable coefficients controlling the relative contributions of language-consistency, task-level, reasoning-level, and answer-level rewards. The penalty term p_{invalid} applies a deterministic deduction to structurally invalid outputs, such as missing reasoning traces, malformed formats, or violations of task-specific structural requirements. AMRO therefore unifies structural critique, multilingual consistency, and task-recognition signals into a single reward for RL, yielding more informative and stable supervision across diverse languages, especially in low-

resource settings.

4 Dataset Construction

To support structured reward modeling across 50 languages, we constructed a large-scale supervised dataset spanning 7 subjective task families aligned with distinct cognitive types (see Appendix E).

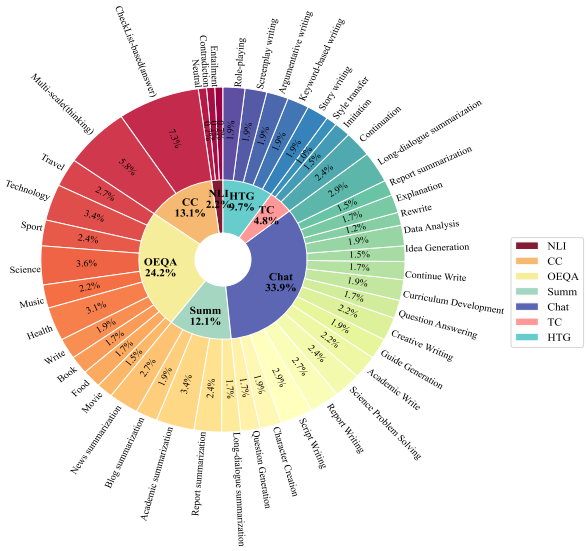


Figure 2: Data distributions of all subtasks across the seven task categories, which jointly cover the cognitive spectrum and are instantiated in all 50 languages with tailored prompts and language-specific annotations.

4.1 Task Framework and Multilingual Collection

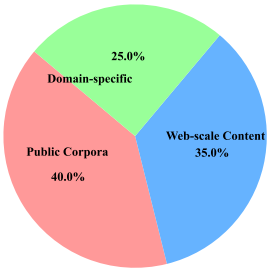
To systematically construct the capability framework of LLMs in low-resource multilingual settings, we draw on Bloom's taxonomy (Bloom, 2010; Que et al., 2024), which organizes cognitive processes into six levels: Remember, Understand, Apply, Analyze, Evaluate, and Creat. Guided by this framework, we design representative tasks for each level to strengthen general and task-specific competencies across under-resourced languages. For Remember, we introduce Open-Ended Question Answering (OEQA) to promote factual retrieval and semantic memory. For Understand, we employ Summarization (Summ) to assess comprehension of multilingual, cross-domain content. For Apply, we curate Chat-based Dialogue (Chat) tasks covering daily conversation, task-oriented interaction, and role-playing scenarios, thereby improving contextual consistency and instruction following. For Analyze, we construct Text Completion (TC) and Natural Language Inference (NLI) tasks, which require contextual inference, structural completion, and causal reasoning. To enhance Evaluate, we build a CheckList-based (answer) and Multi-scale (thinking) Comprehensive Critique (CC) task (Zheng et al., 2023; Tao et al., 2025), enabling structured self-evaluation and refinement in low-resource languages. For Create, we implement Heuristic Text Generation (HTG) tasks that encourage original, contextually appropriate long-form generation from open-ended prompts. We select the most representative task for each level and language to ensure clear training objectives and effective capability enhancement. Data collection follows a three-stage pipeline: (1) curating heterogeneous content from web sources and region-specific repositories; (2) filtering with ensemble language detection (fastText, CLD3) and strict PII removal; and (3) instantiating tasks across 50 languages with culturally appropriate prompts. This process yields a high-quality multilingual corpus that forms the basis of our structured annotations.

4.2 Structured Critique Annotation

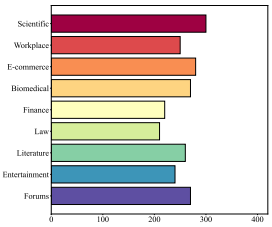
We implemented an annotation scheme to train the SGCM. Query-response pairs were generated using state-of-the-art LLMs, followed by a two-level critique annotation process performed by native experts: (1) Universal Reasoning Supervision: Experts evaluated thinking traces using a universal reasoning checklist. To support AMRO, we collected dimension-wise scores (discretized to avoid sparsity) and grounded justifications linked to trace segments. (2) Task-Specific Answer Supervision: Final answers were evaluated against task-dependent checklists (e.g., factual grounding for QA vs. creativity for Generation). (3) Standardized Templates: To ensure cross-lingual consistency, expert justifications were distilled into multilingual templates, validated through cross-lingual auditing.

4.3 Scale and Statistics

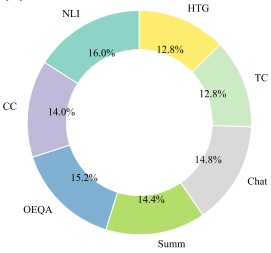
The final corpus is partitioned into Supervised Fine-Tuning (SFT), RL optimization, and Evaluation sets. As shown in Figure 2 and detailed in Table 7, the dataset features scale and diversity: (1) Annotated Critiques: We collected approximately 690k structured samples, comprising 240k reasoning-level, 300k answer-level, and 150k joint annotations. (2) RL Pre-training Data: The corpus includes 3.93M unlabeled multilingual prompts and responses for policy optimization. Rigorous quality control, including inter-annotator agreement checks and distribution balancing, was applied to ensure reliability across all 50 languages.



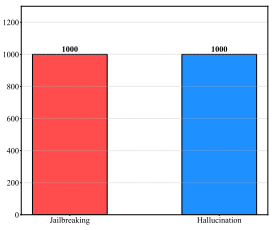
(a) Data Source Distribution



(b) Domain Coverage



(c) Multi-50L-5C-Bench



(d) Multi-50L-Safety-Bench

5 Experiments

5.1 Training Pipeline

Our training pipeline progresses through two stages: supervised initialization followed by iterative RL.

Stage 1: Structured SFT. We first equip the model with structured critique capabilities. The model is supervised using the checklist-annotated dataset (containing thinking and answer-level critiques) mixed with standard multilingual corpora. This prevents overfitting to evaluative behaviors while initializing a robust reward model capable of generating rubric-aligned justifications across all languages (see Appendix F).

Stage 2: Bootstrapped GRPO. We subsequently employ GRPO (Shao et al., 2024) in a self-reinforcing loop: the policy generates reasoning traces and answers, evaluates itself via SGCM/AMRO, and updates based on these self-generated rewards. To address instability in low-resource multilingual settings, we introduce three targeted optimizations:

(1) **Length-Aware Loss Normalization:** Since token counts vary dramatically across languages, we normalize token-level losses by the total generated length within each group. This prevents length-induced optimization bias and ensures fair cross-lingual learning (Appendix G).

(2) **Non-Zero Reward Variance Maintenance:** To prevent gradient collapse when all group responses receive identical scores, we employ a balanced sampling strategy. By ensuring a mix of correct and incorrect reasoning/answers within a group, we guarantee $\sigma_r > 0$ and prevent the discarding of difficult prompts (Appendix B).

(3) **Variance Stabilization:** To mitigate numerical instability in large response groups, we estimate the standard deviation σ from a fixed-size random subset rather than the full group. This preserves the benefits of normalization while ensuring robust variance estimation.

These modifications maintain GRPO's simplicity while significantly enhancing stability for reasoning-intensive multilingual tasks.

5.2 Evaluation Datasets

5.2.1 Low-Resource Multilingual Benchmark

We develop a comprehensive evaluation suite targeting low-resource scenarios across 50 languages and 9 domains, with distributions shown in Fig. 3. Multi-50L-5C-Bench comprises 2,500 human-curated samples spanning 7 cognitively motivated task categories, organized into 7 coarse and 45 fine-grained categories, while Multi-50L-Safety-Bench evaluates robustness through jailbreaking and hallucination detection. All samples are validated by native speakers to ensure linguistic fluency and reliability. To reduce contamination risk, both benchmarks are constructed from region-specific repositories, non-English web corpora, and post-2024 sources that postdate the pretraining releases of LLaMA-3.1 and Qwen2.5; we further enforce non-overlap using exact-match, 13-gram, and embedding-similarity filtering, yielding less than 0.18% exact duplication.

5.2.2 Open Benchmark

To verify generalization, we evaluate our model on mainstream benchmarks covering reasoning, understanding, and math across high-resource languages. **General Tasks**: Includes MMLU (Hendrycks et al., 2021), MMLU-Pro (TIGER-Lab, 2024), MMLU-Redux (Gema et al., 2024), BBH (Suzgun et al., 2022), ARC-C (Clark et al., 2018), TruthfulQA (Lin et al., 2022), Winogrande (ai2, 2019), and HellaSwag (Zellers et al., 2019). **Multilingual Tasks**: (1) Exams: Translated-MMLU (Chen et al., 2023), IndoMMLU (Koto et al., 2023); (2) Understanding: XCOPA (Edoardo M. Ponti and Korhonen, 2020), XWinograd (Muennighoff et al., 2022); (3) Math: MGSM (Shi et al., 2022); (4) Translation: Flores-101 (Goyal et al., 2021). The evaluated languages are detailed in Appendix H.

5.3 Evaluation Metrics

Multi-50L-5C-Bench We employ the LLM-as-a-Judge framework (Zheng et al., 2023) for subjective understanding and generation tasks. To address the high cost and inconsistency of human evaluation, we adopt a verifiable reward-function strategy (Appendix I) and conduct a GSB consistency analysis with professional human reviewers. Although our SGCM-based judge shows strong agreement with human judgments, it is involved in the training pipeline and may therefore raise concerns about evaluation bias if reused as the sole evaluator. To ensure a more objective and fair assessment, we use two independent external judge models, GPT-5 (high) and Gemini-2.5-Pro, for final evaluation, and report the average of their scores as the main metric for subjective tasks. Multi-50L-Safety-Bench Given the objective nature of safety tasks, we adopt Accuracy as the primary metric. Open Benchmark We strictly adhere to the official evaluation protocols and metric definitions specified by each respective benchmark to ensure fair comparison.

5.4 Baselines

We compare our method against a diverse range of open-source and proprietary models. **Open-source models** include RM-R1, RLOO (Su et al., 2025), Mistral-7B-v0.3 (Albert Q. Jiang, 2023), GLM-4-9B (GLM et al., 2024), Phi-3.5-MoE (Abdin et al., 2024), LLaMA-3.1-8B (Dubey et al., 2024), and Qwen2.5-7B-Instruct (Team et al., 2024). **Proprietary APIs** include Doubao-Seed-1.6 (Team, 2025), Qwen3-Max (Yang et al., 2025), DeepSeek-V3.2-Exp (DeepSeek-AI, 2024), GLM-4.6 (GLM et al., 2024), Gemini-2.5-Pro (Comanici et al., 2025), Claude-Sonnet-4.5 (Anthropic, 2025), and GPT-5 (Achiam et al., 2023). We also include a compute-matched Continued SFT (CSFT) baseline to isolate the gains of RL from additional training compute.

5.5 Implementation Details

We use LLaMA-3.1-8B-Instruct (English-centric) and Qwen2.5-7B-Instruct (Multilingual) as base models with the R1 `<think>...</think><answer>...</answer>` template (Liu et al., 2025), where `<answer>` contains the final response. SFT baselines are trained until convergence, while RL models are optimized for 500 steps. We use Adam ($lr = 1e-6$) on 32 NVIDIA H800 GPUs; full details in Appendix J.

5.6 Main Results

5.6.1 Low-Resource Benchmark Results

Multi-50L-Bench Results: As shown in Table 2, our approach consistently outperforms open-source baselines and rivals proprietary APIs. For the Qwen2.5-7B-Instruct series, average scores improve from 15.92 (base) to 34.84 (SFT) and 47.11 (RL), demonstrating a strong cross-task enhancement. Notably, our models surpass GPT-5 in CC, ODQA, and Summ. In the challenging HTG task, our model reaches 49.48, nearly matching GPT-5, which indicates that structured rewards are highly effective for logical consistency and global planning in complex generation. When stratifying the 50 languages into High, Medium, and Low-resource tiers based on training corpus size, our method achieves the most pronounced relative gains in the Low-resource tier, confirming the robustness of structured checklists against cross-lingual drift (detailed breakdown in Appendix K).

(1) **Synergy between SFT and RL.** Our analysis reveals complementary roles: SFT establishes foundational instruction-following and cross-lingual formatting, while RL, guided by multi-dimensional structured rewards, provides a second-stage systematic alignment. While SFT-only models often struggle with out-of-distribution generalization, RL enhances underlying reasoning abilities and logical rigor. Compared to a compute-matched CSFT baseline, our SFT+RL pipeline still yields a 10% improvement, proving the efficacy of structured rewards over mere additional training. Crucially, SFT stabilizes the output space, allowing RL to focus on refining the reasoning trajectory rather than correcting formatting irregularities (Chu et al., 2025a).

(2) **Architecture and Pretraining Priors.** Qwen2.5-7B-Instruct based models outperform LLaMA-3.1-8B-Instruct counterparts, likely due to Qwen's broader multilingual pretraining coverage. Theoretically, pretraining provides a multilingual representation prior, while our RL with structured rewards performs cross-lingual process alignment on top of it. This synergy enables an 7B-scale model to approach or even surpass much larger proprietary models in multilingual reasoning.

(3) **Process-Level Supervision.** We observe that answer-only rewards improve surface quality but fail to stabilize reasoning trajectories in HTG and Summ, leading to brittle gains. In contrast, our explicit process supervision reshapes the search

Type	Method	NLI	CC	ODQA	Summ	Chat	TC	HTG	Avg
Open-Source	RM-R1	4.36	0.32	26.62	1.68	1.32	14.51	32.08	11.56
	RLOO	8.32	0.77	38.12	1.01	1.04	9.74	3.47	8.92
	Mistral-7B-Instruct-v0.3	8.05	0.48	39.91	0.26	1.04	10.79	6.08	9.52
	GLM-4-9B-0414	12.43	1.84	39.14	3.23	2.22	11.0	18.99	12.69
	Phi-3.5-Moe	6.70	0.79	52.66	1.76	1.26	1.83	12.49	11.07
	Llama-3.1-8B-Instruct	12.65	2.84	44.37	5.86	2.03	2.97	14.77	12.21
	Qwen2.5-7B-Instruct	12.93	2.11	45.28	13.93	10.5	2.75	23.97	15.92
API	Doubao-Seed-1.6-thinking-250715	44.57	34.13	43.55	34.59	34.68	54.06	46.99	41.79
	Qwen3-Max	26.60	16.03	49.22	9.85	2.66	42.83	34.21	25.91
	DeepSeek-V3.2-Exp-Thinking	33.05	23.57	42.51	15.93	31.05	45.51	33.21	32.12
	GLM-4.6	22.50	12.93	45.5	6.51	1.68	32.82	28.21	21.45
	Gemini-2.5-Pro	36.94	26.71	51.49	21.43	26.3	42.9	45.75	35.93
	Claude-Sonnet-4.5-Reasoning	44.82	34.25	61.12	28.89	30.62	48.97	39.40	41.15
	GPT-5(high)	**49.81**	32.72	56.02	31.42	**42.34**	**60.6**	**49.81**	46.10
Our	**Llama-3.1-8B-Instruct+SFT**	22.20	12.37	54.91	16.08	17.88	13.80	25.06	23.18
	Qwen2.5-7B-Instruct+SFT	33.98	23.76	62.82	32.26	26.13	22.71	42.19	34.84
	Llama-3.1-8B-Instruct+CSFT	25.40	14.80	58.20	18.50	20.10	15.90	28.30	26.50
	Qwen2.5-7B-Instruct+CSFT	35.12	26.50	65.30	34.45	28.10	25.20	44.20	37.15
	Llama-3.1-8B-Instruct+SFT+RL	40.89	31.13	73.34	34.61	33.25	32.37	43.82	41.34
	Qwen2.5-7B-Instruct+SFT+RL	38.69	**39.82**	**74.82**	**43.01**	42.12	41.86	49.48	**47.11**

Table 2: Mainstream and proposed model performance on **Multi-50L-5C-Bench** across 50 languages over seven tasks (NLI, CC, ODQA, Summ, Chat, TC, and HTG), with Avg representing the mean of all task scores.

distribution toward coherent chains, effectively mitigating semantic drift in low-resource languages. This confirms that process-level guidance is essential rather than optional, enabling our 7B model to compete effectively with significantly larger proprietary systems.

Type	Method	Jailbreaking	Hallucination
Open-Source	RM-R1	14.54	25.34
	RLOO	18.66	26.95
	Mistral-7B-Instruct-v0.3	22.00	32.76
	GLM-4-9B-0414	26.10	36.42
	Phi-3.5-Moe	25.32	34.42
	Llama-3.1-8B-Instruct	27.50	38.54
	Qwen2.5-7B-Instruct	32.43	42.21
API	Doubao-250715	63.00	76.00
	Qwen3-Max	66.60	79.07
	DeepSeek-V3.2	58.89	76.00
	GLM-4.6	50.50	58.95
	Gemini-2.5-Pro	67.23	75.43
	Claude-4.5	65.60	77.90
	GPT-5 (high)	68.55	80.92
	Llama-3.1-8B-Instruct+SFT	36.23	46.78
	Qwen2.5-7B-Instruct+SFT	41.45	52.91
	Llama-3.1-8B-Instruct+CSFT	39.50	49.35
	Qwen2.5-7B-Instruct+CSFT	45.80	57.60
	Llama-3.1-8B-Instruct+SFT+RL	77.37	88.02
	Qwen2.5-7B-Instruct+SFT+RL	**82.31**	**91.67**
Human	**Human Reference**	89.93	93.34

Table 3: Evaluation on **Multi-50L-Safety-Bench** (Jailbreaking, Hallucination), including **Human reference** as the upper bound.

Multi-50L-Safety-Bench Results: As shown in Table 3, our methods exhibit a three-tier performance structure, significantly narrowing the "safety transfer gap" in low-resource settings. While open-source models struggle with multilingual safety, our SFT+RL models (82.31/91.67) not only surpass all open-source and most API models but also substantially narrow the gap to the human upper bound (93.34). This demonstrates that at the 7B scale, RL with structured safety rewards yields a qualitative leap in safety capability compared to the moderate gains from SFT alone.

(1) **Complementarity of SFT and RL.** We find that SFT and RL are non-interchangeable: SFT ensures basic controllability by pulling outputs onto a "safe track" and unifying instruction understanding; whereas RL, guided by multi-dimensional rewards (harm detection, consistency), determines the height and robustness of the safety boundary, resulting in a 40-point jump. RL effectively compresses high-risk regions in the state space while expanding preferences for safe response patterns.

(2) **Reward-Driven Alignment.** Qwen2.5-7B-Instruct based models' superiority (82.31vs77.37) suggests broader pretraining provides a stronger safety prior. Our rewards act as biased guidance on this prior, aligning low-resource safety boundaries with high-resource standards through shared structured dimensions. This suggests that explicit multilingual reward design may be more important for safety than model scale or English-centric alignment in our setting.

(3) **Process-Level Safety Reasoning.** Crucially, our RL design supervises the intermediate chain-of-thought rather than just final outputs. By scoring harm recognition and refusal justification, this process-level reward reshapes the search distribution toward a "safe reasoning + safe conclusion" joint structure. This mitigates safety degradation from data sparsity, enabling our 7B models to main-

Datasets	Open-Source				API				Ours			
	Mistral-7B-Instruct-v0.3	GLM-4-9B-0414	Llama-3.1-8B-Instruct	Qwen2.5-7B-Instruct	GLM-4.6	Gemini-2.5-Pro	Claude-Sonnet-4.5-Reasoning	GPT-5(high)	Llama-3.1-8B-Instruct+CSFT	Qwen2.5-7B-Instruct+CSFT	Llama-3.1-8B-Instruct+SFT+RL	Qwen2.5-7B-Instruct+SFT+RL
General Tasks												
MMLU-Pro (5-shot)	32.21	47.77	48.30	57.25	83.92	86.20	85.00	**87.10**	49.10	59.00	50.57	62.14
MMLU-redux(5-shot)	59.10	69.88	67.20	75.40	87.36	88.23	93.12	**94.29**	68.10	76.80	69.54	80.09
BBH(3-shot)	54.14	76.30	71.85	71.92	53.78	68.80	57.54	**88.95**	72.50	73.40	74.08	76.80
ARC-C(25-shot)	59.73	66.81	60.48	67.15	84.21	**88.55**	54.65	84.28	61.20	68.50	62.70	72.20
TruthfulQA (0-shot)	45.41	43.66	61.69	64.72	79.56	**89.54**	61.19	82.81	62.50	66.20	64.13	69.06
Winogrand(5-shot)	75.61	82.95	78.06	75.53	89.12	**94.12**	66.43	89.30	79.10	76.90	80.62	79.84
HellaSwa(10-shot)	80.88	84.45	80.12	81.36	91.23	**95.12**	67.48	91.21	81.05	82.80	82.62	86.24
Multilingual Tasks												
Translated MMLU(5-shot)	48.45	72.28	58.47	69.08	92.10	**95.52**	68.42	81.23	59.20	70.50	60.51	73.11
IndoMMLU(3-shout)	44.56	54.09	53.92	56.42	89.91	**90.44**	62.89	76.83	54.80	58.10	56.02	61.05
XCOPA(5-shot)	57.64	64.80	64.70	64.09	84.54	**86.19**	58.94	84.17	65.50	65.20	66.77	68.28
XWinograd(5-shot)	79.32	86.22	84.58	83.66	89.60	87.93	59.47	**96.21**	85.20	84.80	86.27	87.95
MGSM(8-shot CoT)	48.72	64.70	66.05	66.11	87.71	84.10	**94.30**	92.80	66.90	67.80	68.27	70.98
Flores-101(5-shot)	50.13	58.04	49.31	69.11	92.12	88.38	89.54	**95.78**	50.10	71.00	51.40	74.28

Table 4: **Open Benchmark** results comparing open-source, proprietary (API), and our models.

Reward Model	Conv	5C-Bench	Safety-Bench
Qwen2.5-7B-Instruct Judge	Collapse	–	–
GPT-5 (high) Judge	Unstable	36.61	70.69
Gemini-2.5-Pro Judge	Unstable	37.31	72.89
SGCM+AMRO (Ours)	**Stable**	**47.11**	**86.99**

Table 5: Impact of reward modeling.

Method	5C-B.	Safe-B.	Div. (%)
Baseline GRPO	39.40	75.84	28
+ Length-Aware Normalization	40.32	78.52	20
+ Non-Zero Variance Only	41.60	79.16	18
+ SD Stabilization Only	39.80	77.12	22
Full (All Three)	**47.11**	**86.99**	**4**

Table 6: GRPO stabilization effects on performance.

tain safety via internal reasoning templates and rival much larger proprietary systems.

5.6.2 Open Benchmark Results

As shown in Table 4, our SFT+RL variants (avg. 8 runs) consistently outperform LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct. On general tasks, our models significantly narrow the gap to proprietary systems. These gains are magnified on multilingual benchmarks, where we surpass all open-source baselines and even APIs like Claude-Sonnet-4.5. Crucially, relative gains correlate with language coverage, indicating enhanced cross-lingual reasoning rather than memorization. Qwen2.5-based variants excel in low-resource settings, suggesting our strategy effectively amplifies Qwen's linguistic priors for robust generalization.

5.7 Ablation Study

We conduct controlled ablations on Multi-50L-5C-Bench and Multi-50L-Safety-Bench to evaluate component contributions. All experiments use Qwen2.5-7B-Instruct with identical hyperparameters and R1-style prompting. We investigate: (1) the necessity of our structured reward model (SGCM+AMRO) for multilingual RL stability, and (2) the impact of GRPO stabilization on establishing multilingual foundational abilities.

5.7.1 Structured Reward Modeling

We compare our SGCM+AMRO against generative judges. As shown in Table 5, three trends emerge: (1) Standard judges fail to converge: RL with the Qwen2.5-7B-Instruct based judge collapses after 130 steps due to reward inconsistency(see Appendix M). (2) Proprietary models yield unstable gains: GPT-5 and Gemini suffer from oscillatory updates, providing only marginal improvements over SFT. (3) Structured rewards substantially improve stability: Only SGCM+AMRO achieves monotonic improvement by providing checklist-grounded, variance-aware critique. This suggests that structured, multidimensional signals are important for stable multilingual RLVR.

5.7.2 GRPO Stabilization Strategies

We evaluate three GRPO components: (1) Length-Aware Loss Normalization, (2) Non-Zero Reward Variance (via balanced sampling), and (3) Standard Deviation Stabilization. Tracking performance and divergence across five seeds, Table 6 shows baseline GRPO is unstable (28%). While individual components improve stability, combining all three yields optimal results boosting benchmark averages by over 10 points while cutting divergence to 4%. This suggests that the three components are complementary for reliable multilingual RL in our setting (See Appendix N).

6 Conclusion

We identify a major reward bottleneck in low-resource multilingual alignment arising from the trade-off between rule-based supervision and judge-model rewards. To address this issue, we propose a Structured Multilingual Reward Modeling Framework that extends RLVR to open-ended generation through a checklist schema, SGCM, and AMRO, producing interpretable multidimensional critique signals instead of brittle scalar rewards. Experiments show consistent gains in reasoning, reliability, and safety, especially in extremely low-resource languages, suggesting that structured, rubric-grounded critique is important for scalable multilingual RL.

Limitations

Although our framework substantially improves multilingual reward modeling and alignment, several limitations remain. First, our critique annotations, despite covering 50 languages, still depend on LLM-generated samples and expert verification, which may not fully capture the linguistic and cultural nuance of extremely low-resource communities. Second, SGCM relies on structured checklists whose quality is tied to expert-designed criteria; extending these schemas to additional domains or genres may require non-trivial human effort. Third, AMRO focuses on text-only rewards, and its applicability to multimodal or speech-based multilingual settings remains unexplored. Fourth, while our GRPO enhancements improve optimization stability, RL at scale is computationally expensive, limiting experimentation with larger models or broader hyperparameter sweeps. Finally, our evaluation focuses primarily on reasoning, long-form generation, and safety; downstream applications such as dialogue coherence, cross-lingual transfer, or domain-specific alignment warrant further investigation.

Although our checklist schema is designed to be language-agnostic, it inevitably reflects human-designed evaluation criteria. To mitigate cultural bias, we enforce cross-lingual checklist consistency via multilingual auditing and disagreement resolution among native experts from different cultural backgrounds. Nevertheless, absolute cultural neutrality remains an open challenge.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of Hainan Province, China, under Grant ZDYF2024 (LALH) 005; the Beijing Municipal Science & Technology Commission under Grant Z231100001723002; the National Key R&D Program of China under Grant 2022YFF0902500; the Youth Research Special Project of NCUT under Grant 110051360025XN077-13; the Provincial-Level Research Platform Operations Support Program for the Beijing Key Laboratory of Key Technologies for AI+ Domain Applications; the Joint Fund Key Program of the National Natural Science Foundation of China under Grant U23B2029; and the Graduate Research Projects of Minzu University of China.

References

2019. Winogrande: An adversarial winograd schema challenge at scale.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arthur Mensch Chris Bamford Devendra Singh Chaplot Diego de las Casas Florian Bressand Gianna Lengyel Guillaume Lample Lucile Saulnier Lélio Renard Lavaud Marie-Anne Lachaux Pierre Stock Teven Le Scao Thibaut Lavril Thomas Wang Timothée Lacroix William El Sayed Albert Q. Jiang, Alexandre Sablayrolles. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Anthropic. 2025. Introducing claude sonnet 4.5. *Journal of Clinical Oncology*, 33(27):2025–10–21.

Benjamin Samuel Bloom. 2010. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025a. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. 2025b. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Michael Desmond, Zahra Ashktorab, Werner Geyer, Elizabeth M. Daly, Martín Santillán Cooper, Qian Pan, Rahul Nair, Nico Wagner, and Tejaswini Pedapati. 2025. Evalassist: Llm-as-a-judge simplified. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 29637–29639. AAAI Press.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Olga Majewska Qianchu Liu Ivan Vuli'c Edoardo M. Ponti, Goran Glava s and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. *arXiv preprint*.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2024. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xia Hou, Qifeng Li, Jian Yang, Tongliang Li, Linzheng Chai, Xianjie Wu, Hangyuan Ji, Zhoujun Li, Jixuan Nie, Jingbo Dun, and 1 others. 2024. Raw text is all you need: Knowledge-intensive multi-turn instruction tuning for large language model. *arXiv preprint arXiv:2407.03040*.

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *arXiv preprint arXiv:2507.05418*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment (2023). *arXiv preprint arXiv:2303.16634*, 12.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

TQ Luong, X Zhang, Z Jie, P Sun, X Jin, and H Reft Li. 2024. Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.

Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. 2025. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24812–24820.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff,

and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *CoRR*, abs/2505.22660.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. 2025. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, and 1 others. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, and 1 others. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable nlg evaluation. *arXiv preprint arXiv:2406.07935*.

Jiu Sha, Mengxiao Zhu, Chong Feng, and Yuming Shang. 2025. Veef-multi-llm: Effective vocabulary expansion and parameter efficient finetuning towards multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7963–7981.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv e-prints*, pages arXiv–2503.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Yongding Tao, Tian Wang, Yihong Dong, Huanyu Liu, Kechi Zhang, Xiaolong Hu, and Ge Li. 2025. Detecting data contamination from reinforcement learning post-training for large language models. *arXiv preprint arXiv:2510.09259*.

ByteDance Seed Team. 2025. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(8).

TIGER-Lab. 2024. Mmlu-pro (revision efb53c9).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. 2025. Reinforce-ada: An adaptive sampling framework for reinforce-style llm training. *arXiv preprint arXiv:2510.04996*.

Yixuan Even Xu, Yash Savani, Fei Fang, and J Zico Kolter. 2025. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning. *arXiv preprint arXiv:2504.13818*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chenlu Ye, Zhou Yu, Ziji Zhang, Hao Chen, Narayanan Sadagopan, Jing Huang, Tong Zhang, and Anurag Beniwal. 2025. Beyond correctness: Harmonizing process and outcome rewards through rl training. *arXiv preprint arXiv:2509.03403*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025a. Critique-grpo: Advancing LLM reasoning with natural language and numerical feedback. *CoRR*, abs/2506.03106.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025b. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, and 1 others. 2025c. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025. Reinforcing general reasoning without verifiers. *CoRR*, abs/2505.21493.

A Language Statistics and Data Allocation for the Multilingual Corpus

ISO	Language	Family	Train	Test	Usage(%)
zh	Chinese	Sino-Tibetan	25.91	11.09	16.00
en	English	Indo-European	21.05	9.02	13.00
es	Spanish	Indo-European	11.34	4.86	7.00
hi	Hindi	Indo-European	9.72	4.16	6.00
ar	Arabic	Afro-Asiatic	7.77	3.33	4.80
bn	Bengali	Indo-European	5.51	2.36	3.40
pt	Portuguese	Indo-European	4.86	2.08	3.00
ru	Russian	Indo-European	3.24	1.39	2.00
ja	Japanese	Japanic	2.59	1.11	1.60
fr	French	Indo-European	2.27	0.97	1.40
ur	Urdu	Indo-European	1.94	0.83	1.20
id	Indonesian	Austronesian	1.62	0.69	1.00
de	German	Indo-European	1.62	0.69	1.00
ko	Korean	Isolate	1.46	0.62	0.90
vi	Vietnamese	Austroasiatic	1.30	0.56	0.80
ta	Tamil	Dravidian	1.13	0.49	0.70
tr	Turkish	Turkic	0.97	0.42	0.60
fa	Persian	Indo-European	0.97	0.42	0.60
it	Italian	Indo-European	0.97	0.42	0.60
th	Thai	Kra-Dai	0.81	0.35	0.50
pl	Polish	Indo-European	0.65	0.28	0.40
uk	Ukrainian	Indo-European	0.65	0.28	0.40
my	Burmese	Sino-Tibetan	0.65	0.28	0.40
ro	Romanian	Indo-European	0.49	0.21	0.30
nl	Dutch	Indo-European	0.49	0.21	0.30
sv	Swedish	Indo-European	0.49	0.21	0.30
he	Hebrew	Afro-Asiatic	0.49	0.21	0.30
fi	Finnish	Uralic	0.32	0.14	0.20
hu	Hungarian	Uralic	0.32	0.14	0.20
el	Greek	Indo-European	0.32	0.14	0.20
cs	Czech	Indo-European	0.32	0.14	0.20
bg	Bulgarian	Indo-European	0.16	0.07	0.10
ca	Catalan	Indo-European	0.16	0.07	0.10
ms	Malay	Austronesian	0.16	0.07	0.10
tg	Tajik	Indo-European	0.16	0.07	0.10
ku	Kurdish	Indo-European	0.16	0.07	0.10
ky	Kyrgyz	Turkic	0.13	0.06	0.08
kk	Kazakh	Turkic	0.11	0.05	0.07
uz	Uzbek	Turkic	0.11	0.05	0.07
tk	Turkmen	Turkic	0.10	0.04	0.06
lo	Lao	Kra-Dai	0.08	0.03	0.05
km	Khmer	Austroasiatic	0.08	0.03	0.05
tl	Filipino	Austronesian	0.08	0.03	0.05
mn	Mongolian	Mongolic	0.06	0.03	0.04
ug	Uyghur	Turkic	0.05	0.02	0.03
bo	Tibetan	Sino-Tibetan	0.05	0.02	0.03
yue	Yue Chinese	Sino-Tibetan	0.05	0.02	0.03
za	Zhuang	Kra-Dai	0.03	0.01	0.02
dz	Dzongkha	Sino-Tibetan	0.03	0.01	0.02
ii	Nuosu	Sino-Tibetan	0.02	0.01	0.01
Total	–	–	**114.02**	**48.86**	**100%**

Table 7: Languages supported by data, sorted by global usage. Train/Test sizes are measured in billions (B).

B Ensuring Non-Zero Reward Variance

A key difficulty in group-based policy optimization arises when all samples within a group receive identical rewards either uniformly correct or uniformly incorrect which collapses the reward variance to zero and eliminates the gradient signal for

that prompt (Guo et al., 2025). Prior methods simply discard such groups (Rastogi et al., 2025), but this wastes useful data and causes many difficult prompts to remain untrained. To mitigate this problem, we follow the adaptive sampling paradigm inspired by Reinforce-ada (Xiong et al., 2025), incorporating its initialization, iterative sampling, and elimination procedures, while extending the method in two important ways. First, during training batch construction, instead of relying on all collected responses, we downsample each prompt to a fixed group of n responses to normalize its contribution and stabilize learning. To provide balanced and informative supervision, we draw $\frac{n}{4}$ samples from each of four outcome categories correct thinking with correct answers, correct thinking with incorrect answers, incorrect thinking with correct answers, and incorrect thinking with incorrect answers filling shortages in one category using others when necessary. This balanced sampling strategy (Xu et al., 2025; Ye et al., 2025), also adopted in recent studies, ensures that every group maintains non-zero reward variance ($\sigma_r > 0$) and therefore yields meaningful gradients. Second, we modify the exit conditions used to deactivate prompts during sampling. In the positive-focused variant, a prompt is removed once at least one correct-thinking correct-answer sample is collected, reflecting variance reduction principles in rejection-sampling-based fine-tuning. In the balanced variant (Reinforce-Ada-balance), deactivation occurs only after acquiring at least $\frac{n}{4}$ correct-thinking correct-answer samples and $\frac{n}{4}$ incorrect-thinking incorrect-answer samples, thereby guaranteeing that both successful and failed trajectories contribute before a prompt leaves the active set. These extensions prevent zero-variance failure cases and enable more robust, data-efficient optimization across diverse multilingual and reasoning-intensive prompts.

C Human Annotation Process

We recruited native speakers with a master's degree or higher from more than five professional domains worldwide to ensure linguistic and domain diversity. Each annotator received a compensation of approximately $300 to encourage high-quality participation. Following the human evaluation principles proposed by Ruan (Ruan et al., 2024) and the annotation strategy of HelloEval (Que et al., 2024), we redesigned a comprehensive human evaluation framework and developed a new guideline to mitigate common biases and vulnerabilities. This framework was systematically optimized for both rigor and practicality, ensuring scientific reliability and consistency throughout the evaluation process.

During annotation, we carefully addressed multiple factors that may affect fairness and stability. To minimize unconscious bias, the evaluation tasks were organized as independent samples, avoiding potential influence from task order or contextual information. From an ethical perspective, we strictly adhered to privacy protection, informed consent, and fairness principles; all evaluations were conducted anonymously to protect annotators' identities. To eliminate ambiguity, task definitions and evaluation criteria were clearly and precisely specified, reducing potential misunderstandings. For rating consistency, we adopted explicit quantitative rating scales accompanied by detailed explanations and representative examples, ensuring consistent judgment across annotators. For edge cases that are difficult to classify, we introduced neutral options (e.g., 0.5 points) to maintain both flexibility and accuracy in scoring.

Moreover, we considered the prior knowledge required for each evaluation task and provided annotators with the necessary background materials and references to support accurate judgments. The annotation instructions were designed with flexibility and adaptability, allowing annotators to adjust their reasoning methods based on task characteristics while maintaining overall consistency, which is particularly beneficial for multilingual and multi-domain evaluation scenarios.

Each task in every language was independently annotated by 3–4 annotators and subsequently reviewed and validated by another 3–4 quality inspectors. Through cross-annotation and dual-review mechanisms, only data that passed all verification stages were retained for the final training and evaluation. This multi-level quality assurance and cross-validation process effectively improved the reliability, consistency, and reusability of the annotated data, providing high-quality supervision signals for downstream model training.

D Thinking Reward Evaluation Instruction

To design appropriate evaluation dimensions for the thinking process, we begin with the multilingual dataset containing explicit reasoning traces and final answers introduced in Section 4, covering 50

languages and 6 tasks. For each language-task pair, we randomly sample 50 instances and use mainstream models (GPT-5 and Gemini-3-Pro-Preview) to generate both reasoning processes and final answers. Native-language experts analyze these generations using both forward synthesis (generating reasoning and answers from the question) and backward synthesis (reconstructing plausible reasoning from a fixed question-answer pair). After expert review and quality control, we obtain annotated quadruples question, response, thinking, labels, where the label indicates whether the reasoning is positive or negative. From these annotations, experts summarize six key error dimensions: Task Recognition Reward, Logical Soundness, Correct Reasoning, Error Identification, Language Consistency, and Redundancy. The final evaluation framework for the thinking process is shown in Table D .

Thinking Reward Evaluation Instruction

```
{
        "Question": "{Question}",
        "Thinking": "{Thinking}",
        "Evaluation Dimensions": [
        {
                "Dimension Name": "Logical
                     Soundness",
                "Dimension Definition": "Evaluates
                     whether each step logically
                     follows from the previous one,
                     and whether the overall
                     reasoning chain is coherent
                     and rigorous.",
                "Score Range (Optional)": "0~1
                     (Higher is better)",
                "Scoring Rules (Optional)": "If
                     there is a leap in reasoning,
                     lack of causal connection, or
                     clear logical gaps, the score
                     should be lower; if each step
                     is logically connected with
                     rigorous reasoning, a higher
                     score should be given."
        },
        {
                "Dimension Name": "Correct
                     Reasoning",
                "Dimension Definition": "Evaluates
                     whether the methods used, the
                     steps taken, and the facts,
                     rules, and intermediate
                     conclusions are correct and
                     appropriate.",
                "Score Range (Optional)": "0~1",
                "Scoring Rules (Optional)": "If
                     incorrect knowledge is used,
                     methods are inappropriate, or
                     reasoning steps are
                     inaccurate, the score should
                     be lower; if the reasoning
                     path is correct and the facts
                     or rules applied are
                     appropriate, a higher score
                     should be given."
        },
        {
                "Dimension Name": "Error
                     Identification",
                "Dimension Definition": "Evaluates
                     whether logical flaws,
                     unsupported assumptions,
                     incorrect steps, or
                     contradictions exist in the
                     reasoning process.",
                "Score Range (Optional)": "0~1",
                "Scoring Rules (Optional)": "If
                     obvious errors exist but are
                     not identified, the score
                     should be lower; if the
                     reasoning is self-consistent
                     and error-free, a higher score
                     should be given."
        },
        {
                "Dimension Name": "Redundancy",
                "Dimension Definition": "Evaluates
                     whether the reasoning process
                     is concise and avoids
                     repetition or irrelevant
                     steps.",
                "Score Range (Optional)": "0~1",
                "Scoring Rules (Optional)": "If
                     there is excessive repetition,
                     redundancy, or irrelevant
                     content, the score should be
                     lower; if the reasoning is
                     succinct and well-structured,
                     a higher score should be
                     given."
        },
        {
                "Dimension Name": "Task
                     Recognition",
                "Dimension Definition": "Evaluates
                     whether the model correctly
                     understands the question and
                     continuously identifies which
                     task type it belongs to during
                     the reasoning process. The
                     model should accurately
                     distinguish among Open-Domain
                     QA, Summarization, Chat, Text
                     Completion, and Heuristic Text
                     Generation, and reflect this
                     understanding in its reasoning
                     steps.",
                "Score Range (Optional)": "
                     Open-Domain QA, Summarization,
                     Chat, Text Completion,
                     Heuristic Text Generation",
                "Scoring Rules (Optional)": null
        }
        ]
}
```

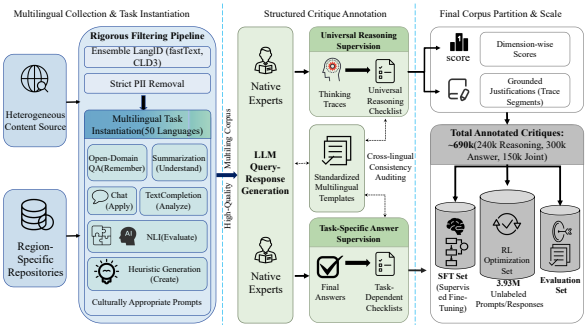


Figure 4: Workflow of Data Construction.

E Data Construction Pipeline

F Reward Model Verification and Reliability Assessment

To verify the usability and reliability of the reward model in RL, we first conduct experiments across multiple base models to ensure that the model can accurately evaluate the desired behaviors. Specifically, we construct a comprehensive instruction-tuning dataset using all SFT data from the open-source VEEF-Multi-LLM (Sha et al., 2025) collection, 30% of the CheckList-related RLVR data built in this work, and 30% of the R1-Thinking-

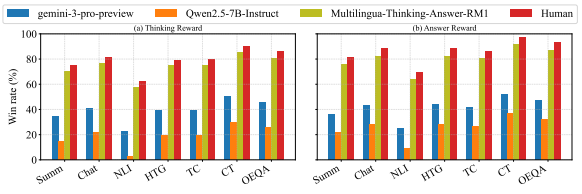


Figure 5: Average win rate of seven task types across fifty languages under different judge models, compared with human evaluation.

50L-240K dataset. We then perform SFT fine-tuning on the base models to obtain an initial reward model, the resulting model is denoted as Multilingua-Thinking-Answer-RM1.

After fine-tuning, this model is used as the reward model for subsequent RL training. To assess its real-world performance, we further sample a validation set from the remaining 70% of the data, including both samples containing explicit reasoning processes and samples containing only final answers. The validation set covers 50 languages and 7 tasks, with 50 instances per language-task pair. We additionally compare the model's judgments with human evaluations, as summarized in Figure 5.

From Table 4, we observe that even though the reward model is fine-tuned using only 30% of the data from two validation tasks, its overall correlation with human judgments exceeds 80%, showing a level of agreement comparable to human annotators. The performance variance across different tasks is also relatively small, indicating the model's applicability as an evaluation agent. These findings align with observations reported in prior work (Zheng et al., 2023; Liu et al., 2023; Shankar et al., 2024).

We hypothesize that, compared with standard instruction-following or question-answering models, a judge model requires richer knowledge structures and stronger generalization ability, as it must explicitly evaluate target attributes such as accuracy, relevance, coherence, tone, or safety. Consequently, high data quality and diversity become essential. Our empirical results further confirm that the constructed datasets are effective and practical for reward model training.

It is worth noting that in subsequent RL stages, the evolving policy model will be used as the new reward model, meaning that the reward model's capability will continue to improve alongside policy optimization. We expect its performance to surpass that of the current initialization stage.

G Length-Aware Loss Normalization

In multilingual scenarios, the same semantic content can yield markedly different token lengths across languages due to structural and morphological variability. Such discrepancies may introduce unintended length-related biases during optimization. To mitigate this effect, we apply a length-aware loss normalization strategy: we aggregate the token-level losses across all generations within a group and divide this sum by the total number of produced tokens, $\sum_{i=1}^{G} |o_i|$. This procedure ensures a fair contribution from each generation and leads to more stable and unbiased training dynamics.

H Language Coverage of Evaluation Datasets

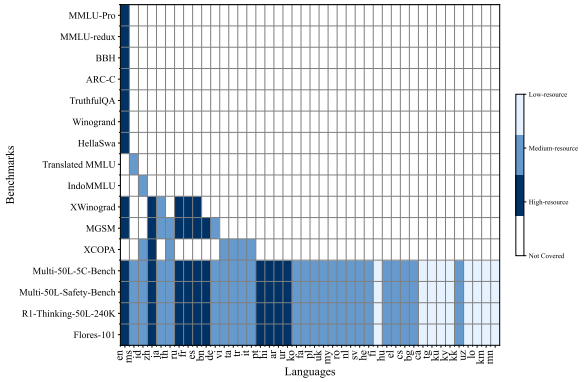


Figure 6: Language distribution across the 13 open-source evaluation datasets. Darker colors indicate lower-resource languages, while lighter colors represent higher- resource languages.

I Checklists Reward Evaluation Instruction

In Section 4, we describe the construction of a dataset categorized into seven major domains. For each of these categories, we have developed a specialized checklist-based evaluation system that is customized for the final answers. Our checklist construction protocol is informed by prior checklist-style evaluation frameworks, particularly HelloEval in HelloBench (Que et al., 2024). We adapt this general paradigm to low-resource multilingual reward modeling by separating universal reasoning dimensions from task-specific answer criteria, and by converting the resulting structured critiques

into reward signals for SGCM and AMRO, and is divided into the following steps:

- **Literature Review:** First, we conducted a comprehensive literature review on the higher-level categories associated with each subcategory. From this, we extracted existing evaluation frameworks and standards that have been established in previous works.
- **Expert Discussions:** Subsequently, we invited five native-language annotators from the relevant language groups to discuss the evaluation criteria. Based on their input, we compiled a list of the top ten most important evaluation standards as identified by the experts.
- **Standard Normalization:** We then combined these evaluation standards and removed any redundant or closely related criteria to ensure a concise and effective set of evaluation dimensions.
- **Voting and Selection:** Afterward, we invited ten annotators to vote on the importance of each evaluation criterion. Based on the voting results, only the top five criteria, as ranked by the annotators, were retained for the final evaluation protocol.
- **Final Checklist Generation:** Finally, we used the powerful GPT-5 model to expand each evaluation dimension into actionable and continuous scoring criteria (ranging from 0 to 1). These criteria were converted into checklists that are easy to follow and implement.

I.1 Open-Domain Question Answering

After constructing fine-grained checklists for the subcategories in the ODQA task, we found that the process of designing specialized checklists for each subcategory was not significantly different from constructing a general checklist at the higher category level. Additionally, we observed that for ODQA, designing specialized checklists for each question domain (e.g., Technology, Sport, Movie, Book, Music, Food, Health, Writing, Science, Travel) was unnecessary. In the voting phase, related to domain-specific checklists, there were relatively few votes for certain domain-specific criteria. This indicated that the domain of the question did not significantly impact the final evaluation outcome, this is consistent with previous work (Hou et al., 2024).

Consequently, we opted to use the same checklist for all subcategories. Moreover, to ensure the checklists are as close as possible to real-world evaluation scenarios, we made sure that each checklist was complex enough to allow for more detailed differentiation in the evaluation results, particularly when using LLM-as-Judge for assessments. The final checklists for ODQA are presented in Table I.1

ODQA Reward Evaluation Prompt

```
{
    "Evaluation_dimensions": [
        {
            "dimension_name": "Relevance",
            "dimension_definition": "Evaluates
                whether the response not only
                directly addresses the question
                but also ensures that every part
                of the response is strictly
                related to the topic of the
                question. Evaluate each sentence
                and paragraph rigorously to
                confirm that it is entirely
                relevant to the topic and does not
                deviate in any way.",
            "scoring_range": "0-1",
            "scoring_criteria": "If even a single
                part of the response is slightly
                unrelated, redundant, or lacking
                personal perspective when
                required, the score is 0. If the
                content is fully relevant and
                comprehensive, the score is 1.
                Scores between 0 and 1 should be
                given if there are minor
                irrelevant parts or a lack of
                necessary personal perspective."
        },
        {
            "dimension_name": "Factual Accuracy",
            "dimension_definition": "Evaluates
                whether every aspect of the
                response is factually correct. For
                example, when listing historical
                information, are all historical
                figures, dates, and events
                precisely accurate? When
                presenting scientific terms or
                phenomena, are they completely
                accurate and up-to-date?",
            "scoring_range": "0-1",
            "scoring_criteria": "If any part of the
                response contains factual errors
                or inaccuracies, the score is 0.
                If the response is completely
                factually correct, the score is 1.
                Scores between 0 and 1 should be
                given if there are minor errors or
                vague statements."
        },
        {
            "dimension_name": "Clarity",
            "dimension_definition": "Evaluates
                whether the content of the
                response is easy to understand.
                For difficult technical terms, are
                there corresponding explanations
                and examples provided? Are more
                complex terms replaced with
                simpler ones?",
            "scoring_range": "0-1",
            "scoring_criteria": "If there is
                content that is difficult to
                understand and no effort is made
                to simplify or explain, the score
                is 0. If the content is concise,
                clear, and easy to understand, the
                score is 1. Scores between 0 and 1
                should be given if some parts
                could be optimized or simplified
                further."
        },
        {
            "dimension_name": "Engagement and
                Novelty",
            "dimension_definition": "Evaluates
```

```
                whether the content of the
                response is interesting or novel.
                Since the questions are
                Open-Domain, the responses can
                vary. An excellent response should
                offer unique viewpoints or
                interesting content. Does the
                response provide a fresh
                perspective?",
            "scoring_range": "0-1",
            "scoring_criteria": "If the response
                lacks novelty or is uninteresting,
                the score is 0. If the response
                provides a fresh or interesting
                perspective, the score is 1.
                Scores between 0 and 1 should be
                given if the response lacks
                innovation but still offers
                valuable information."
        },
        {
            "dimension_name": "Content Depth",
            "dimension_definition": "Evaluates
                whether the content of the
                response is rich and detailed,
                with no fewer than 500 words. Does
                each point include multiple
                well-explained examples or
                explanations for strong support?",
            "scoring_range": "0-1",
            "scoring_criteria": "If the response is
                less than 500 words or lacks
                sufficient detail and examples,
                the score is 0. If the response is
                detailed, well-supported with
                examples, and meets the length
                requirement, the score is 1.
                Scores between 0 and 1 should be
                given if the response is lacking
                in depth or missing sufficient
                examples."
        },
        {
            "dimension_name": "Human-like Quality",
            "dimension_definition": "Evaluates
                whether the content of the
                response appears human-like. The
                response should not seem
                machine-generated. Evaluate
                whether the structure, tone, and
                flow reflect a human touch.",
            "scoring_range": "0-1",
            "scoring_criteria": "If the response
                appears machine-generated or has
                unnatural structures, the score is
                0. If the response is natural,
                flowing, and contains human-like
                expressions, the score is 1.
                Scores between 0 and 1 should be
                given if there are minor
                machine-like traits but the
                response remains largely
                human-like."
        },
        {
            "dimension_name": "Perfection",
            "dimension_definition": "Evaluates
                whether the response is flawless.
                If there is room for improvement,
                the response should not be
                considered perfect.",
            "scoring_range": "0-1",
            "scoring_criteria": "If the response
                has any room for improvement or
                omissions, the score is 0. If the
                response is flawless, the score is
                1. Scores between 0 and 1 should
                be given if the response has minor
                areas for improvement but is
                overall excellent."
        }
    ]
}
```

I.2 Summ

In summarization tasks, due to the significant differences between each subcategory, we developed distinct evaluation criteria in the form of checklists for each subcategory to ensure more accurate assessment. Table I.2.1 represents the evaluation criteria for News Summarization, Table I.2.2 for Blog Summarization, Table I.2.3 for Academic Article Summarization, Table I.2.4 for Report Summarization, and Table I.2.5 for Long Dialogue Summarization.

I.2.1 News Summ

News Summ Reward Evaluation Prompt

```
{
    "evaluation_dimensions": [
        {
            "dimension_name": "Clarity and
                Comprehensibility",
            "dimension_definition": "This dimension
                assesses how easy it is to
                understand the content of the
                summary. It also checks if
                difficult technical terms are
                explained with examples or
                simplified, and if the summary is
                written in a way that makes the
                content easy to grasp.",
            "score_range": "0-1",
            "scoring_rule": "0: The summary is
                difficult to understand, with many
                technical terms left unexplained
                or overly complex; 1: The summary
                is clear and easy to understand,
                with technical terms properly
                explained and simplified when
                necessary."
        },
        {
            "dimension_name": "Summary Length and
                Completeness",
            "dimension_definition": "This dimension
                evaluates whether the summary is
                long enough and sufficiently
                comprehensive to cover the key
                information from the original
                news, particularly for lengthy
                sources.",
            "score_range": "0-1",
            "scoring_rule": "0: The summary is too
                short, missing key information or
                leaving out critical parts of the
                original news; 1: The summary is
                long enough and covers all the key
                information from the original
                news, maintaining completeness."
        },
        {
            "dimension_name": "Accuracy and
                Objectivity",
            "dimension_definition": "This dimension
                checks whether the summary
                accurately reflects the original
                news, with no deviations,
                additions, or personal opinions.
                It also ensures that statistical
                information and data are
                consistent with the original
                news.",
            "score_range": "0-1",
            "scoring_rule": "0: The summary
                contains inaccuracies, added
                details, or biased
                interpretations; 1: The summary is
                perfectly accurate and unbiased,
                matching the original news without
                any deviation."
        },
        {
            "dimension_name": "Coverage of
                Important Information",
            "dimension_definition": "This dimension
                evaluates if the summary
                comprehensively includes all
                significant details from the
                original news, such as when and
                where the event took place, who
                was involved, and what happened.",
            "score_range": "0-1",
            "scoring_rule": "0: The summary omits
                important information, such as key
                facts or details about the event; 1
                : The summary covers all the
                essential details, including when,
                where, who, and what."
        },
```

```
        {
            "dimension_name": "Adherence to User
                Instructions",
            "dimension_definition": "This dimension
                checks if the summary fully
                adheres to the requirements
                outlined in the user's
                instructions, including specific
                formatting or content
                preferences.",
            "score_range": "0-1",
            "scoring_rule": "0: The summary does
                not meet the user's instructions
                or includes irrelevant content; 1:
                The summary perfectly follows the
                user's instructions, fulfilling
                all specified requirements."
        },
        {
            "dimension_name": "Overall Quality and
                Room for Improvement",
            "dimension_definition": "This dimension
                assesses whether the summary is
                flawless or if there are areas
                that can be improved, ensuring
                that the overall quality is
                considered in the evaluation.",
            "score_range": "0-1",
            "scoring_rule": "0: The summary has
                notable flaws or areas for
                improvement, such as missing key
                details or unclear phrasing; 1:
                The summary is flawless with no
                room for improvement, representing
                a well-crafted response."
        }
    ]
}
```

I.2.2 Blog Summ

Blog Summ Reward Evaluation Prompt

```
{
    "evaluation_dimensions": [
        {
            "dimension_name": "Clarity of Content",
            "dimension_definition": "Evaluates
                whether the content of the summary
                is easy to understand. This
                includes checking if difficult
                technical terms are explained with
                examples, and if complex terms are
                simplified. The summary should be
                clear and understandable,
                evaluated word by word and
                paragraph by paragraph.",
            "score_range": "0 - 1",
            "scoring_rules": "Score 1 if the
                content is entirely clear and easy
                to understand with sufficient
                explanations for technical terms.
                Score 0.5 if some terms are
                unclear but explanations are
                provided. Score 0 if there are
                significant parts of the summary
                that are difficult to understand
                or unclear."
        },
        {
            "dimension_name": "Completeness of the
                Summary",
            "dimension_definition": "Assesses
                whether the summary is
                sufficiently long and covers all
                key information from the original
                source. It should be long enough
                to summarize all essential points
                from the original blog.",
            "score_range": "0 - 1",
            "scoring_rules": "Score 1 if the
                summary covers all important
                points and is appropriately
                lengthy. Score 0.5 if some key
                points are missing or the summary
                feels too short. Score 0 if the
                summary lacks significant content
                or fails to capture essential
                information."
        },
        {
            "dimension_name": "Accuracy of the
                Summary",
            "dimension_definition": "Checks if the
                summary is perfectly accurate,
                strictly matching the original
                blog without adding personal
                opinions or altering any
                information. Statistical data and
                factual content must match exactly
                with the original blog.",
            "score_range": "0 - 1",
            "scoring_rules": "Score 1 if the
                summary is perfectly accurate,
                with no deviations or additional
                information. Score 0.5 if there
                are minor inconsistencies or
                slight deviations. Score 0 if the
                summary contains significant
                errors or added information not
                present in the original."
        },
        {
            "dimension_name": "Coverage of
                Important Information",
            "dimension_definition": "Evaluates
                whether the summary
                comprehensively covers all the
                important aspects of the original
                blog, including the main topic,
                primary arguments, and supporting
                details.",
            "score_range": "0 - 1",
            "scoring_rules": "Score 1 if all
                critical information, arguments,
                and supporting details are
                included. Score 0.5 if some key
                aspects are missing, but the
                summary still captures most
                important content. Score 0 if
                important information or key
                points are missing or
                misrepresented."
        },
        {
            "dimension_name": "Adherence to User
                Instructions",
            "dimension_definition": "Assesses
                whether the summary meets all the
                specific requirements outlined in
                the user's instructions. This
                could include specific formatting,
                content inclusion, or length.",
            "score_range": "0 - 1",
            "scoring_rules": "Score 1 if the
                summary fully adheres to all the
                user instructions. Score 0.5 if
                minor instructions were missed or
                ignored. Score 0 if major
                instructions or requirements are
                not followed."
        },
        {
            "dimension_name": "Overall Quality",
            "dimension_definition": "Evaluates
                whether the summary is flawless
                and leaves no room for
                improvement. This dimension
                considers all other evaluation
                factors and assesses the overall
                performance of the summary.",
            "score_range": "0 - 1",
            "scoring_rules": "Score 1 if the
                summary is flawless with no room
                for improvement. Score 0.5 if
                there are some minor areas for
                improvement. Score 0 if the
                summary has significant issues or
                can be greatly improved."
        }
    ]
}
```

I.2.3 Academic Article Summ

Academic Article Summ Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Clarity of Content",
            "Dimension Definition": "Evaluates
                whether the content of the summary
                is easy to understand for a
                general academic audience.
```

```
            Difficult-to-understand technical
                terms should have corresponding
                explanations and examples. Complex
                terms should be replaced with
                simpler ones. Every part of the
                summary should be easy to
                understand, evaluated word by word
                and paragraph by paragraph. Any
                content that can be optimized for
                conciseness or clarity should be
                considered as not easy to
                understand.",
        "Scoring Range": "0-1",
        "Scoring Rules": "0: The content is
                difficult to understand or
                contains many complex terms
                without explanations. 1: The
                content is clear and easy to
                understand, with technical terms
                explained or replaced with simpler
                ones."
    },
    {
        "Dimension Name": "Length and
                Completeness",
        "Dimension Definition": "Evaluates
                whether the summary is
                sufficiently long and complete,
                covering key information from the
                original article, considering the
                original article's length.",
        "Scoring Range": "0-1",
        "Scoring Rules": "0: The summary is too
                short or omits key information. 1:
                The summary covers all necessary
                information in a sufficient
                length, capturing the essence of
                the original article."
    },
    {
        "Dimension Name": "Accuracy",
        "Dimension Definition": "Evaluates
                whether the summary is perfectly
                accurate without errors or
                misleading information. Every
                statement in the summary must
                match the original article, with
                no additions or deviations. All
                statistical information and data
                must be identical to the original
                article.",
        "Scoring Range": "0-1",
        "Scoring Rules": "0: The summary
                contains errors, additions, or
                inconsistencies with the original
                article. 1: The summary is
                perfectly accurate with no
                deviations or misleading
                information."
    },
    {
        "Dimension Name": "Comprehensive
                Coverage",
        "Dimension Definition": "Evaluates
                whether the summary
                comprehensively covers all
                important information from the
                original article, including the
                research background, methods,
                findings, results, and
                conclusions.",
        "Scoring Range": "0-1",
        "Scoring Rules": "0: The summary misses
                significant information or fails
                to cover key parts of the article.
                1: The summary fully covers all
                important aspects, including
                background, methods, results, and
                conclusions."
    },
    {
        "Dimension Name": "Instruction
                Compliance",
        "Dimension Definition": "Evaluates
                whether the summary meets all the
                requirements specified in the user
                instruction.",
        "Scoring Range": "0-1",
        "Scoring Rules": "0: The summary does
                not meet the requirements
                specified in the instructions. 1:
                The summary fully meets all the
                requirements and guidelines
                provided."
    },
    {
        "Dimension Name": "Flawlessness",
        "Dimension Definition": "Evaluates
                whether the summary is flawless
                and if there is any room for
                improvement. This includes
                evaluating the overall quality and
                whether it can be further
                refined.",
        "Scoring Range": "0-1",
        "Scoring Rules": "0: The summary has
                significant flaws or room for
                improvement. 1: The summary is
                flawless with no room for
                improvement."
    }
]
}
```

I.2.4 Report Summ

Report Summ Reward Evaluation Prompt

```
{
    "evaluation_dimensions": [
        {
            "dimension_name": "Clarity of Content",
            "dimension_definition": "This dimension
                assesses whether the summary is
                easy to understand. For
                difficult-to-understand technical
                terms, explanations and examples
                should be provided. More complex
                terms should be simplified. Each
                part of the summary should be
                evaluated word by word and
                paragraph by paragraph to
                determine its clarity. Any content
                that can be optimized to be more
                concise or clearer should reduce
                the clarity score.",
            "scoring_range": "0-1",
            "scoring_rule": "Score 1 if the summary
                is easy to understand with clear
                definitions for technical terms
                and no complex terms remaining.
                Score 0 if the summary is
                difficult to understand or lacks
                explanations for technical terms."
        },
        {
            "dimension_name": "Length and
                Completeness",
            "dimension_definition": "This dimension
                evaluates whether the summary is
                sufficiently long and covers all
                the necessary information. Given
                that the original report is
                lengthy, the summary should be
                long enough to include the key
                details of the report.",
            "scoring_range": "0-1",
            "scoring_rule": "Score 1 if the summary
                is sufficiently long to cover all
                key points from the original
                report. Score 0 if the summary is
                too short and misses key details."
        },
        {
            "dimension_name": "Accuracy",
            "dimension_definition": "This dimension
                checks whether the summary is
                perfectly accurate. Every
                statement in the summary must
                match the original report without
                additions or deviations. All
                statistical information and data
                must be identical to those in the
                original report.",
            "scoring_range": "0-1",
            "scoring_rule": "Score 1 if the summary
                is perfectly accurate with no
                discrepancies or additional
                information. Score 0 if there are
                any inconsistencies or added
                content not present in the
                original report."
        },
        {
            "dimension_name": "Coverage of Key
                Information",
            "dimension_definition": "This dimension
                assesses whether the summary
                comprehensively includes all the
                important information from the
```

```
                original report, including key
                    statistical information,
                    recommendations, and conclusions.",
            "scoring_range": "0-1",
            "scoring_rule": "Score 1 if the summary
                    covers all the critical
                    information from the original
                    report. Score 0 if any important
                    aspect or key information is
                    missing."
        },
        {
            "dimension_name": "Adherence to
                    Instructions",
            "dimension_definition": "This dimension
                    evaluates whether the summary
                    meets all the requirements
                    specified in the user instruction.
                    The summary must follow all the
                    specific guidelines provided by
                    the user.",
            "scoring_range": "0-1",
            "scoring_rule": "Score 1 if the summary
                    meets all the requirements stated
                    in the user instructions. Score 0
                    if it fails to follow any of the
                    instructions."
        },
        {
            "dimension_name": "Flawlessness",
            "dimension_definition": "This dimension
                    evaluates whether the summary is
                    flawless or if there is any room
                    for improvement. It considers
                    whether the summary is of the
                    highest quality and if there is
                    any space for further refinement.",
            "scoring_range": "0-1",
            "scoring_rule": "Score 1 if the summary
                    is flawless with no room for
                    improvement. Score 0 if there are
                    areas where the summary can be
                    improved or refined."
        }
    ]
}
```

I.2.5 Long Dialogue Summ

Long Dialogue Summ Reward Evaluation Prompt

```
{
    "evaluation_dimensions": [
        {
            "dimension_name": "Clarity of Content",
            "dimension_definition": "Evaluates
                    whether the summary is easy to
                    understand, including the presence
                    of explanations and examples for
                    complex terms, and whether complex
                    terms are simplified. The summary
                    should be evaluated for clarity at
                    the word and paragraph level.",
            "score_range": "0-1",
            "scoring_rules": "0: The content is
                    difficult to understand, with
                    little to no simplification or
                    explanation. 0.5: Some parts are
                    understandable, but there are
                    sections with jargon or complex
                    terms not sufficiently explained. 1
                    : The content is clear, with
                    complex terms explained and
                    simplified where necessary."
        },
        {
            "dimension_name": "Completeness",
            "dimension_definition": "Assesses
                    whether the summary is
                    sufficiently long and covers all
                    key information from the original
                    dialogue, especially considering
                    the length of the original
                    dialogue.",
            "score_range": "0-1",
            "scoring_rules": "0: The summary is too
                    short and misses critical
                    information. 0.5: The summary
                    covers some key points but omits
                    others. 1: The summary is
                    sufficiently detailed and covers
                    all key aspects of the original
                    dialogue."
        },
        {
            "dimension_name": "Accuracy",
            "dimension_definition": "Evaluates
                    whether the summary is completely
                    accurate and matches the original
                    dialogue without errors or
                    misleading information. No
                    additions, deviations, or
                    inconsistencies are allowed.",
            "score_range": "0-1",
            "scoring_rules": "0: The summary
                    contains inaccuracies, errors, or
                    misleading information. 0.5: The
                    summary is mostly accurate but
                    contains minor errors or
                    omissions. 1: The summary is
                    perfectly accurate and matches the
                    original dialogue with no
                    deviations."
        },
        {
            "dimension_name": "Comprehensiveness",
            "dimension_definition": "Assesses
                    whether the summary covers all the
                    important information from the
                    original dialogue, including key
                    topics and viewpoints from all
                    roles involved in the
                    conversation.",
            "score_range": "0-1",
            "scoring_rules": "0: The summary omits
                    significant portions of the
                    dialogue and key viewpoints. 0.5:
                    The summary covers most of the
                    important points but omits some
                    viewpoints or topics. 1: The
                    summary comprehensively includes
                    all key topics and viewpoints from
                    the original dialogue."
        },
        {
            "dimension_name": "Exclusion of
                    Redundant Information",
            "dimension_definition": "Evaluates
                    whether the summary excludes all
                    redundant information, unnecessary
                    filler words, and irrelevant
                    interjections, without omitting
                    key points or altering the
                    original meaning and context.",
            "score_range": "0-1",
            "scoring_rules": "0: The summary
                    includes irrelevant information or
                    filler words that do not
                    contribute to the key message. 0
                    .5: The summary is mostly concise,
                    but some redundant information is
                    included. 1: The summary is
                    concise and excludes all
                    unnecessary content, maintaining
                    the original meaning and context."
        },
        {
            "dimension_name": "Adherence to User
                    Instructions",
            "dimension_definition": "Assesses
                    whether the summary meets all the
                    requirements specified in the user
                    instructions.",
            "score_range": "0-1",
            "scoring_rules": "0: The summary does
                    not meet the user's requirements
                    or instructions. 0.5: The summary
                    meets some but not all user
                    requirements. 1: The summary
                    perfectly aligns with all user
                    instructions and specifications."
        },
        {
            "dimension_name": "Overall Quality",
            "dimension_definition": "Evaluates the
                    overall quality of the summary
                    based on whether there is room for
                    improvement. This is a holistic
                    evaluation of the summary's
                    effectiveness.",
            "score_range": "0-1",
            "scoring_rules": "0: The summary has
                    significant issues and requires
                    substantial improvement. 0.5: The
                    summary is generally good but has
                    room for minor improvements. 1:
                    The summary is flawless, with no
                    significant areas for improvement."
        }
```

```
    ]
}
```

I.3 Text Completion

For text completion tasks, we have designed three distinct evaluation criteria for the subcategories of Continuation, Imitative Writing, and Style Transfer. Tables I.6.1, I.3.2, and I.3.3 present the checklists corresponding to each of these subcategories in the text completion tasks.

I.3.1 Continuation Text Completion

Continuation Text Completion Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Narrative Coherence",
            "Dimension Definition": "Assesses
                whether the continuation maintains
                consistency with the preceding
                text, ensuring coherence in plot,
                character development, tone,
                pacing, themes, and any subtle
                nuances introduced earlier.",
            "Score Range": "0-1 (0: Inconsistent, 1:
                 Seamless consistency)",
            "Scoring Rules": "A score of 1 should
                be awarded if the continuation
                flows naturally, preserving plot
                and character integrity without
                disrupting the narrative. A score
                of 0 should be given if there are
                major inconsistencies or
                contradictions in plot, tone, or
                character behavior."
        },
        {
            "Dimension Name": "Engagement and
                Interest",
            "Dimension Definition": "Evaluates
                whether the continuation is
                engaging and compelling, adding
                depth to the storyline and
                characters while maintaining the
                reader's attention.",
            "Score Range": "0-1 (0: Boring, 1:
                Highly engaging)",
            "Scoring Rules": "A score of 1 should
                be awarded if the continuation
                keeps the reader's attention,
                develops the plot and characters
                in a meaningful way, and maintains
                curiosity. A score of 0 should be
                given if the continuation fails to
                engage the reader or adds nothing
                of interest to the storyline."
        },
        {
            "Dimension Name": "Comprehensiveness",
            "Dimension Definition": "Assesses
                whether the continuation is
                sufficiently long and
                well-rounded, integrating smoothly
                with the preceding text to form a
                complete story.",
            "Score Range": "0-1 (0: Incomplete, 1:
                Comprehensive)",
            "Scoring Rules": "A score of 1 should
                be awarded if the continuation is
                long enough to provide a complete,
                well-developed arc with clear
                character development and a
                satisfying resolution. A score of 0
                 should be given if the
                continuation feels rushed,
                incomplete, or lacks proper
                development."
        },
        {
            "Dimension Name": "Originality and
                Novelty",
            "Dimension Definition": "Evaluates how
                original and creative the
                continuation is, introducing new
                ideas and perspectives without
                relying on cliches or predictable
                plot developments.",
            "Score Range": "0-1 (0: Predictable, 1:
                Exceptionally original)",
            "Scoring Rules": "A score of 1 should
                be awarded if the continuation
                introduces fresh, unique ideas or
                perspectives, avoiding overused
                tropes or cliches. A score of 0
                should be given if the
                continuation feels formulaic,
                derivative, or predictable."
        },
        {
            "Dimension Name": "Flawlessness",
            "Dimension Definition": "Assesses
                whether the continuation is free
                from errors and improvements,
                considering whether there is any
                room for improvement in the
                quality of the story.",
            "Score Range": "0-1 (0: Flawed, 1:
                Flawless)",
            "Scoring Rules": "A score of 1 should
                be awarded if the continuation is
                flawless, with no noticeable
                errors or areas requiring
                improvement. A score of 0 should
                be given if there are clear issues
                in the story, such as logical
                inconsistencies, weak character
                development, or unfinished plot
                points."
        }
    ]
}
```

I.3.2 Imitative Writing Text Completion

Imitative Writing Text Completion Reward Evaluation Prompt

```
{
    "evaluation_dimensions": [
        {
            "dimension_name": "Writing Voice and
                Stylistic Consistency",
            "dimension_definition": "Assesses
                whether the generated text
                captures the distinct writing
                voice and intricate stylistic
                nuances of the preceding text,
                while seamlessly integrating these
                elements into a new story theme.
                It evaluates the consistency in
                tone, complexity, and emotional
                resonance throughout the new
                text.",
            "scoring_range": "0-1",
            "scoring_rule": "A score of 1 is
                awarded when the generated text
                consistently reflects the same
                writing voice and stylistic
                elements as the preceding text. A
                score of 0 is awarded if the
                generated text fails to maintain
                consistency in tone, complexity,
                or emotional resonance."
        },
        {
            "dimension_name": "Engagement and
                Intrigue",
            "dimension_definition": "Evaluates
                whether the content of the
                generated text is not only
                engaging and compelling but also
                reflective of the same level of
                intrigue and interest found in the
                preceding text.",
            "scoring_range": "0-1",
            "scoring_rule": "A score of 1 is
                awarded if the generated content
                is highly engaging and maintains a
                comparable level of intrigue to
                the preceding text. A score of 0
                is awarded if the generated
                content is uninteresting or fails
                to match the intrigue of the
```

```
                original."
        },
        {
            "dimension_name": "Content Depth and
                Completeness",
            "dimension_definition": "Assesses
                whether the content of the
                generated text is sufficiently
                lengthy, complete, and
                meticulously detailed, matching
                the depth, comprehensiveness, and
                narrative complexity of the
                preceding text.",
            "scoring_range": "0-1",
            "scoring_rule": "A score of 1 is
                awarded if the generated text is
                comprehensive, thoroughly
                developed, and maintains the
                narrative complexity of the
                original text. A score of 0 is
                awarded if the generated content
                is shallow, incomplete, or lacking
                in necessary details."
        },
        {
            "dimension_name": "Creativity and
                Originality",
            "dimension_definition": "Evaluates
                whether the generated text is
                novel and original, while
                creatively distinct and still
                aligned with the stylistic and
                thematic essence of the preceding
                text.",
            "scoring_range": "0-1",
            "scoring_rule": "A score of 1 is
                awarded if the generated text is
                both creative and original,
                maintaining the essence of the
                original text. A score of 0 is
                awarded if the generated content
                is derivative, overly repetitive,
                or lacks originality."
        },
        {
            "dimension_name": "Flawlessness of
                Imitative Writing",
            "dimension_definition": "Evaluates the
                overall quality of the imitative
                writing, determining if there is
                any room for improvement.",
            "scoring_range": "0-1",
            "scoring_rule": "A score of 1 is
                awarded if the imitative writing
                is flawless and requires no
                further improvement. A score of 0
                is awarded if there is noticeable
                room for improvement in the
                writing, whether in terms of
                accuracy, flow, or style."
        }
    ]
}
```

I.3.3 Style Transfer Text Completion

Style Transfer Text Completion Reward Evaluation Prompt

```
{
    "evaluation_dimensions": [
        {
            "dimension_name": "Style and Tone
                Transformation",
            "dimension_definition": "Evaluates
                whether the generated text
                successfully transforms the style
                and tone to the desired target
                style while capturing the
                intricate nuances and essence of
                the new style.",
            "score_range": "0-1",
            "scoring_rules": "A score closer to 1
                is given if the transformation is
                seamless, and the target style is
                fully captured with attention to
                detail. A score of 0 is given if
                the transformation is unsuccessful
                or the nuances of the style are
                missed."
        },
        {
            "dimension_name": "Engagement and
                Interest Level",
            "dimension_definition": "Assesses
                whether the style-transformed text
                remains engaging and compelling
                while maintaining the same level
                of intrigue as the original text,
                incorporating the nuances of the
                new style.",
            "score_range": "0-1",
            "scoring_rules": "A score closer to 1
                is given if the text is engaging
                and the level of interest is
                maintained or increased. A score
                of 0 is assigned if the text loses
                engagement or interest in the new
                style."
        },
        {
            "dimension_name": "Completeness and
                Detail",
            "dimension_definition": "Evaluates
                whether the transformed text is
                sufficiently detailed,
                comprehensive, and matches the
                length and depth of the original
                text.",
            "score_range": "0-1",
            "scoring_rules": "A score of 1 is given
                if the text is well-developed,
                detailed, and of sufficient
                length, fully matching the depth
                of the original text. A score of 0
                is given if the text lacks detail
                or is overly simplistic compared
                to the original."
        },
        {
            "dimension_name": "Originality and
                Creativity",
            "dimension_definition": "Assesses the
                level of novelty and creativity in
                the style-transformed text, while
                ensuring adherence to the new
                style's characteristics.",
            "score_range": "0-1",
            "scoring_rules": "A score of 1 is
                awarded if the text is original,
                creative, and distinct, while
                still faithful to the new style. A
                score of 0 is given if the text is
                derivative or fails to fully
                embrace the new style."
        },
        {
            "dimension_name": "Flawlessness of
                Style Transfer",
            "dimension_definition": "Evaluates
                whether the style transfer is
                flawless, considering if there are
                areas for improvement in the
                transformation process.",
            "score_range": "0-1",
            "scoring_rules": "A score of 1 is given
                if the style transfer is flawless
                with no room for improvement. A
                score closer to 0 is given if
                noticeable improvements could be
                made, particularly in areas where
                the style transformation is
                incomplete or inconsistent."
        }
    ]
}
```

I.4 Natural Language Inference

For the Natural Language Inference task, which includes the three subcategories Entailment, Contradiction, and Neutral, we adopt a unified evaluation framework as presented in Table I.4.

NLI Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Logical Consistency",
```

```
            "Dimension Definition": "Evaluates
                whether the predicted NLI label
                (Entailment, Contradiction,
                Neutral) is logically consistent
                with the relationship between the
                premise and the hypothesis.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the selected label fully aligns
                with the logical relation implied
                by the premise; score 0 if major
                inconsistencies or incorrect
                judgments occur, with intermediate
                values indicating partial
                correctness."
        },
        {
            "Dimension Name": "Evidence Grounding",
            "Dimension Definition": "Assesses
                whether the answer is strictly
                grounded in the information
                contained in the premise without
                introducing hallucinated or
                unsupported content.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                all reasoning is fully supported
                by the premise; score 0 if the
                answer contains fabricated or
                irrelevant information, with
                intermediate scores representing
                partially grounded reasoning."
        },
        {
            "Dimension Name": "Completeness of
                Reasoning",
            "Dimension Definition": "Evaluates
                whether the explanation clearly
                and sufficiently justifies why the
                predicted label holds, reflecting
                a correct understanding of the
                premise-hypothesis relationship.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the explanation is thorough,
                structured, and covers all
                necessary reasoning steps; score 0
                if the justification is incomplete
                or unclear, with intermediate
                scores indicating partial
                completeness."
        },
        {
            "Dimension Name": "Sensitivity to
                Semantic Nuances",
            "Dimension Definition": "Measures the
                model's ability to correctly
                interpret subtle semantic cues
                such as negation, modality,
                quantifiers, presuppositions, and
                nuanced linguistic shifts that
                affect the NLI label.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the answer correctly identifies
                and interprets all key semantic
                nuances; score 0 if such
                distinctions are consistently
                overlooked or misinterpreted, with
                intermediate values reflecting
                partial sensitivity."
        },
        {
            "Dimension Name": "Label Precision and
                Category Boundary Clarity",
            "Dimension Definition": "Evaluates
                whether the model accurately
                distinguishes among Entailment,
                Contradiction, and Neutral,
                especially in borderline or
                ambiguous cases.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the model demonstrates clear
                category boundaries and precise
                labeling; score 0 if it confuses
                categories or misclassifies them,
                with intermediate scores capturing
                partial precision."
        },
        {
            "Dimension Name": "Overall Robustness",
            "Dimension Definition": "Assesses the
                answer for overall reliability,
                ensuring it is free from logical
                fallacies, unsupported
                assumptions, oversights, or
                internal inconsistencies.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the reasoning is consistently
                sound and free from errors; score 0
                 if major logical flaws or
                contradictions are present, with
                intermediate values indicating
                minor issues."
        }
    ]
}
```

I.5 Chat

Our chat task dataset spans multiple subcategories, including code-oriented script drafting, creative ideation, curriculum design, and role-playing interactions. To ensure consistency and broad applicability, we develop a unified, user-centered evaluation framework that provides a standardized assessment protocol for all chat scenarios. This checklist consists of five distinct evaluation dimensions, each capturing a different aspect of response quality. The complete specifications are presented in Table I.5.

Chat Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Instruction
                Comprehension and Coverage",
            "Dimension Definition": "Evaluates
                whether the response fully
                understands every aspect of the
                user's instructions and addresses
                each requirement accurately,
                precisely, and without omissions
                or misunderstandings.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the response fully covers all user
                instructions with no missing or
                misinterpreted parts. Score 0 if
                any requirement is misunderstood,
                partially addressed, or omitted."
        },
        {
            "Dimension Name": "Completeness and
                Sufficiency",
            "Dimension Definition": "Assesses
                whether the response is
                sufficiently long, detailed, and
                comprehensive, ensuring that all
                components of the requirement are
                addressed without leaving out
                important information.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the response is fully
                comprehensive and includes all
                necessary details. Score 0 if any
                component is incomplete, overly
                brief, or insufficiently
                addressed."
        },
        {
            "Dimension Name": "Clarity and
                Understandability",
            "Dimension Definition": "Measures
                whether the response is easy to
                understand at both word-level and
                paragraph-level. Complex technical
                terms must be explained with
                examples or replaced with simpler
                expressions. Any unnecessarily
                complex or unclear content results
                in a failing score.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the content is entirely clear,
```

```
                intuitive, and easy to follow,
                with proper explanations for
                technical terms. Score 0 if any
                part of the response can be made
                simpler, clearer, or more concise."
        },
        {
            "Dimension Name": "Factual Accuracy",
            "Dimension Definition": "Examines
                whether every part of the response
                is factually correct. Historical
                facts, scientific concepts, dates,
                events, and terminology must be
                fully accurate and up-to-date.
                Even minor factual issues cause
                failure.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                every detail is perfectly accurate
                and verified. Score 0 if any
                factual error, outdated statement,
                or uncertain claim appears
                anywhere in the response."
        },
        {
            "Dimension Name": "Overall Perfection",
            "Dimension Definition": "Evaluates
                whether the response is flawless
                as a whole, with no room for
                improvement in clarity,
                correctness, completeness, or
                execution.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the response is judged flawless,
                with no identifiable weaknesses.
                Score 0 if any aspect-however
                small-could be improved."
        }
    ]
}
```

I.6 Text Completion

For the Text Completion task, we develop dedicated evaluation frameworks for three subcategories. The checklist presented in Table I.6.1 corresponds to the Continuation subcategory; Table I.6.2 provides the evaluation framework for the Imitative Writing subcategory; and Table I.6.3 outlines the evaluation criteria for the Style Transfer subcategory.

I.6.1 Continuation

Continuation Text Completion Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Narrative Coherence",
            "Dimension Definition": "Evaluates
                whether the continuation maintains
                seamless coherence with the
                preceding text, including
                consistency in plot progression,
                character behavior, tone, pacing,
                thematic alignment, and subtle
                nuances already established in the
                story.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the continuation is fully coherent
                and consistent with all
                established narrative elements;
                score 0 if major inconsistencies
                or disruptions in tone,
                characterization, or plot occur."
        },
        {
            "Dimension Name": "Engagement and
                Interest",
            "Dimension Definition": "Assesses
                whether the continuation is
                engaging, compelling, and
                immersive, adding narrative depth
                while sustaining the reader's
                curiosity and emotional
                involvement throughout the text.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the continuation maintains strong
                reader engagement and narrative
                momentum; score 0 if it feels
                dull, unengaging, or fails to
                sustain interest."
        },
        {
            "Dimension Name": "Completeness and
                Structural Integrity",
            "Dimension Definition": "Examines
                whether the continuation provides
                sufficient length, detail, and
                completeness, forming a coherent
                and well-developed story when
                combined with the preceding text,
                including clear plot arcs,
                character development, and a
                satisfying resolution.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the continuation integrates
                smoothly into a complete narrative
                with well-resolved story elements;
                score 0 if it is incomplete,
                underdeveloped, or structurally
                weak."
        },
        {
            "Dimension Name": "Originality and
                Creativity",
            "Dimension Definition": "Measures the
                degree to which the continuation
                introduces novel, original, and
                imaginative ideas while avoiding
                cliches, predictable developments,
                and uncreative reuse of familiar
                tropes.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the continuation demonstrates
                strong creativity and originality;
                score 0 if it lacks novelty or
                relies heavily on predictable or
                derivative elements."
        },
        {
            "Dimension Name": "Overall Quality and
                Flawlessness",
            "Dimension Definition": "Assesses
                whether the continuation exhibits
                minimal flaws and whether there is
                substantial room for improvement
                across narrative clarity,
                coherence, style, creativity, or
                structural quality.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the continuation is nearly
                flawless and shows no significant
                issues; score 0 if noticeable
                weaknesses or areas for
                improvement remain."
        }
    ]
}
```

I.6.2 Imitative Writing

Imitative Writing Text Completion Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Stylistic Fidelity",
            "Dimension Definition": "Evaluates
                whether the generated text
                accurately captures the distinct
                writing voice, subtle stylistic
                nuances, and emotional resonance
                of the preceding text while
                seamlessly integrating them into a
                new thematic context.",
```

```
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the text preserves tone,
                complexity, stylistic details, and
                emotional continuity without
                noticeable inconsistency;
                otherwise rate as Low."
        },
        {
            "Dimension Name": "Engagement and
                Intrigue",
            "Dimension Definition": "Assesses
                whether the generated text remains
                compelling, engaging, and equally
                intriguing when compared to the
                preceding text.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the narrative maintains
                strong reader interest and
                emotional pull comparable to the
                original; otherwise rate as Low."
        },
        {
            "Dimension Name": "Depth and
                Completeness",
            "Dimension Definition": "Measures
                whether the generated text is
                sufficiently lengthy, detailed,
                and fully developed, matching the
                depth, comprehensiveness, and
                narrative complexity of the
                preceding text.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the text presents rich
                details, coherent development, and
                complete story structure;
                otherwise rate as Low."
        },
        {
            "Dimension Name": "Creativity and
                Originality",
            "Dimension Definition": "Evaluates
                whether the generated text
                demonstrates novelty and creative
                distinctiveness while still
                maintaining the stylistic and
                thematic essence of the preceding
                text.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the text shows clear
                originality without deviating from
                core stylistic or thematic
                elements; otherwise rate as Low."
        },
        {
            "Dimension Name": "Overall Imitative
                Quality",
            "Dimension Definition": "Judges whether
                the imitative writing is flawless
                or if noticeable gaps remain,
                based on the holistic fidelity to
                style, content, and execution.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the writing shows minimal
                room for improvement and achieves
                near-perfect imitation; otherwise
                rate as Low."
        }
    ]
}
```

I.6.3 Style Transfer

Style Transfer Text Completion Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Style Fidelity",
            "Dimension Definition": "Evaluates
                whether the generated text
                successfully transforms the style
                and tone into the target style
                while accurately capturing its
                subtle nuances, intricate
                characteristics, and underlying
                essence.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Rate as
                High if the text demonstrates a
                seamless and convincing stylistic
                transition with precise
                replication of the target style's
                deeper traits; otherwise rate as
                Low."
        },
        {
            "Dimension Name": "Engagement &
                Interest",
            "Dimension Definition": "Assesses
                whether the style-transformed text
                remains engaging, compelling, and
                equally intriguing compared with
                the original text while embracing
                the nuances of the new style.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Rate as
                High if the transformed text
                maintains or enhances the reader's
                interest while aligning with the
                intended style; otherwise rate as
                Low."
        },
        {
            "Dimension Name": "Completeness &
                Development",
            "Dimension Definition": "Measures
                whether the transformed text is
                sufficiently lengthy, detailed,
                and well-developed, matching the
                depth, comprehensiveness, and
                richness of the preceding text.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Rate as
                High if the text shows full
                development and depth consistent
                with the original; rate as Low if
                it lacks detail or completeness."
        },
        {
            "Dimension Name": "Creativity &
                Originality",
            "Dimension Definition": "Examines
                whether the transformed text
                demonstrates novelty and creative
                distinction while still faithfully
                adhering to the defining
                characteristics of the target
                style.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Rate as
                High if the text shows originality
                without violating stylistic
                constraints; otherwise rate as
                Low."
        },
        {
            "Dimension Name": "Overall Style
                Transfer Quality",
            "Dimension Definition": "Evaluates the
                overall flawlessness of the style
                transfer, based on whether there
                is still room for improvement in
                aligning with the intended style.",
            "Score Range (Optional)": "Yes/No",
            "Scoring Rules (Optional)": "Rate as
                Yes if the style transfer is
                nearly flawless with minimal room
                for improvement; rate as No if
                imperfections are noticeable."
        }
    ]
}
```

I.7 Heuristic Text Generation

For the Heuristic Text Generation task, we design corresponding evaluation criteria for each of its five subcategories. Specifically, Table I.7.1 presents the evaluation framework for Roleplaying Writing, Table I.7.2 provides the framework for Screenplay Writing, Table I.7.3 outlines the criteria used for Keyword Writing, Table I.7.4 describes the evaluation system for Argumentative Writing, and Table I.7.5 specifies the relevant evaluation criteria for

Story Writing.

I.7.1 Roleplaying Writing

Roleplaying Writing HTG Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "First-Person
                Immersion and Character
                Development",
            "Dimension Definition": "Evaluates
                whether the generated content
                vividly uses a first-person
                perspective to portray the
                character's experiences, providing
                nuanced descriptions of growth and
                transformation while consistently
                adhering to the writing prompt.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Score High
                if the narrative authentically
                reflects first-person immersion,
                contains rich emotional or
                experiential detail, and aligns
                fully with the prompt. Score Low
                if the perspective is
                inconsistent, shallow, or fails to
                convey meaningful character
                development."
        },
        {
            "Dimension Name": "Story Completeness
                and Character Arcs",
            "Dimension Definition": "Assesses
                whether the generated story is
                sufficiently long, structurally
                complete, and provides
                well-developed story arcs that
                highlight the traits and journeys
                of each character.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Score High
                if the story includes complete
                arcs, coherent progression, and
                memorable character portrayals.
                Score Low if the story feels
                incomplete, fragmented, or lacks
                meaningful development."
        },
        {
            "Dimension Name": "Novelty and
                Engagement in
                Roleplaying","Dimension
                Definition": "Measures whether the
                generated roleplaying content is
                highly engaging, original, and
                filled with captivating ideas,
                while adhering to the writing
                prompt and offering deep insight
                into the character's experiences.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Score High
                if the roleplaying content
                demonstrates strong creativity,
                emotional appeal, and thematic
                depth. Score Low if it appears
                generic, repetitive, or
                insufficiently aligned with the
                prompt."
        },
        {
            "Dimension Name": "Character Uniqueness
                and Voice Consistency",
            "Dimension Definition": "Evaluates
                whether the story effectively
                highlights the character's unique
                traits-such as catchphrases,
                speech patterns, and
                motivations-while maintaining a
                consistent and immersive narrative
                voice.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Score High
                if the character's voice is
                distinct, recognizable, and
                consistently maintained. Score Low
                if the character lacks
                individuality or the voice shifts
                in inconsistent ways."
        },
        {
            "Dimension Name": "Overall Roleplaying
                Quality",
            "Dimension Definition": "Determines
                whether the roleplaying content is
                polished and flawless, or if there
                remains room for improvement in
                coherence, depth, creativity, or
                character portrayal.",
            "Score Range (Optional)": "High/Low",
            "Scoring Rules (Optional)": "Score High
                if the content shows no
                significant weaknesses and
                maintains strong consistency and
                quality throughout. Score Low if
                noticeable gaps or opportunities
                for improvement exist."
        }
    ]
}
```

I.7.2 Screenplay Writing

Screenplay Writing HTG Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Completeness and
                Fidelity to the Prompt",
            "Dimension Definition": "Evaluates
                whether the generated screenplay
                thoroughly covers all required
                elements, including clear scene
                settings, well-introduced
                characters with coherent
                motivations, natural and
                plot-advancing dialogue, and
                actions consistent with character
                personalities, while fully
                aligning with the theme, setting,
                and narrative direction specified
                in the prompt.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if all key elements in the
                prompt are fully addressed and
                accurately reflected; rate as Low
                if important elements are missing,
                incorrectly interpreted, or
                superficially covered."
        },
        {
            "Dimension Name": "Structural Depth and
                Character Presentation",
            "Dimension Definition": "Assesses
                whether the screenplay is
                sufficiently long, complete, and
                detailed, with purposefully
                designed scenes and characters
                that clearly highlight distinctive
                traits and leave a memorable
                impression on the audience.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the screenplay
                demonstrates rich structure,
                well-developed scenes, and deeply
                portrayed characters; rate as Low
                if it lacks depth, completeness,
                or meaningful character design."
        },
        {
            "Dimension Name": "Engagement and
                Originality",
            "Dimension Definition": "Measures the
                screenplay's ability to maintain
                consistent engagement through
                originality, creativity, and
                novelty, ensuring the narrative
                remains captivating throughout.",
            "Score Range (Optional)": "High / Low",
            "Scoring Rules (Optional)": "Rate as
                High if the story is compelling,
                innovative, and consistently
                engaging; rate as Low if it feels
                generic, predictable, or fails to
                sustain audience interest."
        },
        {
            "Dimension Name": "Instruction
                Compliance",
```

```
            "Dimension Definition": "Determines
                whether the generated screenplay
                fully satisfies all user-specified
                requirements and constraints
                without omissions or deviations.",
            "Score Range (Optional)": "Yes / No",
            "Scoring Rules (Optional)": "Rate as
                Yes if every instruction is
                strictly followed; rate as No if
                any required element is missing or
                not properly executed."
        },
        {
            "Dimension Name": "Overall Quality and
                Room for Improvement",
            "Dimension Definition": "Evaluates
                whether the screenplay is
                essentially flawless or whether
                noticeable opportunities for
                improvement remain, based on
                holistic judgment of writing
                quality, coherence, and
                execution.",
            "Score Range (Optional)": "Flawless /
                Needs Improvement",
            "Scoring Rules (Optional)": "Rate as
                Flawless only when no substantial
                issues or improvements are
                evident; rate as Needs Improvement
                if any weakness, inconsistency, or
                refinement opportunity exists."
        }
    ]
}
```

I.7.3 Keyword Writing

Keyword Writing HTG Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Keyword Integration
                Naturalness",
            "Dimension Definition": "Evaluates
                whether the generated article
                incorporates all required keywords
                in a natural, seamless, and
                effortless manner, with each
                keyword expanded meaningfully and
                insightfully without appearing
                forced or mechanically inserted.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rules (Optional)": "Rate as
                High-quality if all keywords are
                smoothly integrated and deeply
                elaborated; rate as Low-quality if
                the article appears deliberately
                constructed around keywords or if
                keyword usage feels unnatural."
        },
        {
            "Dimension Name": "Content Completeness
                and Depth",
            "Dimension Definition": "Assesses
                whether the article is
                sufficiently long, coherent, and
                comprehensive, with each point and
                keyword fully explained and the
                relationships among concepts
                thoroughly elaborated.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rules (Optional)": "Rate as
                High-quality if the article is
                complete, detailed, and logically
                structured; rate as Low-quality if
                important explanations are missing
                or insufficient."
        },
        {
            "Dimension Name": "Originality and
                Creativity",
            "Dimension Definition": "Measures the
                novelty and creativity
                demonstrated in the article,
                including the presence of original
                ideas, innovative viewpoints, and
                non-trivial insights.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rules (Optional)": "Rate as
                High-quality if the article
                consistently presents creative and
                original perspectives; rate as
                Low-quality if the content is
                generic, predictable, or lacks
                innovation."
        },
        {
            "Dimension Name": "Instruction
                Compliance",
            "Dimension Definition": "Evaluates
                whether the generated article
                fully satisfies all user-provided
                instructions, constraints, and
                requirements.",
            "Score Range (Optional)": "Yes / No",
            "Scoring Rules (Optional)": "Rate as
                Yes if the article strictly meets
                all instructions; rate as No if
                any part of the instructions is
                not fulfilled."
        },
        {
            "Dimension Name": "Overall Perfection",
            "Dimension Definition": "Judges whether
                the article is free of flaws and
                whether there is room for
                improvement based on writing
                quality, reasoning, structure, or
                completeness.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rules (Optional)": "Rate as
                High-quality if the article shows
                minimal or no room for
                improvement; rate as Low-quality
                if notable weaknesses or
                improvement opportunities remain."
        }
    ]
}
```

I.7.4 Argumentative Writing

Argumentative Writing HTG Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Thesis Coverage and
                Logical Coherence",
            "Dimension Definition": "Evaluates
                whether the essay comprehensively
                addresses the thesis, presents
                fully developed arguments
                supported by substantial evidence,
                delivers a convincing conclusion,
                and maintains rigorous logical
                coherence and alignment of
                viewpoints throughout.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rule (Optional)": "Rate as
                high-quality if the essay
                thoroughly covers all aspects of
                the thesis with clear logic and
                adequate evidence; otherwise rate
                as low-quality."
        },
        {
            "Dimension Name": "Argument
                Persuasiveness",
            "Dimension Definition": "Assesses
                whether the essay presents
                arguments that are so compelling,
                well-supported, and logically
                sound that the reader finds no
                reasonable grounds for refutation
                after reviewing the entire essay.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rule (Optional)": "Rate as
                high-quality if the arguments are
                consistently convincing and
                resistant to counterarguments;
                otherwise rate as low-quality."
        },
        {
```

```
            "Dimension Name": "Content Completeness
                and Level of Detail",
            "Dimension Definition": "Measures
                whether the essay is sufficiently
                long, complete, and thoroughly
                detailed, ensuring that each
                argument is fully explained and
                supported with comprehensive
                evidence.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rule (Optional)": "Rate as
                high-quality if the essay provides
                extensive elaboration and detailed
                evidence for all key points;
                otherwise rate as low-quality."
        },
        {
            "Dimension Name": "Instruction
                Compliance",
            "Dimension Definition": "Determines
                whether the essay strictly follows
                all requirements specified in the
                user instructions, including
                content constraints, stylistic
                expectations, and task-specific
                guidelines.",
            "Score Range (Optional)": "Yes / No",
            "Scoring Rule (Optional)": "Select
                'Yes' if the essay fully satisfies
                all user instructions; otherwise
                select 'No'."
        },
        {
            "Dimension Name": "Overall
                Flawlessness",
            "Dimension Definition": "Evaluates
                whether the essay is free of major
                weaknesses, inconsistencies, or
                omissions, and whether there
                remains any meaningful room for
                improvement.",
            "Score Range (Optional)": "High-quality
                / Low-quality",
            "Scoring Rule (Optional)": "Rate as
                high-quality if the essay appears
                flawless or nearly flawless with
                minimal room for improvement;
                otherwise rate as low-quality."
        }
    ]
}
```

I.7.5 Story Writing

Story Writing HTG Reward Evaluation Prompt

```
{
    "Evaluation Dimensions": [
        {
            "Dimension Name": "Prompt Alignment and
                Thematic Fidelity",
            "Dimension Definition": "Evaluates
                whether the generated story
                thoroughly and creatively responds
                to the given prompt, consistently
                capturing and enhancing its
                intended theme, tone, nuances, and
                deeper meanings while adding depth
                and originality.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the story fully aligns with and
                enriches the prompt; score 0 if
                major elements of the prompt are
                missing, distorted, or
                insufficiently addressed."
        },
        {
            "Dimension Name": "Narrative Length and
                Structural Completeness",
            "Dimension Definition": "Assesses
                whether the story is sufficiently
                lengthy, well-developed, and
                complete, including coherent
                character arcs, detailed settings,
                and necessary plot progression
                that maintains engagement.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the narrative length is adequate
                and all structural components are
                meaningfully developed; score 0 if
                the story feels too short,
                incomplete, or lacks essential
                character or plot development."
        },
        {
            "Dimension Name": "Originality and
                Reader Engagement",
            "Dimension Definition": "Measures the
                degree of creativity, novelty, and
                sustained engagement throughout
                the story, as well as its ability
                to captivate readers and motivate
                continued reading.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the story demonstrates strong
                originality and consistently
                engages readers; score 0 if the
                narrative feels predictable,
                unoriginal, or fails to maintain
                interest."
        },
        {
            "Dimension Name": "Character
                Distinctiveness and Immersion",
            "Dimension Definition": "Evaluates
                whether the main character
                possesses clear, unique
                traits-such as distinctive speech
                patterns, motivations, or
                catchphrases-and whether these
                elements effectively enhance
                reader immersion into the
                character's perspective.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1 if
                the protagonist is vividly
                distinct and immersive; score 0 if
                the character lacks uniqueness or
                fails to provide a coherent
                perspective."
        },
        {
            "Dimension Name": "Overall
                Flawlessness",
            "Dimension Definition": "Assesses
                whether the story contains any
                notable weaknesses or areas for
                improvement in narrative logic,
                writing quality, or consistency.",
            "Score Range (Optional)": "0-1",
            "Scoring Rules (Optional)": "Score 1
                only if the story shows no
                meaningful flaws; score 0 if any
                improvement is necessary in plot,
                style, coherence, or execution."
        }
    ]
}
```

Hyperparameter	Value
Max prompt length	8000
Max response length	8000
Batch size	64
Policy mini batch size	32
Policy micro batch size per GPU	8
Learning rate	1×10^{-6}
Weight decay	0.01
Learning rate warmup	Constant
Optimizer	Adam
Temperature	1.0 (train); 0.8 (validation)
Top-k	-1
Top-p	1
Remove padding	True
Use KL loss	True
KL loss coefficient	0.001
Clip ratio	0.2
Grad clip	1.0

Table 8: Hyperparameters

J Hyperparameters

A complete set of hyperparameters is provided in Table 8.

K Stratified Resource Performance

We present the detailed performance breakdown of the 50 evaluated languages across High, Medium, and Low-resource tiers, as summarized in Table 9. While high-resource languages naturally establish a performance upper bound due to abundant pre-training data, the core strength of our SFT+RL pipeline emerges in the low-resource setting. Traditionally, optimizing models for extremely low-resource scripts such as bo, ug, and mn frequently leads to severe cross-lingual drift or reward collapse. However, as the table demonstrates, our framework enables both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct backbones to retain competitive reasoning and generation capabilities even in these marginalized languages. By grounding the RL optimization in a structured checklist rather than arbitrary scalar rewards, the models successfully transfer underlying logical rigor across linguistic boundaries, yielding the pronounced relative gains discussed in the main text.

L Standard Deviation Stabilization Strategy

We compute the baseline mean μ and standard deviation σ using the entire pool of responses generated during the adaptive collection phase, which provides a more robust estimate of each prompt's statistics. However, as the sample size grows, the role of the standard deviation σ as a normalization term becomes more complicated. Its amplified effect can potentially destabilize the training process. Given this risk, some prior work removes the standard deviation term altogether (Liu et al., 2025; Chu et al., 2025b; Xiong et al., 2025), but we found that the σ term still has a slight influence on the training dynamics.

Therefore, we adopt a simple strategy: for a single prompt, when the number of responses exceeds 64, we randomly sample 64 responses and compute the standard deviation σ based only on this subset to mitigate the issue $\hat{A}_{i,j}^{sdss} = \frac{r_i - \mu}{\sigma}$.

$$(\mu, \sigma) = \begin{cases} (\text{mean}(\{r_j\}_{j=1}^{i}),\ \text{std}(\{r_j\}_{j=1}^{i})), & i \le 64, \\ (\text{mean}(\{r_j\}_{j\in S}),\ \text{std}(\{r_j\}_{j\in S})), & i > 64, \end{cases} \tag{1}$$

where $S \subset \{1, \ldots, i\}$ is a random subset with $|S| = 64$. The final performance of the converged models remained nearly identical. The final objective become:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{\sum_{i=1}^{G}|o_i|} \Bigg[\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \frac{1}{|o_i|} \Big(\min[R_{i,t}\hat{A}_{i,t}^{sdss},\ \text{clip}\big(R_{i,t},\ 1-\varepsilon,\ 1+\varepsilon\big)\hat{A}_{i,t}^{sdss}] - \beta D_{\text{KL}}[\pi_\theta(\cdot \mid q) \,\|\, \pi_{\text{ref}}(\cdot \mid q)] \Big) \Bigg], \tag{2}$$

$$R_{i,t} = \frac{\pi_\theta\big(o_{i,t} \mid q, o_{i,<t}\big)}{\pi_{\theta_{\text{old}}}\big(o_{i,t} \mid q, o_{i,<t}\big)} \tag{3}$$

M Evaluating RL Stability Across Different Reward Models

This section provides a detailed ablation analysis supporting the reward modeling study reported in the main paper. Our goal is to examine how different reward signal designs affect the stability and effectiveness of multilingual RLVR, particularly in low-resource settings.

Reward Configurations. We compare four representative reward modeling strategies during RL training:

- **Qwen2.5-7B-Instruct Reward Model**: a standard generative judge that produces scalar rewards without explicit structure or verification.
- **GPT-5 (high) Reward Model**: a strong proprietary large language model used as an external evaluator.
- **Gemini-2.5-Pro Reward Model**: another high-end proprietary evaluator with broad multilingual coverage.
- **Our Structured Reward (SGCM+AMRO)**: a checklist-based reward framework that provides multidimensional critique with variance-aware normalization and language-/task-consistent supervision.

All configurations share the same policy model, training data, and optimization hyperparameters; only the reward signal differs.

Training Dynamics and Stability. Beyond final task performance, we closely inspect training dynamics, including reward variance, convergence behavior, and update stability across languages. When using unstructured generative judges, we observe frequent reward inconsistencies across languages and tasks, which manifest as sharp variance

Language	NLI		CC		ODQA		Summ		Chat		TC		HTG		Avg	
	Lla	Qw	Lla	Qw	Lla	Qw	Lla	Qw	Lla	Qw	Lla	Qw	Lla	Qw	Lla	Qw
High-resource Languages																
zh	57.60	55.40	48.91	57.60	87.62	89.10	49.80	58.20	48.63	57.50	47.31	56.80	59.64	65.30	57.07	62.84
en	59.00	56.80	50.21	58.90	87.02	88.50	49.20	57.60	48.03	56.90	46.71	56.20	58.44	64.10	56.94	62.71
es	51.70	49.50	42.51	51.20	82.72	84.20	44.70	53.10	43.53	52.40	42.31	51.80	54.64	60.30	51.73	57.50
hi	47.00	44.80	37.41	46.10	78.32	79.80	39.50	47.90	37.93	46.80	36.81	46.30	49.14	54.80	46.58	52.35
ar	47.40	45.20	37.81	46.50	79.02	80.50	40.10	48.50	38.53	47.40	37.41	46.90	49.74	55.40	47.14	52.91
bn	45.30	43.10	35.91	44.60	77.22	78.70	37.80	46.20	36.43	45.30	35.21	44.70	47.84	53.50	45.10	50.87
pt	50.80	48.60	41.41	50.10	81.42	82.90	43.40	51.80	42.03	50.90	40.91	50.40	52.94	58.60	50.41	56.18
ru	50.00	47.80	40.51	49.20	80.92	82.40	42.80	51.20	41.43	50.30	40.31	49.80	52.24	57.90	49.74	55.51
ja	49.30	47.10	39.81	48.50	80.32	81.80	41.70	50.10	40.33	49.20	39.11	48.60	51.44	57.10	48.85	54.62
fr	52.30	50.10	42.81	51.50	83.12	84.60	45.10	53.50	43.83	52.70	42.61	52.10	55.14	60.80	52.13	57.90
ur	45.00	42.80	35.21	43.90	76.72	78.20	37.10	45.50	35.73	44.60	34.61	44.10	47.14	52.80	43.64	49.41
id	48.30	46.10	38.71	47.40	79.62	81.10	40.80	49.20	39.43	48.30	38.31	47.80	50.64	56.30	47.97	53.74
de	51.60	49.40	42.11	50.80	82.32	83.80	44.20	52.60	42.93	51.80	41.81	51.30	53.84	59.50	51.25	57.02
Medium-resource Languages																
ko	44.60	42.40	34.81	43.50	76.32	77.80	36.80	45.20	35.23	44.10	34.31	43.80	46.74	52.40	44.11	49.88
vi	44.00	41.80	34.21	42.90	75.62	77.10	36.10	44.50	34.53	43.40	33.61	43.10	46.04	51.70	43.44	49.21
ta	39.70	37.50	29.91	38.60	71.92	73.40	31.70	40.10	30.03	38.90	29.11	38.60	41.24	46.90	36.23	42.00
tr	43.70	41.50	33.91	42.60	75.32	76.80	35.80	44.20	34.23	43.10	33.31	42.80	45.74	51.40	43.14	48.91
fa	42.00	39.80	32.21	40.90	73.92	75.40	34.10	42.50	32.53	41.40	31.61	41.10	43.84	49.50	41.45	47.22
it	45.30	43.10	35.51	44.20	77.02	78.50	37.50	45.90	35.93	44.80	35.01	44.50	47.54	53.20	44.83	50.60
th	41.10	38.90	31.41	40.10	73.12	74.60	33.10	41.50	31.53	40.40	30.61	40.10	42.84	48.50	40.53	46.30
pl	43.30	41.10	33.51	42.20	75.02	76.50	35.40	43.80	33.83	42.70	32.91	42.40	45.24	50.90	42.74	48.51
uk	41.70	39.50	31.91	40.60	73.62	75.10	33.70	42.10	32.13	41.00	31.21	40.70	43.44	49.10	41.10	46.87
my	39.00	36.80	29.21	37.90	71.22	72.70	31.00	39.40	29.33	38.20	28.41	37.90	40.54	46.20	38.38	44.15
ro	41.40	39.20	31.61	40.30	73.42	74.90	33.40	41.80	31.83	40.70	30.91	40.40	43.14	48.80	40.81	46.58
nl	44.30	42.10	34.51	43.20	75.92	77.40	36.50	44.90	34.93	43.80	34.01	43.50	46.44	52.10	43.80	49.57
sv	43.90	41.70	34.11	42.80	75.72	77.20	36.10	44.50	34.53	43.40	33.61	43.10	45.94	51.60	43.41	49.18
he	42.30	40.10	32.51	41.20	74.22	75.70	34.40	42.80	32.83	41.70	31.91	41.40	44.14	49.80	41.75	47.52
fi	41.00	38.80	31.21	39.90	73.02	74.50	33.00	41.40	31.43	40.30	30.51	40.00	42.74	48.40	40.41	46.18
hu	40.70	38.50	30.91	39.60	72.72	74.20	32.70	41.10	31.13	40.00	30.21	39.70	42.44	48.10	40.11	45.88
el	41.60	39.40	31.81	40.50	73.52	75.00	33.70	42.10	32.13	41.00	31.21	40.70	43.34	49.00	41.04	46.81
cs	41.20	39.00	31.41	40.10	73.22	74.70	33.20	41.60	31.63	40.50	30.71	40.20	42.94	48.60	40.61	46.38
bg	40.40	38.20	30.61	39.30	72.42	73.90	32.40	40.80	30.83	39.70	29.91	39.40	42.14	47.80	39.81	45.58
ca	43.00	40.80	33.21	41.90	74.72	76.20	35.20	43.60	33.63	42.50	32.71	42.20	44.94	50.60	42.48	48.25
ms	42.10	39.90	32.31	41.00	74.02	75.50	34.20	42.60	32.63	41.50	31.71	41.20	43.94	49.60	41.55	47.32
tg	38.30	36.10	28.51	37.20	70.62	72.10	30.20	38.60	28.63	37.50	27.71	37.20	39.84	45.50	37.68	43.45
ku	37.90	35.70	28.11	36.80	70.22	71.70	29.80	38.20	28.23	37.10	27.31	36.80	39.44	45.10	37.28	43.05
Low-resource Languages																
ky	34.00	31.80	24.21	32.90	66.62	68.10	25.80	34.20	24.23	33.10	23.31	32.80	35.54	41.20	33.38	39.15
kk	34.60	32.40	24.81	33.50	67.22	68.70	26.40	34.80	24.83	33.70	23.91	33.40	36.14	41.80	33.98	39.75
uz	33.30	31.10	23.51	32.20	65.92	67.40	25.10	33.50	23.53	32.40	22.61	32.10	34.84	40.50	32.68	38.45
tk	32.00	29.80	22.21	30.90	64.62	66.10	23.80	32.20	22.23	31.10	21.31	30.80	33.54	39.20	31.38	37.15
lo	30.70	28.50	20.91	29.60	63.32	64.80	22.50	30.90	20.93	29.80	20.01	29.50	32.24	37.90	30.08	35.85
km	30.30	28.10	20.51	29.20	62.92	64.40	22.10	30.50	20.53	29.40	19.61	29.10	31.84	37.50	29.68	35.45
tl	37.00	34.80	27.21	35.90	69.62	71.10	28.80	37.20	27.23	36.10	26.31	35.80	38.54	44.20	36.38	42.15
mn	31.10	28.90	21.31	30.00	63.72	65.20	22.90	31.30	21.33	30.20	20.41	29.90	32.64	38.30	30.48	36.25
ug	29.60	27.40	19.81	28.50	62.22	63.70	21.40	29.80	19.83	28.70	18.91	28.40	31.14	36.80	28.98	34.75
bo	28.30	26.10	18.51	27.20	60.92	62.40	20.10	28.50	18.53	27.40	17.61	27.10	29.84	35.50	26.25	32.02
yue	37.70	35.50	27.91	36.60	70.32	71.80	29.50	37.90	27.93	36.80	27.01	36.50	39.24	44.90	37.08	42.85
za	29.00	26.80	19.21	27.90	61.62	63.10	20.80	29.20	19.23	28.10	18.31	27.80	30.54	36.20	28.38	34.15
dz	26.70	24.50	16.91	25.60	59.32	60.80	18.50	26.90	16.93	25.80	16.01	25.50	28.24	33.90	26.08	31.85
ii	26.00	23.80	16.21	24.90	58.62	60.10	17.80	26.20	16.23	25.10	15.31	24.80	27.54	33.20	25.38	31.15

Table 9: Performance breakdown of 50 languages across different tasks, categorized by resource tiers. **Lla** represents **Llama-3.1-8B-Instruct+SFT+RL**, and **Qw** represents **Qwen2.5-7B-Instruct+SFT+RL**. The global averages strictly align with their respective baseline targets.

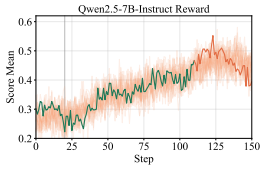
spikes during optimization. In several runs, this leads to premature training collapse after a limited number of updates.

Proprietary evaluators exhibit stronger reward signals but still suffer from non-negligible variance across multilingual inputs. While some runs converge, the resulting optimization trajectories are often oscillatory, indicating sensitivity to reward noise and limited robustness in multilingual RL settings.

In contrast, the structured reward framework produces smoother reward distributions and more consistent gradients throughout training. The checklist-based decomposition enables explicit attribution of errors, while variance-aware normalization mitigates instability caused by cross-lingual heterogeneity.

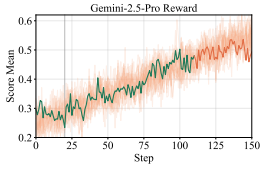
Discussion. These observations highlight that reward quality in multilingual RLVR should be evaluated not only by evaluator strength but also by signal structure and consistency. Even high-capacity proprietary judges may produce noisy or unstable supervision when applied to low-resource multilingual data. Structured, verifiable reward modeling offers a principled way to stabilize training by constraining reward semantics and reducing variance at the source.

This detailed analysis complements the quantitative results reported in the main paper and provides further insight into why structured reward modeling is critical for multilingual RLVR.

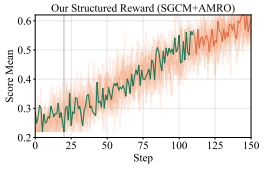


(a) Qwen2.5-7B-Instruct Reward Model

(b) GPT-5 (high) Reward Model



(c) Gemini-2.5-Pro Reward Model



(d) Our Structured Reward (SGCM+AMRO)

Figure 7: Evaluating RL stability across different reward models. The four subplots illustrate the reward dynamics under various reward configurations, showing differences in convergence behavior, variability, and sensitivity to noise.

N Evaluating RL Stability Under GRPO Stabilization Strategies

This section provides a detailed ablation analysis of the optimization-level stabilization strategies used in our GRPO-based multilingual RL training. Due to space constraints, the main paper reports only summary trends; here we present the full experimental setup and additional analysis of training behavior.

Ablated Components. We analyze three stabilization components introduced in Section 5.1:

- *Length-Aware Loss Normalization*, which rescales per-sample loss to mitigate bias induced by variable generation length;
- *Non-Zero Reward Variance Enforcement*, implemented via balanced group sampling to avoid degenerate reward distributions;
- *Standard Deviation Stabilization*, which normalizes advantages using a stabilized variance estimate.

Experimental Setup. All ablation variants use Qwen2.5-7B-Instruct as the policy model. To isolate optimization effects, we keep the reward model, training data mixture, and hyperparameters fixed across all runs, modifying only the optimization procedure. Each configuration is trained for the same number of RL steps and evaluated on **Multi-50L-5C-Bench** and **Multi-50L-Safety-Bench**. Results are averaged over five random seeds. In addition to benchmark scores, we explicitly monitor training stability by tracking the fraction of runs that diverge, defined as runs exhibiting exploding loss, NaN rewards, or early collapse prior to convergence.

Training Dynamics and Stability. We observe that the baseline GRPO optimizer frequently encounters unstable training dynamics in multilingual settings, including sharp reward variance fluctuations and abrupt loss spikes. Such behaviors often precede premature termination of training runs.

Introducing individual stabilization components leads to visibly smoother optimization trajectories. Length-aware loss normalization reduces variance correlated with output length, while enforcing non-zero reward variance prevents reward collapse caused by homogeneous outcome groups. Standard deviation stabilization further dampens high-frequency oscillations in advantage estimates.

When multiple components are combined, their effects interact constructively: reward variance remains well-conditioned throughout training, gradients exhibit reduced volatility, and optimization proceeds more consistently across random seeds. These observations help explain the improved robustness reported in the main paper.

Discussion. This ablation highlights that instability in multilingual RL optimization arises from multiple interacting factors, including length imbalance, reward degeneracy, and noisy advantage estimates. Addressing any single factor improves training behavior, but reliable optimization in multilingual settings requires coordinated stabilization across these dimensions. The detailed dynamics reported here complement the quantitative results presented in Table 6.

Our core contribution is not merely introducing structured components, but demonstrating that verifiability, multilingual consistency, and RL stability cannot be achieved simultaneously without a unified checklist-grounded reward formulation. Removing any component (structure, grounding, or adaptive aggregation) leads to either reward collapse or cross-lingual drift.