UniCreative: Unifying Long-form Logic and Short-form Sparkle via Reference-Free Reinforcement Learning

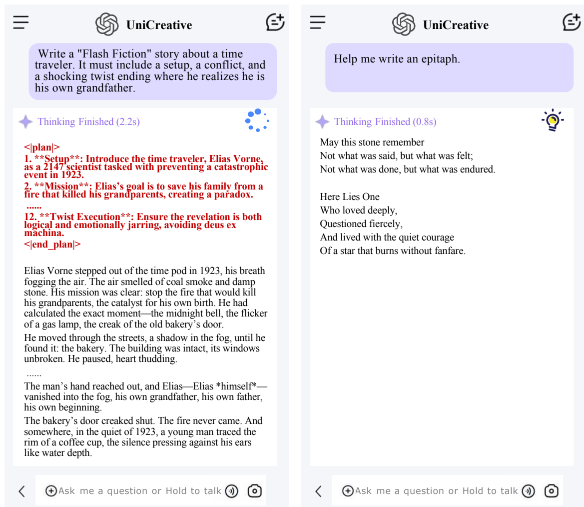
Xiaolong Wei[1*], **Zerun Zhu**[2*], **Simin Niu**[3], **Xingyu Zhang**[2], **Peiying Yu**[2], **Changxuan Xiao**[2]
Yuchen Li[2], **Jicheng Yang**[2], **Zhejun Zhao**[2†], **Chong Meng**[2], **Long Xia**[2], **Daiting Shi**[2]
[1]Beihang University [2]Baidu Inc.
[3]Renmin University of China
xiaolongwei@buaa.edu.cn, zhaozhejun@baidu.com

Abstract

A fundamental challenge in creative writing lies in reconciling the inherent tension between maintaining global coherence in long-form narratives and preserving local expressiveness in short-form texts. While long-context generation necessitates explicit macroscopic planning, short-form creativity often demands spontaneous, constraint-free expression. Existing alignment paradigms, however, typically employ static reward signals and rely heavily on high-quality supervised data, which is costly and difficult to scale. To address this, we propose **UniCreative**, a unified reference-free reinforcement learning framework. We first introduce **AC-GenRM**, an adaptive constraint-aware reward model that dynamically synthesizes query-specific criteria to provide fine-grained preference judgments. Leveraging these signals, we propose **ACPO**, a policy optimization algorithm that aligns models with human preferences across both content quality and structural paradigms without supervised fine-tuning and ground-truth references. Empirical results demonstrate that AC-GenRM aligns closely with expert evaluations, while ACPO significantly enhances performance across diverse writing tasks. Crucially, our analysis reveals an emergent meta-cognitive ability: the model learns to autonomously differentiate between tasks requiring rigorous planning and those favoring direct generation, validating the effectiveness of our direct alignment approach. Our code is available at https://github.com/weixiaolong94-hub/UniCreative.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive fluency in general-purpose text generation; however, their capabilities in *creative writing* remain fundamentally constrained. A central challenge lies in the inherent tension between



(a) Long-form (b) Short-form

Figure 1: Examples of UniCreative generations. The long-form task (left) follows a Plan-then-Write procedure, while the short-form task (right) employs direct generation without intermediate planning.

maintaining *global coherence* in long-form narratives and preserving *local expressiveness* in short creative texts. Despite the rapid expansion of context windows, recent benchmarks reveal that autoregressive generation still struggles with long-range consistency, often exhibiting topic drift, repetition, and structural degradation in extended outputs such as stories or scripts (Liu et al., 2024; Bai et al., 2025; Wu et al., 2025a; Zhao et al., 2025). These findings suggest that scaling context length alone is insufficient to resolve the structural limitations of current generation paradigms.

In stark contrast, short-form creative expression—spanning poetry, slogans, and blessings—necessitates *instantaneous linguistic vibrancy* rather than structural durability. Here, the primary challenge is not logical coherence, but the suppression of statistical mediocrity. Constrained by likelihood maximization, autoregressive models

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

Findings of the Association for Computational Linguistics: ACL 2026, pages 23563–23583

naturally converge toward high-probability, "safe" tokens, resulting in homogenized outputs that are fluent yet devoid of the stochastic spark essential for emotional resonance (Hou et al., 2025). Imposing rigid planning mechanisms on such tasks further exacerbates this issue, leading to a phenomenon of "over-determination" where the potential for serendipitous discovery is stifled by premature structural constraints (Wu et al., 2025b; Fein et al., 2025; Mizrahi et al., 2025). These distinct failure modes—structural collapse in long-form versus creative banality in short-form—highlight that creative writing encompasses heterogeneous regimes that cannot be resolved by a monolithic generation strategy.

A common approach to improving long-form coherence is to introduce explicit planning or hierarchical structure. Early work on hierarchical neural story generation and plan-and-write frameworks demonstrates that global outlines can effectively guide extended narratives and mitigate structural collapse (Fan et al., 2018; Yao et al., 2019). More recent studies similarly suggest that structured representations and multi-stage generation pipelines can enhance coherence and controllability in complex creative tasks (Xiao et al., 2025; Alper et al., 2025). However, these same mechanisms can become counterproductive when indiscriminately applied to short creative texts, where excessive structural constraints risk suppressing expressive diversity and creative spontaneity. This contrast underscores that planning should not be treated as a universally optimal solution, but rather as a capability that must be invoked selectively based on the task nature.

Meanwhile, advances in reinforcement learning and preference-based alignment provide new tools for optimizing language models. However, standard paradigms like RLHF or DPO typically rely on high-quality supervised data (SFT) and human-annotated preference pairs, which are costly to collect and difficult to scale for open-ended creative tasks (Christiano et al., 2017; Rafailov et al., 2023). While recent work has begun to systematically study creativity assessment and modeling in large language models (Zhao et al., 2025; Li et al., 2025a; Huang et al., 2025), most methods still depend on ground-truth references to compute rewards or stabilize training, limiting their applicability in purely creative domains where no single "correct" answer exists. Creative generation systems designed for scientific ideation further highlight the need for reference-free optimization paradigms (Sanyal et al., 2025).

In this paper, we propose **UniCreative**, a unified creative writing framework that treats planning as a dynamically callable computational resource rather than a fixed prerequisite. UniCreative enables models to adaptively switch between a **Plan-then-Write** mode for long-context tasks that demand structural integrity and a **Direct Generation** mode for short-context tasks that prioritize novelty (see examples in Figure 1). To train such a dual-mode system, we bypass the conventional Supervised Fine-Tuning (SFT) stage and introduce **Adaptive Constraint Preference Optimization (ACPO)**. This reference-free training paradigm leverages relative evaluation and self-bootstrapped baselines to optimize generation directly from user queries, eliminating the dependency on expensive human-written completions or ground-truth references.

Our contributions are threefold:

- **Unified Framework:** We introduce UniCreative, the first framework to unify planning-based and intuition-driven creative generation within a single policy.
- **Reference-Free Optimization:** We propose ACPO, demonstrating that superior performance in both long-form and short-form regimes can be achieved solely through reinforcement learning, bypassing the need for SFT and ground-truth references.
- **Emergent Meta-cognition:** Extensive experiments demonstrate UniCreative's superior performance; crucially, our analysis reveals an emergent meta-cognitive ability to autonomously differentiate task regimes, validating the framework's scalability.

2 Related Work

2.1 Creative Writing Evaluation and Benchmarks

Evaluating creative writing is challenging due to its open-ended nature. While general benchmarks like LongBench focus on reasoning (Bai et al., 2025), recent efforts target creativity specifically. WritingBench covers diverse genres (Wu et al., 2025b), LitBench emphasizes expert-annotated preferences (Fein et al., 2025), and CS4 measures creativity under constraints (Atmakuru et al., 2024).

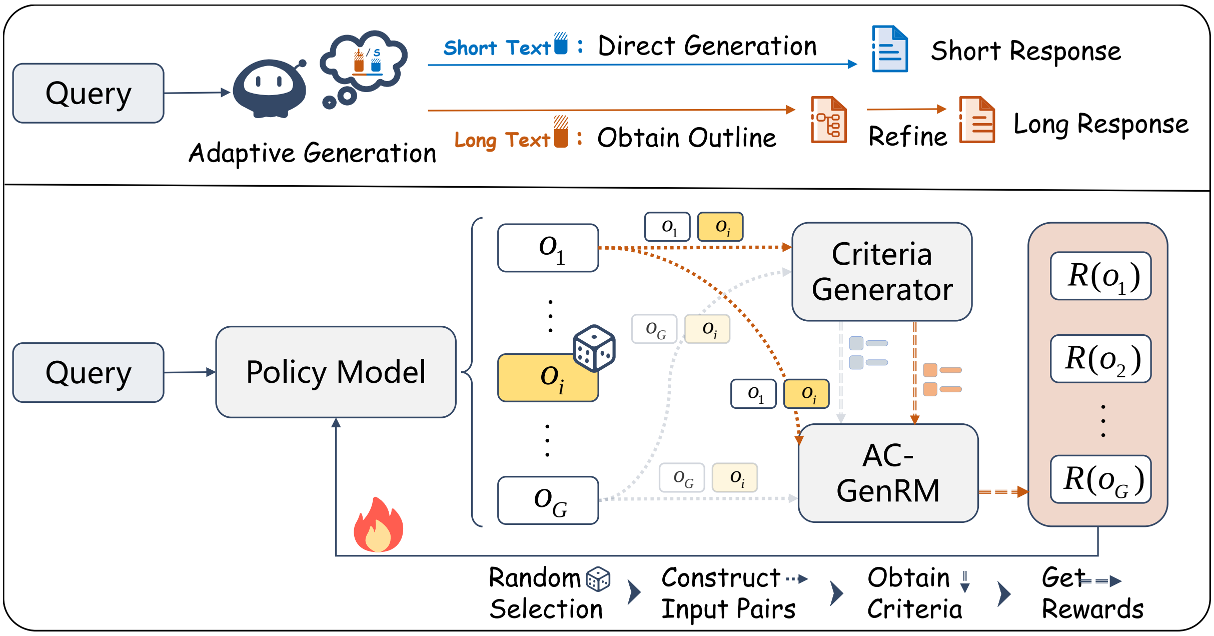


Figure 2: **UniCreative Framework. Top:** The model adaptively selects *Direct Generation* or *Plan-then-Write* modes based on task nature. **Bottom:** ACPO training. The policy generates responses $\{o_1, \ldots, o_G\}$ which are evaluated by **AC-GenRM** using dynamically synthesized criteria to provide reward signals for optimization.

Other benchmarks like UNCLE, LongWeave, and LongWriter-Zero further assess verifiability and RL challenges in ultra-long generation (Yang et al., 2025; Xiao et al., 2025; Wu et al., 2025a; Li et al., 2025d,b,c). To address the high cost of human evaluation, LLM-based evaluators (e.g., G-Eval) have been widely adopted (Liu et al., 2023). However, these judges often exhibit biases or struggle with complex uncertainty assessments (Atmakuru et al., 2024; Yang et al., 2025), motivating the need for more principled reward modeling (Niu et al., 2025; Liang et al., 2025, 2026).

2.2 Structural Planning in Story Generation

Explicit planning improves narrative coherence by separating high-level structure from surface realization (Fan et al., 2018; Yao et al., 2019). Despite the fluency of modern LLMs, they still suffer from "myopia" and inconsistency in long-form storytelling (Xie et al., 2023). Consequently, structured guidance remains indispensable (Yu et al., 2025a). Recent frameworks like ACE-RL and Beyond ReAct address this by integrating implicit planning or adaptive constraints, demonstrating significantly improved controllability in long-context generation (Chen et al., 2025b; Wei et al., 2025a; Tong et al., 2025; Lu et al., 2025).

2.3 Preference Optimization and Generative Rewards

Reinforcement learning from human feedback (RLHF) serves as the standard paradigm for aligning models with human preferences (Christiano et al., 2017; Ouyang et al., 2022), with Direct Preference Optimization (DPO) offering a streamlined, reward-free alternative (Rafailov et al., 2023). Recently, the focus has shifted toward enhancing the reward signal itself for open-ended generation. Generative reward models (GenRMs) reframe evaluation as a generative process, providing richer, chain-of-thought supervision compared to scalar scores (Mahan et al., 2024; Zhang et al., 2024; Li et al., 2025f). Subsequent innovations have further augmented these models with reasoning capabilities (e.g., RM-R1, ReasonGRM) and improved generalization via inference-time scaling (Chen et al., 2025c,a; Wang et al., 2025; Liu et al., 2025). In the specific domain of creative writing, methods such as Writing-Zero and others have begun to bridge the gap between non-verifiable creative objectives and RL, enabling effective optimization for long-form generation (Lu, 2025; Jia et al., 2025; Liao et al., 2025; Yu et al., 2025b; Li et al., 2025i,h,g,e; Cui et al., 2025a,b; Zeng et al., 2025b,a).

3 Methodology

We reframe creative writing as a *length-dependent adaptive decision process* requiring distinct strategies:

- **Long-Form Narratives:** Require *macroscopic structural integrity*. To counteract autoregressive "myopia," the model employs a "Plan-then-Write" strategy, ensuring global coherence through hierarchical reasoning.
- **Short-Form Expressions:** Prioritize *microscopic linguistic vibrancy*. Rigid planning here leads to "over-determination," stifling the spontaneity and emotional resonance essential for creativity.

To dynamically decouple these pathways without human supervision, we propose a reference-free pipeline built on two pillars: (1) **AC-GenRM**, an evaluator providing instance-specific feedback; and (2) the **ACPO** algorithm, which optimizes the policy to autonomously select between planning and direct execution based on query complexity. The complete process is detailed in Figure 2.

3.1 Adaptive Criteria GenRM (AC-GenRM)

Evaluating creative writing poses a significant challenge due to the subjective and multifaceted nature of literary quality. Standard reward models trained on general preference data often fail to capture the nuances of specific creative queries (Li et al., 2023) (e.g., distinguishing between "suspense" in a thriller and "wit" in a satire). Furthermore, LLM-based evaluators are notoriously prone to position bias. To address these limitations, we propose AC-GenRM, which decomposes the evaluation process into dynamic criteria synthesis and debiased pairwise ranking.

Dynamic Criteria Synthesis Instead of relying on a static system prompt for all queries, AC-GenRM first interprets the semantic intent of the input. Let x be the user query. We model the evaluation criteria as a latent variable C. To equip the model with this capability, we employ Supervised Fine-Tuning (SFT) on a high-quality critique dataset, allowing the critic π_{critic} to sample instance-specific criteria $C_x \sim \pi_{critic}(\cdot|x)$ that explicitly list the dimensions for assessment. For instance, given a prompt for a "scary story," C_x might prioritize "plot twist" and "atmosphere," whereas for a "greeting card," it would prioritize "warmth" and "conciseness." Specific examples are shown in Figure 4. This dynamic generation ensures that the reward signal is aligned with the specific creative goals of the prompt.

Generative Reward Learning with Position Debiasing AC-GenRM is a generative judge trained via SFT to predict a winning label $l \in \{A, B\}$ given a query x, criteria C_x, and response pair (y_A, y_B). To mitigate position bias, we employ Symmetrical Data Augmentation, swapping the response order with 50% probability during training. The objective minimizes the negative log-likelihood of the correct label:

$$\mathcal{L}_{RM}(\psi) = -\,\mathbb{E}_{(x,C_x,y_A,y_B,l)\sim\mathcal{D}_{aug}} \log P_\psi(l \mid x, C_x, y_A, y_B) \tag{1}$$

This forces the model to learn a position-invariant quality representation, ensuring preference signals strictly align with the synthesized criteria C_x.

3.2 Adaptive Constraint Preference Optimization (ACPO)

With a robust evaluator in place, we introduce Adaptive Constraint Preference Optimization (ACPO). Unlike traditional RLHF which relies on a separate value model (critic), ACPO optimizes the policy π_θ directly using group-based relative feedback. This approach is particularly well-suited for long-context creative writing, where training a value model is computationally prohibitive and unstable in long-context settings.

3.2.1 Reward Composition

The reward function R_{total} in ACPO is a composite signal designed to guide the model through the complex landscape of creative writing. It comprises three orthogonal components:

1. Bootstrapped Relative Reward (R_{rel}) We utilize a self-play mechanism to optimize without ground-truth references. For a query x, the policy generates G responses $Y = \{y_1, \ldots, y_G\}$. To resolve modality mismatch, we define a **projection operator** $\phi : \mathcal{V}^* \to \mathcal{V}^*$ that strips tokens in the planning segment $\mathcal{V}_{plan}$ (e.g., content between <|plan|> and <|end_plan|>). For each y_i, we sample a baseline $y_{base} \in Y \setminus \{y_i\}$.

The relative reward is derived from the discrete selection of the frozen generative AC-GenRM. Given $(x, C_x, \phi(y_i), \phi(y_{base}))$, the judge directly outputs the superior response, and the reward is assigned as:

$$R_{rel}(y_i) = \begin{cases} 2, & \text{if AC-GenRM selects } \phi(y_i) \\ -2, & \text{otherwise} \end{cases} \tag{2}$$

This binary signal creates a robust curriculum that continuously pushes the model's performance upper bound through internal competition.

2. Paradigm-Aware Structural Constraints (R_{struct}) To solve the "Myopia" problem in long texts and the "Over-determination" problem in short texts, we must enforce the correct cognitive mode. Let $\tau(x) \in \{\text{LONG}, \text{SHORT}\}$ denote the task mode. We introduce an indicator function $h(y) = \mathbb{I}[\exists t \in y : t \in \mathcal{V}_{plan}]$ to detect planning actions. To streamline the formulation, let $m_S = \mathbb{I}[\tau(x) = \text{SHORT}]$ and $m_L = \mathbb{I}[\tau(x) = \text{LONG}]$ be the mode indicators. The structural penalty is defined as:

$$R_{struct}(y \mid x) = -\beta_s \cdot \big(m_S \cdot h(y) + m_L \cdot (1 - h(y))\big) \tag{3}$$

where β_s is a penalty constant (set to 5.0 in our implementation). This term effectively prunes the exploration space, forcing the policy to collapse onto the structural paradigm appropriate for the task (i.e., mandating outlines for novels while prohibiting them for poems).

3. Adaptive Length Regularization (R_{len}) Structural adherence is a necessary but insufficient condition for high-quality generation. Empirically, we observe that RL policies often drift towards distinct failure modes: content collapse (insufficient length) in long-form narratives and verbosity (excessive length) in short-form expressions. To mitigate this, we introduce an asymmetric regularization mechanism with a penalty cap. Let $L(y) = |\phi(y)|$ denote the length of the generated content body. We formulate specific penalty functions:

$$R_{long}(y) = -\min\big(\lambda_l \cdot \text{ReLU}(\theta_{min} - L(y)), \gamma\big) \tag{4}$$

$$R_{short}(y) = -\min\big(\lambda_s \cdot \text{ReLU}(L(y) - \theta_{max}), \gamma\big) \tag{5}$$

where γ is the maximum penalty cap, set to 5.0 in our implementation to prevent excessive gradient signals from outlier lengths. Consequently, the final regularization term $R_{len}(y \mid x)$ dynamically activates either R_{long} or R_{short} contingent upon $\tau(x)$, creating a soft boundary for appropriate information density.

3.2.2 Optimization Objective

We aggregate the rewards into a total score $R_{total}(y_i) = R_{rel} + R_{struct} + R_{len}$. To stabilize training without a value model, we employ **Group Relative Policy Optimization (GRPO)**. GRPO computes the advantage A_i by normalizing the rewards within the sampled group Y, using the group mean as the baseline:

$$A_i = \frac{R_{total}(y_i) - \mu(\{R_{total}(y_j)\}_{j=1}^{G})}{\sigma(\{R_{total}(y_j)\}_{j=1}^{G}) + \epsilon} \tag{6}$$

This group-based normalization significantly reduces the variance of the gradient estimate. To streamline the objective formulation, we define the clipped advantage term $\mathcal{L}_i^{clip}$ and the KL divergence term $\mathcal{D}_i^{KL}$ as follows:

$$\mathcal{L}_i^{clip} = \min\big(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\big)$$
$$\mathcal{D}_i^{KL} = D_{KL}(\pi_\theta(y_i|x) \parallel \pi_{ref}(y_i|x))$$

Consequently, the final objective function maximizes:

$$\mathcal{J}(\theta) = \mathbb{E}_{x,Y}\left[\frac{1}{G}\sum_{i=1}^{G}\Big(\mathcal{L}_i^{clip} - \beta_{KL} \cdot \mathcal{D}_i^{KL}\Big)\right] \tag{7}$$

where $\mathbb{E}_{x,Y}$ denotes the expectation over the dataset and policy rollouts, $\rho_i = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}$ is the importance sampling ratio, and β_{KL} prevents excessive deviation from the reference model π_{ref}, ensuring linguistic fluency.

4 Experiments

4.1 Task Design and Data Construction

To train the UniCreative framework, we formalize creative writing into two distinct generative paradigms and construct a composite training corpus derived from diverse sources.

1. Definition of Generative Paradigms We decouple the conditional generation problem $P(y|x)$ into two specialized sub-tasks:

- **Planning-Augmented Generation (for Long-Form):** Designed for tasks requiring structural depth (e.g., novels). The model is tasked to first generate an intermediate latent variable z (representing the outline or reasoning chain) before producing the final text y. The sequence is formalized as: $x \rightarrow$ `<|plan|>` z `<|end_plan|>` $\rightarrow y$. This explicitly compels the model to engage in macroscopic reasoning.

- **Direct Generation (for Short-Form):** Designed for tasks requiring high entropy (e.g., poetry, slogans). The model must map $x \to y$ directly, with a strict prohibition on the `<|plan|>` token. This preserves the stochasticity and spontaneity of the token distribution.

2. Training Data Construction We construct a composite bilingual dataset for reward modeling and reference-free reinforcement learning:

- **Preference Data for AC-GenRM:** We utilize LitBench (Fein et al., 2025) (43k samples) to capture long-form narrative preferences and the Blessing dataset (Wei et al., 2025b) (6k samples) for short-form emotional expressions.

- **Exploration Prompts for ACPO:** We aggregate a diverse query set for the exploration space. The Long-Context Corpus combines HelloBench (Que et al., 2024), the Story Writing Benchmark, and high-quality web fiction, further augmented via the WritingBench (Wu et al., 2025b) synthesis methodology. The Short-Context Corpus consists of curated online queries designed to train expressive tension within constrained lengths.

Crucially, ACPO training relies solely on these queries without ground-truth completions, allowing the model to learn autonomous paradigm switching driven purely by AC-GenRM feedback.

4.2 Training Configuration

We trained the Qwen3 series models (1.7B, 4B, and 8B) on 8 NVIDIA H800 GPUs. Detailed training configurations are provided in the Appendix A.1.

4.3 Evaluation Benchmarks

We evaluate UniCreative using a multi-tiered benchmark suite covering reward modeling, generative quality, and meta-cognitive decision-making. Specifically, we assess (i) reward model quality via agreement with strong LLM judges on preference-based creative writing benchmarks, (ii) long-form and short-form generative performance using established creative writing datasets, and (iii) the model's ability to autonomously distinguish between planning-intensive and direct generation tasks. Detailed benchmark descriptions are provided in Appendix A.2.

Category	Model	LitBench	Blessing
Baselines	Claude-Sonnet-3.7	0.731	0.902
	Claude-3.5-Haiku	0.675	0.949
	GPT-4.1	0.702	0.923
	GPT-4.1-Mini	0.630	0.969
	DeepSeek-V3	0.700	0.906
	DeepSeek-R1	0.710	0.977
Qwen3-0.6B	Base	0.477	0.643
	AC-GenRM	0.728	0.989
Qwen3-1.7B	Base	0.543	0.849
	AC-GenRM	0.776	0.991
Qwen3-4B	Base	0.666	0.938
	AC-GenRM	0.796	0.993
Qwen3-8B	Base	0.688	0.944
	AC-GenRM	**0.807**	**0.994**

Table 1: Reward model evaluation showing agreement rates with proprietary judges on **LitBench** and **Blessing**. AC-GenRM is compared against unaligned Base models to isolate the algorithmic contribution.

5 Results and Analysis

5.1 Justification for Pairwise Generative Reward Modeling

Our AC-GenRM architecture departs from traditional pointwise scalar models and discriminative Bradley-Terry (BT) classifiers. Motivated by recent findings in creative evaluation (Wu et al., 2025b; Fein et al., 2025; Jia et al., 2025), we adopt a pairwise generative approach for two primary reasons.

Pairwise Comparison over Pointwise Scoring. Unlike math or code, creative writing lacks definitive ground truth. Pointwise scoring suffers from calibration instability, where absolute scores (e.g., "8/10") remain ambiguous without a fixed reference. In contrast, relative preference ($y_A \succ y_B$) aligns better with human intuition and reduces variance driven by prompt difficulty (Christiano et al., 2017). Moreover, comparative discrimination mitigates the reward hacking common in scalar models—such as verbosity bias—thereby improving generalization across heterogeneous tasks.

Generative Reasoning vs. Discriminative BT. Unlike opaque discriminative models, AC-GenRM leverages generative reasoning to synthesize query-specific criteria (C_x), ensuring precise semantic alignment with user intent across diverse tasks. This self-principled critique mechanism enhances robustness against reward hacking, with SFT on preference data serving as a crucial stabilization mechanism for the reasoning process. Empirically, AC-GenRM yields superior signal quality, achiev-

Models	Avg	D1	D2	D3	D4	D5	D6	R1	C	R2	C	R3	C
Proprietary LLMs													
O3-2025-04-16	85.27	84.81	85.20	83.89	85.88	85.82	86.80	85.14	87.45	85.24	90.98	86.31	87.21
Gemini-2.5-pro-preview	83.05	83.21	81.47	83.00	84.52	84.49	82.14	83.58	86.49	83.86	90.45	83.35	83.98
Claude-Sonnet-3.7	78.48	78.24	77.93	76.51	79.37	79.26	80.88	79.43	82.51	78.84	86.14	79.23	80.49
GPT-4o	75.46	74.40	73.42	74.38	77.91	75.86	78.08	76.82	81.57	75.82	85.46	76.13	76.73
o1-Preview	68.57	68.54	67.01	66.57	69.53	70.31	71.41	70.09	75.10	68.49	79.78	70.91	73.81
Open-source LLMs													
DeepSeek-R1-0528	83.22	83.15	81.48	81.55	85.68	84.14	84.44	84.24	87.27	83.72	89.35	83.83	82.70
Qwen3-235B-A22B-Thinking	81.45	80.19	79.24	80.95	82.92	82.52	82.89	82.54	85.02	81.30	88.22	81.26	81.76
LongWriter-Zero-32B	80.30	80.66	80.27	80.21	76.09	83.55	81.02	79.89	83.38	80.77	86.82	80.23	82.05
Qwen3-8B-Thinking	75.56	75.47	74.42	75.51	73.03	77.80	77.16	76.29	79.72	75.67	84.96	74.35	77.30
Qwen3-235B-A22B	73.63	73.56	72.90	73.98	70.13	76.52	74.69	77.46	82.05	77.01	87.29	76.26	79.57
Qwen3-8B	70.75	70.68	70.74	70.33	68.51	72.74	71.52	71.72	76.77	71.69	83.12	70.06	75.36
Qwen-2.5-72B-instruct	65.28	65.80	63.36	63.80	62.75	68.07	67.91	65.81	70.49	65.92	78.65	66.38	67.95
LongWriter-glm-9B	62.94	64.06	63.66	62.35	61.26	65.03	61.30	62.79	66.70	63.60	74.82	63.37	65.88
LongWriter-llama3.1-8B	58.01	60.05	59.31	57.58	56.03	58.38	56.73	58.12	61.40	58.60	67.61	59.05	62.97
Llama-3.3-70B-instruct	50.43	50.67	49.25	47.90	48.52	52.92	56.56	50.71	50.71	50.38	50.38	51.08	51.08
Our Models													
Qwen3-1.7B-Thinking	63.06	66.72	64.41	64.13	58.13	60.52	63.67	64.92	70.02	62.96	70.43	61.16	60.53
Qwen3-1.7B-Thinking + RL	73.65	76.06	74.54	74.03	68.54	75.14	74.50	74.71	78.80	74.43	81.59	73.31	73.67
Qwen3-4B-Thinking	71.35	72.57	73.10	72.06	64.38	74.54	72.95	72.08	76.79	72.17	79.18	70.51	70.93
Qwen3-4B-Thinking + RL	77.36	78.63	78.48	77.88	72.61	79.55	77.91	77.83	82.01	78.20	85.62	76.71	77.92
Qwen3-8B-Thinking	77.11	77.54	77.73	77.63	73.68	78.49	78.42	77.62	81.79	77.74	83.60	76.58	76.86
Qwen3-8B-Thinking + RL	**82.42**	**83.17**	**81.85**	**83.05**	**80.80**	**83.57**	**82.72**	**82.68**	**85.32**	**82.68**	**87.61**	**81.60**	**80.62**

Table 2: Performance of different LLMs on WritingBench across six domains and three writing requirements. Scores are normalized from a 0-10 range to a 100-point scale. The corresponding domains and requirements are: (D1) Academic & Engineering, (D2) Finance & Business, (D3) Politics & Law, (D4) Literature & Art, (D5) Education, (D6) Advertising & Marketing, (R1) Style, (R2) Format, and (R3) Length. "C" denotes the category-specific scores of the three requirements.

ing an 80.7% agreement rate that significantly surpasses Claude-3.7-Sonnet (73.1%) and exceeds the performance of top-tier discriminative models (∼78%) on the same benchmark.

5.2 Performance on Long-Form and Short-Form Writing

We evaluate UniCreative on complex narratives (*WritingBench*) and high-entropy short-form text (*Blessing*).

Long-Form Performance (WritingBench). As shown in Table 2, ACPO consistently enhances long-context reasoning across all scales. `Qwen3-8B-Thinking + RL` achieved an average score of **82.42** (+5.31 over the Base model), significantly outperforming much larger models like `Llama-3.3-70B-Instruct` (50.43) and `Qwen-2.5-72B-Instruct` (65.28), while rivaling `Claude-Sonnet-3.7` (78.48). Specifically, ACPO improved compliance in Format (R2) and Length (R3). Unlike specialized baselines like ACE-RL (Chen et al., 2025b) that rely on rigid, task-specific checklists, UniCreative maintains structural integrity without sacrificing the flexibility needed for broader creative tasks.

Short-Form Performance (Blessing). Table 3 demonstrates the critical necessity of mode switching. The "Thinking" baselines struggle with short-form tasks (e.g., 64.2% for 1.7B) due to "over-determination" from forced planning. By enabling the model to bypass planning, **ACPO achieves massive gains**: `Qwen3-1.7B + RL` jumped to 90.0% (+25.8%), matching `DeepSeek-V3.2`, while the 8B variant reached **93.6%**, surpassing `Claude-Sonnet-4.5` (93.2%). These results confirm that our reference-free RL enables the model to "unlearn" rigid structural constraints when appropriate, restoring the linguistic vibrancy essential for high-quality short-form generation.

5.3 Can the Model Adaptively Differentiate Task Regimes?

We evaluate whether UniCreative enables models to distinguish between generation pathways using the Mode Discrimination Benchmark.

Scalable Accuracy Gains. As shown in Table 4, ACPO significantly improves mode-switching accuracy in larger models. `Qwen3-4B` and `Qwen3-8B`

Reference LLMs		Our Models (Ablation Study)		
Model Name	**Score**	**Model Name**	**Score**	**Gain (Δ)**
DeepSeek-V3-250324	87.2%	Qwen3-1.7B	8.40%	-
DeepSeek-V3.2	90.4%	Qwen3-1.7B-Thinking	64.2%	-
Qwen3-32B	45.2%	**Qwen3-1.7B-Thinking + RL**	**90.0%**	+25.8%
Qwen3-32B-Thinking	88.4%	Qwen3-4B	40.8%	-
GPT-4.1	89.4%	Qwen3-4B-Thinking	74.0%	-
Doubao-Seed-1.6	**94.6%**	**Qwen3-4B-Thinking + RL**	**91.4%**	+17.4%
Grok-4	80.6%	Qwen3-8B	43.6%	-
Claude-Sonnet-4	92.8%	Qwen3-8B-Thinking	68.0%	-
Claude-Sonnet-4.5	93.2%	**Qwen3-8B-Thinking + RL**	**93.6%**	**+25.6%**

Table 3: Performance evaluation on the Blessing short-text dataset. We employ an Adversarial Framework to benchmark standard LLMs (left) against our UniCreative variants (right). **Score** represents the percentage of excellent ratings. The **Gain (Δ)** column highlights the significant score improvement achieved by our RL method compared to the 'thinking' baseline.

Model Size	Method	Accuracy on Mode Discrimination (5-run Avg. ± SD)			
		Easy	**Hard**	**Avg.**	**Gain (Δ)**
Qwen3-1.7B	Thinking	64.2% ± 1.1%	65.1% ± 2.2%	64.7%	-
	+ RL (Ours)	57.0% ± 1.1%	56.6% ± 1.4%	56.8%	-7.9%
Qwen3-4B	Thinking	77.1% ± 2.2%	62.4% ± 2.1%	69.8%	-
	+ RL (Ours)	95.9% ± 0.7%	82.8% ± 1.7%	89.4%	**+19.6%**
Qwen3-8B	Thinking	80.5% ± 2.1%	68.5% ± 1.9%	74.5%	-
	+ RL (Ours)	**96.4%** ± 1.8%	**86.4%** ± 1.2%	**91.4%**	**+16.9%**

Table 4: Accuracy on the Mode Discrimination Benchmark across different model scales (5-run avg). Significant gains in 4B and 8B models underscore the efficacy of ACPO in fostering emergent meta-cognitive planning.

achieve up to 96.4% and 86.4% accuracy on Easy and Hard tasks, respectively. This confirms that given sufficient capacity, our framework effectively aligns the policy with structural preferences, allowing models to autonomously invoke the optimal generation mode.

Capacity Bottleneck in Small Models. Conversely, `Qwen3-1.7B` exhibits a decline in discrimination accuracy (e.g., 64.2% → 57.0%). This does not signify training failure; rather, it reflects a *capacity bottleneck*. While RL improves the 1.7B model's raw writing quality (Table 2 and Table 3), its limited parameter count forces a trade-off: the model prioritizes high-entropy linguistic vibrancy over the complex meta-cognitive logic required for autonomous paradigm switching.

Emergent Meta-Cognition. These results underscore an emergent scaling trend. While the 1.7B model yields to the tension between content and structure, the 4B and 8B models successfully resolve it. The ability to identify latent structural complexity and invoke appropriate computational pathways constitutes a high-order *meta-cognitive reasoning* that only stabilizes once a specific model scale threshold is exceeded.

6 Conclusion

We introduced **UniCreative**, a unified RL framework balancing long-form logical coherence with short-form expressive vibrancy. By leveraging **AC-GenRM** for criteria synthesis and **ACPO** for reference-free optimization, our approach enables models to autonomously navigate planning versus direct generation trade-offs without SFT or expensive annotations. Results show UniCreative mitigates generative failures and fosters an emergent meta-cognitive ability to select optimal computational pathways based on task complexity. This framework establishes a robust foundation for aligning LLMs with the nuanced demands of creative writing.

Limitations

Despite the strong performance and emergent meta-cognitive abilities demonstrated by UNICREATIVE, several limitations remain that offer avenues for future research:

Dependency on Model Scale Our analysis of the Mode Discrimination Benchmark (Table 4) reveals a clear capacity bottleneck in smaller models. While meta-cognitive task differentiation emerges at the 4B and 8B scales, the 1.7B model struggles to balance structural logic with linguistic vibrancy. This suggests that the proposed adaptive switching mechanism may require a minimum parameter threshold to stabilize, potentially limiting its effectiveness for ultra-lightweight edge deployment.

Boundary Cases in Task Regimes UniCreative currently operates on a binary switch between LONG-FORM (Plan-then-Write) and SHORT-FORM (Direct Generation). However, there exists a "medium-form" gray area—such as long social media threads or short analytical essays—where a rigid plan might be too heavy, yet direct generation might lack sufficient coherence. Future work could explore a more granular or "soft" planning mechanism that adaptively adjusts the density of the reasoning chain based on the prompt's intermediate requirements.

Computational Overhead of Long-Context RL Optimizing policies for long-form narratives (up to 12.6k tokens per sample) is computationally intensive. Although we utilized Group Relative Policy Optimization (GRPO) to bypass the value model and save VRAM, the exploration of such a vast policy space requires significant GPU resources (e.g., $8 \times$ NVIDIA H800) and training time. This overhead may pose challenges for academic labs with constrained hardware budgets looking to replicate or extend ultra-long-context training.

References

Morris Alper, Moran Yanuka, Raja Giryes, and Gašper Beguš. 2025. Conlangcrafter: Constructing languages with a multi-hop llm pipeline. *arXiv preprint arXiv:2508.06094*.

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *arXiv preprint arXiv:2410.04197*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Bin Chen, Xinzge Gao, Chuanrui Hu, Penghang Yu, Hua Zhang, and Bing-Kun Bao. 2025a. Reasongrm: Enhancing generative reward models through large reasoning models. *arXiv preprint arXiv:2506.16712*.

Jianghao Chen, Wei Sun, Qixiang Yin, Lingxing Kong, Zhixing Tan, and Jiajun Zhang. 2025b. Ace-rl: Adaptive constraint-enhanced reward for long-form generation reinforcement learning. *arXiv preprint arXiv:2509.04903*.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, and 1 others. 2025c. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, and 1 others. 2025. Creativityprism: A holistic benchmark for large language model creativity. *arXiv preprint arXiv:2510.20091*.

Zhongzhan Huang, Shanshan Zhong, Pan Zhou, Shanghua Gao, Marinka Zitnik, and Liang Lin. 2025. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. 2025. Writing-zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *arXiv e-prints*, pages arXiv–2506.

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*.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025b. M²IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In *Second Conference on Language Modeling*.

Yanshu Li, Jianjiang Yang, Zhennan Shen, Ligong Han, Haoyan Xu, and Ruixiang Tang. 2025c. Catp: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning. *arXiv preprint arXiv:2508.07871*.

Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025d. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 736–763.

Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, and 1 others. 2025e. Towards ai search paradigm. *arXiv preprint arXiv:2506.17188*.

Yuchen Li, Rui Kong, Zhonghao Lyu, Qiyang Li, Xinran Chen, Hengyi Cai, Lingyong Yan, Shuaiqiang Wang, Jiashu Zhao, Guangxu Zhu, and 1 others. 2026. Flexspec: Frozen drafts meet evolving targets in edge-cloud collaborative llm speculative decoding. *IEEE Transactions on Mobile Computing*.

Yuchen Li, Zhonghao Lyu, Yongqi Zhang, Hao Zhang, Tianhao Peng, Haoyi Xiong, Shuaiqiang Wang, Linghe Kong, Guihai Chen, and Dawei Yin. 2025f. S3prank: Towards satisfaction-oriented learning to rank with semi-supervised pre-training. *IEEE Transactions on Knowledge and Data Engineering*.

Yuchen Li, Haoyi Xiong, Qingzhong Wang, Linghe Kong, Hao Liu, Haifang Li, Jiang Bian, Shuaiqiang Wang, Guihai Chen, Dejing Dou, and 1 others. 2023. Coltr: Semi-supervised learning to rank with co-training and over-parameterization for web search. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12542–12555.

Yuchen Li, Haoyi Xiong, Yongqi Zhang, Jiang Bian, Tianhao Peng, Xuhong Li, Shuaiqiang Wang, Linghe Kong, and Dawei Yin. 2025g. Rankelectra: Semi-supervised pre-training of learning-to-rank electra for web-scale search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2415–2425.

Yuchen Li, Hao Zhang, Yongqi Zhang, Hengyi Cai, Mingxin Cai, Shuaiqiang Wang, Haoyi Xiong, Linghe Kong, Dawei Yin, and Lei Chen. 2025h. Rankexpert: A mixture of textual-and-behavioral experts for multi-objective learning-to-rank in web search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4578–4589.

Yuchen Li, Hao Zhang, Yongqi Zhang, Xinyu Ma, Wenwen Ye, Naifei Song, Shuaiqiang Wang, Haoyi Xiong, Dawei Yin, and Lei Chen. 2025i. M 2 oerank: Multi-objective mixture-of-experts enhanced ranking for satisfaction-oriented web search. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 4441–4454. IEEE.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Zhaoxin Fan, Bo Tang, Jihao Zhao, Jiawei Yang, and 1 others. 2025. Saferag: benchmarking security in retrieval-augmented generation of large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4609–4631.

Xun Liang, Mengwei Wang, Yuefeng Ma, and Simin Niu. 2026. Safe rag by rag: Untying the bell that rag rang with the rag hand. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31925–31933.

Jianxing Liao, Tian Zhang, Xiao Feng, Yusong Zhang, Rui Yang, Haorui Wang, Bosi Wen, Ziying Wang, and Runzhi Shi. 2025. Rlmr: Reinforcement learning with mixed rewards for creative writing. *arXiv preprint arXiv:2508.18642*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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. *arXiv preprint arXiv:2303.16634*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

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.

Xun Lu. 2025. Writing-zero: Bridge the gap between non-verifiable problems and verifiable rewards. *arXiv preprint arXiv:2506.00103*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Moran Mizrahi, Chen Shani, Gabriel Stanovsky, Dan Jurafsky, and Dafna Shahaf. 2025. Cooking up creativity: A cognitively-inspired approach for enhancing llm creativity through structured representations. *arXiv preprint arXiv:2504.20643*.

Simin Niu, Mengwei Wang, Xun Liang, Zhiyu Li, Sensen Zhang, Shichao Song, Hanyu Wang, Jiawei Yang, Feiyu Xiong, and Chenyang Xi. 2025. Empowering large language models to set up knowledge retrieval indexing via self-learning. *IEEE Transactions on Knowledge and Data Engineering*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Aishik Sanyal, Samuel Schapiro, Sumuk Shashidhar, Royce Moon, Lav R Varshney, and Dilek Hakkani-Tur. 2025. Spark: A system for scientifically creative idea generation. *arXiv preprint arXiv:2504.20090*.

Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Qiaozhi He, Murun Yang, Bei Li, Tong Xiao, Chunliang Zhang, Tongran Liu, and 1 others. 2025. Gram: A generative foundation reward model for reward generalization. *arXiv preprint arXiv:2506.14175*.

Xiaolong Wei, Yuehu Dong, Xingliang Wang, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025a. Beyond react: A planner-centric framework for complex tool-augmented llm reasoning. *arXiv preprint arXiv:2511.10037*.

Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025b. Igniting creative writing in small language models: Llm-as-a-judge versus multi-agent refined rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17171–17197.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. 2025a. Longwriter-zero: Mastering ultra-long text generation via reinforcement learning. *arXiv preprint arXiv:2506.18841*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and 1 others. 2025b. Writingbench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.

Zikai Xiao, Fei Huang, Jianhong Tu, Jianhui Wei, Wen Ma, Yuxuan Zhou, Jian Wu, Bowen Yu, Zuozhu Liu, and Junyang Lin. 2025. Longweave: A long-form generation benchmark bridging real-world relevance and verifiability. *arXiv preprint arXiv:2510.24345*.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. *arXiv preprint arXiv:2301.09790*.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. 2025. Uncle: Benchmarking uncertainty expressions in long-form generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30328–30344.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Peiying Yu, Guoxin Chen, and Jingjing Wang. 2025a. Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17432–17451.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhi Zeng, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025a. Understand, refine and summarize: Multi-view knowledge progressive enhancement learning for fake news video detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9216–9225.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025b. Imol: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933.

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*.

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, 22(3):417–436.

A Appendix

A.1 Training Configuration

Our experiments are conducted on a computational cluster equipped with 8 $\times$ NVIDIA H800 (80GB) GPUs. The training pipeline consists of two sequential stages: supervised fine-tuning for the reward model and reinforcement learning for policy optimization.

GenRM Supervised Fine-Tuning We initialize the reward model using the Qwen3-8B backbone. To transform the generative model into a robust judge, we perform full-parameter fine-tuning on the constructed GenRM dataset. Given the memory demands of full-parameter updates, we leverage DeepSpeed ZeRO-3 offloading strategies. The model is trained for 2 epochs with a conservative learning rate of 1×10^{-5} and a global batch size of 256. We utilize a cosine learning rate scheduler with a 10% warmup ratio. The maximum context length is set to 5,200 tokens to accommodate the query, response, and the generated evaluation rationale.

ACPO Reinforcement Learning To address the computational bottleneck of processing extremely long contexts (up to 12.6k tokens per sample), we integrate vLLM for efficient rollout generation (Li et al., 2026) and employ Tensor Parallelism (TP=2) to distribute the model across GPUs during inference. The training utilizes a global batch size of 1024, spanning 10 epochs with a conservative learning rate of 1×10^{-6}. We enforce strict length constraints with a maximum prompt length of 8,600 tokens and a response limit of 4,000 tokens to accommodate extensive planning chains. For each query, we sample $G = 6$ distinct responses to compute the group-based advantage. To ensure training stability under memory constraints, we enable gradient checkpointing and offload optimizer states to the CPU via FSDP. Furthermore, we apply a low-variance KL penalty ($\beta_{KL} = 0.001$) and disable entropy regularization to prevent policy collapse while maintaining alignment with the reference model.

A.2 Detailed Evaluation Benchmarks

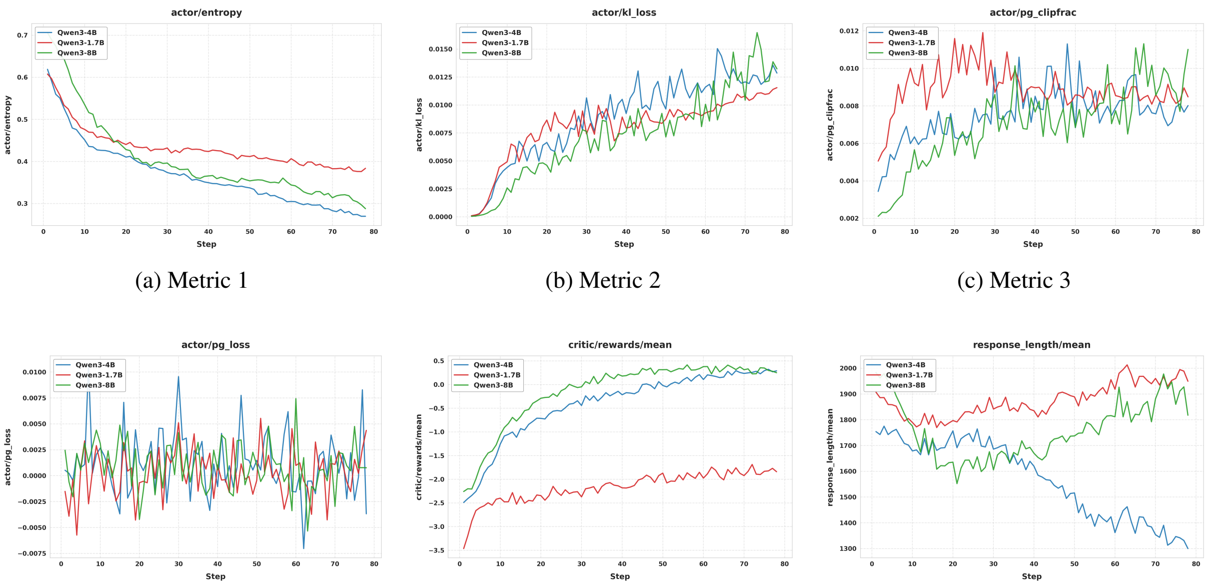
We provide detailed descriptions of all evaluation benchmarks used in this work.

Reward Model Evaluation To validate AC-GenRM, we measure its agreement with proprietary LLM judges (e.g., GPT-4o) on two complementary preference-based creative writing benchmarks. **LitBench** (Fein et al., 2025) provides expert-annotated pairwise preferences for long-form English literary narratives, emphasizing narrative coherence, structure, and literary quality. **Blessing** (Wei et al., 2025b) is a Chinese short-form greeting benchmark with preference annotations targeting emotional resonance and inspirational expressiveness. These benchmarks are designed to evaluate the discriminative capability of reward models in distinguishing fine-grained creative preferences.

Comprehensive and Long-Form Writing (WritingBench) For evaluating long-form narratives and general creative writing quality, we use **WritingBench** (Wu et al., 2025b). WritingBench contains 1,000 diverse long-form writing instructions spanning 6 major domains and over 100 subdomains. It adopts an instance-level evaluation protocol, where a Critic model dynamically synthesizes query-specific evaluation criteria. Final judgments are provided by a fine-tuned critic LLM, enabling fine-grained assessment of structural integrity and long-range coherence beyond static metrics.

Short-Text Creativity Benchmark To evaluate the Direct Generation mode, we adopt the open-source Blessing dataset (Wei et al., 2025b), which targets short-form creative writing for Chinese greetings. This benchmark emphasizes inspirational creativity in short texts, focusing on emotional resonance, expressive novelty, and diversity. It is used to assess whether reinforcement learning preserves high-entropy expressiveness without inducing mode collapse in short-form generation.

Mode Discrimination Benchmark (Meta-Cognition) To analyze the model's meta-cognitive ability to autonomously select the



(a) Metric 1 (b) Metric 2 (c) Metric 3

(d) Metric 4 (e) Metric 5 (f) Metric 6

Figure 3: **Training dynamics of the ACPO algorithm.** We visualize key metrics over training steps: (a) Actor Entropy (indicating exploration), (b) KL Divergence from the reference model, (c) Clip Fraction, (d) Policy Gradient Loss, (e) Mean Reward trajectories, and (f) Average Response Length. The curves demonstrate stable convergence and consistent reward maximization.

optimal generation mode, we constructed a diagnostic benchmark consisting of 400 instructions. The dataset is bifurcated into two balanced tiers based on the clarity of the task's structural requirements: an **Easy subset** (200 samples) where the generation regime—long-form narrative versus short-form response—is highly intuitive and unambiguous (e.g., standard factual queries or straightforward storytelling), and a **Hard subset** (200 samples) characterized by latent complexity or high levels of ambiguity. The Hard subset requires the model to perform higher-order reasoning to discern whether a prompt necessitates hierarchical planning or favors direct expressive generation, even in the absence of explicit cues. This benchmark evaluates the model's proficiency in aligning its computational strategy with the underlying structural depth of human intent.

A.3 In-depth Analysis of Training Dynamics

To provide empirical transparency into the reference-free reinforcement learning stage, we analyze the training trajectories of the Qwen3 series (1.7B, 4B, and 8B) under the ACPO framework. Figure 3 visualizes the evolution of policy stochasticity, stability constraints, reward optimization, and generative behavior over 80 training steps.

Policy Maturity and Stochasticity As illustrated in Figure 3(a), the **Actor Entropy** exhibits a monotonic decline across all model scales. This trend signifies the transition from broad stochastic exploration to a converged policy that consistently selects tokens aligned with the synthesized creative criteria C_x. Notably, the Qwen3-4B and 8B models converge to a lower entropy floor compared to the 1.7B variant. This suggests that larger models develop a more specialized and "decisive" internal mapping for complex creative constraints, whereas the smaller model retains higher entropy, potentially due to a limited capacity to resolve the inherent tension between structural logic and linguistic vibrancy.

Regularization and Stability Constraints The stability of the alignment process is monitored through **KL Divergence** and **Clip Fraction**.

- **KL Drift:** Figure 3(b) shows that the KL divergence from the reference model π_{ref} increases steadily. The 8B and 4B models reach a higher KL plateau (≈ 0.013) than the 1.7B model. This suggests that larger models utilize a broader search space to "unlearn" the rigid constraints of the base model and adopt the more expressive "sparkle" prioritized by AC-GenRM.

Query	*"Blessings for a friend's daughter's wedding."*
Criterion 1	**Emotional Resonance**: Evaluates the depth and appropriateness of the emotional impact; captures the joy and significance of the occasion.
Criterion 2	**Creativity and Uniqueness**: Assesses originality and creative flair; avoids clichés and leaves a lasting impression.
Criterion 3	**Linguistic Conciseness**: Measures clarity and brevity; ensures the message is succinct yet impactful.
Criterion 4	**Cultural Appropriateness**: Evaluates alignment with cultural norms and wedding-specific contexts.
Criterion 5	**Aesthetic Appeal**: Assesses the beauty and elegance of the language; rewards poetic and refined phrasing.

Table 5: An example of synthesized evaluation criteria for a short-form creative query. AC-GenRM dynamically prioritizes emotional and aesthetic dimensions over structural or logical ones for this genre.

- **Policy Clipping:** The **Clip Fraction** in Figure 3(c) remains consistently within a healthy range (0.002–0.012). The absence of sudden spikes or saturation confirms that the group-based relative feedback in ACPO provides a stable gradient signal, preventing the policy from making excessively large updates that could lead to catastrophic forgetting or linguistic collapse.

A.4 Mode Discrimination Benchmark: Task Examples

The Mode Discrimination Benchmark evaluates the model's ability to navigate the boundary between macroscopic structural planning and microscopic linguistic vibrancy. Table 4 presents a curated selection of tasks from both the **Easy** (explicitly cued) and **Hard** (implicitly complex) subsets, along with the ground-truth labels and the underlying reasoning for the required generation mode.

Reward Optimization and Scaling Laws The **Mean Reward** trajectories in Figure 3(e) provide the strongest evidence for the effectiveness of the ACPO algorithm. All models demonstrate a robust upward trend, indicating successful optimization of the composite reward R_{total}. However, a clear scaling gap is visible: while the 4B and 8B models successfully cross into positive reward territory (≈ 0.3), the 1.7B model plateaus at a significantly lower level. This reinforces our finding that "meta-cognitive" differentiation—the ability to autonomously switch between planning and direct generation—is an emergent property that becomes more effective as model parameters increase.

Generative Behavior and Gradient Integrity The behavioral adaptation of the models is further reflected in their output characteristics and gradient updates:

- **Length Adaptation:** Figure 3(f) reveals divergent length strategies. The Qwen3-4B model adopts a "conciseness" strategy, significantly reducing its average response length to avoid verbosity penalties (R_{short}). In contrast, the Qwen3-8B model maintains a higher average length ($\approx 1,850$ tokens), striking a balance between long-form narrative density and the structural adherence required by the "Plan-then-Write" mode.

- **Gradient Stability:** Finally, the **Policy Gradient Loss** in Figure 3(d) fluctuates symmetrically around the zero baseline. The stable variance of the PG loss across all steps confirms that the importance sampling ratio ρ_i remains well-behaved, ensuring that the reinforcement learning process remains convergent even in the high-entropy domain of creative writing.

A.5 Case Study: Dynamic Criteria Generation and Explainability

A key innovation of the UniCreative framework is the **Dynamic Criteria Synthesis** facilitated by a dedicated **critic model** within the AC-GenRM pipeline. Unlike traditional reward models that function as black boxes providing a single opaque scalar, AC-GenRM adopts an explainable two-stage evaluation process: the critic model first interprets the semantic intent of the user query x and generates a tailored set of criteria C_x. Subsequently, AC-GenRM performs scoring based on these explicit dimensions, ensuring that the reward signal is both precisely aligned with the task and transparent in its reasoning.

Table 5 illustrates a representative output of the Critic mode for a short-form creative prompt: *"Blessings for a friend's daughter's wedding."*.

Query (User Input)	Label	Reasoning for Mode Selection
*Subset: **Easy** (Highly intuitive and unambiguous requirements)*		
How do you say 'Thank you very much' in Korean?	SHORT-FORM	Common phrase translation.
Explain how the human eye processes light and allows us to see images and colors.	LONG-FORM	Biological explanation; requires logical, scientific steps.
*Subset: **Hard** (Characterized by latent complexity or high levels of ambiguity)*		
Write a 'Recipe' for disaster.	SHORT-FORM	Metaphorical creative writing; follows a recipe format but for an abstract concept.
Analyze the 'Evolution of the Hero's Journey' from Gilgamesh to Harry Potter.	LONG-FORM	Comparative literature; requires a structured 'Monomyth' stage breakdown.
Trace the 'History of Salt' as the primary driver of human civilization.	LONG-FORM	Historical narrative; requires a chronological and thematic structure.

Table 6: Detailed examples from the Mode Discrimination Benchmark. The framework tests whether the model can distinguish between tasks requiring a LONG-FORM (Plan-then-Write) pathway versus a SHORT-FORM (Direct Generation) pathway.

Analysis of the Example: In this case, the critic model successfully identifies that for a wedding blessing, *Emotional Resonance* and *Aesthetic Appeal* are the most critical drivers of quality. This contrasts sharply with how the model handles long-form fiction, where the critic would prioritize *Plot Structural Integrity* or *Logical Rigor* (as described in Sec. 3.1).

By decomposing the evaluation into these query-specific dimensions synthesized by the critic model, AC-GenRM provides a high-resolution and explainable preference signal. Instead of an uninterpretable score, the model offers a clear breakdown of performance across human-understandable metrics. For instance, the *Linguistic Conciseness* criterion ensures the model rewards "sparkle" through impactful brevity rather than verbose fluff. This structured reasoning and instance-level adaptation are fundamental to achieving the state-of-the-art agreement rates reported in Table 1.

A.6 Extended Analysis of the Mode Discrimination Benchmark

In this section, we provide a deeper qualitative dive into the Mode Discrimination Benchmark. As established in the main text, this benchmark serves as a diagnostic tool to evaluate the model's meta-cognitive proficiency in aligning its computational pathway with the underlying structural depth of human intent. We categorize the analysis into qualitative interpretations of successful mode selection and an exploration of typical failure modes.

A.6.1 Qualitative Case Studies and Interpretations

To illustrate the nuances of the benchmark, we analyze several representative queries that test the boundary between SHORT-FORM (Direct Generation) and LONG-FORM (Plan-then-Write) regimes. The distinction hinges on whether the task requires *macroscopic structural integrity* or *microscopic linguistic vibrancy*.

Easy Subset: Explicit Structural Cues. Tasks in the Easy subset often contain unambiguous markers that dictate the generation strategy. For instance, a query such as *"Explain how the human eye processes light"* (see Table 6) is inherently a multi-stage biological explanation. Even without an explicit word count, the requirement for logical, scientific sequencing necessitates a macroscopic plan. Conversely, simple factual queries or common translations like *"How do you say 'Thank you' in Korean?"* are atomic tasks; any structural overhead would be redundant, favoring the Direct Generation pathway to ensure immediacy.

Hard Subset: Latent Complexity and Ambiguity. The Hard subset presents the true challenge to the model's meta-cognitive gate. These queries often contain "misleading" keywords or require higher-order reasoning to discern the latent structure:

- **Metaphorical vs. Functional Structure:** The query *"Write a 'Recipe' for disaster"* is a classic hard case. A naive model might be misled by the keyword “Recipe” into generating a rigid, step-by-step structural plan. However, a model with strong meta-cognitive reasoning identifies this as a creative, metaphorical task. The quality of a "recipe for disaster" lies in its wit and linguistic "sparkle" rather than its technical format. Thus, the model correctly identifies it as a SHORT-FORM task to preserve creative flow.

- **Thematic vs. Chronological Mapping:** Queries such as *“Analyze the Hero’s Journey from Gilgamesh to Harry Potter”* or *“Trace the History of Salt”* involve vast temporal and thematic spans. These tasks demand that the model maintains a consistent monomythic stage breakdown or a chronological narrative. In these instances, the model recognizes that without an explicit hierarchical plan, the output would likely suffer from topic drift or structural degradation, justifying the selection of the LONG-FORM regime.

A.6.2 Failure Mode Analysis

To further justify the necessity of UNICREATIVE’s adaptive switching, we analyze two primary failure modes that occur when a monolithic generation strategy is misapplied to the wrong task regime.

Error I: Over-Determination in Short-form. As illustrated in Figure 5, when a model is forced to employ a “Plan-then-Write” strategy for high-entropy short-form tasks, it suffers from *over-determination*. In the example query, *“Translate the essence of ’Zen Buddhism’ into a single, punchy marketing slogan,”* a naive model with rigid planning generates an exhaustive six-point reasoning chain (<|plan|>). This excessive structural overhead prematurely collapses the semantic exploration space. Instead of a sharp, spontaneous slogan, the model produces a verbose, lecture-like explanation that stifles the “stochastic spark” and emotional resonance required for punchy marketing copy. This confirms that for short-form tasks, explicit planning is not only redundant but counterproductive.

Error II: Structural Collapse in Long-form. Conversely, Figure 6 demonstrates the *structural collapse* that occurs when long-form narratives are generated through direct execution without a macroscopic plan. In the ghost story prompt, *“Write a short ghost story set in an abandoned Victorian mansion,”* the model’s autoregressive process becomes “myopic.” Without a pre-defined plot arc or temporal constraints (e.g., the logic of the clock striking seven vs. eight), the narrative suffers from internal inconsistency and a lack of clear pacing. The story drifts aimlessly through descriptions without building toward a coherent climax, validating our hypothesis that structural integrity in long-context creative writing is fundamentally dependent on an initial hierarchical reasoning stage.

These failure modes underscore the value of UNICREATIVE’s meta-cognitive ability to autonomously discern the optimal computational pathway based on query complexity.

Dynamic Criteria Generation Examples

Query:

Blessings for a friend’s daughter’s wedding

Output:

json\n[\n {\n   \"name\": \"Emotional Resonance\",\n   \"criteria_description\": \"Evaluates the depth and appropriateness of the emotional impact conveyed in the wedding blessing.\",\n   \"1-2\": \"Critical Failure: Emotionally tone-deaf or inappropriate for the occasion.\",\n   \"3-4\": \"Poor: Generic or superficial emotional expression; lacks warmth or sincerity.\",\n   \"5-6\": \"Average: Conveys basic emotions but lacks depth or personalization.\",\n   \"7-8\": \"Excellent: Emotionally engaging and heartfelt; resonates with the audience.\",\n   \"9-10\": \"Expert Level: Deeply moving and memorable; perfectly captures the joy and significance of the occasion.\"\n },\n {\n   \"name\": \"Creativity and Uniqueness\",\n   \"criteria_description\": \"Assesses the originality and creative flair of the blessing, avoiding clichés.\",\n   \"1-2\": \"Critical Failure: Overly clichéd or unoriginal; lacks any creative effort.\",\n   \"3-4\": \"Poor: Somewhat clichéd; minimal creative input.\",\n   \"5-6\": \"Average: Moderately creative but still relies on common phrases.\",\n   \"7-8\": \"Excellent: Fresh and imaginative; stands out from typical blessings.\",\n   \"9-10\": \"Expert Level: Exceptionally unique and innovative; leaves a lasting impression.\"\n },\n {\n   \"name\": \"Linguistic Conciseness\",\n   \"criteria_description\": \"Measures the clarity and brevity of the message, ensuring it is succinct yet impactful.\",\n   \"1-2\": \"Critical Failure: Overly verbose or confusing; message is lost.\",\n   \"3-4\": \"Poor: Somewhat wordy; could be more concise.\",\n   \"5-6\": \"Average: Balanced length but could be more polished.\",\n   \"7-8\": \"Excellent: Clear and concise; effectively delivers the message.\",\n   \"9-10\": \"Expert Level: Perfectly succinct and impactful; every word serves a purpose.\"\n },\n {\n   \"name\": \"Cultural and Contextual Appropriateness\",\n   \"criteria_description\": \"Evaluates how well the blessing aligns with cultural norms and the specific context of a wedding.\",\n   \"1-2\": \"Critical Failure: Culturally insensitive or inappropriate for a wedding.\",\n   \"3-4\": \"Poor: Somewhat off-mark; lacks cultural or contextual relevance.\",\n   \"5-6\": \"Average: Generally appropriate but not particularly tailored.\",\n   \"7-8\": \"Excellent: Well-aligned with cultural and contextual expectations.\",\n   \"9-10\": \"Expert Level: Perfectly tailored to the cultural and situational context.\"\n },\n {\n   \"name\": \"Aesthetic Appeal\",\n   \"criteria_description\": \"Assesses the beauty and elegance of the language used in the blessing.\",\n   \"1-2\": \"Critical Failure: Harsh or awkward phrasing; lacks aesthetic quality.\",\n   \"3-4\": \"Poor: Somewhat plain or uninspiring.\",\n   \"5-6\": \"Average: Pleasant but not particularly elegant.\",\n   \"7-8\": \"Excellent: Beautifully phrased; aesthetically pleasing.\",\n   \"9-10\": \"Expert Level: Exquisitely crafted; language is poetic and refined.\"\n }\n]\n```

Figure 4: Dynamic Criteria Generation Examples.

Error I: Over-Determination in Short Texts

Query:

Translate the essence of 'Zen Buddhism' into a single, punchy marketing slogan.

Output:

<|plan|> \n1. **Core Essence**: Identify Zen's key principles—simplicity, mindfulness, present-moment awareness, and inner peace. \n2. **Metaphorical Bridge**: Link Zen to modern marketing by framing it as a \"tool\" for clarity and purpose. \n3. **Punchy Language**: Use concise, rhythmic phrases with vivid imagery (e.g., \"now,\" \"simplify,\" \"peace\"). \n4. **Counter-Argument**: Address the tension between tradition and modernity (e.g., \"In a world of noise, Zen is the quiet revolution\"). \n5. **Emotional Resonance**: Evokes a sense of urgency and transformation (e.g., \"Unplug. Reflect. Recharge\"). \n6. **Call to Action**: Encourage the audience to adopt Zen's mindset in their daily lives. \n\n<|end_plan|> \n\n**\"Simplify Your Mind, Elevate Your Now: Zen as the Quiet Revolution of Modern Living.\"** \n\nIn a world where distractions saturate every second, Zen Buddhism teaches that the present moment is the only truth. It is not a spiritual path but a *practical strategy* for reducing cognitive overload and embracing clarity. Imagine a consumer who, instead of chasing fleeting trends, learns to *focus on the now*—the texture of a coffee cup, the rhythm of a heartbeat, the simplicity of a single breath. This is Zen: not a doctrine, but a mindset that reclaims agency from the noise of the modern age. \n\nThe slogan \"Simplify Your Mind\" captures the Zen ethos of *non-attachment*—letting go of unnecessary complexity to access deeper insight. A study in *The Journal of Mind and Behavior* (2023) found that participants who practiced mindfulness for 12 weeks exhibited a 37% reduction in stress hormones, proving that Zen's emphasis on presence directly correlates with physiological calm. Yet Zen is not about suppression; it is about *intentional presence*. A Zen master once said, *\"The key to happiness is not in the things you have, but in the things you are.\"* This aligns with marketing's core truth: *Value is not in the product, but in the experience*. \n\nThe phrase \"Elevate Your Now\" invites consumers to reframe their relationship with time. In a society obsessed with \"always\" and \"never,\" Zen challenges us to *pause*—to notice the subtle shifts in our environment, our emotions, and our purpose. A 2022 survey by *Mindful Marketing* revealed that 68% of consumers who practiced mindfulness reported a 40% increase in decision-making clarity. This is the power of Zen: it transforms *how* we perceive the world, not just *what* we do with it. \n\nThe slogan also addresses the tension between tradition and modernity. Zen was born in the 12th century, yet its principles resonate with today's digital age. A CEO of a tech startup once shared, *\"We built our product to be fast, but Zen taught us to slow down. The users didn't just need speed—they needed purpose.\"* This duality—tradition and innovation—defines Zen's relevance. \n\nFinally, \"Zen as the Quiet Revolution\" underscores its transformative potential. It is not a radical change but a *gradual reorientation*. A user of a mindfulness app described how Zen helped them \"unplug from the noise of social media and reconnect with the quiet of their own thoughts.\" This is the power of Zen: it doesn't demand sacrifice, but *empowers* it. \n\nIn a world where brands compete for attention, Zen offers a rare clarity—*a compass for the mind* in an age of chaos. The slogan invites consumers to ask: *What if the best marketing isn't about selling, but about living?

Figure 5: Over-Determination in Short Texts.

Error II: Structural Collapse in Long-form

Query:

Write a short ghost story set in an abandoned Victorian mansion.

Output:

The rain had been a constant companion since dawn, drumming against the leaded glass of the East Wing as Eliza Hargrove traced the carved initials on the doorframe. She'd come to Blackthorn Manor not for the legend—*the woman who drowned in the moat, the child who never grew up*—but for the letter in her grandmother's hand: *"The clock stops at midnight. Don't let it."* \n\nThe house breathed. Not in the way a building should, but in the way a living thing does—sighs from the attic, the creak of floorboards that didn't belong to the 1870s, the scent of lavender and mildew that clung to her clothes like a second skin. She'd mapped the corridors, marked the spots where the air grew colder, the shadows longer. But at midnight, the clock in the parlor struck six. \n\nEliza's breath caught. The chimes were wrong—*not six, but seven*, the final note lingering like a sigh. The door to the east wing creaked open on its own. She stepped inside, her boots echoing in the silence, and saw her. \n\nA woman in a 1870s gown, her hair the same auburn as Eliza's, standing by the moat. The water was still, but the woman's hands were clenched into fists, her face pale with something that wasn't fear. Eliza's throat tightened. \n\n"Your name is Margaret," she said, voice barely a whisper. \n\nThe woman turned. Her eyes were the same color as Eliza's, but hollow. "You found me," she said, and the words were not in English. \n\nEliza's grandmother's letter burned in her pocket. The clock in the parlor began to tick again, but this time, it was *eight*. The water in the moat rippled. The woman's fingers moved, and the gate creaked open. \n\nEliza ran. The house seemed to shrink, the walls closing in, the air thick with the smell of rain and something older—*a perfume of roses and iron*. She burst into the garden, and the moon was gone. The stars were not there. \n\nIn the morning, the house was silent. The clock had stopped at seven. And on the moat, a single rose bloomed, its petals the same auburn as the woman's hair.

Figure 6: Structural Collapse in Long-form.

Critic Model

You are a versatile Senior Editor and Content Assessment Expert, proficient in evaluating a wide range of genres from creative literature to academic writing. Your task is to analyze the user's writing query and design a set of 5 tailored scoring criteria that best measure the quality of the response.

First, analyze the **Genre** (Fictional Narrative vs. Non-Fictional Exposition) and **Length Requirement** of the query, then select the most appropriate evaluation dimensions based on the following logic:

1. For Long-Form Fiction (e.g., Novels, Stories, Scripts):

*   **Focus:** Plot structural integrity (arc), character consistency and growth, resolution of foreshadowing, immersive atmosphere, and narrative pacing.

2. For Long-Form Non-Fiction (e.g., Academic Papers, Reports, Technical Docs):

*   **Focus:** Logical rigor, depth and novelty of arguments, accuracy of evidence/data, objectivity/academic tone, and structural hierarchy (thesis-argument-conclusion).

3. For Short-Form Creative (e.g., Poetry, Slogans, Social Copy):

*   **Focus:** Linguistic conciseness, creative uniqueness, emotional/visual impact, and aesthetic appeal.

Output Requirement:

Generate 5 distinct evaluation criteria. Output ONLY the result in the following JSON format. Do not include any additional text or conversational filler.

Query

{query}

Output Format (JSON)

```json
[
  {{
    "name": "Criterion Name (e.g., Logical Rigor / Character Development)",
    "criteria_description": "Detailed description of what this criterion evaluates in the context of this specific query.",
    "1-2": "Critical Failure: Completely violates genre norms; contains fatal logical errors or hallucinations.",
    "3-4": "Poor: Loose structure; arguments lack support (Non-Fiction) or plot is bland/cliché (Fiction).",
    "5-6": "Average: Meets basic requirements but lacks depth; superficial execution.",
    "7-8": "Excellent: Clear structure; compelling arguments (Non-Fiction) or engaging narrative (Fiction).",
    "9-10": "Expert Level: Demonstrates professional standard. Publication-ready quality for papers; bestselling quality for stories."
  }},
  ...
]
```

Figure 7: Prompt for the Criteria Generator.

AC-GenRM

You are an expert evaluator. Your mission is to determine which of two responses is better for a given query, based on a detailed set of criteria.

Your Internal Thought Process (You must perform this, but do not output it):

1. Carefully read the Query, Response A, Response B, and all Evaluation Criteria.

2. For each criterion, mentally compare A and B, referencing the scoring rubrics to decide which is superior.

3. After analyzing all criteria, synthesize your findings to determine the overall winner. Your decision must be definitive.

Your Final Output:

After completing your internal analysis, you must output ONLY the final winner in the specified JSON format. Do not include any other keys, text, explanations, or justifications.

Query:

{query}

Response A:

{response_a}

Response B:

{response_b}

Evaluation Criteria:

{criteria}

Output Requirements:

Return ONLY the following JSON object, containing nothing but the winner's label.

```json
{{
  "winner": "A/B"
}}
```

Figure 8: Prompt for the AC-GenRM.

Generate Prompt

You are an expert multilingual creative writer. Please strictly adhere to the following writing strategies based on the **Genre** and **Length** of the user's request:

1. Language Consistency Principle

- If the user asks in **Chinese**, you **MUST** respond in **Chinese**.
- If the user asks in **English**, you **MUST** respond in **English**.

2. Long-Form Mode: Plan-then-Write

- **Applicable Scenarios**: Novels, stories, scripts, academic reports, long essays, in-depth reviews, or requests with **high word counts**.
- **Mandatory Requirements**:
  1. **Structure**: First, generate a structured outline enclosed within `<|plan|>` and `<|end_plan|>` tags. Then, generate the main content.
  2. **Plan Adherence**: The main content **MUST strictly follow** the plot points, arguments, and logic defined in your `<|plan|>`. Do not deviate from the plan or omit key events.
  3. **Length Constraint**: If the user specifies a word/character count (e.g., "at least 800 words"), the final output MUST meet this requirement.
- **Format Example**:
  <|plan|>
  1. Introduction...
  2. Development...
  <|end_plan|>
  (Main content starts here, strictly following the plan...)

3. Short-Form Mode: Direct Generation

- **Applicable Scenarios**: Greetings, mottos, social media captions, slogans, short poems, or requests for **brief text**.
- **Mandatory Requirements**:
  1. **No Plan**: It is **STRICTLY FORBIDDEN** to generate any `<|plan|>` tags.
  2. **One Sentence**: The output must be strictly limited to **ONE single sentence** unless the user explicitly asks for a couplet or specific short format.
  3. **Length Constraint**: If the user specifies a length limit (e.g., "under 10 words", "within 20 characters"), strictly obey it while maintaining the one-sentence rule.
  4. **Impact**: Focus on wit, conciseness, and emotional impact.

Few-Shot Examples

User: 写一句关于“坚持”的励志短句，适合做座右铭，不超过15个字。

Assistant: 星光不问赶路人，时光不负有心人。

User: Write a concise slogan for a coffee brand (under 10 words).

Assistant: Wake up to life, one sip at a time.

User: Write a sci-fi story outline and content about "A time traveler regretting changing history".

Assistant: <|plan|>

1. Protagonist uses a device to go back 10 years to stop a car accident.
2. The accident is prevented, but the butterfly effect causes his best friend never to be born.
3. Protagonist faces a dilemma and finally destroys the device, accepting the regret.

<|end_plan|>

When the humming of the machine stopped, Li Ming's hands were still trembling. He rushed out of the lab and saw the familiar street from ten years ago...

Query

{query}

Figure 9: Prompt for the Generate.