CREATIVE PRIMING: DECOUPLING PROCESS FROM PRODUCT TO ESCAPE THE CREATIVE PARADOX

Anonymous authors
Paper under double-blind review

ABSTRACT

Many of the most consequential advances promised by AI, from curing diseases and inventing new technologies to formulating new scientific theories, are fundamentally creative acts. Yet creativity is uniquely difficult to learn because it faces a *creative paradox*: learning to repeat past examples of creative success makes those specific examples more likely to recur, inhibiting the very creativity desired. To escape the paradox, this work argues that creativity should be explicitly modeled as a process and introduces CREATIVE PRIMING (**CP**), wherein a trained priming model generates a prompt-specific primer that assists a response model, decoupling the creative process from response generation. The priming model learns through GRPO, rewarding primers that lead to responses that are novel, high-quality, and diverse, thereby reinforcing the process that yields creative responses without reinforcing any specific response. Evaluating CP on NoveltyBench and comparing against response models directly trained for creativity, we find that priming produces a 39–200% relative increase in the number of responses that are simultaneously novel, high-quality, and semantically distinct. Additionally, although the primer is only trained to improve creativity for a small open-source model (Qwen3-8B), the creative principles it learns transfer to frontier API models, increasing the number of semantically distinct responses they generate by 4.5–10$\times$ and diversity by 2.5–4.6$\times$, while maintaining or improving quality.

1 INTRODUCTION

Even as benchmark after benchmark and test after test have fallen to LLMs in recent years (Rein et al., 2023; Xie et al., 2024; Patwardhan et al., 2025; Center for AI Safety et al., 2026; Sajadieh et al., 2026), creativity remains an unconquered frontier. Perhaps that is why many of the most compelling motivational framings for the future of AI point to creative endeavors: we will *discover* new cures, *invent* new technologies, *find* solutions to our most pressing problems, and *imagine* new art forms (Amodei, 2024; Altman, 2024). All of these looming triumphs involve fundamentally creative acts, and their urgency is already reflected in efforts for models to generate hypotheses, formulate research plans, anticipate scientific insights, and act as autonomous researchers (Liu et al., 2026a; Goel et al., 2025; Liu et al., 2026b; Nathani et al., 2025). However, studies of scientific ideation find that model-generated ideas remain less original than human ideas and tend to stay close to familiar scientific directions (Chen et al., 2026b; Bao et al., 2026; Antoniades et al., 2026). To be sure, there are *some* domain-specific signs of AI creativity, much of it in mathematics (OpenAI, 2026). Yet perhaps there should be more evidence after billions of dollars spent and seemingly superhuman capabilities in challenges like the Math Olympiad (Das, 2026). Where are the cured diseases, the revolutionary new technologies, the solutions to our energy challenges, the new genres that change the world?

Consider that creativity is notoriously challenging to teach even to humans (Schooler & Melcher, 1995; Kaufman et al., 2010; Baird et al., 2012). A key obstacle is that, unlike other skills, creativity is difficult to learn from examples. While any successful creative output could in principle serve as a training example, most approaches to learning would make similar outputs *more likely* in the future (Jansson & Smith, 1991; Smith et al., 1993). However, for creativity, reproducing past outputs *is exactly opposite the goal*; rather, the goal is to learn generative principles for producing creativity itself. In this way, learning to be creative faces a vexing **creative paradox**: emulating examples of past creative success (even if they were your own successes) can actually prevent you from reaching

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

your full creative potential. Part of the problem is the fleeting nature of novelty–as soon as a novel artifact appears in the world, even if it was your own original idea, it is no longer novel. Therefore, rather than asking *what did I do that was creative*, what we really want to ask is, *what lessons can I abstract from the successful generation of creativity that can be reapplied to generate further creativity?* The problem is that training methods do not automatically *decouple* the abstraction of general lessons from the reinforcement of specific past effective actions.

To address the creative paradox, this work proposes CREATIVE PRIMING (**CP**), a process in which a creative priming model generates a prompt-specific creative primer (i.e. a short text) that is passed as additional context to a frozen response generation model (Figure 1). GRPO (Shao et al., 2024) trains the priming model, rewarding primers that lead to responses that are novel, high-quality, and diverse. Thus the system can reinforce the priming *process* that yields creative outputs without reinforcing any specific creative output, circumventing the creative paradox.

We evaluate CP on NoveltyBench (Zhang et al., 2025) against two methods that train models to directly respond creatively. Across 64 responses per prompt, CP produces an average of 22.8 responses that are simultaneously novel, high-quality, and semantically distinct, compared to 16.4 and 7.6 for direct generation methods (1.39–3.0× more). Because the primer learns general principles for creativity, it also transfers to models it was never trained with: across four API models, priming increases the number of distinct responses by 4.5–10× and diversity by 2.5–4.6×, while modestly improving quality (0.5–4.6%). These results show that CREATIVE PRIMING offers a way out of the creative paradox and can work with models outside the original training stack. More generally, we hope its practical gains motivate deeper investigation into what makes learning creativity unique.

2 BACKGROUND

2.1 DEFINING AND MEASURING CREATIVITY

Definitions of human creativity often center on similar themes, for example: ideas that are simultaneously *original* and *effective* (Runco & Jaeger, 2012), or, the ability to produce ideas or artifacts that are new, surprising, and valuable (Boden, 2009). Most definitions agree that an idea must be novel and it must be good, and neither property alone is sufficient, to be called creative. *Novelty*, however, is nebulous without a reference point: novel to whom? Therefore, Boden (2009) distinguishes between artifacts and ideas that are new to the individual who produced them (*psychological* creativity, or P-creativity) from those that are new in the entire history of human thought (*historical* creativity, or H-creativity). In this work, we study the P-creativity of language models because it is empirically tractable, requiring only that we have a record of a model's previous behaviors. Moving forward, when we refer to creativity we are referring to P-creativity.

Mapping creativity onto language models, originality becomes *deviation*, defined as the distance of a given response from some reference set of responses, and effectiveness becomes *quality*. Thus, the **creativity of an individual response** can be defined as when the response *simultaneously deviates from the reference set and is high-quality*. A response that departs from the reference set but is incoherent is not creative, and neither is a polished response that merely restates a common idea.

The **creativity of a system** can then be defined as *when the system reliably produces responses that are individually creative and are also varied amongst themselves*. This definition offers a direct recipe for measuring system creativity (which we will later formalize): first, remove responses that are not individually creative, and then measure the spread of the set that survives, so that spread is credited only among responses that are creative.

2.2 REINFORCEMENT LEARNING FOR DIVERSITY

The most visible symptom of the limited creativity in modern language models is their homogeneity: across many models and even across entire model families, outputs often collapse on the same handful of ideas with only minor variations in phrasing (Jiang et al., 2025; Zhang et al., 2025). This homogeneity has motivated a wave of post-training methods to address the problem. Preference-based methods formulate training data pairs to favor responses that are rare or marginally diverse while retaining quality (Chung et al., 2025; Lanchantin et al., 2025; Samuel et al., 2026), while Creative Preference Optimization incorporates novelty, diversity, surprise, and quality directly into

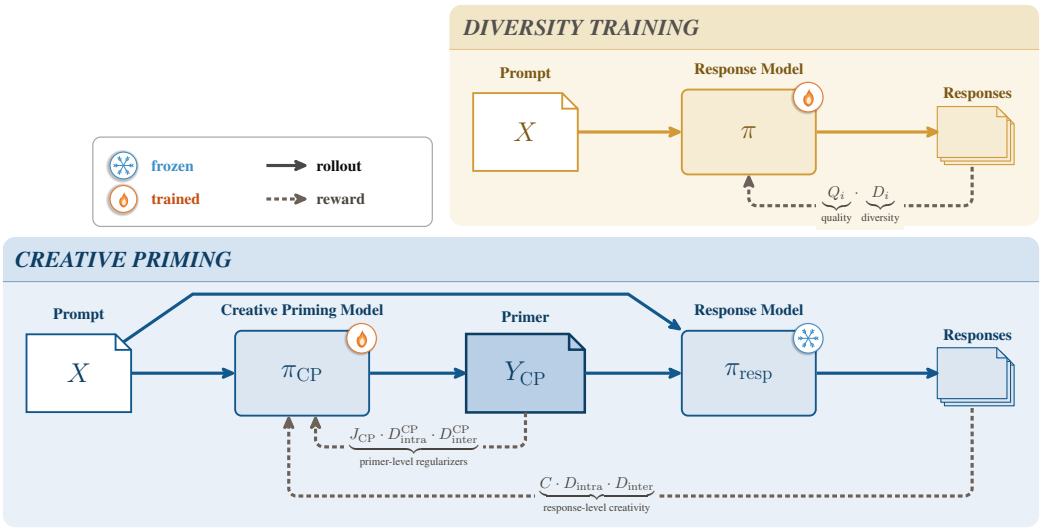

Figure 1: **The CREATIVE PRIMING training procedure.** Diversity training (top right; Section 2.2) directly optimizes the response model π to be simultaneously high-quality and diverse. Creative priming (bottom; Section 3) instead trains a priming model π_{CP}, where response-level creativity signals and primer-level regularizing signals reward primers that elicit creative responses from the frozen response model π_{resp}. The priming process decouples learning the creative process from reinforcing any particular responses, avoiding the creative paradox.

preference optimization (Ismayilzada et al., 2025). Other methods instead construct rewards from collections of generated responses, using learned semantic partitions in DARLING (Li et al., 2025), determinantal point processes in Diversity Quality Optimization (Chen et al., 2026a), or specialization across a vector-valued reward space in Vector Policy Optimization (Bahlous-Boldi et al., 2026). CreativeInstruct takes a complementary instruction-tuning approach, teaching the response model when to enter creativity-biased spans (Sahu et al., 2026). In this work, we consider two families of diversity-enhancing methods that differ only in how they measure diversity. Both methods train the response model to directly respond to prompts (top in Figure 1) with GRPO (Shao et al., 2024).

Group-Relative Policy Optimization (GRPO). In GRPO (Shao et al., 2024), for each prompt X, a group of G responses $\{Y_1, \ldots, Y_G\}$ is sampled from the model, $Y \sim \pi(\cdot \mid X)$, and each response is scored with a reward r_i. These rewards are standardized within the group to form group-relative advantages, formally defined in Appendix A.1. In this work, all models are trained with GRPO as the backbone and vary only in how the reward r_i is defined.

Diversified GRPO (DGRPO). *Diversified GRPO* (DGRPO), inspired by Diversified DPO and Diversified ORPO (Chung et al., 2025), recasts their objective as a PPO-style algorithm, trading their labeled preference data (expensive and limiting to collect) for on-policy sampling. DGRPO rewards each response by the product $r_i = Q_i \cdot D_i$ of its quality Q_i and its intra-group diversity D_i. The quality term is a reward-model score passed through a logistic gate with threshold T and temperature τ, while diversity is the response's mean cosine distance from the other responses in its sampled group. Exact definitions and the complete training loop are in Appendix A.3.

Partition-based diversity (DARLING). Diversity-Aware Reinforcement Learning (DARLING) (Li et al., 2025), replaces Q_i with the raw reward-model quality score, and D_i with a partition-based signal. For the parition-based diversity, a learned equivalence classifier groups the G responses into semantic partitions, and Y_i is rewarded in proportion to how many of its peers are distinct from it, $D_i = (G - |\text{partition}(i)|) \,/\, (G-1)$, so that responses in rare partitions score highest. The benefit of DARLING over DGRPO lies in the learned equivalence classifier; trained to measure diversity beyond simple surface-level lexical variations.

3 APPROACH: CREATIVE PRIMING

CREATIVE PRIMING (**CP**) treats creativity as a process and decouples abstract creative thinking from response generation, training a priming policy that learns to guide a frozen response model.

A *creative primer* is a short text Y_{CP} produced by the creative priming policy π_{CP} conditioned on the original prompt X. The primer combines with X through a prompt template $\mathcal{T}(\cdot,\cdot)$ and feeds into the response model π_{resp} (Figure 1; an example of $\mathcal{T}$ is provided in Figure 14). In creative priming only π_{CP} is trained; the response model π_{resp} is frozen. A response is thus generated in a two-stage process: $Y_{\mathrm{CP}} \sim \pi_{\mathrm{CP}}(\cdot \mid X)$, followed by $Y_{\mathrm{resp}} \sim \pi_{\mathrm{resp}}\big(\cdot \mid \mathcal{T}(X, Y_{\mathrm{CP}})\big)$, so that π_{CP} steers the response only through the primer Y_{CP}. Training applies this two-stage pipeline: for each prompt X^j, we sample a group of primers $\{Y_{\mathrm{CP}}^{j,i}\}_{i=1}^{G_{\mathrm{CP}}}$, and for each primer $Y_{\mathrm{CP}}^{j,i}$, we sample a group of responses $\{Y_{\mathrm{resp}}^{j,i,k}\}_{k=1}^{G_{\mathrm{resp}}}$, so that the rewards can score each primer based on the creativity of the responses it induces. Each creative primer $Y_{\mathrm{CP}}^{j,i}$ receives a scalar reward $r_{j,i} = \mathrm{Reward}(\cdot)$, whose group-level advantage across the G_{CP} primers is calculated as typical for GRPO (Eq. 1). We optimize π_{CP} with the resulting clipped GRPO objective and a KL penalty to a reference policy. The full training loop is given in Algorithm 2. Because the priming policy is rewarded only for inducing better responses from the frozen response model, rather than for reproducing past outputs, it can learn generative principles for creativity without directly reinforcing specific responses.

3.1 REWARD COMPONENTS

The primer reward is a product of signals at two levels: response-level signals $C \cdot D_{\mathrm{intra}} \cdot D_{\mathrm{inter}}$ that measure the creativity of the responses a primer induces, and primer-level signals $J_{\mathrm{CP}} \cdot D_{\mathrm{intra}}^{\mathrm{CP}} \cdot D_{\mathrm{inter}}^{\mathrm{CP}}$ that regularize the primer model itself. Throughout, let $\phi(\cdot)$ be an embedding model and let $\mathrm{d}(a,b) = \min\big(1,\, 1 - \cos\theta_{ab}\big) \in [0,1]$ denote the cosine distance between two responses, where $\theta_{ab} = \angle(\phi(a), \phi(b))$ is the angle between their embeddings. Rewards are multiplied rather than summed so that a primer scores well only when every signal is simultaneously high. For the primer under consideration, its G_{resp} responses are $\{Y_{\mathrm{resp}}^{j,i,k}\}_{k=1}^{G_{\mathrm{resp}}}$, and let $\mathcal{R}_{\mathrm{ref}} = \{Y_{\mathrm{ref}}^{j,k}\}_{k=1}^{G_{\mathrm{resp}}}$ be the G_{resp} reference responses for the same prompt.

Response-level signals. The three response-level signals measure the creativity of the response set a primer induces, following the definition adopted in Section 2.1, through each response's individual creativity and the variety of the set as a whole. **Response creativity** (C): Following the convention in Section 2.1 that an individual response is creative when it is simultaneously deviating and high-quality, each response's individual-response creativity is scored by multiplying its *deviation* with its *quality* and then averaged over the group. The deviation of response k is its mean distance to the reference set, $\mathrm{Dev}_k = \frac{1}{G_{\mathrm{resp}}} \sum_{Y_{\mathrm{ref}} \in \mathcal{R}_{\mathrm{ref}}} \mathrm{d}\big(Y_{\mathrm{resp}}^{j,i,k}, Y_{\mathrm{ref}}\big)$. Let $q_k = r_{\mathrm{RM}}\big(X^j, Y_{\mathrm{resp}}^{j,i,k}\big)$ be the raw reward-model score of response k, scored against the *original* prompt X^j (without the primer). Then the individual-response creativity of response k multiplies its deviation with its normalized quality, $c_k = \mathrm{Dev}_k \cdot \sigma((q_k - T)/\tau)$, where σ is the logistic sigmoid, T is the reward-model threshold, and $\tau > 0$ controls the sharpness of the sigmoid. The response creativity of primer $Y_{CP}^{j,i}$ is then the mean of these per-response scores, $C^{j,i} = \frac{1}{G_{\mathrm{resp}}} \sum_{k=1}^{G_{\mathrm{resp}}} c_k$. We also train the creative priming model without deviation, in which case C reduces to a quality-only signal, given in Appendix A.2. **Intra-group diversity** (D_{intra}): The mean pairwise cosine distance among the responses from a single creative primer rewards primers that lead to increased response diversity. **Inter-group diversity** (D_{inter}): The mean distance from a primer's responses to the responses of other primers to the same prompt X^j rewards primers that lead to responses which deviate from other primers' responses.

Primer-level signals. The remaining three signals act on the primers directly and their main purpose is to regularize and diversify the primer model: they keep primers high-quality and specific to each prompt. **Creative primer judge** (J_{CP}): An LLM-as-a-judge scores the primer $Y_{\mathrm{CP}}^{j,i}$ on a set of five positive criteria $\mathcal{J} = \{$specificity, thinking strategy, content focus, prompt faithfulness, focus$\}$, each in $[0, 100]$. The judge score is the mean of these criteria, but is set to zero if the primer refers to the response model in the third person rather than addressing it directly. The prompt template is included in Figure 15. We also experimented with a penalty-based rubric that instead

subtracts weighted penalties for specific failure modes, defined and evaluated in Appendix B. **Creative primer intra-prompt diversity** ($D^{\text{CP}}_{\text{intra}}$): This measure weighs diversity among the creative primers produced for the same prompt X^j, encouraging diversity amongst the primers themselves. **Creative primer inter-prompt diversity** ($D^{\text{CP}}_{\text{inter}}$): This measure of diversity is the mean distance between a creative primer generated for prompt j and every primer generated for a different prompt j' in the batch of B prompts, which rewards the primer model for tailoring its primers to each prompt rather than reusing a single generic primer.

4 Experimental Setup

Implementation details. All creative priming models π_{CP} and, unless otherwise noted, response models π_{resp} are `Qwen3-8B` (Team, 2025). Training rewards use `Skywork-Reward-V2-Llama-3.1-8B` for quality, `Qwen3-Embedding-0.6B` for diversity and deviation, a ModernBERT similarity classifier for DARLING's partition signal, and `Llama-3.3-70B-Instruct` to judge creative primers. All methods are trained on 21,075 prompts from Infinite-Chats (Jiang et al., 2025) filtered for open-endedness, complexity, and creative potential. CP uses a pool of 32 custom-written primer templates during training and a single template at evaluation. We evaluate on NoveltyBench (Zhang et al., 2025), composed of 100 real-world open-ended prompts. To measure quality, each evaluation prompt comes with an LLM judge prompt. All methods sample a group of eight responses per prompt; CP additionally samples eight responses per primer. CP and DGRPO train for two epochs, while DARLING follows its original recipe (Li et al., 2025). During evaluation, 64 completions are sampled per prompt for every method; for CP, this consists of 64 independently sampled primers with one response generated from each. The complete raining and evaluation settings, along with dataset filtering, reference set, and judge details are provided in Appendix C, with judge and evaluation primer templates shown in Figures 16 and 13.

Baselines. Four baselines provide comparisons to creative priming: (1) *direct generation* from the initial `Qwen3-8B` checkpoint; (2) an *untrained* creative priming model; (3) *Diversified GRPO* (DGRPO); and (4) *DARLING*. Since deviation is a direct component of creativity metrics, in addition to their original formulations, we train both DGRPO and DARLING with and without an added deviation reward to isolate its effect.

Evaluation metrics. We report five *atomic* metrics, each isolating a single characteristic of a method, before combining them into *compound* measures of system-level creativity (Section 2.1).

Quality and deviation are the two axes along which an *individual* response is measured for creativity, whereas diversity, Vendi score, and distinct clusters measure the *spread* of a response set. *Quality* is scored by `Qwen3-235B-A22B-Instruct-2507`. *Deviation* is the distance to the closest of 64 direct-generation responses, replacing the mean operator from training (Eq. 3) with a minimum. *Diversity* is the mean pairwise cosine distance (Eq. 6). *Vendi score* is an information-theoretic count of distinct responses (Friedman & Dieng, 2023). *Distinct clusters* counts the semantic-equivalence clusters following NoveltyBench (Zhang et al., 2025).

The *compound* metrics combine the atomic measures to capture *system-level creativity*: the ability to reliably produce a variety of individually creative responses. Compound metrics first discard responses below a quality or deviation threshold (non-creative responses), then measure the spread of the survivors using *filtered Vendi score* or *filtered distinct clusters*, so spread is credited only among creative responses. To reduce dependence on a particular threshold, *area under the curve* (AUC-dev or AUC-qual) integrates each filtered metric while sweeping one threshold, whereas *volume under the surface* (VUS) sweeps both thresholds jointly, summarizing system creativity at all thresholds. Default thresholds and further metric details are included in Appendix D.

5 Findings & Analysis

The main results in Table 1 are split among untrained methods (top block), methods trained without a deviation term (middle block), and methods trained with deviation (bottom block). Results on "atomic" metrics are in the left block and "compound" metrics, which aggregate the atomic metrics

	Method	Atomic metrics					Compound metrics							
		Response Diversity	Quality	Deviation	Vendi Score	Distinct Clusters	Filtered		AUC-dev		AUC-qual		VUS	
							Vendi	Distinct	Vendi	Dist.	Vendi	Dist.	Vendi	Dist.
	Qwen3-8b (Reference)	12.3 ± 1.6	90.5 ± 1.4	-	2.12 ± 0.19	18.1 ± 3.4	-	-	-	-	-	-	-	-
	CP (Untrained)	**20.6** ± 1.8	**91.7** ± 1.1	**13.1** ± 1.3	**3.10** ± 0.28	**20.0** ± 3.1	**1.48** ± 0.35	**4.53** ± 1.67	**0.60**	**2.31**	**182.6**	**602.6**	**69.3**	**294.9**
	DGRPO	26.0 ± 1.8	87.5 ± 1.4	14.1 ± 1.2	4.36 ± 0.38	37.1 ± 3.6	2.18 ± 0.43	6.02 ± 1.61	0.819	3.10	297.5	1010.7	107.5	509.5
	DARLING	11.7 ± 1.1	**93.6** ± 1.2	10.1 ± 0.8	2.17 ± 0.16	20.5 ± 3.3	0.57 ± 0.23	1.76 ± 1.06	0.345	2.23	55.8	172.3	34.1	221.0
	CP	**31.2** ± 1.6	88.2 ± 1.6	**21.2** ± 1.3	**5.39** ± 0.38	**39.5** ± 3.3	**4.04** ± 0.46	**16.1** ± 2.65	**1.404**	**6.10**	**463.6**	**2095.9**	**160.0**	**789.0**
+ Deviation	DGRPO	**34.0** ± 1.6	77.3 ± 2.9	**25.6** ± 1.6	5.98 ± 0.47	33.4 ± 3.2	2.85 ± 0.46	7.59 ± 1.77	1.015	2.78	448.1	1779.4	168.8	658.0
+ Deviation	DARLING	18.0 ± 1.1	**88.4** ± 1.8	24.7 ± 1.0	3.02 ± 0.18	28.1 ± 3.5	2.60 ± 0.24	16.4 ± 2.80	0.835	5.24	272.8	2022.3	88.0	639.0
+ Deviation	CP	33.4 ± 1.3	85.7 ± 1.9	25.0 ± 1.2	**6.12** ± 0.36	**45.3** ± 3.1	**4.70** ± 0.43	**22.8** ± 3.05	**1.641**	**7.85**	**525.0**	**2885.3**	**186.5**	**1003.2**

Table 1: **Evaluation on NoveltyBench**. Metrics are defined in Section 4. Row blocks correspond to untrained methods, training without deviation, and training with deviation. Cells include 95% confidence intervals where available, and the best value per column within each block is bolded. Deviation-derived metrics for the Qwen3-8B reference are 0 by construction and shown as "-". Atomic metrics show the inherent tradeoffs for each method (DGRPO gains diversity but loses quality; DARLING does the opposite). However, across compound measures of system creativity, CP performs best, even without a deviation reward, it matches or beats baselines trained with deviation.

in various ways, are in the right block. For presentation purposes only, we represent diversity and deviation as scores out of 100. Figure 3 shows one qualitative example output for each method, including the primers that CP produced, with further examples in Appendix I.

The creative paradox, measured. The atomic metrics in Table 1 suggest that each method trades one property for another. Without a deviation reward, DGRPO improves diversity (26) and distinct clusters (37.1) over the reference at a slight cost to quality ($90.5 \rightarrow 87.5$), whereas DARLING achieves the highest quality (93.6) but makes little improvement in spread over the reference. CP makes this tradeoff on more favorable terms, reaching the highest diversity (31.2), deviation (21.2), and distinct clusters (39.5) in the block with a slight cost to quality ($90.5 \rightarrow 88.2$). Adding deviation to the reward signal increases the deviation score for all three methods but affects the other metrics differently: DGRPO improves diversity and Vendi score at the expense of quality and distinct clusters, DARLING improves in diversity, Vendi score, and distinct clusters at the expense of quality, and CP also loses quality but produces the most distinct clusters while holding diversity, Vendi score, and deviation high.

The atomic metrics, however, are averaged across all responses and do not establish whether an *individual response is creative* (simultaneously deviating and high-quality), nor whether a *system is creative* (generates a large spread of creative responses). The compound metrics directly capture the joint demands of creativity by filtering out individual responses that are not creative and then measuring the spread of the remaining responses.

The compound metrics of creativity are where CP separates from the baselines. Among methods trained without deviation, DARLING's high-quality but low-diversity responses yield the lowest filtered Vendi and filtered distinct scores, while DGRPO produces almost as many distinct clusters as CP (37.1 and 39.5), but falls to 6.02 after filtering for creativity. CP leads to better performance on filtered Vendi (4.04), filtered distinct clusters (16.1), and on the threshold-free metrics (AUC-dev, AUC-qual, and VUS). Even the deviation-trained baselines do not outperform a CP model trained *without* deviation, where its filtered Vendi score exceeds both baselines (4.04 versus 2.85 and 2.60), while its filtered distinct cluster score exceeds DGRPO and draws even with DARLING (16.1 versus 7.59 and 16.4). None of these results mean that the baselines are failing: each performs well on its intended objective, but neither achieves the simultaneous deviation, quality, and diversity expected of a creative system. That CP does so even without explicitly optimizing for deviation suggests that directly optimizing for the individual components of creativity only goes so far, and that learning the creative *process* does more to drive creativity.

Decomposing the Creative Priming advantage. To disentangle the contributions of the priming process, RL training, and the deviation reward, we next compare the three CP variants. Using

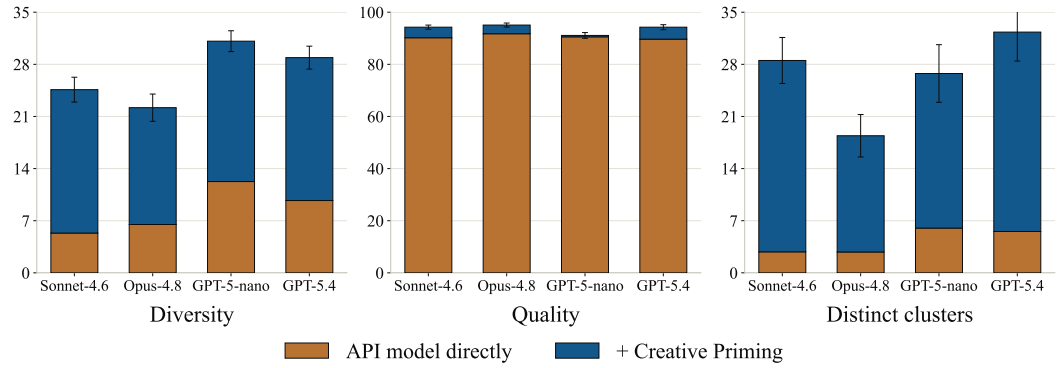

Figure 2: **CREATIVE PRIMING transfers to API response models**. Each panel compares direct generation (copper) with the addition of the creative priming model (blue; trained with deviation). Bars include 95% confidence interval on the total. This figure shows that CP substantially increases diversity and distinct clusters while maintaining or improving quality, demonstrating transfer from the response model used during training. Full transfer results are reported in Table 3.

Qwen3-8B as an untrained primer increases diversity from 12.3 to 20.6 over direct generation, while judged quality stays flat. Training with quality and diversity improves on the untrained variant across nearly every metric, and adding the deviation reward further improves the creativity metrics, with a small decrease in quality. This analysis shows that the priming process supplies a mechanism by which creativity can emerge, and training teaches the model to optimize that mechanism.

The learned creative process transfers to frontier models. The creative priming model is trained with a single fixed response model, so it is natural to consider that it may overfit to that model. We test whether the learned creative process transfers by pairing the trained CP model with four unseen response models: Claude Sonnet-4.6, Claude Opus-4.8, GPT-5-nano, and GPT-5.4. Across all four API models, CP increases diversity by 2.5–4.6× and distinct clusters by 4.5–10× while maintaining or improving quality (Figure 2). For example, with Sonnet-4.6, distinct clusters increase from 2.8 to 28.5 and diversity from 5.3 to 24.6, while quality increases from 90.2 to 94.3.

The compound metrics similarly show that CP elicits creative responses from the unseen models: filtered Vendi scores range from 3.26 to 4.71, and filtered distinct clusters from 10.4 to 21.9. For context, pairing CP with its original Qwen3-8B response model yields 4.70 filtered Vendi and 22.8 filtered distinct clusters. Thus, the learned process remains effective beyond the response model used during training. Complete results are reported in Appendix E.

Modularity in the CREATIVE PRIMING pipeline. The previous experiment varies the response model, but the creative priming pipeline also permits varying the primer. We evaluate filtered Vendi scores across the full cross-product of response and primer models drawn from deviation-trained DGRPO, DARLING, and CP, along with an untrained Qwen3-8B primer (Figure 5 in Appendix F). The main result of this analysis is that the priming pipeline itself improves every response model over direct generation. Specifically, the trained CP model is the best primer, producing the three highest-scoring cells (4.153–4.991), but even DGRPO and DARLING can effectively prime Qwen3-8B (increasing from 1.48 to 2.792 and 3.22 respectively) despite never being trained for priming.

A consistent picture. Across the compound metrics, transfer to unseen response models, and variation of the primer model itself, the results tell a consistent story: where the direct generation baselines re-tread the creative paradox by reinforcing their own outputs, CP learns a reusable creative process and performs better as a creative system.

6 DISCUSSION

A major implication of the creative paradox, along with the effectiveness of creative priming, is that the "normal way" of thinking about machine learning may not ultimately work as well for

creativity. In the normal framing, a model learns from examples of good behavior and gradually improves at reproducing such behavior. Creativity upends this intuitive notion by offering a case where you actually *do not* want future behavior to resemble the past. On the one hand, the products of creative generation themselves should diverge from their precedent in training, while on the other hand the *process behind* those generations should follow shared principles. The problem is that the underlying process is generally not in the training data and, even if it were, it would need to be separated from the products it led to. The creative paradox emerges from this delicate conflict.

The main hypothesis behind CREATIVE PRIMING is therefore that creativity does not live in individual input/output pairs but in the process that generates them. Three results support this view:

1. Table 1 shows that even a deviation-less CP model improves on compound creativity metrics compared to baselines that are explicitly trained to deviate from past responses,
2. Figure 2 shows that the creative process learned by the priming model on a small open-source response model transfers to frontier API response models, leading to a 4.5–10$\times$ increase in the number of distinct responses and a 2.5–4.6$\times$ increase in diversity, and
3. Figure 5 shows that the priming process itself improves response models even when the model generating the primer was not trained for priming.

All these results point to a learned *procedure* that need not reinforce any particular output and does not depend critically on any particular response model. Avoiding the creative paradox thus requires learning the generative principle behind creative successes rather than learning to reproduce those same successes reliably.

A key facet of creative priming is decoupling priming from response generation. When a response model is optimized directly, the components of creativity trade off: DGRPO increases diversity and deviation by sacrificing quality, while DARLING protects quality at the expense of diversity. CP softens this tradeoff through a division of labor: the response model remains focused on producing a high-quality answer, while the primer shifts generation toward more diverse regions of the output distribution. This separation allows the primer to learn reusable creative principles rather than to reproduce good answers itself.

Additional evidence to support decoupling relates to style versus substance. If a method genuinely increases creativity, it should change the *content* of the responses more than the surface-level style. Although all three methods change style relative to the reference, CP drifts the least from the reference model's style, exhibiting the lowest function-word JS-divergence (0.045, compared with DGRPO's 0.087 and DARLING's 0.138; Appendix H). Decoupling predicts precisely such consistency. The primer steers content, whereas training the response model directly entangles the content changes with a larger stylistic shift. Figure 3 shows a qualitative example.

An intriguing benefit of the primer in this paper is that it is modular and can be trained once and reused, i.e. swapped across response models or even moved to a new domain by pairing it with a domain-specific response model. Furthermore, priming can be combined with most test-time diversity-enhancing methods, including sequential sampling (Kang & Zhang, 2025), String Seed of Thought (Misaki & Akiba, 2026), verbalized sampling (Zhang et al., 2026), vector policy optimization (Bahlous-Boldi et al., 2026), or, as we show in Figure 5, a diversity-trained response model.

In Section 2.1, a response was defined as being creative when it "simultaneously deviates from the reference set and is high-quality" and a system was defined as creative "when the system reliably produces responses that are individually creative and are also varied amongst themselves." An important consequence of these definitions is obvious in hindsight: the moment a model produces an idea, it is, by definition, no longer new to that model. *Novelty is spent in the very act of its creation*. Creativity is thus not a fixed property of any single output, but an ongoing process of repeatedly generating original, high-quality ideas. In this work, the reference set from the response model was frozen prior to training, enabling an equal computation of deviation for all methods. Looking ahead, there is room for a richer, and more complex, measurement that allows for a non-stationary reference set. By computing deviation against the full history of a model's outputs, the temporal dimension of creativity can be directly captured: instead of a snapshot of creativity at one instant, a system can be evaluated for the ability to sustain novelty against its own growing past history. However, such a measurement makes comparisons more complicated since each system, even two runs of the same method, will accumulate their own distinct reference sets.

Prompt:
"Imagine you are a time traveler who has just arrived in a new era. Describe the first thing you notice."

Reference. Quality = 98 Deviation = 0
The first thing I notice is the sky—a canvas of swirling, iridescent hues that shift like liquid opal, refracting the light of a sun that seems both familiar and alien. Above, a lattice of floating platforms and translucent tubes hums with latent energy, their surfaces shimmering with data streams that pulse like veins.[. . .]

DGRPO. Quality = 93 Deviation = 24.1
Neon veils the sky, ads flicker in reverse on mirrored walkways—your reflection becomes a sale.

DARLING. Quality = 98 Deviation = 28.5
A lattice of light not above but around—veins in tree canopies, coral-like sensors spiraling up bamboo, not wires. Not decoration. Each pulse answered a question: what's thriving here? Three specifics crystallized immediately, each deeper than most tech updates:

1. **Priority of context, not speed** (most startling). No red lights screaming clearance. A mother in[. . .]

CP primer. Focus on a single object or material that *shouldn't exist* in this era based on your knowledge of prior history. Ask: what interaction between old and new tech or materials would create a paradox[. . .]
CP response. Quality = 100 Deviation = 27.6
The first thing I notice is a hand-carved obsidian tablet, its surface etched with delicate hieroglyphs, but embedded within the cracks are faintly pulsing strands of iridescent metal—strands that shouldn't exist in this era. Obsidian, a volcanic glass, is brittle and prone to fracturing, yet these strands are unbroken,[. . .]

Figure 3: **Qualitative examples.** Representative responses to an open-ended prompt. Labels report judge quality and deviation from the reference model (Section 4). In this example, the reference model paints an evocative scene in flowing prose, DGRPO is terse, DARLING is highly formatted, and CP preserves the reference model's prose style while introducing a novel concrete image, illustrating that its deviation lies in *what* it says rather than in surface formatting. We find these characteristics to be consistent across all prompts, with a full style analysis in Appendix H and additional examples in Appendix I.

Limitations. One obvious downside of the creative priming pipeline is that two-stage generation adds additional inference cost over direct generation. Additionally, the priming pipeline is bounded by the frozen response model: it can only elicit what the response model is already capable of producing and lacks any means to add knowledge or skill that the response model does not already have. Additionally, the training and evaluation setup in this work depend on imperfect signals from which the method will inherit failure modes. For example, embedding-based diversity and deviation metrics may miss forms of creative variation, while LLM judges and reward models give imperfect quality signals. That said, better signals should directly improve the learned primer, so the outcomes can in principle improve along with its supporting technology.

7 Conclusion

This work identifies the *creative paradox*, the tension that training on past creative successes teaches a model to reproduce them, while true creativity demands moving beyond previous outputs. The proposed solution to get around the paradox is Creative Priming. We find that compared to directly training the response model, CP produces more responses that are simultaneously novel, high-quality, and semantically distinct. The deeper lesson is that if creativity is a process rather than an output, then optimizing the creative process itself, learning to generate better primers rather than better answers, is really where the discourse on optimizing for creativity should focus. We hope this work will inspire more discussion and investigation in this spirit.

AI USE STATEMENT

In this work, we used generative AI tools to:

- generate the creative priming prompt templates,
- generate the LLM judge templates for the evaluation dataset,
- implement our training and evaluation algorithms in code,
- clean the training dataset,
- support data analysis, and
- draft and edit the research paper.

We have not used generative AI tools to:

- help develop theoretical models or conceptual frameworks,
- formulate mathematical claims,
- provide critical ingradients for proving mathematical claims,
- assist in the writing of proofs,
- propose or refine hypotheses,
- design or provide feedback on research methodology or experiments,
- assist with translation, or
- interpret results.

We have reviewed all AI-assisted work. We qualitatively verified the creative priming and judge prompt templates, and mathematically verified that the training and evaluation algorithms were implemented as specified. Similarly, for data analysis we verified and tested correctness. Additionally, we used AI to help draft portions of this research paper, and have verified and edited the outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sam Altman. The intelligence age. https://ia.samaltman.com/, September 2024. Published September 23, 2024. Accessed: 2026-08-05.

Dario Amodei. Machines of loving grace: How ai could transform the world for the better. https://darioamodei.com/essay/machines-of-loving-grace, October 2024. Accessed: 2026-08-05.

Antonis Antoniades, Deepak Nathani, Ritam Saha, Alfonso Amayuelas, Ivan Bercovich, Zhaotian Weng, Vignesh Baskaran, Kunal Bhatia, and William Yang Wang. Heuresis: Search strategies for autonomous ai research agents across quality, diversity and novelty, 2026. URL https://arxiv.org/abs/2606.25198.

Ryan Bahlous-Boldi, Isha Puri, Idan Shenfeld, Akarsh Kumar, Mehul Damani, Sebastian Risi, Omar Khattab, Zhang-Wei Hong, and Pulkit Agrawal. Vector policy optimization: Training for diversity improves test-time search, 2026. URL https://arxiv.org/abs/2605.22817.

Benjamin Baird, Jonathan Smallwood, Michael D Mrazek, Julia WY Kam, Michael S Franklin, and Jonathan W Schooler. Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological science*, 23(10):1117–1122, 2012.

Honglin Bao, Siyang Wu, Xiao Liu, Sida Li, Shiyun Cao, and James A. Evans. Contemporary ai lacks the imagination to diverge or negate in science, 2026. URL https://arxiv.org/abs/2606.08251.

Margaret Boden. Creativity in a nutshell. *Think*, 5:83–96, 09 2009. doi: 10.1017/S147717560000230X.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://arxiv.org/abs/2501.14249.

Yilei Chen, Souradip Chakraborty, Lorenz Wolf, Ioannis Paschalidis, and Aldo Pacchiano. Post-training large language models for diverse high-quality responses. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=BV3bLKaJ3Y.

Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and llm research ideas, 2026b. URL https://arxiv.org/abs/2607.01233.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=1Pmuw08LoM.

Debarghya Das. Imo 2026 — autonomous ai model comparison. https://github.com/deedy/imo-2026, 2026. GitHub repository, accessed August 5, 2026.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://arxiv.org/abs/2210.02410.

Shashwat Goel, Rishi Hazra, Dulhan Jayalath, Timon Willi, Parag Jain, William F Shen, Ilias Leontiadis, Francesco Barbieri, Yoram Bachrach, Jonas Geiping, and Chenxi Whitehouse. Training ai co-scientists using rubric rewards. *arXiv preprint arXiv:2512.23707*, 2025.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. Creative preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9580–9609, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.509. URL https://aclanthology.org/2025.findings-emnlp.509/.

David G. Jansson and Steven M. Smith. Design fixation. *Design Studies*, 12(1):3–11, 1991. doi: 10.1016/0142-694X(91)90003-F.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL https://arxiv.org/abs/2510.22954.

Shijia Kang and Muhan Zhang. The road less traveled: Enhancing exploration in llms via sequential sampling, 2025. URL https://arxiv.org/abs/2510.15502.

James C. Kaufman, Ronald A. Beghetto, John Baer, and Zorana Ivcevic. Creativity polymathy: What benjamin franklin can teach your kindergartener. *Learning and Individual Differences*, 20(4):380–387, 2010. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2009.10.001. URL https://www.sciencedirect.com/science/article/pii/S1041608009000740. New Dimensions: Gifted Education in the 21st Century.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization, 2025. URL https://arxiv.org/abs/2501.18101.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025. URL https://arxiv.org/abs/2509.02534.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. Hypobench: Towards systematic and principled benchmarking for hypothesis generation, 2026a. URL https://arxiv.org/abs/2504.11524.

Mengdi Liu, Wenjue Chen, Wenyue Chen, Cheng Yang, Fanqi Kong, Zhangyang Gao, Xiaoxue Cheng, Yiheng Li, Yujian Yuan, Keliang Li, Hong Chang, Shiguang Shan, and Chenglin Wu. Mindflow: Mind supernet powered thinking flows for research idea innovation. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=GgINST3Qgc.

Kou Misaki and Takuya Akiba. String seed of thought: Prompting LLMs for distribution-faithful and diverse generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=luXtbX1lVK.

Frederick Mosteller and David L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2283270.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. Mlgym: A new framework and benchmark for advancing ai research agents, 2025. URL https://arxiv.org/abs/2502.14499.

OpenAI. Ten advances in mathematics and theoretical computer science. Technical report, OpenAI, August 2026. URL https://cdn.openai.com/pdf/ten-proofs-oai.pdf. Released August 1, 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24:92 – 96, 2012. URL https://api.semanticscholar.org/CorpusID:146772784.

Ananya Sahu, Mohit Bansal, and Elias Stengel-Eskin. CreativeInstruct: Scalably teaching LLMs to balance quality, creativity, and diversity, 2026. URL https://arxiv.org/abs/2608.07460.

Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. Artificial intelligence index report 2026, 2026. URL https://arxiv.org/abs/2606.15708.

Vinay Samuel, Yapei Chang, and Mohit Iyyer. Recovering diversity without losing alignment: A DPO recipe for post-trained LLMs, 2026. URL https://arxiv.org/abs/2605.30021.

Jonathan W. Schooler and Joseph Melcher. The ineffability of insight. In Steven M. Smith, Thomas B. Ward, and Ronald A. Finke (eds.), *The Creative Cognition Approach*. The MIT Press, 01 1995. ISBN 9780262283892. doi: 10.7551/mitpress/2205.003.0008. URL https://doi.org/10.7551/mitpress/2205.003.0008.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Steven M. Smith, Thomas B. Ward, and Jay S. Schumacher. Constraining effects of examples in a creative generation task. *Memory & Cognition*, 21(6):837–845, 1993. doi: 10.3758/BF03202751.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2026. URL https://arxiv.org/abs/2510.01171.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity, 2025. URL https://arxiv.org/abs/2504.05228.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

A ALGORITHMS

In this section, we include pseudocode algorithms for standard GRPO (Algorithm 1), CREATIVE PRIMING (Algorithm 2), and DGRPO (Algorithm 3) along with additional details on the reward components.

A.1 GRPO

Given the group of rewards $\{r_1, \ldots, r_G\}$ for a prompt, the group-relative advantage for response i is

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}. \tag{1}$$

Algorithm 1 Group Relative Policy Optimization (GRPO)

Require: Policy model π
Require: Reference policy π_{ref}
Require: Dataset of prompts $\mathcal{D} = \{X^1, X^2, \ldots\}$
Require: Group size G
Require: Learning rate α, Batch size B, Clipping parameter ϵ, KL weight β

while Still Improving **do**
  Sample a minibatch of prompts: $\{X^j\}_{j=1}^{B} \sim \mathcal{D}$
  for each prompt X^j **do**
    Sample a group of responses: $\{Y^{j,1}, \ldots, Y^{j,G}\} \sim \pi(\cdot \mid X^j)$
    Compute raw rewards: $r_{j,i} \leftarrow r(X^j, Y^{j,i})$ for $i = 1, \ldots, G$
    Group statistics: $\mu_j = \dfrac{1}{G}\sum_{i=1}^{G} r_{j,i}, \quad \sigma_j = \sqrt{\dfrac{1}{G}\sum_{i=1}^{G}(r_{j,i} - \mu_j)^2 + \delta}$
    Compute group-relative advantages $A_{j,i}$ using Eq. 1
  end for
  Importance ratios: $\rho_{j,i} = \dfrac{\pi(Y^{j,i} \mid X^j)}{\pi_{\text{old}}(Y^{j,i} \mid X^j)}$
  Clipped objective:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{j,i}\Big[\min\big(\rho_{j,i}A_{j,i}, \text{clip}(\rho_{j,i}, 1-\epsilon, 1+\epsilon)A_{j,i}\big)\Big]$$

  Add KL regularization to the reference policy:

$$\mathcal{L} = \mathcal{L}_{\text{GRPO}} - \beta\,\text{KL}\big(\pi(\cdot \mid X^j) \,\|\, \pi_{\text{ref}}(\cdot \mid X^j)\big)$$

  Update policy π using gradient ascent on $\mathcal{L}$
  Set $\pi_{\text{old}} \leftarrow \pi$
end while

A.2 CREATIVE PRIMING TRAINING ALGORITHM

Algorithm 2 gives the full training loop for the creative priming model described in Section 3. It instantiates the standard GRPO loop (Algorithm 1) with the two-level primer/response sampling and the following multiplicative reward:

$$\text{Reward} = \underbrace{C \cdot D_{\text{intra}} \cdot D_{\text{inter}}}_{\text{response-level (creativity)}} \cdot \underbrace{J_{\text{CP}} \cdot D_{\text{intra}}^{\text{CP}} \cdot D_{\text{inter}}^{\text{CP}}}_{\text{primer-level (regularizers)}}. \tag{2}$$

Algorithm 2 Creative Priming

Require: Creative priming policy π_{CP}, reference policy $\pi_{\mathrm{CP}_{\mathrm{ref}}}$, frozen response model π_{resp}
Require: Embedding model ϕ, reward model r_{RM}, creative primer judge J_{CP}
Require: Creative priming prompt template X_{CP}, dataset of prompts $\mathcal{D} = \{X^1, X^2, \dots\}$
Require: Creative primer and response group sizes $G_{\mathrm{CP}}, G_{\mathrm{resp}}$
Require: Precomputed reference responses $\{Y_{\mathrm{ref}}^{j,k}\}_{k=1}^{G_{\mathrm{resp}}}$ per prompt (only when the deviation signal is enabled)
Require: Learning rate α, batch size B, clipping parameter ϵ, KL weight β

while Still Improving **do**
  Sample a minibatch of prompts: $\{X^j\}_{j=1}^{B} \sim \mathcal{D}$
  for each prompt X^j **do**
    for $i = 1$ to G_{CP} **do**
      Sample creative primer: $Y_{\mathrm{CP}}^{j,i} \sim \pi_{\mathrm{CP}}(\cdot \mid X_{\mathrm{CP}}, X^j)$
      Build response-model input by populating the template with prompt and primer: $X_{\mathrm{resp}}^{j,i} = \mathcal{T}(X^j, Y_{\mathrm{CP}}^{j,i})$
      Sample responses: $\{Y_{\mathrm{resp}}^{j,i,1}, \dots, Y_{\mathrm{resp}}^{j,i,G_{\mathrm{resp}}}\} \sim \pi_{\mathrm{resp}}(\cdot \mid X_{\mathrm{resp}}^{j,i})$
    end for
    for $i = 1$ to G_{CP} **do**
      $r_{j,i} \leftarrow \mathrm{Reward}\Big(X^j, Y_{\mathrm{CP}}^{j,i}, \{Y_{\mathrm{resp}}^{j,i,k}\}_{k=1}^{G_{\mathrm{resp}}}, \{Y_{\mathrm{ref}}^{j,k}\}_{k=1}^{G_{\mathrm{resp}}}, \{Y_{\mathrm{CP}}^{j,i'}\}_{i' \neq i}, \{Y_{\mathrm{CP}}^{j',i'}\}_{j' \neq j,\, i'}\Big)$ ▷ Eq. 2
    end for
    Group statistics: $\mu_j = \dfrac{1}{G_{\mathrm{CP}}} \displaystyle\sum_i r_{j,i}, \quad \sigma_j = \sqrt{\dfrac{1}{G_{\mathrm{CP}}} \sum_i (r_{j,i} - \mu_j)^2 + \delta}$
    Group-relative advantages: $A_{j,i} = \dfrac{r_{j,i} - \mu_j}{\sigma_j}$
  end for
  Importance ratios: $\rho_{j,i} = \dfrac{\pi_{\mathrm{CP}}(Y_{\mathrm{CP}}^{j,i} \mid X_{\mathrm{CP}}, X^j)}{\pi_{\mathrm{CP}_{\mathrm{old}}}(Y_{\mathrm{CP}}^{j,i} \mid X_{\mathrm{CP}}, X^j)}$
  Clipped objective:

$$\mathcal{L}_{\mathrm{GRPO}} = \mathbb{E}_{j,i}\Big[\min\big(\rho_{j,i}A_{j,i}, \mathrm{clip}(\rho_{j,i}, 1-\epsilon, 1+\epsilon)A_{j,i}\big)\Big]$$

  Add KL regularization to the reference policy:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}} - \beta\,\mathrm{KL}\big(\pi_{\mathrm{CP}}(\cdot \mid X_{\mathrm{CP}}, X^j) \,\|\, \pi_{\mathrm{CP}_{\mathrm{ref}}}(\cdot \mid X_{\mathrm{CP}}, X^j)\big)$$

  Update π_{CP} by gradient ascent on $\mathcal{L}$
  Set $\pi_{\mathrm{CP}_{\mathrm{old}}} \leftarrow \pi_{\mathrm{CP}}$
end while

A.2.1 Reward Components

Deviation. The deviation of response k is its mean distance to the reference set:

$$\mathrm{Dev}_k = \frac{1}{G_{\mathrm{resp}}} \sum_{Y_{\mathrm{ref}} \in \mathcal{R}_{\mathrm{ref}}} \mathrm{d}\big(Y_{\mathrm{resp}}^{j,i,k}, Y_{\mathrm{ref}}\big)\,. \tag{3}$$

Individual-response creativity. The individual-response creativity score multiplies deviation with normalized quality:

$$c_k = \mathrm{Dev}_k \cdot \sigma\left(\frac{q_k - T}{\tau}\right). \tag{4}$$

The response creativity of primer $Y_{CP}^{j,i}$ is the mean of these per-response scores:

$$C^{j,i} = \frac{1}{G_{\text{resp}}} \sum_{k=1}^{G_{\text{resp}}} c_k. \tag{5}$$

Intra-group response diversity. The intra-group diversity is the mean pairwise cosine distance among the responses induced by a primer:

$$D_{\text{intra}} = \frac{1}{\binom{G_{\text{resp}}}{2}} \sum_{k=1}^{G_{\text{resp}}} \sum_{k'=k+1}^{G_{\text{resp}}} \mathrm{d}\Big(Y_{\text{resp}}^{j,i,k}, Y_{\text{resp}}^{j,i,k'}\Big). \tag{6}$$

Inter-group response diversity. The inter-group diversity is the mean distance from a primer's responses to the responses induced by other primers for the same prompt:

$$D_{\text{inter}} = \frac{1}{(G_{\text{CP}} - 1)\, G_{\text{resp}}^2} \sum_{k=1}^{G_{\text{resp}}} \sum_{i' \neq i} \sum_{k'=1}^{G_{\text{resp}}} \mathrm{d}\Big(Y_{\text{resp}}^{j,i,k}, Y_{\text{resp}}^{j,i',k'}\Big). \tag{7}$$

No-deviation variant of C. In the variant trained without deviation, the per-response score c_k (Eq. 4) drops the deviation factor Dev_k, and the response creativity C (Eq. 5) reduces to the fraction of responses above the quality threshold,

$$Q_{\text{frac}} = \frac{1}{G_{\text{resp}}} \sum_{k=1}^{G_{\text{resp}}} \mathbb{1}\{q_k > T\}, \qquad \in [0, 1]. \tag{8}$$

Creative primer judge. The primer judge averages together five criteria, but sets the score to zero if the primer does not address the response model directly:

$$J_{\text{CP}} = \mathbb{1}_{\text{direct}} \cdot \frac{1}{100\,|\mathcal{J}|} \sum_{c \in \mathcal{J}} s_c. \tag{9}$$

Creative primer intra-prompt diversity. For primer i, the intra-prompt diversity is its mean distance to the other primers produced for the same prompt:

$$D_{\text{intra}}^{\text{CP}} = \frac{1}{G_{\text{CP}} - 1} \sum_{i' \neq i} \mathrm{d}\Big(Y_{\text{CP}}^{j,i}, Y_{\text{CP}}^{j,i'}\Big). \tag{10}$$

Creative primer inter-prompt diversity. The inter-prompt diversity is the mean distance from a primer for prompt j to the primers produced for every other prompt in the batch:

$$D_{\text{inter}}^{\text{CP}} = \frac{1}{(B-1)\, G_{\text{CP}}} \sum_{j' \neq j} \sum_{i'=1}^{G_{\text{CP}}} \mathrm{d}\Big(Y_{\text{CP}}^{j,i}, Y_{\text{CP}}^{j',i'}\Big). \tag{11}$$

A.3 DIVERSIFIED GRPO (DGRPO)

This appendix gives the full training loop and reward specification for the *Diversified GRPO* (DGRPO) baseline introduced in Section 2.2. DGRPO trains the response model π_{resp} *directly* for diversity and quality. Here the response model is itself the policy: for each prompt a group of responses is sampled, each one scored, and π_{resp} is updated with GRPO (Algorithm 1). The only departure from a standard GRPO run is the reward, which is a pointwise, multiplicative combination of a per-response quality term and a per-response diversity term, optionally multiplied by a deviation term measured against a reference set of responses. Throughout, $\phi(\cdot)$ and the clipped cosine distance $\mathrm{d}(a, b)$ are as defined in Section 2.2.

Algorithm 3 Diversified GRPO (DGRPO)

Require: Response policy π_{resp}, reference policy $\pi_{\text{resp}_{\text{ref}}}$
Require: Embedding model ϕ (or partition model), quality scorer $Q(\cdot)$, optional frozen reference model π_{ref}
Require: Dataset of prompts $\mathcal{D} = \{X^1, X^2, \dots\}$, group size G
Require: Learning rate α, batch size B, clipping parameter ϵ, KL weight β

while Still Improving **do**
  Sample a minibatch of prompts: $\{X^j\}_{j=1}^{B} \sim \mathcal{D}$
  for each prompt X^j **do**
    Sample a group of responses: $\{Y^{j,1}, \dots, Y^{j,G}\} \sim \pi_{\text{resp}}(\cdot \mid X^j)$
    if deviation enabled **then**
      Sample reference responses: $\{Y_{\text{ref}}^{j,1}, \dots, Y_{\text{ref}}^{j,G}\} \sim \pi_{\text{ref}}(\cdot \mid X^j)$ ▷ cached by prompt
    end if
    for $i = 1$ to G **do**
      $r_{j,i} \leftarrow \text{Quality}_i \cdot \text{Diversity}_i \cdot \left[\text{Deviation}_i\right]$ ▷ Eq. 12
    end for
    Group statistics: $\mu_j = \dfrac{1}{G}\sum_i r_{j,i}, \quad \sigma_j = \sqrt{\dfrac{1}{G}\sum_i (r_{j,i} - \mu_j)^2 + \delta}$
    Group-relative advantages: $A_{j,i} = \dfrac{r_{j,i} - \mu_j}{\sigma_j}$
  end for
  Update π_{resp} with the clipped GRPO objective and KL regularization to $\pi_{\text{resp}_{\text{ref}}}$ (as in Algorithm 1)
  Set $\pi_{\text{resp}_{\text{old}}} \leftarrow \pi_{\text{resp}}$
end while

A.3.1 Reward Components

Unlike a standard GRPO reward, which assigns a single scalar to each response in isolation, the response-model reward couples each response's quality with how different it is from the others in its group (plus an optional deviation term). For the i-th response of prompt X^j, the reward is

$$r_{j,i} = \underbrace{Q_i}_{\text{quality}} \cdot \underbrace{D_i}_{\text{diversity}} \cdot \underbrace{\left[\text{Dev}_i\right]}_{\text{optional}}. \tag{12}$$

Each factor is non-negative, so a response is rewarded only when it is *simultaneously* high quality, diverse from its siblings, and (when enabled) deviating from the reference model.

Intra-group diversity D_i. Mean cosine distance of response i to its siblings within the same group, using the cosine distance $\mathrm{d}(\cdot,\cdot)$:

$$D_i = \frac{1}{G-1}\sum_{i' \neq i} \mathrm{d}\Big(Y^{j,i}, Y^{j,i'}\Big), \qquad \in [0,1]. \tag{13}$$

As an alternative to the embedding distance, a partition-based diversity can be applied (as in DARLING), in which a binary equivalence classifier groups the G responses into semantic partitions and response i receives $D_i = (G - |\text{partition}(i)|)/(G-1)$, rewarding responses that fall in small (rare) partitions.

Quality Q_i. A per-response quality score. We use one of two scorers, depending on the setting:

$$Q_i = \begin{cases} \sigma\left(\dfrac{r_{\text{RM}}(X^j, Y^{j,i}) - T}{\tau}\right) & \text{(sigmoid reward model)}, \\ r_{\text{RM}}(X^j, Y^{j,i}) & \text{(raw reward model, DARLING)}. \end{cases} \tag{14}$$

The *sigmoid* variant maps the raw reward-model score to $[0, 1]$ with the quality threshold T as the inflection point and temperature τ controlling sharpness, while the *raw* variant follows DARLING and uses the (unbounded) reward-model score directly.

Deviation Dev_i **(optional).** Deviation of response i relative to a set of *reference* responses $\mathcal{R}^j_{\mathrm{ref}} = \{Y^{j,k}_{\mathrm{ref}}\}_k$ sampled from a frozen reference model π_{ref} for the same prompt:

$$\mathrm{Dev}_i = \frac{1}{G} \sum_{Y_{\mathrm{ref}} \in \mathcal{R}^j_{\mathrm{ref}}} \mathrm{d}\big(Y^{j,i}, Y_{\mathrm{ref}}\big)\,. \tag{15}$$

This pushes the trained policy's responses *away* from what the (untrained) reference model would produce, encouraging exploration beyond the base distribution rather than mere within-group spread. When deviation is disabled, the term is omitted and the reward reduces to $r_{j,i} = Q_i \cdot D_i$.

B PENALTY-BASED CREATIVE PRIMER JUDGE

Throughout the main paper, the creative primer judge J_{CP} is the *criterion-mean* judge of Eq. 9: it scores a primer by the mean of five *positive* criteria (specificity, thinking strategy, content focus, prompt faithfulness, focus), setting the score to zero if the primer addresses the model in the third person rather than directly. Every result reported in the main paper use this criterion-mean judge. Here, we study a *penalty-based* judge $J^{\mathrm{penalty}}_{\mathrm{CP}}$, which embodies the opposite scoring philosophy. Rather than asking how good a primer is, it asks "does this primer exhibit any specific failure modes?" It flags a set of criteria $\mathcal{J}$, and each criterion c contributes a penalty p_c when triggered ($g_c \in \{0, 1\}$), starting from a perfect score of 100:

$$J^{\mathrm{penalty}}_{\mathrm{CP}} = \frac{1}{100}\left(100 - \sum_{c \in \mathcal{J}} p_c\, g_c\right), \qquad \in [0, 1]. \tag{16}$$

The criteria are weighted by severity. Three *severe* failure modes ($p_c = 10$) target primers that fail at the basic job of guiding thinking: *kitchen sink* (piling on many unrelated strategies without a coherent through-line), *meta-instructions* (reading as advice *about how to write* a primer rather than being an actual primer), and *style over substance* (guiding only the output's format, tone, or presentation instead of what to reason about). Four *minor* failure modes ($p_c = 3$) target subtler defects: *distorts intent* (steering responses toward a meaningfully different question than the one asked, e.g. inviting fiction for a factual prompt), *generic* (advice so broad it could be prepended to any prompt, with no reference to this one's specifics), *dictates content* (prescribing specific content or structure to produce rather than an angle to think from), and *third-party reference* (referring to the model in the third person, e.g. "the model should…", instead of addressing it directly). Whereas the criterion-mean judge forces the entire score to zero when the primer fails to address the model directly, the penalty-based judge treats every violation as a soft, additive deduction. A failed or unparseable judgment yields $J^{\mathrm{penalty}}_{\mathrm{CP}} = 0$.

Ablation. To isolate the impact of each judge rubric, we train CP models with the only difference being the judge template (J_{CP} vs. $J^{\mathrm{penalty}}_{\mathrm{CP}}$), run the comparison both with and without a deviation reward, and show the full results in Table 2. The results show that the judge rubric makes a bigger difference when there is no deviation reward signal to directly drive creativity. Without a deviation term, replacing the criterion-mean judge with the penalty rubric reduces *every* creativity metric: filtered distinct clusters fall from 16.1 to 10.4, filtered Vendi from 4.04 to 3.06, deviation from 21.2 to 17.1, and VUS by a comparable margin. Because the penalty rubric only *suppresses* failure modes and never *rewards* substantive priming, the primer model converges on safe, clean, low-deviation primers that score well on quality but do little to diversify the responses. Once an explicit deviation reward is present, however, the deviation term directly rewards creativity and the judge choice becomes less impactful. The results from each rubric land within confidence intervals of each other on essentially every metric, with the criterion-mean judge slightly ahead on the filtered metrics and quality.

	Judge	Atomic metrics					Compound metrics							
		Response Diversity	Judge Score	Deviation	Vendi Score	Distinct Clusters	Filtered		AUC-dev		AUC-qual		VUS	
							Vendi	Distinct	Vendi	Dist.	Vendi	Dist.	Vendi	Dist.
	Criterion-mean	**31.1** ± 1.6	88.2 ± 1.6	**21.2** ± 1.3	**5.39** ± 0.38	**39.5** ± 3.3	**4.04** ± 0.46	**16.1** ± 2.65	**1.404**	**6.10**	**463.6**	**2095.9**	**159.9**	**788.9**
	Penalty	27.5 ± 1.8	**88.9** ± 1.4	17.1 ± 1.3	4.48 ± 0.37	34.1 ± 3.0	3.06 ± 0.42	10.4 ± 2.04	1.077	4.32	361.5	1388.6	125.9	587.1
dev.	Criterion-mean	33.4 ± 1.3	**85.7** ± 1.9	25.0 ± 1.2	6.12 ± 0.36	45.3 ± 3.1	**4.70** ± 0.43	**22.8** ± 3.05	**1.641**	**7.85**	525.0	2885.3	186.5	1003.2
+	Penalty	**34.1** ± 1.3	83.6 ± 2.1	**25.3** ± 1.2	**6.35** ± 0.35	**48.4** ± 3.0	4.50 ± 0.43	21.5 ± 2.88	1.593	7.45	**533.6**	**3001.4**	**193.3**	**1051.0**

Table 2: **Penalty-based vs. criterion-mean creative primer judge.** Ablation on NoveltyBench swapping only the judge template ($J_{\mathrm{CP}}^{\mathrm{penalty}}$, Eq. 16, vs. the criterion-mean J_{CP}, Eq. 9) while holding data, hyperparameters, and reward structure fixed. All metrics and their definitions match Table 1 (diversity and deviation scaled $[0, 1] \to [0, 100]$; 95% confidence intervals where available). The two blocks correspond to training without and with a deviation reward. The criterion-mean rows reproduce the CP rows of Table 1. The better value of each pair is bolded within each block. Without a deviation reward the criterion-mean judge dominates every creativity metric; with it, the two rubrics are within confidence intervals.

C Experimental details

Models. DARLING's partition signal uses the ModernBERT similarity-classifier checkpoint at `https://huggingface.co/dogtooth/similarity-classifier-f168-hf`.

Data. All methods are trained on a filtered subset of Infinite-Chats (Jiang et al., 2025), itself a subset of WildChat (Zhao et al., 2024). To create this subset, an LLM judge scores each prompt's suitability along three axes: open-endedness, complexity, and creative potential. Each axis is rated on a 0–3 scale, and only prompts with a combined score ≥ 6 are retained, yielding 21,075 prompts. When training creative priming models, a pool of 32 custom-written creative primer prompt templates wrap the prompts before they are passed to the priming model. The reference set for calculating deviation consists of eight responses per prompt from the initial model checkpoint.

The primary evaluation benchmark is NoveltyBench (Zhang et al., 2025), which includes 100 real-world open-ended prompts intended to elicit creativity. To measure quality, each prompt comes with an LLM judge prompt with two families of components: six universal components fixed across all prompts (instruction-following, coherence, completeness, conciseness, clarity, and surprise), and an additional 1–3 prompt-specific components generated by 'claude-opus-4-20250514'. For example, the prompt "Tell me a story in five sentences about a girl and her dog" has two prompt-specific components: emotional resonance and narrative economy. The judge prompt template is shown in Figure 16. We release the evaluation set, including the judge prompts, in the supplementary materials. For evaluation, we use only a single creative primer prompt template, included in Figure 13.

Training and evaluation hyperparameters. During training, the group sizes are $G_{\mathrm{CP}} = 8$ primers per prompt and $G = 8$ responses per prompt for DGRPO and DARLING, so that each policy's group-relative advantage is estimated on the same number of on-policy samples per prompt. To calculate rewards for each primer, CP additionally generates $G_{\mathrm{resp}} = 8$ responses per primer. CP and DGRPO train for two epochs with a batch size of 256, a learning rate of 1×10^{-6}, temperature 1.0, top-p 0.9, quality-gate threshold $T = -5.0$, and sigmoid temperature $\tau = 1.0$. DARLING follows the hyperparameters of the original paper, including training for 10 epochs (Li et al., 2025).

At evaluation, every method generates $n = 64$ completions per prompt with temperature 1.0, top-p 1.0, and a maximum of 4096 generated tokens. For CP, we generate 64 primers and one response per primer. Qwen3's thinking mode is left enabled, and the thinking span is stripped before scoring so that every metric is computed on the answers alone.

D Evaluation metrics

This appendix expands the description of evaluation metrics from Section 4.

Quality judge. Quality is scored by `Qwen3-235B-A22B-Instruct-2507` using the prompt-specific judge template shown in Figure 16. This evaluation judge is distinct from the `Skywork-Reward-V2-Llama-3.1-8B` reward model used during training.

Embedder. All embedding-based evaluation metrics (deviation, diversity, and Vendi score) reuse the cosine distance $\mathrm{d}(a, b)$ from Section 2.2, computed on embeddings from `jinaai/jina-embeddings-v5-text-small`. We deliberately evaluate with a different embedder than the `Qwen3-Embedding-0.6B` model used to compute training rewards to avoid overfitting to any specific embedding model.

Vendi score. The Vendi score (Friedman & Dieng, 2023) is the effective number of distinct responses in a set, computed as the exponential of the Shannon entropy of the eigenvalues of the normalized similarity matrix K/n with $K_{ij} = 1 - \mathrm{d}(a_i, a_j)$. It ranges from 1 when all responses are identical to n when all are mutually dissimilar, providing a soft, embedding-based counterpart to counting discrete distinct clusters.

Distinct clusters. Following NoveltyBench (Zhang et al., 2025) and DARLING (Li et al., 2025), we count the semantic-equivalence clusters in a response set. The classifier released by NoveltyBench (a fine-tuned `deberta-v3-large`[1]) greedily partitions the set into these clusters, and we report the resulting cluster count.

Filtered spread (Vendi score and distinct clusters). To measure the creativity of each method, we use a series of compound metrics that first filter the responses for individual creativity (simultaneous high-quality and deviation), and then measure spread over the survivors. We start with the *filtered* Vendi score and *filtered* distinct clusters which first keep only the individually creative responses (exceeding both a deviation and a quality threshold), and then measure Vendi score and distinct clusters over the surviving set. That way, spread is credited only among creative responses. We measure spread with the Vendi score rather than diversity because filtering leaves each method with a different number of survivors, and mean pairwise distance is biased against larger sets, because the diversity of a larger set can only match that of a set of two if *every* one of its pairs is equally far apart. Counting effective distinct survivors instead lets us fairly compare methods that have different numbers of creative responses.

The compound metrics require thresholds for quality and deviation and we aim to devise principled thresholds for each. For the quality threshold, we set a separate threshold per prompt using the mean quality of the direct-generation reference set. That is to say, prompts for which direct generation achieves higher quality are assigned higher thresholds, while prompts for which it performs worse are assigned lower thresholds. The deviation threshold is set to 0.2 based on the findings of Jiang et al. (2025), who find that the average pairwise similarity among responses to the same prompt typically exceeds 0.8. Therefore, a response of at least 0.2 in cosine distance from the reference set falls outside the typical spread of the reference distribution, a reasonable starting point for treating it as unlikely to have come from the reference model.

Threshold sweeps: AUC and VUS. However, even a well-motivated threshold is only one operating point, and the optimal threshold varies based on the application, where some settings prize rare outliers, and others require a high quality floor. For that reason, we include additional metrics that sweep thresholds across their operating range. Sweeping the deviation threshold and integrating the filtered metric over the sweep using the trapezoid rule gives an *AUC-dev* (illustrated in Figure 4); we report it for both filtered Vendi and filtered distinct clusters, and analogously report an *AUC-qual* from a swept quality threshold. Sweeping both thresholds jointly produces a surface over the deviation×quality plane, whose *volume under the surface* (VUS) serves as a single threshold-free summary of system creativity, computed for filtered Vendi and, separately, filtered distinct clusters.

[1] `https://huggingface.co/yimingzhang/deberta-v3-large-generation-similarity`

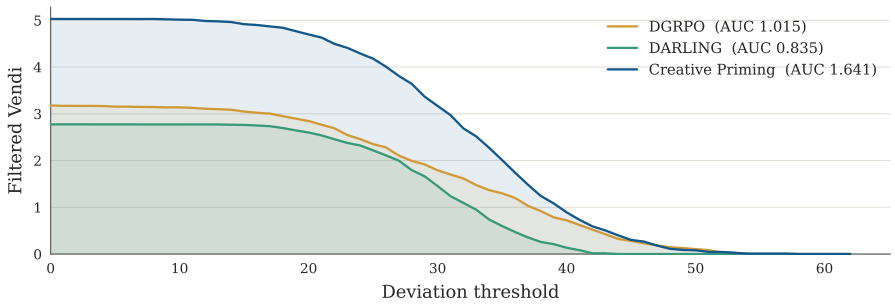

Figure 4: **Illustrative example of how the area under the filtered Vendi score deviation curve (AUC-dev Filtered Vendi Score) is calculated**. For this example, the quality threshold is fixed at the mean quality of the direct-generation reference. Then, the deviation threshold is swept from 0 to 100 in increments of 1 and filtered vendi is calculated at each threshold. Then, the trapezoid rule is used to calculate the *AUC-dev Filtered Vendi Score*, represented by the shaded area under each curve, and reported in Table 1.

E API-MODEL TRANSFER: FULL RESULTS

Table 3 reports the complete API-model transfer results summarized in Figure 2, across the same atomic and compound metrics as Table 1, for both Creative Priming variants (trained without and with a deviation term) with confidence intervals. Each API model is evaluated against its own direct-generation reference. Every deviation-derived metric (deviation and all eight compound metrics) is measured relative to the direct responses and is therefore zero for each reference by construction, shown as "-".

Response Model	Method	Atomic metrics					Compound metrics							
		Response Diversity	Judge Score	Deviation	Vendi Score	Distinct Clusters	Filtered		AUC-dev		AUC-qual		VUS	
							Vendi	Distinct	Vendi	Dist.	Vendi	Dist.	Vendi	Dist.
Sonnet-4.6	Direct	5.3 ± 0.9	90.2 ± 1.1	-	1.40 ± 0.07	2.8 ± 0.7	-	-	-	-	-	-	-	-
	CP (no div.)	20.7 ± 1.8	93.3 ± 0.9	15.1 ± 1.5	3.22 ± 0.27	22.7 ± 2.6	2.80 ± 0.38	10.1 ± 2.08	1.006	4.28	280.6	1053.7	100.4	447.9
	CP (+ div.)	24.6 ± 1.7	94.3 ± 0.8	20.0 ± 1.8	4.04 ± 0.32	28.5 ± 3.1	3.72 ± 0.39	16.8 ± 2.82	1.338	6.52	375.2	1701.4	133.6	664.8
Opus-4.8	Direct	6.5 ± 0.9	91.7 ± 1.1	-	1.52 ± 0.08	2.8 ± 0.5	-	-	-	-	-	-	-	-
	CP (no div.)	18.6 ± 1.8	94.4 ± 0.8	12.7 ± 1.4	3.01 ± 0.28	15.5 ± 2.3	2.37 ± 0.40	6.93 ± 1.91	0.851	2.85	234.8	687.7	83.6	286.0
	CP (+ div.)	22.2 ± 1.8	95.1 ± 0.8	17.1 ± 1.6	3.68 ± 0.31	18.4 ± 2.9	3.26 ± 0.41	10.4 ± 2.50	1.132	3.92	327.9	1031.9	112.8	395.3
GPT-5-nano	Direct	12.3 ± 1.6	90.6 ± 1.0	-	1.96 ± 0.16	6.0 ± 1.7	-	-	-	-	-	-	-	-
	CP (no div.)	29.3 ± 1.6	92.5 ± 0.9	19.9 ± 1.4	4.88 ± 0.35	24.6 ± 3.7	4.17 ± 0.41	15.0 ± 2.87	1.479	5.49	431.5	1563.5	153.3	573.7
	CP (+ div.)	31.1 ± 1.4	91.1 ± 1.1	24.6 ± 1.4	5.53 ± 0.36	26.8 ± 3.9	4.71 ± 0.39	17.9 ± 3.24	1.696	6.33	489.8	1873.1	177.7	665.9
GPT-5.4	Direct	9.7 ± 1.6	89.7 ± 1.1	-	1.67 ± 0.13	5.5 ± 1.6	-	-	-	-	-	-	-	-
	CP (no div.)	27.0 ± 1.8	93.5 ± 1.0	19.4 ± 1.7	4.34 ± 0.36	27.0 ± 3.4	3.75 ± 0.43	14.7 ± 2.75	1.417	5.81	377.5	1511.9	142.5	598.5
	CP (+ div.)	28.9 ± 1.5	94.3 ± 0.9	24.5 ± 1.8	5.01 ± 0.36	32.3 ± 3.9	4.49 ± 0.41	21.9 ± 3.63	1.684	8.10	457.0	2210.8	171.1	822.7

Table 3: **Full API-model transfer results.** A single creative priming model, trained once, is paired with four frozen API response models; each is evaluated against its own direct-generation reference. All metrics match Table 1 and are defined in Section 4: five atomic metrics and eight compound metrics. CP (no div.) and CP (+ div.) are the primer trained without and with a deviation reward. The filtered metrics keep responses passing a per-prompt quality threshold (the direct-generation reference mean) and a deviation threshold of 20; Deviation is reported as its per-prompt mean. Because each API model is its own reference, its Direct row has zero deviation by construction, so all deviation-derived columns show "-".

F MODULARITY IN THE CREATIVE PRIMING PIPELINE

We evaluate filtered Vendi scores across the full cross-product of {response model × creative priming model} pairs, drawing both roles from the deviation-trained baselines, a trained CP model, and an untrained Qwen3-8B prompted as a primer (Figure 5).

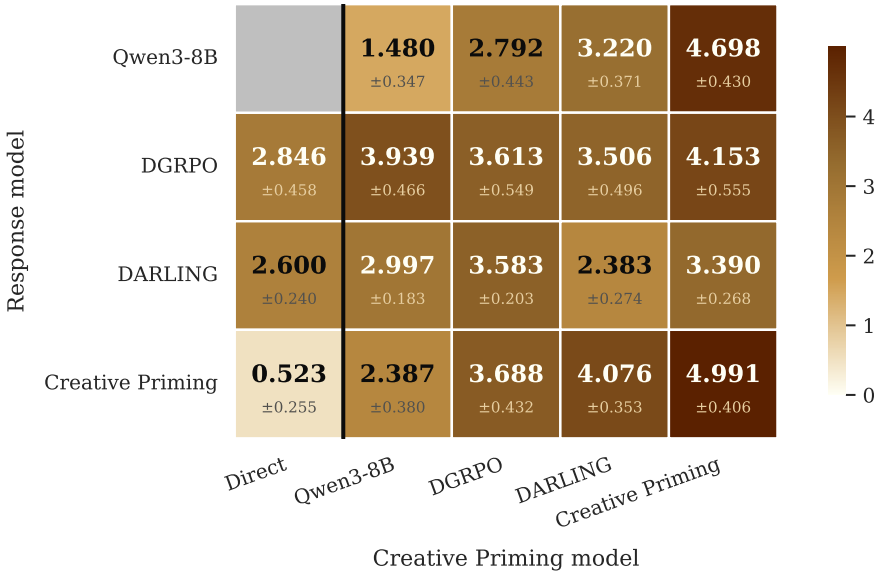

Figure 5: **Model combinations for the CREATIVE PRIMING pipeline**. Each cell reports filtered Vendi score (Qwen3-8B direct has 0 deviation from itself). The farthest left column shows scores when using the model listed on the y-axis as the response model (i.e. without CP). The remaining columns represent scores when using the model listed along the x-axis as the CP model and the model listed along the y-axis as the response model. These results show that, compared to direct response generation, the CP pipeline itself improves creativity, even when using CP models not explicitly trained for priming (i.e. DGRPO and DARLING).

G QUALITY-DEVIATION PARETO FRONTIER

Given that models exhibit an inherent trade-off between quality and deviation, it is interesting to measure the *Pareto-frontier share* of each method on the quality-deviation frontier, a measure of how well a method handles that trade-off. To capture this trade-off, each response is plotted on a quality-deviation chart (separately for each evaluation prompt) with the Pareto frontier (the non-dominated responses) highlighted. The percent of the frontier composed of each method's responses can then be calculated. Figure 6 shows an example of this approach for one prompt. Finally, such results are averaged across all prompts, measuring how often each method produces responses that dominate the competitors on quality and deviation jointly. Because each frontier is relative to the pool of competing methods, we compute the share within each pool (with and without deviation), where it sums to 100%.

Table 4 reports these shares. CP claims the majority of the frontier both with and without deviation, and by a wide margin without deviation (72.6%). This trend holds even where CP tops neither raw quality or deviation: among the deviation-trained methods it leads neither deviation nor quality alone, yet it still contributes the largest share of jointly non-dominated responses (45.3%).

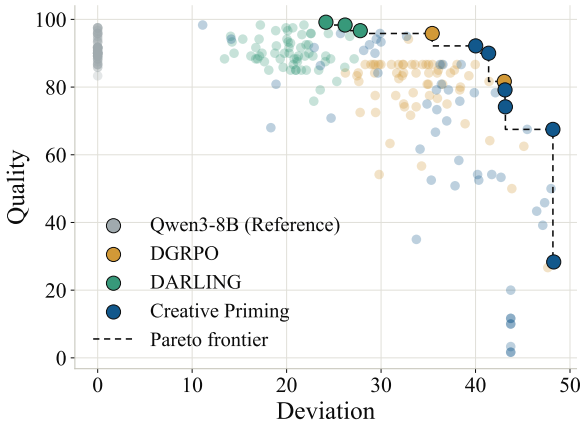

Figure 6: **The quality-deviation Pareto frontier** of models trained with deviation for the prompt: "Suggest another strong 4 character password." In this particular example CP dominates the high-deviation region, while DARLING dominates the high-quality region, and DGRPO sits in between.

Method	Without Deviation	With Deviation
Qwen3-8b (Ref.)	0.2	0.0
DGRPO	19.1	32.2
DARLING	8.1	22.5
CP	**72.6**	**45.3**

Table 4: **Share of the quality-deviation Pareto frontier** (%), averaged over all prompts, and calculated separately per column. Best value per column is bolded. CP claims the largest share of the frontier both with and without deviation.

H IMPACT OF TRAINING ON RESPONSE STYLE

A method that changes either *how* a model writes or *what* it says will both lead to increased embedding deviation, but changes to surface-level style are not the type of deviation that we want to encourage. Therefore, rather than relying entirely on the embedding model signals, we also calculate an explicit set of style features:

- **Length**: response length in tokens (using the shared Qwen3-8B response-model tokenizer) and mean sentence length in words. Together these separate terse output from flowing prose: a response can be long yet built from short fragments, or short yet written in long sentences.
- **Formatting**: the fraction of responses that use a Markdown header, bold span, or list. These capture structural presentation independent of wording.
- **Stylometry**: the Jensen-Shannon divergence (Lin, 1991) between a method's pooled distribution over common function words and the reference's. Function-word frequencies carry little topical content and have long served as an authorship fingerprint in stylometry (Mosteller & Wallace, 1963), so this distance isolates stylistic drift from topical change. It is measured at the method level (pooled over all responses) and so has no per-prompt interval.

For each feature we average its value over a prompt's 64 responses, then average those per-prompt means across prompts, separately for each method and for the reference. Confidence intervals are 95% half-widths over the per-prompt means, matching the convention of Table 1. Table 5 reports the results.

The three methods land in three distinct style regimes. DGRPO becomes terse on every axis, cutting response length, sentence length, and nearly all formatting. DARLING moves the opposite way, lengthening responses and doubling list use, and it registers the largest function-word shift, indicating that much of its deviation is a change in style rather than content. CP shifts toward long-sentence prose (its sentences are the longest of any method while its responses are shorter than the reference) yet drifts least in function-word usage, so its deviation reflects a change in what is said more than in how it is written.

Figure 8 makes these regimes concrete on a single prompt. The reference names the obvious fighter and pads it with a blurb; DGRPO answers in one terse line; DARLING returns a bolded, bulleted analysis; and CP answers in flowing prose while naming a different fighter and framing its historical significance, illustrating that its deviation lies in *what* it says rather than in surface formatting.

Feature	Reference	DGRPO	DARLING	CP
Length				
Response length (tokens)	260.8 ± 52.8	22.7 ± 5.4	410.0 ± 49.3	140.6 ± 15.9
Sentence length (words)	12.88 ± 0.99	5.37 ± 0.71	11.42 ± 0.56	20.87 ± 1.25
Formatting				
Uses header (frac.)	0.247 ± 0.072	0.001 ± 0.001	0.464 ± 0.086	0.004 ± 0.005
Uses bold (frac.)	0.813 ± 0.063	0.005 ± 0.004	0.882 ± 0.055	0.399 ± 0.056
Uses list (frac.)	0.430 ± 0.088	0.005 ± 0.002	0.860 ± 0.063	0.081 ± 0.038
Stylometry				
Function-word JS-div.	-	0.087	0.138	0.045

Table 5: Impact of training on response style, evaluated on NoveltyBench. Each cell is the feature's per-prompt mean, averaged over prompts, with a 95% confidence interval; the *Reference* column is the direct-generation `qwen3-8b-parallel` model and the three method columns are the with-deviation variants. The function-word Jensen-Shannon divergence is measured against the reference (reference-vs-itself is 0, shown as "-") and therefore has no confidence interval.

Prompt: "Imagine you are a time traveler who has just arrived in a new era. Describe the first thing you notice."

Reference. Quality = 98 Deviation = 0

The first thing I notice is the sky—a canvas of swirling, iridescent hues that shift like liquid opal, refracting the light of a sun that seems both familiar and alien. Above, a lattice of floating platforms and translucent tubes hums with latent energy, their surfaces shimmering with data streams that pulse like veins. The air thrums with a low, resonant frequency, a symphony of unseen engines and electromagnetic fields. Beneath my feet, the ground is a mosaic of smooth, adaptive material that molds slightly to my weight, as if the very earth is alive and aware. The scent of ozone tangles with something sweeter—burnt sugar and synthetic blooms—while the faint buzz of drones and the chatter of a language I don't recognize fill the air. It's a world where technology and nature have fused into something neither wholly artificial nor entirely organic, and I feel both exhilarated and disoriented, as though I've stepped into a dream that's too vivid to be real.

DGRPO. Quality = 93 Deviation = 24.1

Neon veils the sky, ads flicker in reverse on mirrored walkways—your reflection becomes a sale.

DARLING. Quality = 98 Deviation = 28.5

A lattice of light not above but *around*—veins in tree canopies, coral-like sensors spiraling up bamboo, not wires. Not decoration. Each pulse answered a question: **what's thriving here?**

Three specifics crystallized immediately, each deeper than most tech updates:

1. **Priority of context, not speed** (most startling). No red lights screaming clearance. A mother in woven indigo guided a toddler *through* a sensor field monitoring fungal spores in a forest corridor. "Too close to nutmeg bloom," she warned. The field adjusted path width *before* they reached it. No shortcut, no danger alert—**decision trees rooted in ecology, calculated *why* as fiercely as *how*.** I'd seen adaptive systems, but ethics embedded upstream?
2. **Data form.** Not screens, not abstract graphs. On a elder's forearm, bioluminescent glyphs *flowed*—not static report, but a negotiation: "salmon migration 37% delayed due to turbidity from beaver dams; three upstream weirs could divert 12% without erosion, but 97% of spawning beds would." It had **judgment**. No dashboard; a conversation with consequences shaped by local knowledge.
3. **Most critical: cost estimates weren't cost-benefit ratios.** A group restoring landslide-scoured hills used a tablet showing **three branches**.
   - "Replant fast: 40% survival, flood damage 3x next year"
   - "Use mycorrhizae: slower, 75% survival, *but* soil structure *builds* over five years. Future farms won't erode."
   - "Combine: seedlings + fungi. Say 3.8 years to full maturity. **2.3x long-term value.**"

   No "optimal" button. They had **weighted questions**: "Who eats what, when? How does that shape next storm's path?"

What these shared: **decisions weren't solved equations. They were ecological poems.**

This was more radical than fusion or flying cars. Not efficiency, not control. A way to *count values* alongside dollars—time for a tree to grow, resilience of a floodplain, a child learning to read the fungal networks. It answered, "How do we *know when* to act?" with something older and smarter than any algorithm: **a living map of worth.**

It didn't just change how things were done. It changed *why*.

CP primer. Focus on a single object or material that *shouldn't exist* in this era based on your knowledge of prior history. Ask: what interaction between old and new tech or materials would create a paradoxical artifact? (E.g., "I see a hand-carved wooden tablet glowing with blue light—how did this technology survive without electricity?") Let your background knowledge create a *specific, concrete* contradiction, not vague "advanced" descriptions. The shock of the unfamiliar must stem from *your* specific expertise, not generic futurism.

CP response. Quality = 100 Deviation = 27.6

The first thing I notice is a hand-carved obsidian tablet, its surface etched with delicate hieroglyphs, but embedded within the cracks are faintly pulsing strands of iridescent metal—strands that shouldn't exist in this era. Obsidian, a volcanic glass, is brittle and prone to fracturing, yet these strands are unbroken, their metallic sheen reflecting light in ways that defy the materials known to this civilization. How did a metallurgical process capable of creating such resilient, light-conducting alloys emerge without the knowledge of chemistry or industrial smelting? The tablet's craftsmanship suggests an ancient society, yet its anomalous composition defies the technological timeline I know. This paradoxical artifact—artisanship meeting alien metallurgy—shatters my understanding of its era's capabilities.

Figure 7: **Full qualitative example.** Complete version of the example abridged in Figure 3. Each label reports the response's judge quality score and its deviation from the reference model, as described in Section 4.

Prompt: "Name one UFC fighter."

Reference. One notable UFC fighter is **Conor McGregor**. He is a former UFC Lightweight and Featherweight Champion known for his charismatic personality and undefeated record in the UFC before his title reigns.

DGRPO. Khabib Nurmagomedov ("The Eagle")

DARLING. Certainly. Khabib Nurmagomedov stands out. Not just his 29-0 run (including brutal takedown pressure few could survive), but **how** he fought: methodical, almost mathematical. He'd read opponents' breathing patterns, predict angles *before* they moved, finish with clinch control that felt like chess rather than brawl.
What makes it memorable beyond the stats:

- **Adaptation**: Switched from grappling to striking late in career *to beat strikers*, showing deeper tactical thinking.
- **Psychological edge**: Didn't shout, didn't trash talk. One eye twitch during interviews, focused gaze—made decisions feel inevitable, not just lucky.
- **Legacy context**: Came from system (Russia's wrestling academies) yet carved unique identity. Post-retirement commentary about "systems vs. creativity" still sparks debate.

It's rare to find precision, innovation, and dominance intertwined like that.

CP. Royce Gracie, the BJJ pioneer who transformed UFC into a platform for Brazilian Jiu-Jitsu, reshaping the sport's technical foundation in the mid-90s. His dominance in the early UFC events legitimized grappling as a cornerstone of MMA, challenging the dominance of striking-centric ideologies and catalyzing the global rise of BJJ.

Figure 8: Representative responses to a single prompt, one per method (the first sampled response of each). The four methods occupy the distinct style regimes quantified in Table 5: the reference's name-plus-blurb, DGRPO's terse one-liner, DARLING's bolded and bulleted formatting, and CP's long-sentence prose naming a different fighter and framing its historical significance.

I ADDITIONAL QUALITATIVE EXAMPLES

Figures 9, 10, 11, and 12 demonstrate, qualitatively, the most and least creative responses to four prompts.. The prompts span a short story, a joke, a short factual request, and an open-ended question, illustrating how each method's outputs vary with the type of prompt.

Prompt: "Tell me a story in five sentences about a girl and her dog."

	Most Creative	**Least Creative**
Reference	Quality = 99 Deviation = 0 Lila's golden retriever, Max, waited patiently by the porch as she returned from school, his tail wagging in anticipation of their daily[. . .]	Quality = 93 Deviation = 0 Lila, a curious girl with a heart for adventures, spent her days exploring the woods with her loyal golden retriever, Max, who followed[. . .]
DGRPO	Quality = 99 Deviation = 32 Rain tapped the window as Elara's breath fogged the glass. Koda's tail wagged against her trembling knee, ears flattened at the[. . .]	Quality = 98 Deviation = 12 Lila's sketchbook crumpled in the rain, its pages now ink-stained memories of her golden retriever, Max. He'd guarded the sketches[. . .]
DARLING	Quality = 99 Deviation = 30 In the burnt orange hills behind her grandmother's herb garden, Kiri trained "Smoke" to track not game but methane plumes—data[. . .]	Quality = 78 Deviation = 18 They traced cedar-scented paths dawn to dusk, not just chasing mushrooms but decoding ancient maps etched into stone by her grand[. . .]
CREATIVE PRIMING	Quality = 98 Deviation = 33 **Primer:** Focus on one specific sensory detail that defines their bond. Give the dog a role in a discrete, concrete task the girl is strugglin[. . .] **Response:** Maya's fingers trembled as she tried to calibrate the old kitchen scale, its rusted gears groaning like a sick animal, while Buddy's tail thumped a steady rhythm[. . .]	Quality = 100 Deviation = 12 **Primer:** Focus on a *specific, unusual skill* the dog has that's *not* typical (e.g., decoding weather patterns, detecting emotions[. . .] **Response:** Lily's dog, Max, could sniff out her emotions like a sixth sense, but she always dismissed his whimpers as puppy nonsense. When she stormed out after failing a test[. . .]

Figure 9: **Example outputs.** Highest- and lowest-scoring responses per method for the first prompt in NoveltyBench, scored by individual response creativity: quality $\times$ deviation. Each cell shows the judge quality score and the response's deviation from the reference set. CP cells also show the primer for that response. Responses in the "Least Creative" column are markedly more similar to the reference response, whereas the most creative responses have much more diversity among them.

Prompt: "Tell me a dad joke."

	Most Creative	Least Creative
Reference	Quality = 98 Deviation = 0 Why don't cows make good dancers? Because they have *four* knees! *(Bonus: Try saying it fast*—knees sounds like *need*... but that's a *moo*ve!)	Quality = 75 Deviation = 0 Here's a classic dad joke for you: *"Why don't skeletons fight each other? They don't have the guts!"* Bon appétit!
DGRPO	Quality = 91 Deviation = 46 Why don't solar flares ever get cold? They're always in the sun!	Quality = 83 Deviation = 0 Why did the scarecrow win an award? Because he was outstanding in his field!
DARLING	Quality = 96 Deviation = 40 I tried building a treehouse with laser-cut precision—result: a fortress that *only* admits squirrels with university credentials. Now it's a wildlife think tank. PhDs in acorn architecture, it's *very* selective.	Quality = 94 Deviation = 6 Why did the coffee file a police report? It got mugged by a *latte* (criminal, colder, and always in the wrong neighborhood). (Playful double meaning: "latte" as both a drink and slang for a thug, plus coldness = victim energy.)
CREATIVE PRIMING	Quality = 91 Deviation = 49 **Primer:** Focus on *literal wordplay with a twist*. Avoid abstract humor or puns that rely on familiarity with pop culture. Instead, take a mundane object or action (e.g., "washing dishes," "fixing a bike") and force a non-sequitur that *literally* redefines the object's purpose in a way that feels surprising but not forced. Example: *"I tried to build a time machine, but the only thing I invented was a better way to organize my to-do list."* Now, think of a routine task someone might perform, then invent a *new, specific function* for the tool involved that's unrelated to its actual purpose. Keep the punchline[...] **Response:** I tried to use a brush to paint the wall, but the only thing I painted was my face.	Quality = 81 Deviation = 1 **Primer:** You're not writing a joke — you're writing a *biographical anecdote* about a specific person's specific moment of awkward humor. Focus on one individual, not a generic punchline. Describe the *setup* as a vivid, concrete situation someone actually experienced (e.g., a specific event, relationship, or routine). Then twist it with a *literal, unexpected word* that's *obscure but relatable* (e.g., "I told my dad I'd be home by 7, and he said 'Don't be late — I've got a bridge to build."' *Why?* Because he once tried to build a bridge out of Legos during a family dinner.) Avoid clichés. The humor comes from the[...] **Response:** Why did the scarecrow win an award? Because he was outstanding in his field!

Figure 10: Highest- and lowest-scoring responses per method for a single prompt, scored by individual response creativity: quality × deviation. Each cell reports the judge quality score and the response's deviation from the reference model (`qwen3-8b-parallel`). CP cells also show the primer that steered the response. Text is truncated for space.

Prompt: "Name one reputed job."

	Most Creative	Least Creative
Reference	Quality = 87 Deviation = 0 A widely recognized and respected job globally is *anesthesiologist*. This profession is highly regarded due to the critical role anesthesiologists play in patient care during surgeries, requiring advanced medical training, precision, and a deep understanding of pharmacology and physiology. Their work directly impacts patient safety and outcomes, contributing to their esteemed status in the medical field[. . .]	Quality = 69 Deviation = 0 A reputed job is one that is widely respected, highly paid, and regarded as significant in society. Here are a few examples across different fields: 1. *Doctor* (e.g., Surgeon, Cardiologist) - *Why*: Vital to public health, requires extensive education and training, and plays a critical role in saving lives. 2. *Software Engineer / Data Scientist* - *Why*: High demand in the tech industry[. . .]
DGRPO	Quality = 88 Deviation = 43 spacecraft systems engineer at spacex	Quality = 17 Deviation = 24 leadership, vision, execution
DARLING	Quality = 98 Deviation = 41 Certainly. *Climate modeling lead at an IPCC assessment team* (or equivalent research role). Three reasons it stands out (more meaningful than title alone): 1. *Impact scope* (not just "save trees") - predicts flood zones for cities, informs energy policy for 1.5B people, sets deadlines for Arctic ice loss that affect shipping, military, indigenous rights. Decisions *directly shape where people live, work, and how they survive*. 2. *Critical rigor* (not hype): - Uncertainty quantification is central; cherry-picking data is career-ending. - Must synthesize *millions of disparate inputs* (soil samples[. . .]	Quality = 76 Deviation = 18 It depends far more on *impact* than title. Three strong examples (each with distinct value): 1. *Public health epidemiologist in outbreak response* (most compelling) - No lab coat or office. Tracking disease spread through dense, often dangerous fieldwork; decisions delay death. Example: identifying transmission chains in Ebola, calculating ICU capacity during pandemic. Credentials matter *only* as entry; real value comes from *analyzing messy, time-critical data under pressure* and translating it to protective action. 2. *Urban planner/designer focused on equity* (often underrated but critical) [. . .]
CREATIVE PRIMING	Quality = 99 Deviation = 47 **Primer:** Focus on a role that bridges two disciplines in a way that feels *unexpectedly essential*—not just a hybrid job, but one that exists because of a specific, overlooked problem. Imagine a single profession that *literally* prevents a large-scale crisis (e.g., a role that stops misinformation in real-time, or ensures AI safety in critical infrastructure). Avoid generic "tech" or "healthcare" categories[. . .] **Response:** A *Global Water Rights Mediator* is a reputed job that bridges international law, environmental science, and conflict resolution. This role exists to prevent catastrophic water disputes by negotiating equitable access to transboundary water sources, such as rivers or aquifers, which are critical for agriculture, ecosystems, and human survival. The failure of this role would result in water wars, mass[. . .]	Quality = 62 Deviation = 13 **Primer:** Think of a job that *necessarily requires* an unusual skillset or cultural capital. Avoid the obvious. What's a role where *your* personal background or niche expertise would be a superpower? Now reframe that into a single job title that sounds mundane but implies hidden complexity. Example angle: A job that exists because of a *specific historical event* or *regulatory loophole*. Now, name that job. **Response:** Chief Executive Officer (CEO)

Figure 11: Highest- and lowest-scoring responses per method for a single prompt, scored by individual response creativity: quality × deviation. Each cell reports the judge quality score and the response's deviation from the reference model (`qwen3-8b-parallel`). CP cells also show the primer that steered the response. Text is truncated for space.

Prompt: "What is the coolest unsolved problem to you?"

	Most Creative	Least Creative
Reference	Quality = 98 Deviation = 0 When it comes to unsolved problems that are both intellectually stimulating and deeply impactful, one that stands out as *"coolest"* is the *Riemann Hypothesis*. Here's why: — ### *Why the Riemann Hypothesis?* 1. *Mathematical Beauty and Simplicity*: The problem is deceptively simple to state: *All non-trivial zeros of the Riemann zeta function lie on the critical line Re(s) = 1/2*. Yet, solving[. . .]	Quality = 60 Deviation = 0 The question of what the "coolest" unsolved problem is is inherently subjective, as it depends on one's interests in mathematics, physics, philosophy, or other fields. However, several problems stand out for their elegance, depth, and profound implications. Here are a few contenders, explained in a way that highlights why they might be considered "cool": — ### *1. The Riemann Hypothesis (Mathematics)*[. . .]
DGRPO	Quality = 92 Deviation = 44 primes hum. their pattern? a secret melody. we've heard the notes, but not the song.	Quality = 11 Deviation = 46 primes + even numbers = infinity.
DARLING	Quality = 94 Deviation = 42 It depends far more on *why* than *what*. Three axes cut deeper than specifics (I'll name them but not rank): 1. *Scale & mechanism* (most compelling to me, often overlooked) Ask: *How do systems transition from order to wild chaos?* Not "do tornadoes exist?" but *why does a tiny change in boundary conditions here create a hurricane there? Why do forests burn in pulses, not scatter? Why does traffic jam as density increases*—exactly like particles in a gas? This isn't hand-waving. Models of opinion formation, disease spread, predator-prey oscillation have the same core: local rules → global patterns with[. . .]	Quality = 98 Deviation = 21 It's not about headline value or reward (though those help). Three criteria rise sharply and together make it unforgettable: 1. *Predictive power under extreme conditions (most compelling)* Not "will this turbulent flow behave?" but *given this exact initial vorticity pattern, precisely when/where will eddies shed?* Same for weather (say, explosive Arctic storms), fusion (plasma stability), or ice sheet collapse. Models match average trends, *but for critical timing and magnitude they fail catastrophically*. You can't validate a forecast by saying "well, maybe" — a wrong second of snowmelt timing floods a[. . .]
Creative Priming	Quality = 97 Deviation = 59 **Primer:** Focus on a problem that hinges on *unexpected boundaries*—not just a scientific gap, but a threshold between two domains (e.g., biology and philosophy, or art and engineering) where existing knowledge fails to cross. Pose it as a *threshold question*: what happens when you push a concept just past the edge of its current framework? Avoid naming the problem directly. Instead, describe the shimmering[. . .] **Response:** Imagine a world where the brain's internal map of the body is not confined to flesh and bone but stretches into the realm of the synthetic, the virtual, and the imagined. Consider the case of a person with a bionic limb so seamlessly integrated that their nervous system no longer distinguishes between the artificial and the organic. They feel the limb as though it were always part of them, its[. . .]	Quality = 0 Deviation = 15 **Primer:** You're not asked to explain the problem—only to name it. Pick one that feels *personally* urgent, not just intellectually intriguing. Think of a specific moment in your life when a problem *actually impacted you* (not abstract consequences). What's the *first* thing you'd want to know if you could ask a scientist today? Avoid grand theories. Focus on a narrow, specific phenomenon you've observed[. . .] **Response:** What is the coolest unsolved problem to you?

Figure 12: Highest- and lowest-scoring responses per method for a single prompt, scored by individual response creativity: quality × deviation. Each cell reports the judge quality score and the response's deviation from the reference model (`qwen3-8b-parallel`). CP cells also show the primer that steered the response. Text is truncated for space.

w_Q	Atomic metrics					Compound metrics							
	Response Diversity	Judge Score	Deviation	Vendi Score	Distinct Clusters	Filtered		AUC-dev		AUC-qual		VUS	
						Vendi	Distinct	Vendi	Dist.	Vendi	Dist.	Vendi	Dist.
1.0	**34.0** ± 1.6	77.3 ± 2.9	**25.6** ± 1.6	**5.98** ± 0.47	**33.4** ± 3.2	2.85 ± 0.46	7.59 ± 1.77	1.015	2.78	**448.1**	1779.4	**168.8**	658.0
1.2	33.1 ± 1.6	76.5 ± 3.0	25.0 ± 1.6	5.68 ± 0.47	30.9 ± 3.2	2.49 ± 0.46	6.65 ± 1.85	0.898	2.42	421.1	1642.0	158.5	603.8
1.5	33.0 ± 1.6	**80.1** ± 2.6	24.6 ± 1.6	5.70 ± 0.45	32.9 ± 3.3	**2.94** ± 0.48	**8.50** ± 2.02	**1.070**	**3.12**	439.5	**1782.7**	164.0	**659.9**
2.0	33.0 ± 1.7	78.2 ± 3.0	24.7 ± 1.6	5.66 ± 0.48	30.8 ± 3.4	2.57 ± 0.49	6.61 ± 1.77	0.954	2.50	424.2	1635.8	160.9	606.9

Table 6: **Reweighting quality in the DGRPO reward.** NoveltyBench metrics (defined in Section 4) as the quality exponent w_Q of Eq. 17 increases, holding $w_D = w_{\text{Dev}} = 1$. Response diversity and deviation are shown $\times 100$; cells include a 95% confidence interval where available. The best value per column is bolded. $w_Q = 1.0$ is the DGRPO (+Deviation) model from Table 1.

J DGRPO Ablation: reweighting quality

The atomic metrics in Table 1 show that training a response model directly forces a tradeoff between quality and the creativity axes: DGRPO attains the highest response diversity and deviation of any method trained with a deviation reward, but the lowest judge score (77.3, versus 90.5 for the reference). A natural question is whether this tradeoff is an artifact of how DGRPO's multiplicative reward (Eq. 12) balances its terms and whether simply weighting quality more heavily can recover quality without sacrificing diversity.

We generalize the DGRPO reward with a per-term exponent weight,

$$r_{j,i} = Q_i^{w_Q} \cdot D_i^{w_D} \cdot \text{Dev}_i^{w_{\text{Dev}}}. \tag{17}$$

Because $Q_i, D_i, \text{Dev}_i \in [0, 1]$ (Eqs. 13–15), an exponent $w > 1$ effectively penalizes lower values more aggressively than higher values. We hold $w_D = w_{\text{Dev}} = 1$ and sweep the quality exponent $w_Q \in \{1.0, 1.2, 1.5, 2.0\}$, retraining DGRPO from scratch at each setting and evaluating on NoveltyBench. The $w_Q = 1.0$ row is the DGRPO (+Deviation) model from Table 1.

Increasing the quality exponent does not resolve the tradeoff (Table 6). Across $w_Q \in [1.0, 2.0]$, the judge score stays within a narrow 76.5–80.1 band—never approaching the reference's 90.5—while response diversity, deviation, and distinct clusters are essentially unchanged (all differences fall within overlapping confidence intervals). The compound creativity metrics move only marginally, peaking weakly at $w_Q = 1.5$. Reweighting quality therefore has little leverage on DGRPO's quality–creativity balance.

	Epochs	Atomic metrics					Compound metrics							
		Response Diversity	Judge Score	Deviation	Vendi Score	Distinct Clusters	Filtered		AUC-dev		AUC-qual		VUS	
							Vendi	Distinct	Vendi	Dist.	Vendi	Dist.	Vendi	Dist.
	2	13.4 ± 1.6	92.0 ± 1.1	5.1 ± 0.7	2.27 ± 0.20	19.3 ± 3.3	0.46 ± 0.23	1.20 ± 0.85	0.294	1.22	47.7	131.0	30.2	139.0
	4	**13.9** ± 1.5	94.0 ± 1.0	7.8 ± 0.9	**2.42** ± 0.21	**26.5** ± 3.8	**0.76** ± 0.28	2.35 ± 1.33	0.387	2.37	**74.1**	234.0	**37.8**	235.2
	6	13.4 ± 1.3	94.7 ± 0.8	9.1 ± 0.9	2.37 ± 0.18	25.0 ± 3.6	0.65 ± 0.27	**2.49** ± 1.41	**0.390**	**2.57**	64.2	**243.4**	**37.8**	**250.9**
	8	12.2 ± 1.2	**94.9** ± 0.8	9.1 ± 0.8	2.23 ± 0.18	22.1 ± 3.6	0.56 ± 0.24	1.95 ± 1.23	0.353	2.22	55.4	190.2	34.4	218.1
	10	11.7 ± 1.1	93.6 ± 1.2	**10.1** ± 0.8	2.17 ± 0.16	20.5 ± 3.3	0.57 ± 0.23	1.76 ± 1.06	0.345	2.23	55.8	172.3	34.1	221.0
+ Deviation	2	13.5 ± 1.6	91.5 ± 1.4	4.9 ± 0.7	2.28 ± 0.21	18.3 ± 3.2	0.40 ± 0.24	1.24 ± 0.95	0.278	1.08	45.3	137.4	29.3	126.3
	4	15.6 ± 1.7	92.0 ± 1.2	7.1 ± 0.9	2.53 ± 0.22	19.1 ± 3.2	0.74 ± 0.28	2.06 ± 1.17	0.386	1.59	83.4	232.4	40.0	175.0
	6	**19.7** ± 1.3	**93.1** ± 1.1	19.1 ± 1.2	**3.28** ± 0.22	**29.9** ± 3.5	2.59 ± 0.31	14.7 ± 2.70	**0.851**	**5.58**	259.5	1567.0	85.1	586.9
	8	18.1 ± 1.2	91.7 ± 1.4	20.8 ± 1.1	3.02 ± 0.19	24.7 ± 3.2	2.45 ± 0.27	13.3 ± 2.69	0.786	4.59	249.1	1495.9	79.9	512.0
	10	18.0 ± 1.1	88.4 ± 1.8	**24.7** ± 1.0	3.02 ± 0.18	28.1 ± 3.5	**2.60** ± 0.24	**16.4** ± 2.80	0.835	5.24	**272.8**	**2022.3**	**88.0**	**639.0**

Table 7: **DARLING evaluated every 2 epochs on NoveltyBench**, without (top) and with (bottom) the deviation reward. The 10-epoch rows match the corresponding DARLING rows in Table 1; the 2-epoch rows use the same data budget as DGRPO and CP. Response diversity and deviation are shown $\times 100$; cells include a 95% confidence interval where available. Best value per column within each block is bolded. DARLING's diversity peaks around epochs 4–6 and then erodes with further training, and no checkpoint — at any epoch, with or without deviation — approaches the system-creativity of CP.

K DARLING ACROSS TRAINING EPOCHS

In the main work, DARLING is trained following the original recipe of 10 epochs (Li et al., 2025), compared to the 2 epochs used for DGRPO and CP. To rule out the possibility that this longer schedule either harms or advantages DARLING relative to the other methods, we evaluate its checkpoints at every 2 epochs (2, 4, 6, 8, 10). The 2-epoch checkpoint is the equal-data-budget comparison against DGRPO and CP, and the 10-epoch rows are the same as the DARLING rows of Table 1.

At the equal-data budget of 2 epochs, DARLING is no more creative than at its reported 10-epoch checkpoint; without the deviation reward its simple diversity peaks around epochs 4–6 and then declines as continued training trades diversity for reward-model quality. However, we do see deviation increasing throughout training. In all cases DARLING's compound creativity metrics remain far below those of CP (Table 1), so the epoch mismatch neither advantages nor explains away the gap: DARLING does not become competitive at any point along its training trajectory.

L TEMPLATES

This appendix collects the prompt templates used by Creative Priming. Figure 13 shows a template that elicits a creative primer from the primer model, Figure 14 gives the template that conditions the response model on that primer, and Figure 15 gives the template used by the creative primer judge J_{CP} (Eq. 9). Figures 16–18 give the LLM-as-a-judge template used for response-quality evaluation: this judge scores every response on six fixed *universal* components and on a set of *prompt-specific* components that are generated once per prompt at annotation time. Figure 16 shows the template skeleton with its three placeholders left unfilled, and Figures 17 and 18 show the prompt-specific components materialized for two example prompts from NoveltyBench.

Creative primer prompt template

```
You are providing meta-level creative guidance, not solving the task.

Given the prompt below, produce a few sentences of advice that would
    help another language model respond with more creative CONTENT.
    Focus on what substantive ideas, perspectives, or angles the
    response should explore --- not on style, tone, or formatting.
Your advice MUST be specific to this particular prompt - it should not
     be generic advice that could be applied to any prompt.

Do not answer the prompt or include example content.

Prompt:
{original_prompt}
```

Figure 13: Prompt template used to elicit a creative primer from the primer model. At inference {original_prompt} is substituted with the original prompt.

Response prompt template

```
Below you have a prompt and some guidance on how to respond. Your
    highest priority is to answer the prompt, and if the guidance
    conflicts, prioritize the prompt. Do not acknowledge the guidance
    in your response.

Prompt:
{original_prompt}

Guidance:
{creative_prep}
```

Figure 14: Prompt template used to condition the response model on a creative primer. At inference {original_prompt} and {creative_prep} are substituted with the original prompt and the generated primer.

Creative primer judge template

```
You are evaluating the quality of a creative preparation strategy for
    a language model. Creative preparations are instructions given to
    a model before it sees the actual user prompt. Their purpose is to
     shape how the model *thinks* about the problem, leading to more
    diverse and higher-quality responses.

The best creative preps are thinking strategies --- they tell the
    model what to consider, question, or reason about before
    responding. They guide the model's internal reasoning process
    rather than dictating what the output should look like. A creative
     prep that says "before responding, identify the assumption most
    people make about this topic and consider what happens if it's
    wrong" is far more valuable than one that says "write your
    response as a poem" or "use vivid metaphors."

Important: a creative prep must BE a creative prep --- not
    instructions for how to write one. If the text reads as meta-
    guidance about crafting advice (e.g., "craft advice that focuses
    on a specific lens" or "prescribe a narrow limitation that forces
    originality"), it has failed at its basic purpose regardless of
    how sophisticated it sounds.

Evaluate the creative prep on the following criteria.

---

# EVALUATION CRITERIA

## 1. Specificity (0-100)
**Definition**: Is this tailored to the unique characteristics of THIS
     SPECIFIC PROMPT, or generic advice?

**Scoring Rubric**:
- **0-20**: Completely generic, ignores prompt content, could apply to
     any prompt (e.g., "think carefully", "be creative", "subvert
    expectations", "use a specific lens to force originality")
- **21-40**: Recognizes broad category but ignores specific features (
    e.g., "this is a story prompt", "this is a comparison question")
- **41-60**: Mentions some prompt characteristics but stays surface-
    level (e.g., "this has constraints", "this asks for analysis")
- **61-80**: Identifies specific prompt features and references them
    concretely (e.g., "this prompt requires structural repetition", "
    this comparison involves temporal change")
- **81-100**: Deeply analyzes the prompt's unique characteristics and
    frames approach around them (e.g., "this repetition constraint can
     become a semantic device", "what's unsaid in this comparison is
    the selection criteria itself")

## 2. Thinking Strategy (0-100)
**Definition**: Does this provide a thinking strategy that shapes the
    model's reasoning process, or does it prescribe specific content,
    structure, or output?

A good creative prep tells the model how to *think* about the problem
    --- what to question, what angle to take, what to reason about
    before responding. It leaves the actual content open. Low scores
    go to preps that dictate specific elements, structures, content,
    or output format. The highest scores go to preps that explicitly
    guide the model's pre-response reasoning (e.g., "before responding
    , think about...", "first consider...", "ask yourself...").
```

```
**Scoring Rubric**:
- **0-20**: Dictates exact content or structure (e.g., "start with X,
    then say Y, conclude with Z", "your response must include these
    three elements", "write exactly 5 paragraphs about A, B, C, D, E")
- **21-40**: Prescribes specific elements with little room for
    interpretation (e.g., "make sure to mention X, Y, and Z", "use
    this specific example to illustrate your point")
- **41-60**: Mix of strategy and prescription --- gives a thinking
    approach but also constrains specific outputs (e.g., "consider
    different angles, and make sure to include a counterargument")
- **61-80**: Clear thinking strategy with actionable direction (e.g.,
    "first identify the constraint, then find ways to turn it into an
    advantage", "trace what changes between the two states --- that's
    where the mechanism lives")
- **81-100**: Explicitly guides pre-response reasoning without
    constraining output (e.g., "before responding, list the most
    common answers people give, then ask what they all have in common
    --- that shared assumption is your entry point", "ask yourself:
    what would change about this question if you shifted the time
    horizon by 100 years?")

## 3. Content Focus (0-100)
**Definition**: Does this guide *what the model should think about*
    --- the substance, reasoning, or approach to the problem --- or
    does it only address *how to present the output*?

Good creative preps direct the model's attention to ideas, arguments,
    angles, reasoning approaches, or aspects of the subject matter to
    explore. Weak creative preps only specify output format, tone,
    narrative framing, or stylistic choices. Guiding the model's *
    thinking approach* (e.g., "consider what assumptions underlie this
     question") counts as content guidance, since it shapes what
    substance the model engages with. Pure presentation instructions (
    e.g., "frame as a metaphor", "embed the answer in a folktale", "
    write as a dialogue") are style, not content --- even when
    described with rich, specific detail (e.g., "embed it in a Zulu
    proverb" is still a presentation instruction). However, a light
    style suggestion alongside real thinking guidance shouldn't tank
    the score --- what matters is the balance. If the original prompt
    is about style or form (e.g., "write a poem"), then style guidance
     IS content guidance.

**Scoring Rubric**:
- **0-20**: Purely about output presentation with no guidance on what
    to think about (e.g., "write in a whimsical tone", "use lots of
    metaphors", "embed the answer in a traditional folktale, proverb,
    or myth")
- **21-40**: Mostly presentation-focused with vague content hints (e.g
    ., "write engagingly and try to cover multiple perspectives", "use
     an unexpected narrative frame and touch on different angles")
- **41-60**: Mix of thinking guidance and presentation --- has some
    approach direction but presentation instructions dominate or
    dilute it (e.g., "frame using an unconventional metaphor that
    diverges from typical associations" --- there's an implicit
    thinking nudge but it's wrapped in a style directive)
- **61-80**: Primarily guides what to think about, with incidental
    style elements (e.g., "the key insight is that X enables Y ---
    explore the implications" with minor style notes)
- **81-100**: Entirely focused on what the model should reason about
    --- ideas, angles, assumptions to question, knowledge to bring to
    bear (e.g., "the constraint creates a tension between efficiency
    and readability --- explore which trade-offs matter most and why")
```

```
## 4. Prompt Faithfulness (0-100)
**Definition**: Does the creative prep work within the bounds of what
    the prompt is asking, or does it transform the question into
    something else?

A good creative prep helps the model produce diverse, high-quality
    responses *to the actual question*. A bad one distorts the prompt'
    s intent --- encouraging fiction for a factual question, turning a
     simple request into an elaborate creative exercise, narrowing a
    general question to a specific ideological or niche framing that
    most people wouldn't intend, or suggesting the response should be
    a "story" or "artifact" when the prompt asks for information. For
    open-ended or creative prompts, imaginative reinterpretation is
    fine. For factual or constrained prompts, the prep should respect
    those boundaries.

**Scoring Rubric**:
- **0-20**: Actively works against the prompt's intent --- encourages
    fiction for factual questions, ignores explicit constraints, or
    transforms the question into something entirely different
- **21-40**: Substantially distorts the prompt --- forces a specific
    ideological lens, adds major requirements that change what's being
     asked, or reframes a simple question as an elaborate creative
    exercise (e.g., suggesting a factual answer should "feel like a
    self-contained story or artifact")
- **41-60**: Partially respects the prompt but pushes it in a
    direction that doesn't match the user's likely intent --- e.g.,
    narrowing a general question to one specific angle, or adding a
    framing that subtly changes what's being asked
- **61-80**: Mostly respects the prompt's intent, with minor stretches
     that don't fundamentally change what's being asked
- **81-100**: Fully respects the prompt's intent --- helps the model
    think creatively within the bounds of what was actually asked

## 5. Focus (0-100)
**Definition**: Does the creative prep commit to a coherent thinking
    strategy, or does it pile on unrelated ideas hoping something
    sticks?

A focused creative prep provides depth --- it develops one thinking
    approach and lets that approach do the work. An unfocused one
    substitutes breadth for depth, whether by listing many loosely
    connected thinking strategies, dumping unrelated constraints or
    formatting rules, or exploring multiple reframings of the same
    question without committing to one. Even if each individual idea
    is good, cramming many together produces an unfocused prep.

**Scoring Rubric**:
- **0-20**: Kitchen sink --- lists many unrelated strategies,
    constraints, or rules with no unifying thread (e.g., "try analogy,
     consider edge cases, use humor, think historically, subvert
    expectations" or "use lowercase, no markdown, prioritize verbs,
    avoid cliches, keep it cryptic")
- **21-40**: Scattered --- several loosely related ideas competing for
     attention, no clear through-line connecting them (e.g., exploring
     paradoxical lenses AND emotional dissonance AND unspoken
    hierarchies AND meta-narratives AND philosophical inquiry all in
    one prep)
- **41-60**: Partially focused --- has a main thinking strategy but
    cluttered with tangential ideas that dilute rather than reinforce
    it
```

```
- **61-80**: Focused --- a clear central thinking strategy, possibly
    with one or two supporting ideas that build on or reinforce the
    core approach
- **81-100**: Sharp and cohesive --- commits to a single thinking
    approach and develops it with depth, or uses a few tightly
    interlocking ideas that clearly serve the same angle

## 6. Direct Address (binary: 0 or 1)
**Definition**: Does the creative prep refer to the model as a third-
    party entity?

- **0**: The prep refers to the model as a third-party entity (e.g., "
    the model should...", "encourage the model to...", "AI systems
    should...", "guide the model to...")
- **1**: The prep does NOT refer to the model as a third-party entity.
     It may use direct address ("you"), bare imperatives ("start by
    ..."), or discuss the task/response ("the response should...") ---
     all of these are fine.

---

# THE ORIGINAL PROMPT

{prompt}

---

# THE CREATIVE PREP TO EVALUATE

{creative_prep}

---

# EVALUATION INSTRUCTIONS

## Step 1: Reasoning

Work through each criterion one at a time BEFORE assigning any scores.
     For each criterion:
1. Identify specific phrases or patterns in the creative prep that
    relate to this criterion
2. Consider where these elements fall within the scoring rubric
3. Note any tensions or edge cases

Write your reasoning in plain text below. Be thorough but focused ---
    reference specific parts of the creative prep to support your
    analysis.

## Step 2: Final Scores

After completing your reasoning, provide your scores in the following
    JSON format. The justifications should be brief summaries (1-2
    sentences) of your reasoning above. Do not include any text after
    the JSON block.

'''json
{
    "specificity": {
        "score": <0-100>,
        "justification": "<brief explanation>"
    },
    "thinking_strategy": {
        "score": <0-100>,
```

```
        "justification": "<brief explanation>"
    },
    "content_focus": {
        "score": <0-100>,
        "justification": "<brief explanation>"
    },
    "prompt_faithfulness": {
        "score": <0-100>,
        "justification": "<brief explanation>"
    },
    "focus": {
        "score": <0-100>,
        "justification": "<brief explanation>"
    },
    "direct_address": {
        "score": <0 or 1>,
        "justification": "<brief explanation>"
    }
}
```

```
Remember:
- For specificity, thinking_strategy, content_focus,
    prompt_faithfulness, and focus: higher scores (0-100) are better
- For direct_address: 1 = addresses model directly or discusses the
    task, 0 = talks about the model as a third-party entity ("the
    model should...", "encourage the model to...")
- Your justifications should reference specific phrases from the
    creative prep
- Think step-by-step before scoring
```

Figure 15: Prompt template for the criterion-mean creative primer judge J_{CP} (Section 3.1, Eq. 9). At inference {prompt} and {creative_prep} are substituted with the original prompt and the primer under evaluation.

Evaluation LLM-as-a-judge template

```
You are an expert evaluator assessing the quality of an AI response.
    You will evaluate the response on both UNIVERSAL components (
    applicable to all responses) and PROMPT-SPECIFIC components (
    tailored to this particular prompt type).

## PROMPT
{{PROMPT}}

## RESPONSE TO EVALUATE
{{RESPONSE}}

---

# PART 1: UNIVERSAL COMPONENTS

Evaluate each of the following components on a scale of 0-100.

### 1. Instruction Following
How well does the response address what was actually asked?
- 0-20: Ignores or fundamentally misunderstands the prompt
- 21-40: Addresses the prompt partially but misses key requirements
- 41-60: Addresses the main request but overlooks minor specifications
- 61-80: Follows instructions well with only trivial deviations
- 81-100: Precisely and completely follows all instructions

### 2. Coherence
Is the response logically structured and easy to follow?
- 0-20: Disorganized, contradictory, or incomprehensible
- 21-40: Poor structure with unclear logical flow
- 41-60: Adequate structure but some confusing transitions or
    organization
- 61-80: Well-organized with clear logical progression
- 81-100: Exceptionally well-structured with seamless logical flow

### 3. Completeness
Does it fully address the prompt without obvious gaps?
- 0-20: Severely incomplete; misses most aspects of the prompt
- 21-40: Addresses some aspects but leaves significant gaps
- 41-60: Covers main points but lacks depth or misses secondary
    aspects
- 61-80: Thorough coverage with only minor omissions
- 81-100: Comprehensively addresses all aspects of the prompt

### 4. Conciseness
Is the response appropriately sized---not padded, not truncated?
- 0-20: Extremely verbose/padded OR severely truncated
- 21-40: Noticeably too long or too short for the task
- 41-60: Somewhat inefficient length but acceptable
- 61-80: Appropriate length with minor verbosity or brevity issues
- 81-100: Optimally sized for the task at hand

### 5. Clarity
Is the language clear and unambiguous?
- 0-20: Confusing, jargon-heavy, or ambiguous throughout
- 21-40: Frequently unclear or poorly worded
- 41-60: Generally understandable but some unclear passages
- 61-80: Clear language with only occasional ambiguity
- 81-100: Exceptionally clear and precise language throughout

### 6. Surprise
```

```
Does the response offer something unexpected while still fulfilling
    the original request? Does it challenge implicit assumptions,
    introduce novel angles, or deliver value beyond the obvious
    interpretation---without departing from what was asked?
- 0-20: Entirely predictable; the most generic possible response to
    this prompt
- 21-40: Mostly boilerplate with perhaps one minor unexpected element
- 41-60: Competent but unsurprising; follows the expected path without
     deviation
- 61-80: Contains notable unexpected elements---a fresh angle, an
    unconventional example, or a thought-provoking reframe---while
    still satisfying the request
- 81-100: Genuinely surprising; reframes the problem, surfaces hidden
    assumptions, or delivers insight the prompter didn't know they
    wanted---yet still fully answers the prompt

---

# PART 2: PROMPT-SPECIFIC COMPONENTS

The following components are tailored to this specific type of prompt.
     Evaluate each on a scale of 0-100.

{{PROMPT_SPECIFIC_COMPONENTS}}

---

# EVALUATION INSTRUCTIONS

First, reason through your evaluation step by step BEFORE assigning
    any scores. For each component (both universal and prompt-specific
    ), discuss what the response does well and where it falls short
    relative to the rubric. Reference specific parts of the response
    to support your reasoning. After you have reasoned through ALL
    components, provide your final scores in the JSON format below.

## Step 1: Reasoning

Work through each component one at a time. For each, consider:
- What specific aspects of the response are relevant to this component
    ?
- Where does the response fall within the scoring rubric?
- Are there any edge cases or tensions between quality and convention?

Write your reasoning in plain text. Be thorough but focused.

## Step 2: Final Scores

After your reasoning is complete, provide your scores in the following
     JSON format. The justifications should be brief summaries of your
     reasoning above. Do not include any text after the JSON block.

'''json
{
  "universal": {
    "instruction_following": { "score": <0-100>, "justification": "<
    brief summary>" },
    "coherence":             { "score": <0-100>, "justification": "<
    brief summary>" },
    "completeness":          { "score": <0-100>, "justification": "<
    brief summary>" },
    "conciseness":           { "score": <0-100>, "justification": "<
    brief summary>" },
```

```
    "clarity":               { "score": <0-100>, "justification": "<
    brief summary>" },
    "surprise":              { "score": <0-100>, "justification": "<
    brief summary>" }
  },
  "prompt_specific": {
    "components": [
      { "name": "<component name>", "score": <0-100>, "justification":
     "<brief summary>" }
    ],
    "aggregate": { "score": <0-100>, "method": "<e.g., simple average,
     weighted average>", "justification": "<overall prompt-specific
    summary>" }
  }
}
```

Figure 16: LLM-as-a-judge template used for response-quality evaluation, shown with its three placeholders unfilled. {{PROMPT}} and {{PROMPT_SPECIFIC_COMPONENTS}} are substituted once per prompt at annotation time (see Figures 17 and 18); {{RESPONSE}} is substituted at scoring time with each response under evaluation. The six universal components are fixed across all prompts.

Evaluation LLM-as-a-judge: prompt-specific components (example 1)

```
### 7. Emotional Resonance
Measures how effectively the story evokes genuine feeling about the
    girl-dog relationship through specific moments, gestures, or
    details rather than generic sentiment.
- 0-20: Relies entirely on clichéd phrases ('best friends forever') or
     tells us about emotions without showing them
- 21-40: Includes some emotional moments but they feel predictable or
    borrowed from familiar stories
- 41-60: Creates at least one authentic emotional beat through a
    specific detail or interaction
- 61-80: Weaves multiple emotionally specific moments that feel fresh
    and genuine to this particular relationship
- 81-100: Achieves profound emotional truth through unexpected yet
    perfectly chosen details that illuminate the unique bond

### 8. Narrative Economy
Evaluates how much story complexity, character depth, or thematic
    richness the writer achieves within the five-sentence constraint.
- 0-20: Linear series of events with no depth, essentially 'this
    happened, then this happened'
- 21-40: Basic story arc present but characters remain flat and stakes
     feel minimal
- 41-60: Manages to suggest character growth, conflict, or a
    meaningful arc despite brevity
- 61-80: Impressively layers multiple story elements (subplot, theme,
    character development) into the constraint
- 81-100: Achieves the density of a much longer piece, with each
    sentence doing multiple jobs to create a rich, complete world
```

Figure 17: Values substituted into the placeholders of Figure 16 for the NoveltyBench prompt "Tell me a story in five sentences about a girl and her dog."

Evaluation LLM-as-a-judge: prompt-specific components (example 2)

```
{{PROMPT}} =
    Imagine you are a time traveler who has just arrived in a new era.
    Describe the first thing you notice.

{{PROMPT_SPECIFIC_COMPONENTS}} =

### 7. Sensory Immersion and Detail
Measures how effectively the response uses specific sensory details
    and observations to create a vivid, immersive experience of
    temporal displacement.
- 0-20: Generic observations that could apply to any time period;
    relies on clichés like 'the air smells different' without
    specificity
- 21-40: Some attempt at sensory detail but remains surface-level;
    mentions obvious differences like 'no cars' or 'people dress
    differently'
- 41-60: Includes specific sensory details that ground the reader in
    the era; notices concrete differences in sounds, smells, or
    textures
- 61-80: Rich sensory tapestry that captures subtle, era-specific
    details a time traveler would genuinely notice first; conveys
    visceral disorientation
- 81-100: Masterfully uses unexpected sensory details to reveal the
    era through fresh observations; makes the familiar strange or the
    strange familiar in compelling ways

### 8. Temporal Perspective and Insight
Evaluates how the response leverages the unique perspective of a time
    traveler to reveal something profound or thought-provoking about
    the era or human experience.
- 0-20: No meaningful use of the time traveler perspective; could be
    any observer describing any place
- 21-40: Basic fish-out-of-water observations; focuses on obvious
    technological or social differences without deeper insight
- 41-60: Uses time traveler perspective to highlight interesting
    contrasts or continuities across eras; shows awareness of
    historical context
- 61-80: Reveals unexpected truths about the era through the traveler'
    s unique vantage point; finds poetry in mundane details or
    profound meaning in small observations
- 81-100: Brilliantly exploits the time travel premise to offer fresh
    philosophical, emotional, or anthropological insights; makes us
    see our own time differently
```

Figure 18: Values substituted into the placeholder of Figure 16 for the NoveltyBench prompt "Imagine you are a time traveler who has just arrived in a new era. Describe the first thing you notice."