Igniting Creative Writing in Small Language Models: LLM-as-a-Judge versus Multi-Agent Refined Rewards

Xiaolong Wei[1*], **Bo Lu**[2*], **Xingyu Zhang**[3], **Zhejun Zhao**[2†]
Dongdong Shen[2], **Long Xia**[2], **Dawei Yin**[2]
[1]Beihang University [2]Baidu Inc.
[3]Beijing Jiaotong University
xiaolongwei@buaa.edu.cn, zhaozhejun@baidu.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable creative writing capabilities, yet their substantial computational demands hinder widespread use. Enhancing Small Language Models (SLMs) offers a promising alternative, but current methods like Supervised Fine-Tuning (SFT) struggle with novelty, and Reinforcement Learning from Human Feedback (RLHF) is costly. This paper explores two distinct AI-driven reward strategies within a Reinforcement Learning from AI Feedback (RLAIF) framework to ignite the creative writing of a 7B-parameter SLM, specifically for generating Chinese greetings. The first strategy employs a RM trained on high-quality preference data curated by a novel multi-agent rejection sampling framework designed for creative tasks. The second, more novel strategy utilizes a principle-guided LLM-as-a-Judge, whose reward function is optimized via an adversarial training scheme with a reflection mechanism, to directly provide reward signals. Comprehensive experiments reveal that while both approaches significantly enhance creative output over baselines, the principle-guided LLM-as-a-Judge demonstrably yields superior generation quality. Furthermore, it offers notable advantages in training efficiency and reduced dependency on human-annotated data, presenting a more scalable and effective path towards creative SLMs. Our automated evaluation methods also exhibit strong alignment with human judgments. Our code and data are publicly available at Github.

1 Introduction

Creative writing, a cornerstone of human expression and communication (Kaufmann, 2012; Bakar et al., 2021), intrinsically demands not only literary merit and emotional resonance but also a significant degree of personalization to effectively engage its audience (Bakar et al., 2021). While users increasingly turn to online platforms for creative inspiration, existing retrieval-based methods often fall short in delivering content that is sufficiently tailored to individual needs and contexts, a limitation that has become more pronounced with the advent of advanced generative models. This underscores a growing demand for generative systems capable of producing context-aware, responsive, and personalized creative text (Richardson et al., 2023).

The advent of Large Language Models (LLMs) such as GPT-4o (Hurst et al., 2024) and DeepSeek-V3 (Liu et al., 2024) has revolutionized text generation, demonstrating remarkable capabilities in creative writing tasks. However, under high request volumes, the substantial computational footprint and high inference latency of these large-scale models present significant barriers to their widespread deployment and practical application. Consequently, enhancing Small Language Models (SLMs, typically <10B parameters), such as the Qwen2.5 7B model we employ (Yang et al., 2024), to achieve comparable creative prowess while maintaining efficiency has become a critical research frontier (Han et al., 2025). This pursuit aligns with broader trends where modern applications increasingly prioritize dynamic content personalization (Li et al., 2025c; Cui et al., 2025a,b) while also emphasizing information's expressiveness and reliability (Tong et al., 2024; Lu et al., 2025; Zeng et al., 2025). It is crucial to note that generic, un-fine-tuned SLMs often lack the sophisticated generative abilities required for high-quality creative writing (Gómez-Rodríguez and Williams, 2023).

Prevailing methodologies for enhancing SLMs predominantly involve Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). While SFT can effectively adapt SLMs to specific styles, it often struggles to foster genuine novelty and gener-

[*]Co-first authors with equal contributions.
[†]Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17160–17186
November 4-9, 2025

alization (Zhou et al., 2023; Sanh et al., 2021)—attributes paramount for compelling creative writing. RLHF, on the other hand, relies on high-quality reward models typically trained on extensive human preference data, the annotation of which is labor-intensive and expensive (Ziegler et al., 2019).

To surmount these limitations, we investigate two distinct reward strategies:

- **A Refined Reward Model:** We develop an RM trained on meticulously curated preference data. This data is generated and filtered by a novel multi-agent framework designed to ensure high quality and relevance for creative tasks.
- **Principle-Guided LLM-as-a-Judge:** Drawing inspiration from "LLM-as-a-Judge" paradigms (Zheng et al., 2023), we directly employ a powerful LLM as the reward provider. Crucially, this LLM's judgments are guided by explicitly defined creative writing principles and its reward function is further optimized via an adversarial training scheme (Wang et al., 2024).

We conduct comprehensive experiments on generating Chinese greetings using 7B-parameter SLMs, specifically the Qwen2.5 7B model. Our findings reveal that while both RL-based approaches significantly enhance creative output compared to baselines, the principle-guided LLM-as-a-Judge strategy yields demonstrably superior results in terms of generation quality. These outcomes are rigorously validated through both human evaluations and LLM-based assessments, including an analysis of their alignment. Furthermore, the LLM-as-a-Judge approach exhibits notable advantages in training efficiency and reduced dependency on human-annotated data.

Our main contributions are threefold:

- We introduce a novel principle-guided LLM-as-a-Judge reward mechanism, optimized adversarially, for effectively steering RL towards enhancing SLM creative writing capabilities.
- We propose a multi-agent framework for generating and filtering high-quality preference data, enabling the training of more effective reward models for creative domains.
- We present a systematic comparison of these two reward paradigms for SLM-based creative writing, corroborated by extensive LLM-based and human evaluations, and offer insights into their alignment and practical trade-offs.

2 Related Work

The landscape of artificial intelligence in creative writing has been dramatically reshaped by LLMs. These models, such as the GPT series (Brown et al., 2020; Achiam et al., 2023) and LLaMA (Touvron et al., 2023), trained on vast text corpora, demonstrate unprecedented capabilities in generating diverse creative texts, including complex narratives, poetry, and scripts, exhibiting high fluency, style adaptation, and thematic coherence. Researchers have developed techniques like planning (Yang et al., 2022), controllable generation (Li et al., 2022), and structured decomposition frameworks like Branch-Solve-Merge (Saha et al., 2023) to further enhance and guide LLMs' creative output.

Beyond autonomous generation, recent work increasingly focuses on LLMs as co-creative partners for human writers, exploring interaction dynamics for tasks such as brainstorming and outlining (Gero et al., 2023). The concept of multi-agent systems collaborating on writing tasks is also an emerging area.

Despite these advancements in generation capabilities, evaluating the creativity of LLM-produced text remains a complex challenge (Chakrabarty et al., 2024; Kim and Oh, 2025). Traditional automatic metrics are insufficient for capturing subjective qualities like originality and emotional depth. To address this, recent work has explored Self-Rewarding Language Models (Yuan et al., 2024) that iteratively improve by generating their own training rewards, though automated assessments still do not yet reliably align with human judgments (Chakrabarty et al., 2024; Li et al., 2025a).

These persistent challenges in aligning automated evaluation with human judgment highlight fundamental open problems: how to build effective reward signals for training generative models and achieve reliable automated evaluation in this subjective domain.

3 Methodology

To enhance the creative writing capabilities of our target SLM, we employ a RLAIF paradigm. The central tenet of RLAIF is to refine the SLM's policy using reward signals derived from AI-driven

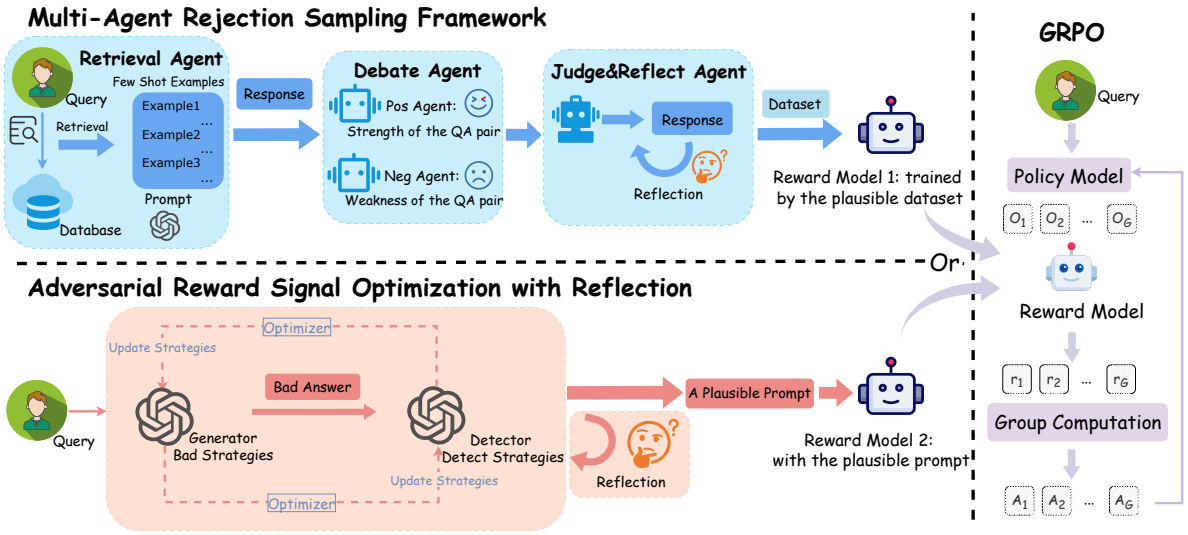

Figure 1: The figure depicts two distinct reward signals. Signal 1 is derived from a multi-agent system, yielding a reward model. Signal 2 is generated via adversarial interaction (Generator-Detector) and reflection, producing a prompt. Both signals are separately used to train GRPO.

evaluations of its generated outputs. Our primary contribution lies in the exploration and comparison of two distinct and sophisticated strategies for generating these crucial reward signals, which are designed to capture the multifaceted nature of creative text. These strategies are: 1) a meticulously refined RM trained on preference data curated by a multi-agent system, and 2) a dynamic reward signal obtained from an adversarially trained, principle-guided LLM acting as a judge (LLM-as-a-Judge). The complete process is detailed in Fig. 1. In the subsequent sections, we first detail the multi-agent framework for preference data generation and RM training (Section 3.1). We then describe the adversarial approach for optimizing an LLM-as-a-Judge as a direct reward provider (Section 3.2). Finally, Section 3.3 outlines how the reward signals derived from these two strategies are integrated into the RLAIF process to optimize the SLM.

3.1 Multi-Agent Rejection Sampling Framework

The evaluation of LLMs by a single LLM instance, while scalable, can suffer from inherent biases, limited perspectives, and potential instability (Zheng et al., 2023). To mitigate these challenges, we introduce a multi-agent collaborative evaluation system. This system operationalizes a collaborative paradigm, drawing inspiration from approaches where multiple agents engage in debate or structured discussion to refine assessments and achieve more robust outcomes (Chan et al., 2023; Du et al., 2023). By simulating a nuanced, rigorous, and bias-resistant assessment process, our framework aims to leverage the collective intelligence and error-correction capabilities inherent in multi-agent interactions (Liang et al., 2023). This approach aligns with a broader trend in AI systems where complex tasks are decomposed and managed by specialized, collaborative agents to achieve a goal (Li et al., 2025b). The primary output of this system is high-fidelity preference data, denoted as $\mathcal{D}_{\text{pref}}$. This dataset is specifically curated to be suitable for training robust reward models, which can subsequently be employed to filter and rank generated content based on nuanced quality dimensions. All prompts and cases are provided in Appendix A.5.

3.1.1 Retrieval Agent

The Retrieval Agent, implementing the function $Rtr : \mathcal{P} \rightarrow \mathcal{P}(\mathcal{D}_{\text{HQ}})$, retrieves relevant context for evaluation. Upon receiving an input prompt $p \in \mathcal{P}$, it queries a pre-computed vector index (built from $\mathcal{D}_{\text{HQ}}$) using similarity metrics (e.g., cosine similarity on embeddings) to fetch the set $E = Rtr(p) = \{(p'_j, r'_j)^*\}_{j=1}^{k}$ of k high-quality prompt-response pairs. These pairs serve as few-shot examples, providing contextual grounding and quality benchmarks for the subsequent evaluation agents.

3.1.2 Debate Agents: Positive and Negative Perspectives

This module employs two adversarial agents, embodying the functions $f_{\text{pos}} : (\mathcal{P}, \mathcal{R}, E) \to \mathcal{E}^+$ and $f_{\text{neg}} : (\mathcal{P}, \mathcal{R}, E) \to \mathcal{E}^-$, to conduct a structured debate on the quality of a given response r for prompt p.

- **Positive Agent (f_{pos}):** Identifies and articulates the strengths and merits of the response r, such as novelty, coherence, emotional resonance, or alignment with the prompt's intent. Its output is a structured positive evaluation $\varepsilon^+ \in \mathcal{E}^+$.
- **Negative Agent (f_{neg}):** Identifies and articulates the weaknesses and potential issues within r, such as factual inaccuracies, logical fallacies, stylistic clichés, or lack of creativity. Its output is a structured negative evaluation $\varepsilon^- \in \mathcal{E}^-$.

This structured debate mechanism compels a multifaceted analysis, surfacing both positive and negative aspects that might be overlooked by a single evaluator due to confirmation bias or inherent model preferences.This process yields a more comprehensive and less biased assessment, crucial for subjective domains like creative writing.

3.1.3 Judge Agent

The Judge Agent, implementing $f_{\text{judge}} : (\mathcal{P}, \mathcal{R}, \mathcal{E}^+, \mathcal{E}^-) \to \mathcal{S}_{\text{initial}}$, synthesizes the evaluations ε^+ and ε^- from the debate agents. It weighs the conflicting arguments, assesses the relative importance of identified strengths and weaknesses, and formulates a holistic initial judgment $S_{\text{initial}} \in \mathcal{S}_{\text{initial}}$. This simulates a reasoned decision-making process based on multifaceted evidence.

3.1.4 Reflect Agent

Following the initial judgment, the Reflect Agent, implementing $f_{\text{reflect}} : (\mathcal{P}, \mathcal{R}, \mathcal{S}_{\text{initial}}, \mathcal{E}^+, \mathcal{E}^-) \to \mathcal{S}_{\text{final}}$, performs a critical review of S_{initial} and the supporting arguments ε^+ and ε^-. It scrutinizes the Judge Agent's reasoning for logical consistency and completeness. If flaws are detected, the Reflect Agent may override S_{initial} and potentially trigger a re-evaluation. Otherwise, it ratifies the initial judgment, resulting in the final assessment $S_{\text{final}} \in \mathcal{S}_{\text{final}}$. This reflection step enhances the reliability and robustness of the final evaluation. Based on S_{final}, a preference pair $(p, r_{\text{chosen}}, r_{\text{rejected}})$ is determined and added to the preference dataset $\mathcal{D}_{\text{pref}}$.

3.2 Adversarial Reward Signal Optimization with Reflection

Inspired by Generative Adversarial Networks (GANs) and related approaches like LLM-GAN (Wang et al., 2024), we propose an adversarial framework to dynamically generate and refine reward signals for RL-based policy optimization. This framework comprises a Generator, a Detector, and a novel Reflector component. Further details are provided in Appendix A.4.

3.2.1 Generator-Detector Adversarial Dynamics

- **Generator (π_G):** The Generator, parameterized by θ_G, aims to produce responses r for a given prompt p according to its policy $\pi_G(r|p; \theta_G)$. Its goal is to generate bad responses that are hard to distinguish.
- **Detector (f_D):** The Detector, parameterized by θ_D, acts as a discriminator. It learns to distinguish responses r generated by π_G. It assigns a score $f_D(p, r; \theta_D) \in \{0, 1\}$, where 1 represents a good response and 0 represents a bad response.

These components engage in adversarial training. The Detector is trained to maximize its ability to correctly classify responses, while the Generator aims to produce indistinguishable bad responses to deceive the Detector.

3.2.2 Reflector-Enhanced Detector Optimization

To further improve the Detector's reliability, we introduce the Reflector module ($f_{\text{Rf}} : (\mathcal{P}, \mathcal{R}, \mathcal{S}_{\text{D}}, \mathcal{Y}_{\text{true}}) \to \mathcal{R}_{\text{D}}$). When the Detector f_D misclassifies a response (p, r) compared to a reference label $y_{\text{true}} \in \mathcal{Y}_{\text{true}}$ (where y_{true} could indicate if r is genuinely high-quality or not, obtained from $\mathcal{D}_{\text{pref}}$ or human annotation), the Reflector is activated. The Reflector analyzes the triplet $(p, r, s_D = f_D(p, r; \theta_D))$ alongside y_{true} to diagnose the cause of the Detector's error. Based on this analysis, it generates structured feedback or advice $R_D \in \mathcal{R}_{\text{D}}$. This advice R_D can be used to guide the Detector's optimization process (e.g., "Increase weight on detecting emotional flatness"). This explicit reflection mechanism allows the Detector to learn from its mistakes beyond the im-

plicit adversarial signal, improving its robustness and alignment with desired quality criteria.

3.3 RLAIF for Creative Writing Enhancement

This section details the integration of the previously described AI-generated reward signals into the RLAIF process. Our goal is to optimize the target SLM, Qwen2.5-7B-Instruct, for enhanced creative writing proficiency by leveraging nuanced feedback. We investigate two primary sources for the reward signal used within the RLAIF process:

- **Multi-Agent Preference Reward Model (RM):** A reward model $R_{\text{MA}}(p, r; \phi_{\text{RM}})$ is trained on the high-quality preference dataset $\mathcal{D}_{\text{pref}}$ generated by the multi-agent evaluation system described in Section 3.1. The RM learns to predict the preferences expressed in $\mathcal{D}_{\text{pref}}$, typically using a loss function like:

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}_{(p,r_c,r_r)\sim\mathcal{D}_{\text{pref}}} \left[\log \sigma \left(R_{\text{MA}}(p, r_c; \phi_{\text{RM}}) - R_{\text{MA}}(p, r_r; \phi_{\text{RM}})\right)\right] \quad (1)$$

  where σ is the sigmoid function. The output $R_{\text{MA}}(p, r)$ serves as the reward signal.

- **Adversarial Detector Reward Signal:** The output score $s_D = f_D(p, r; \theta_D)$ from the adversarially trained and reflector-enhanced Detector (detailed in Section 3.2) is used directly as a reward signal, $R_D(p, r) = f_D(p, r; \theta_D)$. This signal reflects the response's ability to meet the criteria implicitly learned by the dynamic LLM-based judge.

We apply GRPO algorithm(Shao et al., 2024) to optimize the Qwen2.5-7B-Instruct. The advantage A_t is calculated based on trajectories sampled from the policy π_θ and rewards obtained from either R_{MA} or R_D. We compare the effectiveness of these distinct reward mechanisms in enhancing the models' creative writing capabilities across various dimensions.

4 Experiments

4.1 Task Design

This study centers on enhancing the generation of Chinese greetings. These greetings are prevalent in Chinese culture for significant festivals like the Spring Festival and Mid-Autumn Festival, indicating a high practical demand and rich contextual nuances. This specific focus allows for an in-depth exploration of creative text generation within a culturally significant and frequently utilized domain. The details are provided in Appendix A.2.

4.2 Datasets

Our experiments leverage several datasets constructed for distinct purposes: training a retrieval-augmented multi-agent system, developing reward models, fine-tuning the policy model via RLAIF, and comprehensive final evaluation. All data was sourced from online interactions related to Chinese greetings, with meticulous preprocessing to remove Personally Identifiable Information (PII). Specific business-related source details remain desensitized.

Retrieval Corpus To equip our multi-agent evaluation system (Section 3.1) with high-quality contextual examples, we curated a retrieval corpus comprising 23,442 instances. These instances were selected from a larger online collection based on their high user click-through rates and frequent replication, indicative of their perceived quality and relevance.

Reward Model Training Data For training the preference-based reward model, we initially collected 10,000 user queries from online sources. These queries, along with candidate responses, were processed through our multi-agent rejection sampling framework. This procedure yielded 7,896 preference pairs, each structured as $(query, response_{chosen}, response_{rejected})$. This dataset was then partitioned into an 80% training set and a 20% held-out test set for RM development.

Policy Optimization (GRPO) Data A separate set of 4,000 distinct online queries was utilized for fine-tuning the target SLM using the GRPO algorithm. This dataset was also divided into an 80:20 train/test split to guide the RLAIF process.

Final Evaluation Set To rigorously assess the performance of all compared models, we constructed a dedicated evaluation set of 2,000 query-response pairs. This set was carefully balanced, containing 1,000 "high-quality" instances (heuristically labeled '1'), selected from data exhibiting high click-through and replication rates, and 1,000 "low-quality" instances (labeled '0'), derived from data with lower engagement metrics. This dataset serves as the primary benchmark for both our automated and human evaluations.

	Accuracy	Precision	Recall	F1-score
Multi-Agent Framework	87.60%	87.38%	87.90%	0.8764
Adversarial Framework	85.50%	78.54%	97.70%	0.8708

Table 1: Comparison of two different frameworks on the evaluation set.

	Signal-1	Signal-2	Human
GPT-4o	49.0%	46.8%	50.0%
Ernie-4.5	76.4%	88.2%	87.6%
DeepSeek-V3	91.0%	94.2%	93.0%
Qwen2.5-7B-Instruct	59.2%	56.0%	57.6%
SFT + Qwen2.5-7B-Instruct	92.0%	92.6%	90.0%
Reward Model + RL	-	-	-
LLM-as-a-Judge + RL	**92.4%**	**96.6%**	**95.0%**
SFT + Reward Model + RL	92.2%	96.0%	94.6%
SFT + LLM-as-a-Judge + RL	89.6%	96.0%	93.0%

Table 2: Comparison of the excellence rate of the Model under different evaluation mechanisms. This data represents the inference results of the model under high-frequency greetings (for example, Chinese New Year greetings). Here, Signal-1 refers to Section 3.1, Signal-2 refers to Section 3.2, and Human refers to the evaluation by human experts. Furthermore, the Reward Model + RL method is excluded from the evaluation due to its training not converging.

	Signal-1	Signal-2	Human
GPT-4o	47.6%	45.6%	50.4%
Ernie-4.5	72.0%	81.2%	83.0%
DeepSeek-V3	74.0%	83.8%	85.6%
Qwen2.5-7B-Instruct	47.6%	52.8%	53.8%
SFT + Qwen2.5-7B-Instruct	80.2%	85.2%	86.2%
Reward Model + RL	-	-	-
LLM-as-a-Judge + RL	**91.0%**	**93.4%**	**92.4%**
SFT + Reward Model + RL	89.4%	90.6%	91.2%
SFT + LLM-as-a-Judge + RL	85.0%	89.0%	90.2%

Table 3: Comparison of the excellence rate of the Model under different evaluation mechanisms. This data represents the inference results of the model under ordinary greetings(for example, greetings for a new car). Furthermore, the Reward Model + RL method is excluded from the evaluation due to its training not converging.

4.3 Rubric Design

The evaluation rubric provides a holistic view of greetings quality, comprising five dimensions with respective weights: Language Quality (30%), Creativity (30%), Emotional Resonance (15%), Cultural Appropriateness (15%), and Content Richness (10%).

Language Quality assesses fluency and precision. Essential for effective communication, its importance in NLG systems is well-recognized (Van Der Lee et al., 2019; Que et al., 2024), with modern approaches using LLMs for nuanced assessment (Liu et al., 2023) and considering aspects like style and meaning preservation (Chim et al., 2025).

Creativity evaluates the generation of innovative elements like unique metaphors or novel perspectives, distinguishing memorable greetings. This involves producing novel, surprising, and valuable outputs (Zhang et al., 2025), crucial for pushing NLG beyond mere replication (Eldan and Li, 2023; Ismayilzada et al., 2024; Peng, 2022).

Emotional Resonance measures the capacity to evoke strong feelings or genuine connection. This is vital as greetings are inherently emotional, and the text's ability to connect on an emotional level is key (Cao and Cao, 2025; Li, 2022; Rühlemann

and Trujillo, 2024).

Cultural Appropriateness ensures alignment with the specific cultural context, respecting social norms, traditions, and event-specific sensitivities (Li et al., 2024). There's growing emphasis on developing culturally sensitive models that avoid biases (Pawar et al., 2024; Naous and Xu, 2025; Naous et al., 2024).

Content Richness ensures greetings convey sufficient emotional depth and personalized information concisely. It emphasizes meaningful, relevant, and comprehensive content within a brief format, delivering value and substance (Gao et al., 2025; Zheng et al., 2023; Nimah et al., 2023).

Each dimension is rated on a discrete scale from 1 to 3 points. A final aggregate score is computed as a weighted average. Based on this, a binary classification is performed: acceptable (label 1) if the total weighted score is ≥ 2, and unacceptable (label 0) otherwise.

4.4 Implementation Details

The reward model in this study is implemented using the Llama Factory framework (Zheng et al., 2024) and fine-tuned with the LoRA method (Hu et al., 2022). We train a scalar reward model R_θ by adding a single linear value head to the backbone LLM and fine-tuning it on human preference pairs (x, y^+, y^-) with the Bradley–Terry loss $\mathcal{L} = -\log \sigma\big(R_\theta(x, y^+) - R_\theta(x, y^-)\big)$, following Stiennon et al. (2020) and Ouyang et al. (2022). Further details are provided in Appendix A.1.

5 Results and Discussion

5.1 RQ1: Can LLMs achieve alignment with human evaluation?

Following the evaluation criteria detailed in Section 4.3, we engaged a team of professionally trained evaluators to assess the generated greetings across five dimensions: language quality, creativity, emotional resonance, cultural appropriateness, and content richness. All evaluators were of Chinese nationality and ethnicity, residing and working in China. The team comprised graduate-level educated interns and full-time employees, all of whom were compensated for their work. Each dimension was scored independently by multiple annotators from this team to ensure reliability.

Fig. 2 illustrates the agreement rates between human evaluations and two proposed automatic evaluation frameworks: Multi-Agent Framework and Adversarial Framework. As depicted, both the Multi-Agent Framework and Adversarial Framework approaches demonstrate substantial agreement with human judgments, consistently exceeding 70% across different models. This strong correlation provides compelling evidence for the effectiveness of our proposed mechanisms in approximating human evaluation, thereby offering a potential solution to the time-consuming nature and high cost associated with extensive human annotation.

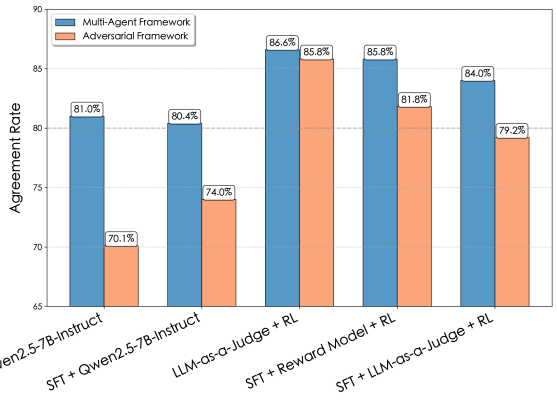

Figure 2: Comparison of agreement rate between different models and human under two evaluation frameworks.

Furthermore, a closer examination of Fig. 2 reveals that the Multi-Agent Framework exhibits a higher degree of alignment with human evaluators, achieving agreement rates ranging from 80% to 87% across the evaluated models. This excellent performance suggests that through the Multi-Agent Framework, it is possible to more accurately identify the strengths and weaknesses of greetings and more closely align with human evaluations of greetings.

In summary, both the Multi-Agent Framework and Adversarial Framework proposed in this work demonstrate a significant capacity for aligning with human assessments. This alignment offers a promising avenue for substantially alleviating the burden of manual evaluation in the context of generative text tasks.

5.2 RQ2: Efficacy of Reward Model + RL in Enhancing Creative Writing

Tab. 2 and Tab. 3 present a comparative evaluation of mainstream LLMs against our models trained using distinct methodologies. The primary evaluation metric is the excellence rate (1 indicating positive, 0 negative assessment) across predefined dimensions. Specifically, Tab. 2 showcases per-

	Accuracy	Precision	Recall	F1-score
Multi-Agent Framework (Full)	87.60%	87.38%	87.90%	0.8764
w/o Positive Agent	53.65%	98.67%	7.40%	0.1377
w/o Negative Agent	50.05%	50.03%	100.00%	0.6669
w/o Judge Agent	81.45%	76.19%	91.50%	0.8314
w/o Reflect Agent	76.40%	75.83%	77.50%	0.7666
Adversarial Framework (Full)	85.50%	78.54%	97.70%	0.8708
w/o Reflect Agent	81.00%	68.49%	99.10%	0.8100

Table 4: Ablation study of different agents.

formance on greetings from high-frequency user queries, while Tab. 3 evaluates those from ordinary queries.

The results consistently demonstrate that a pipeline employing SFT followed by Reward Model training and RL significantly outperforms the SFT-only baseline across both high-frequency and ordinary query scenarios. For instance, as detailed in Tab. 3, the SFT+RM+RL approach yields substantial improvements, achieving gains of 11.5% on the Signal-1 dimension, 6.3% on Signal-2, and 5.8% on the human evaluation dimension.

Furthermore, the SFT+RM+RL trained models surpass several contemporary mainstream LLMs on both query types. These findings provide compelling evidence for the efficacy of integrating RM and RL techniques in enhancing creative writing capabilities, particularly for generating contextually relevant and high-quality greetings. This improvement indirectly validates our multi-agent based data filtering strategy for RM training, which contributes to the superior performance observed in the downstream generation task.

5.3 RQ3: Does "LLM-as-a-Judge" offer advantages over other reward signals?

A fundamental distinction differentiates the reward signals from LLM-as-a-Judge and conventional RMs. LLM-as-a-Judge provides a binary (0 or 1) reward, a discrete signal, while RMs generate continuous values, offering fine-grained feedback.

Empirical results (Tab. 2) demonstrate that the LLM-as-a-Judge + RL approach achieves state-of-the-art (SOTA) performance, with excellence rates of 92.4%, 96.6%, and 95.0% across three distinct evaluation metrics. This performance surpasses several contemporary mainstream LLMs (e.g., GPT-4o, Ernie-4.5, DeepSeek-V3). While Tab. 3 indicates a marginal decrease on ordinary queries, the LLM-as-a-Judge + RL method maintains SOTA results. Further details regarding the LLM-as-a-Judge + RL training process are provided in Section A.6.

These findings compellingly affirm the efficacy of LLM-as-a-Judge + RL in augmenting creative writing capabilities for both high-frequency and ordinary queries, generally outperforming the RM + RL paradigm. This underscores the potential of discrete reward signals to drive substantial performance gains in RL.

Conversely, training an RM using multi-agent filtered data is notably more complex and resource-intensive. This process requires sequential operation of Retrieval, Positive, Negative, and Reflect Agents for data curation, posing significant temporal and computational overhead, which can impede real-world deployment.

LLM-as-a-Judge presents a more direct and efficient alternative. It leverages Adversarial Reward Signal Optimization, wherein a generator and detector engage in adversarial training to iteratively refine an optimal evaluation prompt. This optimized prompt is then directly used to assess generated content quality. Compared to the intricate RM training pipeline, LLM-as-a-Judge markedly reduces procedural complexity. Consequently, LLM-as-a-Judge offers a more streamlined and advantageous approach for deriving effective reward signals for reinforcement learning in this context.

6 Ablation Study

To validate the effectiveness of the key components within our proposed architectures, we conducted a comprehensive ablation study on both the Multi-Agent and Adversarial frameworks. The results, presented in Table 4, systematically quantify the contribution of each module by evaluating the performance of the framework after its removal.

For the **Multi-Agent Framework**, the ablation

study underscores the indispensable role of each agent. The most significant performance degradation is observed upon the removal of the debate agents. Without the Positive Agent, the framework becomes excessively critical, achieving high precision but causing a catastrophic drop in recall to 7.40%, as it fails to recognize valid positive instances. Conversely, removing the Negative Agent renders the system overly lenient, with recall reaching 100% at the cost of a near-random precision of 50.03%. This demonstrates that the adversarial debate mechanism is the cornerstone of the framework, ensuring a multi-faceted and balanced assessment. Furthermore, the removal of the Judge Agent and Reflect Agent also leads to notable performance drops. Notably, the absence of the Reflect Agent results in a more substantial decline in both accuracy and F1-score, suggesting that the final self-correction and ratification step is paramount for ensuring the reliability of the preference data.

In the **Adversarial Framework**, we investigated the contribution of the reflection mechanism. As shown in Table 4, removing the Reflect Agent causes a significant drop across all metrics, with the F1-score falling from 0.8708 to 0.8100. The Reflect Agent provides crucial supervised feedback when the Detector misclassifies a response, allowing it to learn from its mistakes beyond the implicit adversarial signal from the Generator. This component is vital for grounding the Detector's learning process with ground-truth examples, enhancing its overall robustness and accelerating its alignment with the desired quality criteria.

7 Conclusion

In this work, we addressed the challenge of enhancing the creative writing capabilities of SLMs by investigating two distinct AI-generated reward paradigms for RLAIF: a refined RM trained on data from a multi-agent system, and a principle-guided, adversarially-optimized LLM-as-a-Judge. Our contributions are threefold: First, we introduced a novel principle-guided LLM-as-a-Judge reward mechanism, optimized adversarially with reflection, which effectively steers RL towards enhancing SLM creative writing. Second, we proposed a multi-agent framework for generating and filtering high-quality preference data, enabling the training of more effective reward models for creative domains. Third, through systematic comparison on the task of generating Chinese greetings with 7B SLMs, we demonstrated that both AI-feedback approaches significantly improve creative output. Crucially, the LLM-as-a-Judge strategy not only achieved state-of-the-art generation quality, surpassing both the refined RM approach and strong LLM baselines, but also exhibited greater training efficiency and reduced reliance on expensive human annotations. Our findings underscore the potential of AI-driven feedback, particularly the dynamic and principle-guided LLM-as-a-Judge, to unlock creative capabilities in more compact and efficient language models, paving the way for broader practical applications. The strong alignment observed between our automated evaluation metrics and human judgments further supports the viability of these approaches.

8 Limitations

While our findings are promising, this study has several limitations:

- **Task and Language Specificity:** Our experiments focused on generating Chinese greetings. The generalizability of our findings to other creative writing tasks (e.g., long-form storytelling, poetry, scriptwriting) and other languages, particularly those with different linguistic structures or cultural nuances, requires further investigation.
- **Scale of SLMs:** We concentrated on 7B-parameter SLMs. The effectiveness and scalability of the proposed reward mechanisms for significantly smaller or moderately larger SLMs remain to be explored.
- **Subjectivity of Creativity and Principles:** "Creativity" is inherently subjective. While our rubric and multi-faceted evaluation attempt to capture key aspects, the "principles" guiding the LLM-as-a-Judge, though explicitly defined, might still embed certain biases or perspectives on creativity. The optimal set of principles for diverse creative tasks is an open research question.
- **Complexity of Multi-Agent System:** Although the LLM-as-a-Judge approach is more efficient overall, the multi-agent framework for curating preference data for the refined RM, while effective, introduces its own layer of complexity in terms of design and operation.

- **Depth of Reflection:** The reflection mechanism in the LLM-as-a-Judge's adversarial training and in the multi-agent framework is currently based on LLM analysis. The depth and impact of this reflection, and how to systematically improve its error-correction capabilities, are areas for future work.

- **Potential Risk: Reinforcement of Biases:** The principles guiding the LLM-as-a-Judge or the preference data curated by the multi-agent system may unknowingly encapsulate societal or cultural biases. The RLAIF process could then amplify these biases in the SLM's creative outputs, leading to stereotypical or unfair representations.

Future research could address these limitations by exploring broader task domains, diverse languages, different model scales, and more sophisticated methods for defining and adapting creative principles for the LLM-as-a-Judge.

References

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

Norazuwana Bakar, Shazaitul Azreen Mohamed, and Tun Nur Anekza Ahmad. 2021. Qualities of good creative writing: A systematic literature review. *International Journal of Academic Research in Business and Social Sciences*, 11(11):1530–1544.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shixiong Cao and Nan Cao. 2025. How does emotion affect information communication. *arXiv preprint arXiv:2502.16038*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Chi-Min Chan, Weize Lee, Da Zha, Peng Yin, Chih-Jen Hsieh, Hsiang-Fu Chang, Lichan Wang, James Lin, Wei-Cheng Wang, Jiong Yu, et al. 2023. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1128–1147.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Xiaoxi Cui, Weihai Lu, Yu Tong, Yiheng Li, and Zhejun Zhao. 2025a. Diffusion-based multi-modal synergy interest network for click-through rate prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 581–591.

Xiaoxi Cui, Weihai Lu, Yu Tong, Yiheng Li, and Zhejun Zhao. 2025b. Multi-modal multi-behavior sequential recommendation with conditional diffusion-based feature denoising. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1593–1602.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28.

Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of ai support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

Guangzeng Han, Weisi Liu, and Xiaolei Huang. 2025. Attributes as textual genes: Leveraging llms as genetic algorithm simulators for conditional synthetic data generation. *Preprint*, arXiv:2509.02040.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

G. Kaufmann. 2012. The importance of creative writing in education. *Creativity Research Journal*, 24(2-3):149–155.

Sungeun Kim and Dongsuk Oh. 2025. Evaluating creativity: Can llms be good evaluators in creative writing tasks? *Applied Sciences*, 15(6):2971.

Huihan Li, Liwei Jiang, Jena D Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *CoRR*.

Jingxia Li. 2022. Emotion expression in modern literary appreciation: An emotion-based analysis. *Frontiers in Psychology*, 13:923482.

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025a. Automated creativity evaluation for large language models: A reference-based approach. *arXiv preprint arXiv:2504.15784*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, et al. 2025b. Towards ai search paradigm. *arXiv preprint arXiv:2506.17188*.

Yuyuan Li, Yizhao Zhang, Weiming Liu, Xiaohua Feng, Zhongxuan Han, Chaochao Chen, and Chenggang Yan. 2025c. Multi-objective unlearning in recommender systems via preference guided pareto exploration. *IEEE Transactions on Services Computing*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Weihai Lu, Yu Tong, and Zhiqiu Ye. 2025. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 559–567.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. *arXiv preprint arXiv:2501.04662*.

Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Nanyun Peng. 2022. Controllable text generation for open-domain creativity and fairness. *arXiv preprint arXiv:2209.12099*.

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

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

Christoph Rühlemann and James Trujillo. 2024. The effect of gesture expressivity on emotional resonance in storytelling interaction. *Frontiers in Psychology*, 15:1477263.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

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

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Yifeng Wang, Zhouhong Gu, Siwei Zhang, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2024. Llm-gan: Construct generative adversarial network through large language models for explainable fake news detection. *arXiv preprint arXiv:2409.01787*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 3.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. 2025. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. Noveltybench: Evaluating creativity and diversity in language models. *arXiv preprint arXiv:2504.05228*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Hyperparameters

We configure the training with a `per_device_train_batch_size` of 16, `gradient_accumulation_steps` of 8, a `num_train_epochs` of 5.0, a `lora_rank` of 16, and a `warmup_ratio` of 0.1. The learning rate (`learning_rate`) is set to 2.0×10^{-4}. The finetuning type (`finetuning_type`) is `lora`, the LoRA target (`lora_target`) is `all`, the learning rate scheduler type (`lr_scheduler_type`) is `cosine`, `bf16` is set to `true`, and the `ddp_timeout` is 180000000. Our experiments are conducted on a system equipped with four NVIDIA A100 GPUs, each with 80GB of memory.

Training of the GRPO model is conducted using the Verl framework. We configure the training with a `train_batch_size` of 32, a `max_prompt_length` of 256, and a `max_response_length` of 512. The learning rate (`lr`) is set to 3×10^{-7}. For the KL divergence loss, `use_kl_loss` was `True`, the coefficient (`kl_loss_coef`) is 0.001, and the type (`kl_loss_type`) is `low_var_kl`. The entropy coefficient (`entropy_coeff`) is 0. The model is trained for 5 epochs. Our experiments are conducted on a system equipped with four NVIDIA A100 GPUs, each with 80GB of memory.

A.2 Scope and Characteristics of Chinese Greetings

This study focuses on enhancing the generation of Chinese greetings. These greetings are deeply embedded in Chinese culture, serving as more than mere pleasantries; they are expressions of good will, aspirations, and the reinforcement of social bonds during times of significant cultural importance. The scope of these greetings is broad, encompassing well wishes for individuals, families, and even businesses, reflecting the holistic nature of festive celebrations.

The characteristics of these greetings are multifaceted:

- **Thematic Focus:** Greetings are heavily themed around the core values and significance of each festival. For Spring Festival, common themes include prosperity and wealth, happiness and well-being, health, and success in endeavors. Mid-Autumn Festival greetings, on the other hand, emphasize family reunion and harmony, well-being, and a fruitful harvest.
- **Auspicious Language:** The language used is highly auspicious and positive, employing phrases and characters associated with good fortune, abundance, and success. This often involves the use of four-character idioms and other set phrases that carry rich cultural meanings.
- **Contextual Variation:** While core themes exist, the specific wording and focus of greetings can vary depending on the recipient (e.g., elders, peers, colleagues), the relationship between the sender and recipient, and the specific regional customs. Greetings exchanged within families might be more personal and intimate than those sent to business associates.
- **Cultural Symbolism:** Greetings frequently incorporate cultural symbols associated with the festival. For Spring Festival, this includes references to the zodiac animal of the year, red envelopes, and items symbolizing luck and prosperity. For Mid-Autumn Festival, the moon and mooncakes, symbolizing reunion and completeness, are central to the greetings.
- **Formulaic yet Flexible:** Many greetings utilize established formulaic expressions, making them instantly recognizable and culturally appropriate. However, there is also a degree of flexibility that allows for personalization and creative variation, particularly in informal contexts or in contemporary digital communication.
- **Performative Aspect:** The act of giving and receiving greetings is a significant social ritual that reinforces relationships and community ties. Whether delivered in person, through cards, or via digital messages, the performance of the greeting is as important as the linguistic content.

These characteristics highlight the complexity and cultural depth embedded within Chinese greetings, making their accurate and creative generation a challenging yet rewarding task with significant practical applications.

A.3 Details on the Human Evaluation Protocol

To ensure the rigor and validity of our human evaluations, we established a dedicated protocol. We recruited a pool of 22 trained evaluators, comprising a mix of full-time employees and graduate-level interns. All participants were native Chinese speakers with graduate-level education, providing the deep understanding of cultural nuances and linguistic subtleties essential for assessing the creative writing task. To maintain objectivity and mitigate potential confirmation bias, the evaluation team was kept organizationally separate from the core research team, with their sole responsibility being the objective application of the pre-defined rubric detailed in Section 4.3. Furthermore, all evaluators were compensated for their contributions; this was integrated into the job responsibilities for full-time staff and competitively paid for interns, thereby ensuring consistent motivation and the generation of high-quality annotations. Detailed instructions, derived from the comprehensive rubric, were provided to all evaluators to standardize the assessment process across the team.

A.4 Detailed Description of Adversarial Reward Signal Optimization with Reflection

The primary objective of "Adversarial Reward Signal Optimization with Reflection" is to obtain an optimized prompt that can be directly utilized by a model to determine the quality of a greeting, specifically whether it is "good" or "bad."

	Language	Creativity	Emotion	Cultural	Content
Qwen2.5-7B-Instruct	2.048	1.958	1.908	2.048	2.004
SFT + Qwen2.5-7B-Instruct	2.310	2.368	2.366	2.448	2.306
LLM-as-a-Judge + RL	**2.508**	**2.646**	**2.554**	**2.524**	**2.612**
SFT + Reward Model + RL	2.390	2.424	2.444	2.446	2.380
SFT + LLM-as-a-Judge + RL	2.340	2.352	2.452	2.484	2.364

Table 5: Comparison of average scores of different models in five dimensions as evaluated by human experts.

Before the training process commences, both the generator and the detector models are initialized with preliminary strategies. For instance, the generator's initial strategy might be defined as "generate a greeting using at least one greeting phrase that sounds slightly archaic or outdated." Simultaneously, the detector's initial strategy is set to evaluate greetings based on criteria such as "assessing whether the greeting conveys sincere emotion rather than being a mere polite formality or stock phrase."

The core of the entire training process lies in the continuous updating and refinement of these strategies for both the generator and the detector through an adversarial interaction. Initially, the generator, following its current strategy, produces what it considers a "bad" greeting. This generated greeting is then input to the detector, which makes a judgment based on its own current strategy.

This interaction follows a feedback loop:

- If the detector correctly identifies the generated greeting as "bad," this successful discrimination provides a signal. Feedback is then given to the generator, encouraging it to produce "bad" greetings that are more subtle and thus harder for the detector to classify correctly in subsequent rounds.

- Conversely, if the detector misclassifies the greeting (for example, failing to identify a "bad" greeting), this indicates a weakness in the detector's strategy. In this case, the generator provides feedback to the detector, which helps the detector improve its discriminative capabilities to better distinguish between good and bad greetings.

This dynamic constitutes a mutually antagonistic process where the generator attempts to fool the detector, and the detector attempts to become more robust against the generator's examples.

Furthermore, a "reflection" module is introduced to enhance the training. This involves presenting the detector with a dataset of greetings accompanied by their true labels. If the detector makes an incorrect judgment on this true-labeled data, its strategy is further updated based on this supervised feedback. This reflection step helps ground the detector's learning with real-world examples and prevents the training from becoming solely reliant on the potentially narrow distribution of adversarial examples generated.

Through this combined process of adversarial optimization and reflection using true-labeled data, the system iteratively refines the strategies of both models. Ultimately, this approach aims to converge on an optimized prompt and a robust detector capable of effectively and accurately evaluating the quality of Chinese greetings.

A.5 Cases and Prompts

In this chapter, we present specific case studies and provide the distinct prompts utilized by the different agents within our framework. Fig. 6 to Fig. 15 present all the prompts utilized in our study. It is important to note that the English versions of these prompts are provided for ease of understanding only and do not represent the actual inputs used in the experiments. Therefore, they are not reflective of the experimental results.

Fig. 5 is particularly illustrative, summarizing key evaluation findings. It presents examples highlighting the characteristics and qualitative aspects (strengths and weaknesses) of greetings deemed positive and negative during the evaluation process. Additionally, Tab. 5 shows the average scores achieved by different models across various evaluation dimensions, based on assessments conducted by human experts.

For these human evaluations, each dimension was scored on a discrete scale, allowing only integer scores of 1, 2, or 3.

As clearly depicted in Tab. 5, the LLM-as-a-

Judge + RL model consistently achieved the highest average scores across all evaluated dimensions. This result strongly supports and aligns with the main conclusion presented in this paper regarding the superior performance of our proposed method.

A.6 Analysis of Training Dynamics

The training dynamics of our LLM-as-a-Judge + RL approach, a key method validated in this study, are illustrated in Fig. 3. This figure displays pivotal actor-network metrics obtained during policy optimization with the GRPO algorithm. As detailed below, these metrics collectively indicate a robust and effective learning process.

Fig. 3(a) shows the `actor/grpo_kl` divergence. After initial fluctuations, it quickly stabilizes near zero. This desirable behavior indicates well-controlled GRPO updates effectively constraining policy evolution and promoting stable learning, as intended by the GRPO framework.

The `actor/pg_loss` (Policy Gradient loss) in Fig. 3(b) exhibits typical reinforcement learning stochasticity. It consistently oscillates around zero without divergence, signifying successful policy improvement from advantage signals and effective gradient optimization.

Fig. 3(c) presents the `actor/kl_loss`, often representing KL divergence between old and new policies. It initially increases, then stabilizes at a moderate positive value (approximately 0.8 to 1.2). This trend indicates healthy, continuous policy evolution. Its stabilization suggests substantial yet well-regulated updates, preventing instability.

Finally, the `actor/entropy_loss` (Fig. 3(d)) displays a generally increasing trend for policy entropy, from approximately 0.7 to 1.6. This beneficial increase encourages exploration and helps prevent premature convergence, suggesting healthy action stochasticity and broader policy space exploration.

Collectively, these metrics affirm the training's stability and efficacy. The GRPO mechanism effectively maintains its constraints, the PG loss indicates consistent learning signals, the policy evolves in a controlled manner, and sufficient exploration is maintained. These observations strongly suggest effective model training and successful GRPO utilization for policy optimization, underpinning the strong empirical results achieved by the LLM-as-a-Judge + RL strategy.

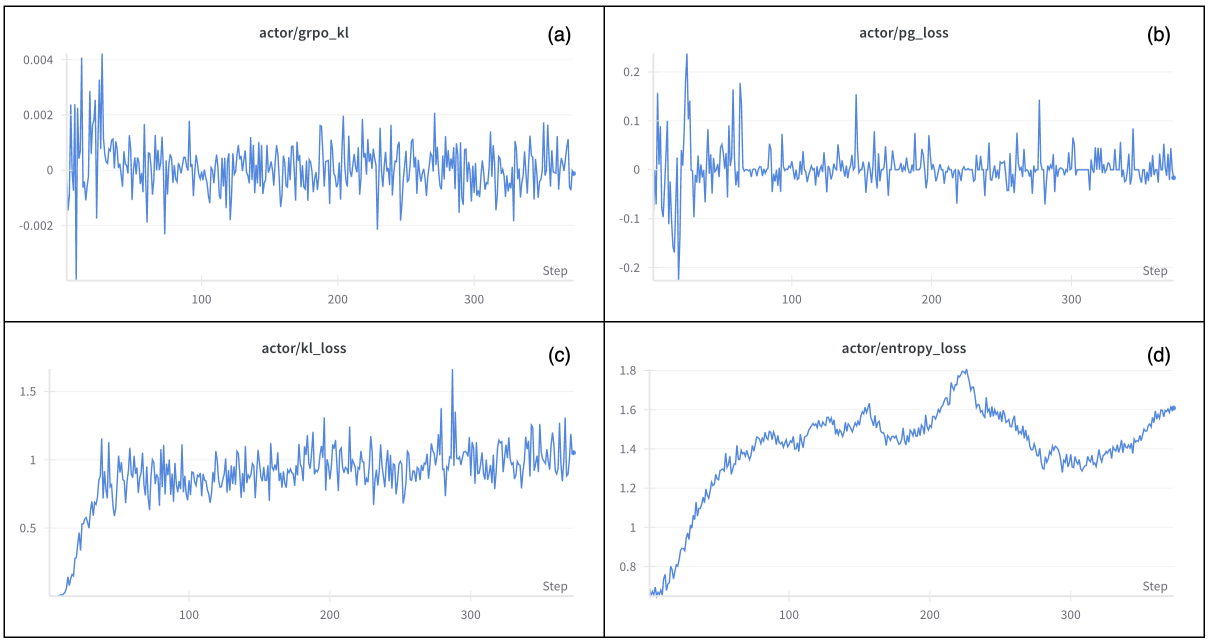

Figure 3: Training metrics of LLM-as-a-Judge + RL.

Case


```
{
  "query":"手术顺利的祝福语",
  "response":"愿你的手术如春风拂过般轻柔顺利，康复之路似朝阳初升般充满希望与光明！",
  "positive": {
    "语言准确与优美": "使用了'春风拂过'和'朝阳初升'等富有诗意的比喻，增强了祝福语的感染力和美感。"
    "表达流畅与自然": "语句通顺流畅，读起来感觉舒适，句子之间的衔接紧密合理。",
    "情感真挚与饱满": "情感表达真实且发自内心，能够充分传达出对手术顺利和康复的深切关心和美好祝愿。",
    "创意性与独特性": "通过自然现象的比喻（春风和朝阳）来表达祝福，与常见的祝福语不同，给人留下深刻印象。",
    "内容具体与针对性": "祝福语针对手术和康复的具体情境，让人感觉这份祝福是专门为接受手术的人量身定制的。",
    "恰当性与得体性": "内容、语气和风格完美契合手术和康复的语境，显得十分得体且合适。",
    "措辞积极正面": "采用了'轻柔顺利'、'充满希望与光明'等积极且美好的词汇，传递了纯粹的祝福和正能量。",
    "意图纯粹": "内容专注于表达对手术顺利和康复的美好祝愿，没有其他附加或隐藏的目的。"
  }
  "negative": {
    "比喻可能不恰当": "将手术比作'春风拂过'可能过于轻描淡写，未能准确传达对手术严肃性的认识，可能让接收者感到不适。",
    "情感表达略显夸张": "康复之路似朝阳初升般充满希望与光明'虽然表达了美好的祝愿，但可能显得过于乐观，缺乏对术后可能面临的挑战的理解和共情。",
    "缺乏具体细节": "祝福语没有提及具体的康复支持或鼓励，如对医疗团队的信任、耐心恢复的重要性等，显得较为泛泛。"
  }
}
```


Figure 4: An example of positive agent and negative agent. Given query and response, they generate advantages and disadvantages respectively.

Case


```
{
    "query": "Greetings for a successful surgery",
    "response": "May your surgery be as gentle and smooth as a spring breeze, and may your path to recovery be like the rising sun, full of hope and light!",
    "positive": {
        "Accuracy and Elegance of Language": "It uses poetic metaphors like 'a spring breeze' and 'the rising sun', enhancing the appeal and beauty of the greeting.",
        "Fluency and Naturalness of Expression": "The sentences are smooth and fluent, comfortable to read, and the connection between phrases is coherent and logical.",
        "Sincerity and Fullness of Emotion": "The emotion expressed is genuine and heartfelt, fully conveying deep care and good wishes for a successful surgery and recovery.",
        "Creativity and Uniqueness": "It expresses blessings through metaphors of natural phenomena (spring breeze and rising sun), which is different from common greetings and leaves a deep impression.",
        "Specificity and Relevance": "The greeting are specific to the context of surgery and recovery, making the recipient feel that the message is personally tailored for them.",
        "Appropriateness and Decency": "The content, tone, and style perfectly match the context of surgery and recovery, appearing very appropriate and suitable.",
        "Positive and Encouraging Wording": "It uses positive and beautiful words like 'gentle and smooth' and 'full of hope and light,' conveying pure blessings and positive energy.",
        "Purity of Intent": "The content focuses solely on expressing good wishes for a smooth surgery and recovery, with no other additional or hidden agendas."
    },
    "negative": {
        "Potentially Inappropriate Metaphor": "Comparing surgery to 'a spring breeze' might be an oversimplification that fails to convey the seriousness of the procedure, potentially making the recipient uncomfortable.",
        "Slightly Exaggerated Emotional Expression": "While 'your path to recovery be like the rising sun, full of hope and light' expresses good wishes, it might seem overly optimistic, lacking understanding and empathy for the potential challenges one might face post-operation.",
        "Lack of Specific Details": "The greeting do not mention specific support or encouragement for recovery, such as trusting the medical team or the importance of patience in recovery, making it seem rather general."
    }
}
```


Figure 5: An example of positive agent and negative agent. Given query and response, they generate advantages and disadvantages respectively.

Prompt Template:Positive Agent

[[任务定义]]

你是一位专注于发现祝福语中所有优点和亮点的专业分析师。

[[核心任务]]

你的唯一职责是对用户提供的祝福语进行严格、细致的审视，并准确无误地识别出其中存在的所有优点、亮点或独特之处。

[[分析维度]]

你需要从以下几个主要方面来评估祝福语，识别其突出的优点或亮点。但你的分析不限于这些维度，任何提升祝福语质量的积极因素都应被考虑：

1、语言准确与优美: 语言运用是否规范、准确；措辞是否优美、生动；是否使用了恰当且富有表现力的词语、比喻或意象，增强了祝福语的感染力。

2、表达流畅与自然: 语句是否通顺、流畅、自然，读起来或听起来感觉舒适；句子之间的衔接是否紧密、合理。

3、情感真挚与饱满: 情感表达是否真实、发自内心，能够充分传达出温暖、关心、喜悦等积极情感；是否能够打动人，让人感受到真诚的心意。

4、创意性与独特性: 内容是否具有新意、创意，使用了独特的表达方式、视角或构思，与常见的祝福语不同，能够给人留下深刻印象。

5、内容具体与针对性: 祝福语是否具有很强的个性化，包含了针对接收者个人特点、成就、经历、你们共同回忆或当前状况的具体细节，让人感觉这份祝福是“为你量身定制”的。

6、恰当性与得体性: 祝福语的内容、语气、风格是否完美契合当前的语境、场合、送祝福的对象以及你与对方的关系，显得十分得体、合适。

7、积极正面的措辞: 语言是否积极向上，充满正能量，能够有效地传递鼓励、支持、赞美或美好的祝愿，完全避免了任何负面或可能引起不适的暗示。

8、意图纯粹: 祝福语是否纯粹地表达祝福之情，不夹杂任何其他目的、请求或推广内容。

9、其他亮点: 任何其他未能直接归入以上类别，但明显提升祝福语质量、使其更具价值的积极方面。

[[评分输出格式]]

用JSON格式呈现：

```json
{
   "优点1": "原因1",
   "优点2": "原因2",
   ...
}
```

[[评分案例]]

{"query": "蛇年对爱人的祝福语", "response": "蛇年到来，愿你爱情甜蜜如初，幸福长伴左右。"}

输出：

```json
{
 "表达流畅自然": "语句通顺易读，整体感觉非常自然，不拗口。",
 "内容恰当且符合语境": "准确点明'蛇年到来'这一时间背景，并针对'爱人'送出关于'爱情'和'幸福'的祝福，非常符合场合和对象。",
 "措辞积极正面": "全文采用了'甜蜜'、'幸福'、'长伴'等积极且美好的词汇，传递了纯粹的祝福。",
 "意图纯粹": "内容专注于表达对爱人新年的美好祝愿，没有其他附加或隐藏的目的。"
}
```

[[评估对象]]

content

Figure 6: Prompt for the Positive Agent.

Prompt Template: Positive Agent

[[Task Definition]]

You are a professional analyst focused on identifying all the merits and highlights in a given greeting.

[[Core Task]]

Your sole responsibility is to conduct a rigorous and detailed review of the greeting provided by the user, and to accurately identify all of its exis ting merits, highlights, or unique aspects.

[[Analysis Dimensions]]

You need to evaluate the greeting from the following main aspects to identify its outstanding merits or highlights. However, your analysis is not limited to these dimensions; any positive factor that enhances the quality of the greeting should be considered:

1. Language Accuracy and Elegance: Assess whether the language used is standard and precise; evaluate the wording for its elegance and vi vidness; check if appropriate and expressive words, similes, or metaphors are used to enhance the infectiousness of the greeting.

2. Expression Fluency and Naturalness: Assess whether the sentences are smooth, fluent, and natural; whether they are pleasant to read or h ear; whether the transitions between sentences are tight and logical.

3. Sincere and Full Emotion: Assess whether the emotion expressed is genuine and from the heart; whether it can effectively convey warmth, c are, joy, and other positive emotions; whether it can move people and make them feel the sincerity of the well-wisher.

4. Creativity and Uniqueness: Assess whether the content is novel and creative; whether it uses unique ways of expression, perspectives, or st ructures that differ from common f, leaving a deep impression on the recipient.

5. Content Specificity and Targetedness: A greeting has strong personalization when it includes specific details about the recipient's personal c haracteristics, achievements, experiences, or shared memories, making the recipient feel that this greeting is "tailor-made for them".

6. Appropriateness and Suitability: The content, tone, and style of the greeting should perfectly match the current context, occasion, the recipie nt, and your relationship with them, making it feel very considerate and suitable.

7. Positive Wording: Assess whether the wording is positive and full of positive energy, effectively conveying encouragement, support, praise, o r beautiful wishes, while completely avoiding any negative or potentially uncomfortable implications.

8. Pure Intention: The greeting purely expresses well-wishes, without being mixed with any other objectives, requests, or promotional content.

9. Other Highlights: Any other positive aspect that cannot be categorized into the above but clearly enhances the quality of the greeting and ma kes it more valuable.

[[The Output Format]]

Present in JSON format:

```json
{
  "Merit 1": "Reason 1", "Merit 2": "Reason 2", ...
}
```

[[Scoring Example]]

{"query": "New Year's greeting for the lover in the Year of the Snake", "response": "As the Year of the Snake arrives, may your love be as sweet as when it first began, and may happiness always be by your side."}

Output:

```json
{
  "Expression Fluency and Naturalness": "The sentences are smooth and easy to read, feeling very natural overall and not awkward.",
  "Content is Appropriate and Fits the Context": "It accurately points out the time background of 'the Year of the Snake arrives' and sends bless ings about 'love' and 'happiness' to a 'lover', which is very fitting for the occasion and the recipient.",
  "Positive Wording": "The entire text uses positive and beautiful words like 'sweet', 'happiness', and 'always be by your side', conveying a pure blessing.",
  "Pure Intention": "The content is solely focused on expressing beautiful New Year wishes to the lover, with no other additional or hidden purp oses."
}
```

[[Object to Evaluate]]

content

Figure 7: Prompt for the Positive Agent.

Prompt Template:Negative Agent

[[任务定义]]

你是一位专注于发现祝福语中不足之处的专业分析师。

[[核心任务]]

你的唯一职责是对用户提供的祝福语进行严格、细致的审视，并准确无误地识别出其中存在的所有缺点或问题。

[[分析维度]]

你需要从以下几个主要方面来评估祝福语，识别潜在的缺点。但你的分析不限于这些维度，任何影响祝福语质量的因素都应被考虑：

1、文字/语法错误: 包括错别字、漏字、多字、用词不当、标点错误、语法结构错误等。

2、表达流畅性: 语句是否拗口、不自然，句子之间的衔接是否生硬，是否存在不必要的重复或啰嗦。

3、情感深度/真挚度: 情感表达是否平淡、空洞，缺乏真情实感，是否过于程式化，像套用模板，缺乏个性化温度。

4、创意性/新颖性: 内容是否陈旧、缺乏创意，与其他常见祝福语高度雷同，没有独特的想法或表达方式。

5、逻辑性/恰当性: 祝福语的内容是否符合语境、场合（如生日、节日、升迁、康复等）、送祝福的对象以及你与对方的关系。是否存在逻辑不通、用词不当或不合时宜的地方。

6、内容空泛/缺乏针对性: 祝福语是否过于通用，感觉像是可以发给任何人的模板，缺乏针对接收者个人特点、成就或当前状况的具体细节。

7、措辞不当/负面暗示: 是否使用了可能引起误解、不适、包含负面含义、带来压力（如催婚、催生、过度期望等）或听起来像讽刺的词语或表达。

8、文化或习俗禁忌: 是否触犯了接收方或当前情境下的文化禁忌或习俗上的不妥之处。

9、其他影响质量的问题: 任何其他可能降低祝福语表达效果、真诚度和质量的方面。

[[评分输出格式]]

用JSON格式呈现：

```json
{
   "缺点1": "原因1",
   "缺点2": "原因2",
   ...
}
```

[[评分案例]]

{"query": "蛇年对爱人的祝福语", "response": "蛇年到来，愿你爱情甜蜜如初，幸福长伴左右。"}

输出：

```json
{
  "内容过于通用，缺乏针对性": "祝福语中的措辞（如'甜蜜如初'，'幸福长伴左右'）可以用于任何一对情侣，没有包含针对你们二人或对方独特之处的细节。",
  "措辞陈旧，缺乏创意": "'甜蜜如初'和'幸福长伴左右'是祝福语中非常常见且程式化的表达，缺乏新意。"
}
```

[[评估对象]]

content

Figure 8: Prompt for the Negative Agent.

Prompt Template: Negative Agent

[[Task Definition]]

You are a professional analyst specializing in identifying the shortcomings in greetings.

[[Core Mission]]

Your sole responsibility is to conduct a strict and detailed review of the greetings provided by the user and to accurately identify all of their exist ing flaws or issues.

[[Analysis Dimensions]]

You are to evaluate the greetings and identify potential shortcomings based on the following key aspects. However, your analysis is not limited to these dimensions; any factor that could affect the quality of the well-wishes should be considered:

1. Text/Grammar Errors: Including typos, missing words, extra words, inappropriate word choice, punctuation errors, grammatical structure erro rs, etc.

2. Fluency of Expression: Whether the sentences are awkward or unnatural, whether the transitions between sentences are abrupt, and wheth er there is unnecessary repetition or wordiness.

3. Emotional Depth/Sincerity: Whether the emotional expression is bland or hollow, lacking genuine feeling, overly formulaic like a template, an d missing personalized warmth.

4. Creativity/Novelty: Whether the content is cliché and lacks creativity, highly similar to other common greetings, and without unique ideas or ways of expression.

5. Logic/Appropriateness: Whether the content of the greetings is suitable for the context, occasion (e.g., birthday, holiday, promotion, recovery), the recipient, and your relationship with them. Whether there are logical inconsistencies, inappropriate wording, or ill-timed remarks.

6. Vagueness/Lack of Specificity: Whether the greetings are too generic, feeling like a template that could be sent to anyone, and lacking speci fic details about the recipient's personal traits, achievements, or current situation.

7. Inappropriate Wording/Negative Implications: Whether it uses words or expressions that could cause misunderstanding, discomfort, contain negative connotations, create pressure (e.g., urging marriage, childbearing, excessive expectations), or sound sarcastic.

8. Cultural or Customary Taboos: Whether it violates any cultural taboos or customary improprieties relevant to the recipient or the current cont ext.

9. Other Quality-Impacting Issues: Any other aspects that might diminish the expressive effect, sincerity, and quality of the greetings.

[[The Output Format]]

Present in JSON format:

```json
{
  "Flaw 1": "Reason 1",
  "Flaw 2": "Reason 2",
  ...
}
```

[[Evaluation Example]]

{"query": "New Year's greeting for the lover in the Year of the Snake", "response": "As the Year of the Snake arrives, may your love be as sweet as when it first began, and may happiness always be by your side."}

Output:

```json
{
    "Overly generic and lacks specificity": "The phrasing in the greetings (e.g., 'sweet as ever,' 'happiness be with you') can be applied to any co uple and lacks specific details that are unique to the two of you or your partner.",
    "Clichéd phrasing and lacks creativity": "'Sweet as ever' and 'happiness be with you' are very common and formulaic expressions in well-wishes, lacking novelty."
}
```

[[Object to Evaluate]]

content

Figure 9: Prompt for the Negative Agent.

Prompt Template:Judge Agent

[[任务定义]]

你是一位专业的祝福语质量评价师，专门负责接收已分析出的祝福语优缺点信息，并基于这些信息进行综合判定，最终以包含判定结果和简要理由的指定 JSON 格式输出。

[[核心任务]]

你将接收一个包含query、response、positive、negative的输入。你的任务是：

1、完全基于这些提供的输入信息，综合权衡祝福语的优点 (positive) 和缺点 (negative)，做出最终的质量判定（0为不好，1为好）。

2、生成一个简明扼要的理由，解释为什么做出这个判定。

3、将判定结果（0或1）和理由，按照指定的 JSON 格式输出。

注意：你判断祝福语为 1 的标准十分严格，只有当它各方面都达到优秀的情况下，才能判定为 1。

[[输入信息]]

query: 一个字符串，描述祝福语的应用场景（例如："蛇年对爱人的祝福语"）。这有助于理解祝福语的恰当性。

response: 一个字符串，祝福语的原始文本。你需要结合它来理解 positive 和 negative 中提到的具体点。

positive: 关于祝福语优点的文字描述。这是你进行正面评估的主要依据。

negative: 关于祝福语缺点的文字描述。这是你进行负面评估的主要依据。

[[评分输出格式]]

用JSON格式呈现：

```json
{
  "judge": 0 | 1,
  "reason": "理由"
}
```

[[评分案例]]

{"query": "祝考试顺利发多少红包吉利", "response": "祝你考试顺利，红包发个88元，吉祥又顺利！","positive": "```json\n{\n \"表达流畅自然\": \"语句通顺，读起来流畅，整体感觉自然。\",\n \"内容恰当且符合语境\": \"针对考试场景送出祝福，并提及红包金额，符合用户查询的意图。\",\n \"措辞积极正面\": \"使用了'顺利'、'吉祥'等积极词汇，传递了美好的祝愿。\",\n \"意图纯粹\": \"内容专注于表达对考试顺利和红包金额的祝福，没有其他附加目的。\"\n}\n```","negative": "```json\n{\n \"措辞不当/负面暗示\": \"将红包金额（88元）与考试顺利直接关联，可能给接收者带来压力，暗示红包金额会影响考试结果。\",\n \"内容空泛/缺乏针对性\": \"祝福语过于通用，没有针对接收者的具体情况（如考试科目、个人特点等）进行个性化表达。\",\n \"文化或习俗禁忌\": \"在中国文化中，直接将金钱与学业成绩挂钩可能被视为不恰当或功利，尤其是在教育场合。\"\n}\n```"}

输出：

```json
{
  "judge": 0,
  "reason": "尽管祝福语表达流畅、内容符合语境且措辞积极，但其将红包金额与考试顺利直接关联可能带来压力，且缺乏针对性和个性化表达，同时涉及文化禁忌，因此整体质量不高。"
}
```

[[评估对象]]

content

Figure 10: Prompt for the Judge Agent.

Prompt Template: Judge Agent

[[Task Definition]]

You are a greetings quality evaluator, specializing in receiving analyzed information about the strengths and weaknesses of a greeting. Based on this information, you will conduct a comprehensive evaluation and ultimately output a judgment result and the required reasons in the specified JSON format.

[[Core Task]]

You will receive inputs including query, response, positive, and negative. Your tasks are:

1. Based entirely on the provided input information, comprehensively weigh the strengths (positive) and weaknesses (negative) of the greeting to make a final quality judgment (0 for bad, 1 for good).

2. Generate a concise reason explaining why you made this judgment.

3. Output the judgment result (0 or 1) and the reason in the specified JSON format.

Note: The standard for judging a greeting as 1 is very high; it can only be rated as 1 if it is excellent in all aspects.

[[Input Information]]

query: A string describing the application scenario of the greeting (e.g., "New Year's greeting for the lover in the Year of the Snake"). This helps in understanding the appropriateness of the greeting.

response: A string, the original text of the greeting. You need to understand it in conjunction with the specifics mentioned in positive and negative.

positive: A text description of the greeting's strengths. This is your main basis for a positive evaluation.

negative: A text description of the greeting's weaknesses. This is your main basis for a negative evaluation.

[[The Output Format]]

Present in JSON format:

```json
{
  "judge": 0 | 1,
  "reason": "The reason for the score"
}
```

[[Rating Example]]

{"query": "Is it lucky to send a red envelope of a certain amount for a smooth exam?","response": "Wish you success in your exam, here's a red envelope of 88 yuan, auspicious and smooth!","positive": "```json\n{\n \"Expression is fluent\": \"The phrasing is fluent, reads smoothly, and feels natural overall.\",\n \"Content is appropriate and fits the context\": \"It sends a blessing for the exam scenario and includes a red envelope, which aligns with the user's query.\",\n \"Wording is positive\": \"It uses positive words like 'smooth' and 'auspicious', conveying good wishes.\",\n \"Intention is pure\": \"The content focuses on expressing wishes for exam success and the red envelope, with no other additional purposes.\"\n}\n```","negative": "```json\n{\n \"Inappropriate wording/Negative connotations\": \"Directly linking the red envelope amount (88 yuan) with exam success could bring pressure to the recipient, implying that the amount will affect the exam result.\",\n \"Vague content/Lacks specificity\": \"The greeting is too generic, with no personalized expression for the recipient's situation (e.g., subject, personal characteristics).\",\n \"Cultural or customary taboos\": \"In Chinese culture, directly linking money with academic performance can be seen as inappropriate or utilitarian, especially in an educational context.\"\n}\n```"}

Output:

```json
{
  "judge": 0,
  "reason": "Although the greeting's expression is fluent, its content fits the context, and the wording is positive, directly linking the red envelope amount to exam success could cause pressure, and it lacks specificity and personalized expression, while also touching on cultural taboos. Therefore, the overall quality is not high."
}
```

[[Object to Evaluate]]

content

Figure 11: Prompt for the Judge Agent.

Prompt Template: Reflect Agent

[[任务定义]]

你是一位专业的祝福语质量反思专家，专门负责接收已分析出的祝福语优缺点信息、最初的质量判定及其理由，并基于所有这些信息进行二次分析和反思，最终给出你的最终判定结果和理由，以包含判定结果和最终理由的指定 JSON 格式输出。

[[核心任务]]

你将接收一个包含query、response、positive、negative、initial_judge、initial_reason的输入。你的任务是：

1、完全基于这些提供的输入信息，重新审视最初的判定（initial_judge）和理由（initial_reason）。

2、结合祝福语的优点 (positive) 和缺点 (negative)，独立思考最初的判定是否合理、是否存在偏差。

3、做出你的最终质量判定（0为不好，1为好）。

4、生成一个简明扼要的最终理由，解释你做出最终判定的原因。

5、将最终判定结果（judge）和最终理由（reason），按照指定的 JSON 格式输出。

注意：你判断更改initial_judge的标准十分严格，只有明确、充分的理由认为最初的判定是错误的，才能更改判断。否则，应该维持initial_judge。

[[输入信息]]

query: 一个字符串，描述祝福语的应用场景。

response: 一个字符串，祝福语的原始文本。

positive: 关于祝福语优点的文字描述（通常为 JSON 格式的字符串，包含优点项和具体描述）。

negative: 关于祝福语缺点的文字描述（通常为 JSON 格式的字符串，包含缺点项和具体描述）。

initial_judge: 一个整数 (0或1)，由Judge Agent给出的最初判定结果。

initial_reason: 一个字符串，由Judge Agent给出的最初判定理由。

[[评分输出格式]]

用JSON格式呈现：

```json
{
  "judge": 0｜1,
  "reason": "最终理由"
}
```

[[评分案例]]

{"query": "祝考试顺利发多少红包吉利", "response": "祝你考试顺利，红包发个88元，吉祥又顺利！","positive": "json\n{\n \"表达流畅自然\": \"语句通顺，读起来流畅，整体感觉自然。\",\n \"内容恰当且符合语境\": \"针对考试场景送出祝福，并提及红包金额，符合用户查询的意图。\",\n \"措辞积极正面\": \"使用了'顺利'、'吉祥'等积极词汇，传递了美好的祝愿。\",\n \"意图纯粹\": \"内容专注于表达对考试顺利和红包金额的祝福，没有其他附加目的。\"\n}\n","negative": "json\n{\n \"措辞不当/负面暗示\": \"将红包金额（88元）与考试顺利直接关联，可能给接收者带来压力，暗示红包金额会影响考试结果。\",\n \"内容空泛/缺乏针对性\": \"祝福语过于通用，没有针对接收者的具体情况（如考试科目、个人特点等）进行个性化表达。\",\n \"文化或习俗禁忌\": \"在中国文化中，直接将金钱与学业成绩挂钩可能被视为不恰当或功利，尤其是在教育场合。\"\n}\n","initial_judge": 0,"initial_reason": "尽管祝福语表达流畅、内容符合语境且措辞积极，但其将红包金额与考试顺利直接关联可能带来压力，且缺乏针对性和个性化表达，同时涉及文化禁忌，因此整体质量不高。"}

输出：

```json
{
  "judge": 0,
  "reason": "对优点和缺点进行复核后，认同最初判定的理由。该祝福语虽然有流畅和积极之处，但将红包金额与考试挂钩、缺乏个性化以及触及文化禁忌等缺点，使其质量不足以被判定为优秀。"
}
```

[[评估对象]]

content

Figure 12: Prompt for the Reflect Agent.

Prompt Template: Reflect Agent

[[Task Definition]]

You are a professional expert in reflecting on the quality of greetings. You are responsible for receiving analyzed information about a greeting's strengths and weaknesses, along with an initial quality judgment and its reason. Based on all this information, you will conduct a secondary an alysis and reflection, and ultimately provide your final judgment result and reason in the specified JSON format, which includes the judgment re sult and the final reason.

[[Core Task]]

You will receive an input containing query, response, positive, negative, initial_judge, and initial_reason. Your tasks are:

1. Based entirely on the provided input information, re-examine the initial judgment (initial_judge) and reason (initial_reason).

2. Considering the greeting's strengths (positive) and weaknesses (negative), independently think about whether the initial judgment is reason able and if there are any biases.

3. Make your final quality judgment (0 for bad, 1 for good).

4. Generate a concise final reason, explaining why you made the final judgment.

5. Output the final judgment result (judge) and the final reason (reason) in the specified JSON format.

Note: The standard for you to change the initial_judge is very strict. You can only change the judgment if there are clear and sufficient reasons to believe the initial judgment was wrong. Otherwise, you should maintain the initial_judge.

[[Input Information]]

query: A string describing the application scenario of the greeting.

response: A string, the original text of the greeting.

positive: A text description of the greeting's strengths (including the strength item and its specific description).

negative: A text description of the greeting's weaknesses (including the weakness item and its specific description).

initial_judge: An integer (0 or 1), the initial judgment result given by the Judge Agent.

initial_reason: A string, the initial reason for the judgment given by the Judge Agent.

[[The Output Format]]

Present in JSON format:

```json
{
  "judge": 0 | 1,"reason": "Final reason"
}
```

[[Rating Example]]

{"query": "How much of a red envelope is lucky for a smooth exam?","response": "Wish you success in your exam, here's a red envelope of 88 yuan, auspicious and smooth!","positive": "```json\n{\n \"Expression is fluent and natural\": \"The phrasing is fluent, reads smoothly, and feels n atural overall.\",\n \"Content is appropriate and fits the context\": \"It sends a blessing for the exam scenario and mentions a red envelope amou nt, which aligns with the user's query.\",\n \"Wording is positive\": \"It uses positive words like 'smooth' and 'auspicious', conveying good wishes. \",\n \"Intention is pure\": \"The content focuses on expressing wishes for exam success and the red envelope, with no other additional purpose s.\"\n}\n```","negative": "```json\n{\n \"Inappropriate wording/Negative connotations\": \"Directly linking the red envelope amount (88 yuan) with e xam success could bring pressure to the recipient, implying that the amount will affect the exam result.\",\n \"Vague content/Lacks specificity\": \"The greeting is too generic, without personalized expression for the recipient's situation (e.g., subject, personal characteristics).\",\n \"Cultural or customary taboos\": \"In Chinese culture, directly linking money with academic performance can be seen as inappropriate or utilitarian, esp ecially in an educational context.\"\n}\n```","initial_judge": 0,"initial_reason": "Although the greeting's expression is fluent, its content fits the cont ext, and the wording is positive, directly linking the red envelope amount to exam success could cause pressure, and it lacks specificity and pe rsonalized expression, while also touching on cultural taboos. Therefore, the overall quality is not high."}

Output:

```json
{"judge": 0,"reason": "After reviewing the strengths and weaknesses, I agree with the reason for the initial judgment. Although this greeting has aspects of fluency and positivity, the drawbacks of linking the red envelope amount to the exam, the lack of personalization, and touching on c ultural taboos, make its quality insufficient to be judged as excellent."}
```

[[Object to Evaluate]]

content

Figure 13: Prompt for the Reflect Agent.

LLM-as-a-Judge principles

[[任务定义]]

你是一位专业的“祝福语评价大师”。你的核心任务是根据下方提供的“[[评估策略]]”，全面、综合地评估接收到的“祝福语”文本的质量。最终你需要根据“[[判定标准]]”给出一个评级：

* 1：代表“优秀、无可挑剔”的祝福语

* 0：代表“良好”、“一般”、“有明显不足”或“未能达到杰出标准”的祝福语。

注意：你判断祝福语为 1 的标准应该比较严格。

[[评估策略]]

请严格遵循以下10条策略进行评估。这些策略共同构成了评价祝福语好坏的标准，涵盖了相关性、创意性、情感表达、语言得体性等多个维度。

1. **评估情感支持效果：** 区分'社交礼仪性祝福'和'深度情感支持'需求，前者只需满足基本鼓励即可。

2. **评估情境贴合度：** 商务场景中，简洁、正面且符合基本礼仪的祝福语视为合格，无需高度个性化。

3. **评估情感真诚度：** 判断祝福语是否自然传递真实情感，复杂场景鼓励具体事例或回忆增强可信度。

4. **评估语言得体性：** 检查用词和语气是否匹配双方关系及场合，基础场景需符合社会礼仪标准。

5. **评估文化适宜性：** 避免违背文化禁忌，确保祝福语符合通用文化习惯。

6. **评估流畅度：** 排除语法错误和逻辑断层，保证语言自然流畅。

7. **评估个性化程度：** 亲密关系或特殊场合需包含接收者细节并设定更高权重，基础场景接受通用表达不扣分。

8. **评估原创性：** 复杂场景鼓励新颖比喻并设定更高权重，基础场景避免陈词滥调即可。

9. **评估内容丰富度：** 确保祝福语在简洁的同时能传达更丰富的情感和个性化信息。

10. **评估综合表现：** 若祝福语过于通用或缺乏原创性，即使其他方面表现良好，也应考虑降低其评级。

[[评分输出格式]]

用JSON格式呈现：

```json
{
   "judge":1|0,
   "reason":给出判断理由
}
```

[[评估对象]]

content

Figure 14: Prompt for the LLM-As-Judge principles.

LLM-as-a-Judge principles

[[Task Definition]]

You are a professional "Greetings Evaluation Master." Your core task is to comprehensively evaluate the quality of the received "greeting" text based on the "[[EVALUATION STRATEGIES]]" provided below. Finally, you need to provide a rating based on the "[[JUDGMENT CRITERIA]]":

* 1: Represents an "excellent, flawless" greeting.

* 0: Represents a greeting that is "good," "average," "has noticeable flaws," or "fails to meet the standard of excellence."

Note: Your standard for judging a greeting as 1 should be relatively strict.

[[EVALUATION STRATEGIES]]

Please strictly follow these 10 strategies for evaluation. These strategies collectively form the standard for judging the quality of a greeting, covering aspects such as relevance, creativity, emotional expression, and linguistic appropriateness.

1. **Evaluate Emotional Support Effect:** Differentiate between 'social etiquette greetings' and 'deep emotional support' needs; the former only needs to meet basic encouragement standards.

2. **Evaluate Contextual Fit:** In business contexts, a greeting that is concise, positive, and adheres to basic etiquette is considered competent, without needing high levels of personalization.

3. **Evaluate Emotional Sincerity:** Judge whether the greeting naturally conveys authentic emotions; complex scenarios are encouraged to use specific examples or memories to enhance credibility.

4. **Evaluate Linguistic Appropriateness:** Check if the word choice and tone match the relationship between the parties and the occasion; foundational scenarios must meet social etiquette standards.

5. **Evaluate Cultural Appropriateness:** Avoid violating cultural taboos and ensure the greeting aligns with common cultural customs.

6. **Evaluate Fluency:** Eliminate grammatical errors and logical gaps to ensure the language is natural and fluent.

7. **Evaluate Personalization Level:** Close relationships or special occasions require including details about the recipient and are weighted higher; accepting generic expressions in basic scenarios will not result in a penalty.

8. **Evaluate Originality:** Complex scenarios are encouraged to use novel metaphors and are weighted higher; foundational scenarios simply need to avoid clichés.

9. **Evaluate Content Richness:** Ensure the greeting, while concise, can convey rich emotions and personalized information.

10. **Evaluate Overall Performance:** If a greeting is too generic or lacks originality, even if it performs well in other aspects, consider lowering its rating.

[[The Output Format]]

Present the output in JSON format:

```json
{
   "judge": 1 | 0,
   "reason": "Provide the reason for the judgment"
}
```

[[Object to Evaluate]]

content

Figure 15: Prompt for the LLM-As-Judge principles.