Forty Shades of Blue: Quality-Diversity Alignment via Mode-Conditioned Reinforcement Learning

Jiayi Yuan [*1] **Hangoo Kang** [*2] **James Jihao Liu** [*2] **Yejin Choi** [2] **Vikram Iyer** [1] **Liwei Jiang** [†1] **Natasha Jaques** [†1]

Abstract

A notable byproduct of LLM alignment training is *mode collapse*: the progressive loss of output diversity that narrows a model's expressivity at inference time. This degradation is especially limiting for applications requiring open-ended exploration and pluralistic perspectives, such as scientific ideation and creative writing. We present **MODA** (**MO**de-conditioned **D**iversity **A**lignment), an online post-training RL algorithm that jointly optimizes generation quality and diversity, inspired by the coordination perspective in multi-agent reinforcement learning (MARL). MODAtrains a single shared LLM policy conditioned on abstract numbered mode tokens, where each mode acts as an agent that competes to produce outputs distinct from the others. This MARL-inspired formulation encourages mode-conditioned agents to explore complementary regions of the high-quality output space without requiring hand-crafted personas or architectural modifications. MODA employs a prompt-adaptive quality gating mechanism that calibrates quality thresholds against a frozen reference policy and grants diversity rewards only to responses that meet them, preventing reward-hacking behaviors such as language switching and verbosity. To study quality-diversity tradeoffs, we evaluate MODAon a comprehensive benchmark spanning seven general capability tasks and four domain-specific diversity tasks in scientific ideation and creative writing. MODA improves SBERT diversity by **29.1%** while simultaneously increasing general capability performance by **17.3%** over the base model. Overall, MODAprovides a drop-in alternative to standard post-training methods that preserves and expands the model's expressive output space while improving quality.

1. Introduction

Post-training alignment is essential for improving the capability and safety of large language models (LLMs), but it also introduces a concerning side effect: *mode collapse*. Aligned models increasingly favor a narrow set of responses (the "mode") over the broader space of plausible outputs, resulting in a substantial loss of generation diversity (Jiang et al., 2025; West and Potts, 2025a; Huang et al., 2024; Sourati et al., 2026). Yet in many domains, diversity is not merely desirable but essential. In AI-assisted scientific discovery, progress relies on exploring a broad and plausible hypothesis space rather than prematurely converging on a single line of inquiry (Gruver et al., 2023; Romera-Paredes et al., 2024). Likewise, generative diversity is critical for applications requiring creativity and open-ended exploration (Shumailov et al., 2024; Abdulhai et al., 2026). While inference-time interventions (Misaki and Akiba, 2026b; Zhang et al., 2025a) can partially mitigate this loss of diversity, a more fundamental solution is to redesign post-training algorithms so that expressivity is preserved throughout the alignment process by construction.

We introduce MODA (**MO**de-conditioned **D**iversity **A**lignment), an online post-training RL algorithm that jointly optimizes the quality and diversity of generated responses by formulating alignment through the lens of multi-agent reinforcement learning (MARL), where agents develop diverse behaviors through competition and coordination (Eysenbach et al., 2018; Liang et al., 2024; Li et al., 2021). MODA introduces *role conditioning* to instantiate multiple lightweight agents within a single shared policy. Specifically, we prepend abstract role tokens to the system prompt, enabling each role to generate a response to the same user query while independently optimizing a group-relative diversity reward. This creates competitive reward

[*]Equal first author. [†]Equal senior author.

Code: https://github.com/yuanjiayiy/mode-conditioned-diversity-alignment

[1]University of Washington [2]Stanford University. Correspondence to: Jiayi Yuan <yuancarrieyjy@cs.washington.edu>, Hangoo Kang <hangook@stanford.edu>.

ICML'26 Workshop on Human-AI Co-Creativity, Seoul, South Korea. Copyright 2026 by the author(s).

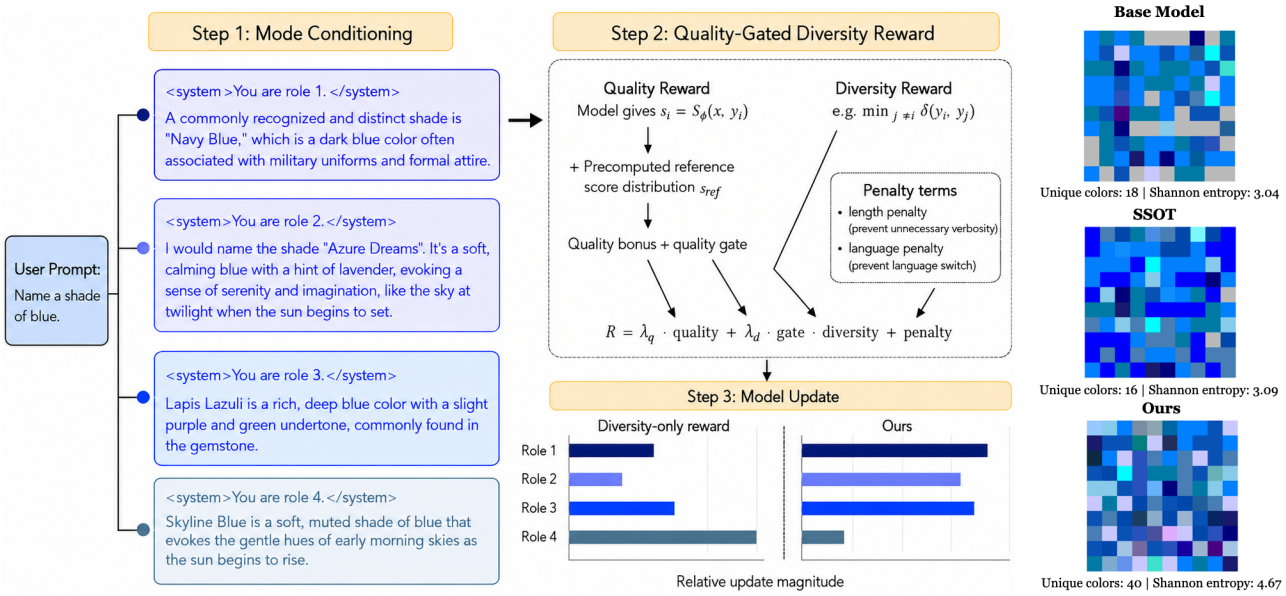


Figure 1. **Left**: Overview of MODA. MODA uses mode conditioning to elicit diverse candidate responses, then applies a quality gated diversity reward with penalty terms to promote high-quality diversity while suppressing low-quality or reward-hacking outputs. **Right**: 100 responses to the prompt **"Name a shade of blue"** from Qwen-3-8B (base) , SSOT , and MODA under diverse mode inference. MODA produces 40 distinct shades of blue, 122% more than the base model and 150% more than SSOT, with 100% valid answers, versus 20% invalid answers from the base model.

dynamics among roles, encouraging them to specialize in distinct regions of the high-quality response space and collectively produce a diverse set of outputs.

A key design choice of MODA is a *prompt-adaptive quality gate* that grants diversity rewards only to responses exceeding a prompt-specific quality threshold computed from a frozen reference policy. Among responses that satisfy the gate, diversity is rewarded according to semantic distance from the nearest neighboring response, encouraging exploration within the space of sufficiently high-quality generations. By coupling diversity rewards with quality, this mechanism prevents reward hacking, in which models maximize diversity by producing unusual but irrelevant outputs (Li et al., 2025b; Wan et al., 2025).

Assessing quality-diversity tradeoffs requires evaluating both standard capabilities and applications that benefit from diverse outputs. To this end, we curate a comprehensive 11-task evaluation suite that jointly measures general capability retention and diversity-focused applications where multiple high-quality generations are useful or necessary. The suite includes seven general-capability benchmarks (GSM8K, MMLU, GPQA, BoolQ, HellaSwag, TruthfulQA, and IFEval) and four open-ended domain application benchmarks spanning scientific ideation and creative writing such as HypoBench , PreScience , NoveltyBench , and a held-out split of Infinity-Chat . We evaluate output diversity using complementary *lexical* (1-Self-BLEU), *semantic* (SBERT), *concept-level* (E-Vendi), and *learned pairwise discriminator metrics*. Under the same thinking-disabled setting, MODA improves SBERT diversity and E-Vendi by 29.0% and 47.3%, respectively, over the Qwen3-8B-Instruct initial policy on the diversity-focused benchmark suite, while simultaneously increasing average pass@1 by 17.3% over the initial policy and by 12.2% over the strongest DARLING baseline on general capability benchmarks.

Compared with prior diversity-oriented post-training methods (Li et al., 2025b; Lanchantin et al., 2025b; Chung et al., 2025; Chen et al., 2025; Slocum et al., 2025; Puri et al., 2026), MODA conditions diversity on an explicit mode token rather than baking it unconditionally into the model. As a result, it naturally supports two inference modes: supplying mode tokens yields a diverse set of responses, while omitting them recovers the base model's standard behavior, preserving single-response capability. More broadly, we hope MODA motivates future work that brings richer MARL principles into LLM post-training.

Method	Mechanism	Code
INFERENCE-TIME		
Verbalized Sampling (Zhang et al., 2025a)	Prompting	✓
String Seed of Thought (Misaki and Akiba, 2026a)	Prompting	✓
DIPPER (Lau et al., 2025)	Prompting	✗
Multi-Novelty (Hu et al., 2025)	Prompting	✗
Multilingual Prompting (Wang et al., 2025)	Prompting	✓
Nucleus Sampling (Holtzman et al., 2020)	Decoding	✓
min-p Sampling (Nguyen et al., 2024)	Decoding	✓
G2 (Ruan et al., 2025)	Decoding	✗
Adaptive Decoding (LPO) (Dhuliawala et al., 2024)	Decoding	✓
TRAINING-TIME		
DARLING (Li et al., 2025b)	RL	✓
DivPO (Lanchantin et al., 2025a)	DPO	✗
DQO (Chen et al., 2025)	RL	✓
Multi-answer RL (Puri et al., 2026)	RL	✓
Soft Preference Learning (Slocum et al., 2025)	Entropy	✗
MODA(ours)	**Mode-Conditioned MARL**	✓

Table 1. **Diversity-oriented methods.** MODA conditions diversity on explicit mode tokens, inspired by MARL.

2. Related work

Why Does Generation Diversity Matter? Language models often collapse toward a narrow set of high-probability responses, resulting in an "Artificial Hivemind" characterized by substantial intra- and inter-model homogeneity on open-ended prompts (Jiang et al., 2025). Zhang et al. (Zhang et al., 2025a) further show that post-training alignment reduces diversity through biases in preference data, limiting performance in domains that require diverse outputs. In scientific reasoning, automated research agents benefit from a wider exploration breadth (Zou et al., 2025), and mode collapse causes reduced exploration in solution spaces (Yuan et al., 2026). In creative tasks, reduced diversity in LLM outputs also diminishes the creativity of LLM-assisted writing (Abdulhai et al., 2026; Anderson et al., 2024; Padmakumar and He, 2023). In mathematical proofs, strategy diversity is critical since repeated sampling is only useful when samples explore distinct reasoning paths (Wu et al., 2025; Cao et al., 2025). These findings motivate methods that preserve quality while expanding the range of model outputs.

Measuring Diversity of LLMs. Diversity of LLM-generated content has been measured along several dimensions. Lexical metrics include distinct n-grams (Ippolito et al., 2019), Self-BLEU (Zhu et al., 2018), Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010), and compression-based homogeneity scores (Shaib et al., 2024). Semantic metrics build on sentence embeddings (Reimers and Gurevych, 2019; Wieting and Gimpel, 2018) and information-theoretic constructions such as the Vendi Score (Friedman and Dieng, 2023) and its conditional variant (Jalali et al., 2024), as well as gradient-based measures like G-Vendi (Jung et al., 2025). A complementary line of work studies how diversity collapses through the training pipeline (Kirk et al., 2024; O'Mahony et al., 2024; Padmakumar and He, 2024; West and Potts, 2025b; Dang et al., 2025), and recent benchmarks evaluate human-like or effective semantic diversity directly (Zhang et al., 2025b; Shypula et al., 2025; Guo et al., 2025b; Shahid et al., 2025). However, most evaluations assess diversity or quality in isolation, whereas we jointly evaluate capability and diversity across real-world applications to ensure models preserve multiple plausible answers without mode collapse (Misaki and Akiba, 2026a; Puri et al., 2026) or catastrophic forgetting (Luo et al., 2025).

Improving Diversity during Training and Inference. Recent work has explored improving diversity through prompting (Zhang et al., 2025a; Lau et al., 2025; Hu et al., 2025; Wang et al., 2025; Kim et al., 2026; Wu et al., 2025), decoding (Holtzman et al., 2020; Nguyen et al., 2024; Ruan et al., 2025; Dhuliawala et al., 2024), and training (Li et al., 2025b; Lanchantin et al., 2025b; Chung et al., 2025; Chen et al., 2025; Slocum et al., 2025; Puri et al., 2026) (Table 1). At inference time, Verbalized Sampling (Zhang et al., 2025a) elicits a distribution over responses and String Seed of Thought (Misaki and Akiba, 2026a) injects random-string entropy, while prompt ensembles (Lau et al., 2025; Hu et al., 2025) and multilingual or chain-of-thought prompting (Wang et al., 2025; Wu et al., 2025) broaden outputs. At decoding time, min-p sampling (Nguyen et al., 2024), G2 (Ruan et al., 2025), and adaptive temperature (Dhuliawala et al., 2024) steer generation toward more varied continuations. In post-training, DARLING (Li et al., 2025b) jointly optimizes quality and semantic diversity but relies on a learned partition classifier; DivPO (Lanchantin et al., 2025a) applies DPO to responses that are both diverse and exceed a quality threshold; and DQO (Chen et al., 2025) uses a determinant-based group diversity reward that is shared across the group. Other RLVR methods address the Pass@1–Pass@k tension (He et al., 2025; Cheng et al., 2025; Li et al., 2025a), while Multi-answer RL (Puri et al., 2026) outputs multiple diverse answers within one response and Soft Preference Learning decouples entropy from the KL penalty (Slocum et al., 2025).

Different from prior training-based works, MODA approaches quality-diversity alignment from a MARL-inspired perspective, treating diverse generation as a competition-and-coordination problem among role-conditioned agents within a shared policy. This formulation enables explicit per-role credit assignment under quality constraints, encouraging specialization across the high-quality response space. In contrast to prompt- or decoding-based methods, MODA learns diversity through post-training, while avoiding the coarse set-level optimization used by prior RL-based approaches.

3. MODA: Mode-Conditioned Diversity Alignment for LLMs

3.1. Preliminaries: LLM RL Post-Training and Mode Collapse

Recent RL post-training has substantially improved the utility and safety of LLMs (Shao et al., 2024; Guo et al., 2025a). However, optimizing only the expected reward often leads to *mode collapse*, where the learned policy concentrates on a small set of high-reward responses while suppressing equally valid alternatives. As a result, generated responses become repetitive despite maintaining high quality. **Our goal is to learn policies that generate responses that are simultaneously high-quality, diverse, and non-redundant.**

Let $\mathcal{S}$ denote the space of natural language token sequences. Given a prompt $x \in \mathcal{S}$, a language model $\pi(\cdot \mid x)$ defines a probability distribution over responses in $\mathcal{S}$, where $\pi(y \mid x)$ denotes the probability of generating response $y \in \mathcal{S}$. For each prompt, we independently sample k responses, $y_1, \ldots, y_k \sim \pi(\cdot \mid x)$, forming a *response group* $\mathcal{Y} = \{y_i\}_{i=1}^{k}$.

Conventional RL post-training maximizes the expected reward

$$\max_{\theta} \; J_{\mathrm{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} r(x,y)\right], \qquad (1)$$

where $r(x, y)$ is the reward assigned to response y for prompt x. Optimizing this objective alone progressively concentrates the policy on a small set of high-reward responses, reducing the policy entropy

$$H(\pi_\theta(\cdot \mid x)) = -\sum_{y\in\mathcal{S}} \pi_\theta(y \mid x) \log \pi_\theta(y \mid x), \qquad (2)$$

and causing the model to repeatedly generate only a few dominant responses despite the existence of many alternatives with comparable rewards. We refer to this phenomenon as *mode collapse*.

3.2. MARL-Inspired Mode Conditioning

A rich body of work in multi-agent reinforcement learning (MARL) has explored how to incentivize agents to develop diverse behaviors through information sharing, KL divergence maximization, or adversarial learning (Eysenbach et al., 2018; Liang et al., 2024; Li et al., 2021). We transplant this idea into the LLM post-training. Using the system prompt, we condition a single model on different roles, and we treat each role as an independent agent. Each agent competes to maximize the uniqueness of its response within the response group, subject to a shared quality objective.

This motivates our core mechanism, *mode conditioning*: whereas standard alignment exposes a single default interface for learning high-reward behavior, mode conditioning creates a richer interface that lets the model accommodate a broader range of learned behaviors. We instantiate mode conditioning with roles. In the primary setting, the role instruction is deliberately abstract: `You are role` i. We use numbered roles rather than crafted role descriptions to avoid committing domain-specific personas or introducing human bias about which perspectives should exist. While our ablation experiments show than hand-crafted roles such as "innovative thinker," "passionate educator," or "methodical experimenter" can be easier to learn, we nevertheless choose to learn emergent roles purely through maximizing the diversity objective, giving the model more flexibility with less bias.

We now formalize this setup. Let $x \in \mathcal{X}$ denote a user prompt and let $\mathcal{Z} = \{z_1, \ldots, z_k\}$ denote a fixed set of numbered roles. The shared policy is $\pi_\theta(y \mid x, z_i)$ where all roles share the same parameters θ. During training, we sample responses from each role, which forms a response group for each prompt $\mathcal{Y}_x = \{y_{i,k} : i = 1, \ldots, N,\ k = 1, \ldots, K\}$. We reward each $y_{i,k}$ inside the $\mathcal{Y}_x$ using the Quality Gated Diversity Reward, detailed in Section 3.3, and optimize π_θ using Group Relative Policy Optimization (GRPO; (Shao et al., 2024)). We provide brief pseudocode in Alg 1, with detailed pseudocode is included in Appendix 2.

Dual Inference Modes. Not all queries require diverse responses. Factual questions such as "*What is Albert Einstein's birthday?*" have a single correct answer. A key advantage of role conditioning is that when diversity is not required, removing the role instruction recovers the base model's behavior exactly. This allows MODA to operate seamlessly in both diverse and standard inference modes. In standard mode, no role code is supplied and the model produces one ordinary response. In diverse mode, the same prompt is evaluated under several role codes, producing a candidate set $\{y_1, \ldots, y_N\}$. In practice, the response set can be returned directly, ranked by a quality model, or passed to an aggregator model that synthesizes a final answer from several distinct candidates.

3.3. Quality Gated Diversity Reward

While our goal is to improve the model's ability to generate diverse responses to the same prompt, quality is a non-sacrificeable bottom-line. To achieve an effective balance between quality and diversity, we introduce **Quality Gated Diversity Reward (QGDR)** which gates the diversity term by a prompt-adaptive quality threshold, so that diversity is rewarded only when the response is sufficiently high quality. In initial experiments, we found that this reward was

critical to preventing the model from learning to hack the diversity reward, for example through language mixing or nonsensical answers.

Diversity scoring. In our default implementation, the diversity score of a response is its nearest-neighbor semantic distance within the same response group:

$$d(x, y_i, \mathcal{Y}) = \min_{j \neq i} \delta(y_i, y_j), \tag{3}$$

where $\delta(y_i, y_j) = 1 - \cos(\mathbf{e}_i, \mathbf{e}_j)$ is the cosine distance between the sentence embeddings $\mathbf{e}_i, \mathbf{e}_j$ of responses y_i and y_j. This formulation ensures that a response receives a high diversity reward only if it is far from its closest alternative. By discouraging multiple responses from collapsing to the same answer while a single outlier carries the group-level diversity score, this formulation introduces competitive multi-agent dynamics among the roles, which must specialize to a unique style of response in order to maximize diversity among the group.

Quality Scoring. We only want to reward diversity if the quality does not degenerate from the initial policy, and reward for quality as well. We therefore design a prompt-adaptive quality reward and quality gating mechanism.

Reference Score Instantiations. RLHF-trained reward models are often prompt-dependent and calibrated primarily for relative comparison. Therefore, instead of using s_i directly as a reward, we calibrate it against scores from a fixed reference policy. For each prompt x, let $\mathcal{Y}_{\text{ref}}(x)$ denote a set of responses generated by the reference base model. Let $s^{(x)}_{\max}$, $s^{(x)}_{\min}$, and $\bar{s}^{(x)}$ denote the max, min, and mean absolute scores over $\mathcal{Y}_{\text{ref}}(x)$, respectively. We define a prompt-adaptive quality target and scale as

$$\begin{aligned} \tau_q^{(x)} &= s^{(x)}_{\min} + \alpha\big(s^{(x)}_{\max} - \bar{s}^{(x)}\big), \\ \sigma_q^{(x)} &= \gamma\big(\bar{s}^{(x)} - s^{(x)}_{\min} + \epsilon\big), \end{aligned} \tag{4}$$

where α shifts the quality target above the reference minimum by a fraction of the reference score spread, γ controls the sharpness of the scoring transition, and $\epsilon > 0$ ensures numerical stability. The quality reward for a generated response y_i is then

$$r_{\text{qual}}(x, y_i) = \mu \tanh\left(\frac{s_i - \tau_q^{(x)}}{\sigma_q^{(x)}}\right), \tag{5}$$

where μ controls the magnitude of the quality reward. Responses near the prompt-adaptive target $\tau_q^{(x)}$ receive near-zero quality reward, while responses above the target receive a positive bonus. The bounded $\tanh$ transformation prevents the quality term from dominating optimization once the response already exceeds the target by a large margin.

Quality Gating. We further define a binary quality gate

$$g(x, y_i) = \mathbf{1}\Big[s_i \geq \tau_q^{(x)}\Big]. \tag{6}$$

This gate ensures that the model receives diversity reward only when the response meets the prompt-adaptive quality threshold. As a result, responses that are merely unusual, off-topic, or malformed do not receive diversity credit, preventing spurious solutions which hack the diversity reward.

Reward-shaping Axillary Term. In preliminary experiments, we observed that the algorithm maximizes its diversity reward by engaging in language mixing and producing longer responses, a phenomenon also observed by DeepSeek-R1(Guo et al., 2025a). We therefore introduced a penalty term to suppress degenerate reward-hacking behaviors that can artificially increase measured diversity without improving response usefulness. Specifically, we include penalties for excessive length and language mismatch.

Let $d(x, y_i, \mathcal{Y})$ denote a diversity score measuring the novelty of y_i relative to the response group $\mathcal{Y} = \{y_1, \ldots, y_K\}$. The final QGDR reward function is

$$\begin{aligned} r_{\text{QGDR}}(x, y_i, \mathcal{Y}) = \lambda_q r_{\text{qual}}(x, y_i) + \lambda_d g(x, y_i) d(x, y_i, \mathcal{Y}) \\ + r_{\text{penalty}}(x, y_i), \end{aligned} \tag{7}$$

where λ_q and λ_d balance quality and diversity. The penalty term r_{penalty} explicitly discourages reward-hacking behaviors such as language switching, excessive verbosity and formatting failures, so that diversity gains reflect genuine semantic novelty.

3.4. Policy Optimization

We optimize the shared policy with on-policy reinforcement learning. For each prompt, all role-conditioned samples form one reward group. We optimize π_θ using Group Relative Policy Optimization (GRPO; (Shao et al., 2024)), which estimates advantages from within-group reward comparisons without a separate value network:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}} = \mathbb{E}_{x, \{y_i\}}\Big[\frac{1}{k}\sum_{i=1}^{k} \min\big(\rho_i \hat{A}_i,\ \text{clip}(\rho_i,\ 1{-}\varepsilon,\ 1{+}\varepsilon)\ \hat{A}_i\big) \\ - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big] \end{aligned} \tag{8}$$

where $\rho_i = \pi_\theta(y_i \mid x)/\pi_{\text{ref}}(y_i \mid x)$ is the importance ratio (Kloek and Van Dijk, 1978), $\hat{A}_i = (r_i - \text{mean}(\mathbf{r}))/\text{std}(\mathbf{r})$ is the group-normalized advantage over $\mathbf{r} = \{r(x, y_i, \mathcal{Y})\}_{i=1}^k$ from Eq. 7, and $\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ controls divergence from the base language model (Jaques et al., 2017). The reward is computed over the set of role-conditioned generations, but the trainable model remains a single shared policy.

Algorithm 1 MODA: Mode-Conditioned Diversity Alignment

Require: Dataset $\mathcal{D}$ with precomputed $(s_{\min}, s_{\max}, \mu_{\text{ref}})$; policy π_θ; reference policy π_{ref}; reward model S_ϕ; abstract roles $\mathcal{M} = \{m_i\}_{i=1}^K$; diversity metric δ; weights λ_q, λ_d; constants $\alpha, \gamma, \epsilon, \eta$

for training iteration $t = 1, \ldots, T$ **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{D}$
  for all $x \in \mathcal{B}$ **do**
    Retrieve $(s_{\min}(x), s_{\max}(x), \mu_{\text{ref}}(x))$
    $q_{\text{tar}} \leftarrow s_{\min} + \alpha(s_{\max} - \mu_{\text{ref}})$, $\tau_q \leftarrow \gamma(\mu_{\text{ref}} - s_{\min} + \epsilon)$
    Generate mode-conditioned responses with abstract roles $\mathcal{Y}(x) = \{y_i \sim \pi_\theta(\cdot \mid x, m_i)\}_{i=1}^K$
    for $i = 1, \ldots, K$ **do**
      $q_i \leftarrow S_\phi(x, y_i)$, $R_{\text{qual}} \leftarrow \mu \tanh((q_i - q_{\text{tar}})/\tau_q)$
      $G_{\text{qual}} \leftarrow \mathbf{1}\{q_i \geq q_{\text{tar}}\}$, $R_{\text{div}} \leftarrow \min_{j \neq i} \delta(y_i, y_j)$
      $R_i \leftarrow \lambda_q R_{\text{qual}} + \lambda_d G_{\text{qual}} R_{\text{div}} + R_{\text{pen}}(x, y_i)$
    end for
    $A_i \leftarrow \big(R_i - \text{mean}(R_{1:K})\big)/\big(\text{std}(R_{1:K}) + \eta\big)$ for $i = 1, \ldots, K$
  end for
  Update π_θ with GRPO using advantages $\{A_i\}$ and KL regularization to π_{ref}
end for
return trained policy π_θ

4. Experiments

We implement MODA using the verl codebase (Sheng et al., 2024), using vLLM (Kwon et al., 2023) for inference and FSDP (Zhao et al., 2023) for training. We use Qwen3-8B-Instruct (Qwen/Qwen3-8B) (Yang et al., 2025) as the base model. We sample 5 responses using the initial policy for each prompt, precompute the reference responses using the reward model `Skywork-Reward-V2-Llama-3.1-8B-40M` (Liu et al., 2025), and store them as part of the dataset. During training, we sample 6 role-response pairs from the model for each prompt as a response group, and embed them using `all-MiniLM-L6-v2` (Wang et al., 2020). Our training has two generation settings: thinking-enabled and thinking-disabled. During reward computation, both quality and diversity rewards are computed only over the final answer, with the thinking trace masked out. We include more implementation details in Appendix A.

4.1. Baselines

We compare our method against the following baselines. **Repeated Sampling** samples the base model k times per prompt without role conditioning, testing whether stochastic decoding alone suffices. **String Seed of Thought** (Misaki and Akiba, 2026a) prepends a random string to guide diverse generation; quality and diversity metrics are computed on the final answer span only, excluding any seed or scratchpad text. **DARLING** is an online RL baseline that jointly optimizes response quality and semantic diversity using a learned diversity signal (Li et al., 2025b). We train two versions of DARLING: one using our mixture data specified in Sec 4.2 and one using their specified recipe (10k prompts subsampled from WildChat).

4.2. Training Data

The training set contains 10K prompts, constructed from a 4:1 mixture of uniformly sampled Tulu3-SFT-Mixture prompts (Lambert et al., 2024) and Infinity-Chat dataset (Jiang et al., 2025). We use this mixture to balance quality and diversity: Tulu3-SFT-Mixture provides instruction-following prompts that help preserve response quality, while Infinity-Chat encourages open-ended exploration and promote response diversity.

4.3. Generative Diversity Evaluation

To evaluate generative diversity, we focus on two application domains where multiple distinct high-quality outputs are desirable: scientific ideation and creative writing. We evaluate all methods under the diverse mode, with the role code injected into the system prompt. Notably, none of the evaluation datasets below overlap with our training data, so strong performance here reflects genuine transfer rather than in-domain memorization. In the science domain, we adapt HypoBench (Liu et al., 2026) and PreScience (Ajith et al., 2026) for our cases. In the creative writing domain, we evaluate on NoveltyBench (Zhang et al., 2025b) and a held-out set of Infinity-Chat (Jiang et al., 2025). Below we detail HypoBench and PreScience, whose adaptation to our setting warrants further explanation. We use a comprehensive set of diversity metrics: SBERT (Reimers and Gurevych, 2019), self-BLEU (Zhu et al., 2018), and E-Vendi (Friedman and Dieng, 2023), as well as a discriminator trained in-house.

HypoBench. We adopt HypoBench (Liu et al., 2026), a benchmark designed to evaluate LLMs on hypothesis generation across diverse domains. In each task, the model is given a dataset and asked to propose a plausible hypothesis. Although HypoBench was originally designed to evaluate a model's capacity for inductive reasoning, it also fits naturally within the broader setting of research ideation, in which theories are formed from empirical observations. Diversity is especially important in this setting because the same observation can often support multiple plausible explanations, and generating varied hypotheses increases the chance of uncovering non-obvious patterns and alternative mechanisms for further investigation.

PreScience. We adopt the Contribution Generation task from PreScience (Ajith et al., 2026), where the model is

given a set of prior works and asked to generate a plausible title–abstract pair for a future scientific contribution. This setup mirrors one mode of scientific collaboration: a team of scientists, each bringing expertise from their prior work, collaborates to develop a new research idea grounded in those foundations. Diversity is crucial for this task because scientific progress often depends on exploring multiple possible combinations of prior ideas.

4.4. General Capability Retention.

For general capability retention, we evaluate each trained model in the standard single-response setting on a suite of seven widely used benchmarks under the standard mode, reporting both pass@1 and pass@5. Specifically, we use GSM8K (Cobbe et al., 2021) to evaluate mathematical reasoning; MMLU (Hendrycks et al., 2021), GPQA (Rein et al., 2024), and BoolQ (Clark et al., 2019) to evaluate broad knowledge and question answering; HellaSwag (Zellers et al., 2019) to evaluate commonsense inference; TruthfulQA-MC1 (Lin et al., 2022) to evaluate truthfulness; and IFEval (Zhou et al., 2023) to evaluate instruction following. Together, these benchmarks allow us to access whether the model preserves its general capabilities.

5. Quantitative Results

Domain Applications. As shown in Table 2, our model achieves the best performance across all four diversity metrics. We provide a breakdown by individual diversity metric in Appendix B. Notably, our model outperforms all baselines on HypoBench and PreScience, both of which are scientific ideation tasks that are out of domain relative to our training set, which primarily consists of standard SFT data and open-ended questions. These results suggest that MODA enhances general exploration capability, even on out-of-domain topics.

General Capability Retention. As shown in Table 3 and Appendix 6, with thinking disabled, our model achieves the best overall average score for general capability retention, outperforming the base model by **17.3%** in relative improvement. It matches or exceeds the base model on 6 out of 7 benchmarks and achieves the highest score on 5 of them. With thinking enabled, our model also improves over the base model, achieving a **2.4%** gain in average score. These results suggest that our method is an effective alignment approach that preserves, and in several cases improves, the model's general capabilities. We used a context window of 16,384 tokens for the MMLU and GPQA tasks, and a context window of 8,192 tokens for the remaining tasks.

We report qualitative analysis, including embedding-space visualizations and generation examples, in Appendix B.4.

5.1. Ablation Studies.

We ablate four design choices of MODA: mode conditioning, the explicit diversity reward, the quality gate, and abstract versus hand-crafted roles. Full combined results (general capability and diversity) are reported in Table 5, Appendix B.

Role Injection. We investigate the effect of mode conditioning by training a variant with the same procedure as MODA, but without injecting role instructions into the system prompt. As shown in Table 5, this variant achieves some improvement in diversity compared to base model, suggesting that RL training alone can partially broaden the response distribution. However, MODA achieves higher diversity compared to no role injection, indicating that explicit mode conditioning provides a useful mechanism for separating generations into distinct behavioral modes.

Quality-only Reward. We then examine whether the diversity gains are driven by the explicit diversity reward or arise naturally from repeated sampling and RL optimization. We train a variant using only the quality reward while keeping the training setup unchanged. As shown in Table 5, the quality-only model achieves no diversity gains over the base model, suggesting that the explicit quality-gated diversity objective is necessary to steer the model toward distinct high-quality responses.

Gating Mechanism vs. Additive Aggregation. Many existing works on quality-diversity optimization aggregate quality and diversity objectives using a weighted sum (Mouret and Clune, 2015; Chen et al., 2025). We compare our method with variants trained using a purely additive reward objective. We find that additive aggregation is highly sensitive to the diversity weight λ_d. In particular, diversity gains emerge only when $\lambda_d = 1000$, at which point diversity dominates quality. This makes additive aggregation difficult to tune, especially when changing the quality model or diversity metric. In contrast, MODA separates quality control from diversity optimization through the quality gate, making the reward less sensitive to coefficient scaling and enables faster iteration. Additionally, although additive aggregation shows a significant increase in diversity metrics, its general capability suffers catastrophically.

Crafted personas prompting. Prior work has shown that manually designed prompts can induce different model behaviors (Park et al., 2024; Argyle et al., 2023). We therefore craft six personas and inject them into the system prompt instead. This ablation tests whether hand-designed personas provide a stronger diversity prior than the abstract numbered modes used in the main MODA setting. The re-

Model Variant	Disc Div. ↑	SBERT Div. ↑	E-Vendi ↑
NO THINKING MODE			
Qwen3-8B	0.466 ± 0.048	0.299 ± 0.113	2.026 ± 0.155
DARLING (Mixture)	0.488 ± 0.056	0.351 ± 0.109	2.579 ± 0.123
DARLING (WildChat-10k)	0.473 ± 0.048	0.306 ± 0.107	2.095 ± 0.160
MODA	$\mathbf{0.500 \pm 0.041}$	$\mathbf{0.386 \pm 0.107}$	$\mathbf{4.531 \pm 0.858}$
THINKING MODE			
Qwen3-8B	0.460 ± 0.038	0.275 ± 0.072	2.299 ± 0.140
SSoT	0.447 ± 0.031	0.322 ± 0.058	2.589 ± 0.285
MODA	$\mathbf{0.512 \pm 0.042}$	$\mathbf{0.457 \pm 0.127}$	$\mathbf{3.496 \pm 0.963}$

Table 2. Aggregate domain-application diversity across HypoBench, PreScience, NoveltyBench, and Infinite-Chat. Metrics are averaged over domain-level scores and reported with standard errors; higher is better, and bold indicates the best model in each column under each mode.

Model	G8K	MMLU	GPQA	BQ	HS	TQA	IF	Avg.
NO THINKING MODE								
Qwen3-8B	84.8	80.0	42.9	83.6	48.3	21.7	76.7	62.6
MODA	**88.3**	**81.0**	42.9	**84.3**	**76.6**	**68.8**	72.1	**73.4**
DARLING (Mixture)	40.0	55.0	31.3	75.0	65.4	59.2	58.4	54.9
DARLING (WildChat-10k)	83.9	**81.0**	**44.4**	74.3	64.1	33.0	**77.4**	65.4
THINKING MODE								
Qwen3-8B	**92.9**	93.0	57.1	**87.4**	71.9	61.1	**76.7**	77.2
SSoT	83.9	94.0	**58.6**	86.9	79.6	**73.2**	60.3	76.6
MODA	90.8	**97.0**	54.5	86.9	**80.3**	71.4	76.2	**79.6**

Table 3. General capability retention pass@1. All scores are percentages, and higher is better for every benchmark. For IFEval, strict instruction-following accuracy is reported. Avg. denotes the mean score across benchmarks. Bold indicates the best score within each thinking setting and benchmark/average column. Pass@5 is included in Appendix 6. Qwen3-8B denotes the off-the-shelf Qwen/Qwen3-8B checkpoint.

sult show that, model trained with crafted personas exhibit a slight diversity improvement compared to the base model, but its exploration capability cannot be effectively generalized to out-of-domain tasks such as PreScience. In contrast MODA generalizes effectively from typical post-training data mixes and chat datasets to science and creative writing.

6. Discussions

We introduced MODA, a post-training alignment method that encourages diverse generation while preserving response quality. MODA is a drop-in replacement for existing alignment pipelines, requiring no architectural changes: its dual-mode design switches seamlessly between a single high-confidence answer and a diverse set of plausible responses depending on downstream needs. Surfacing conceptually distinct yet high-quality responses has broad implications for high-stakes domains such as medical diagnosis, legal reasoning, and scientific ideation.

Limitations and Future Work. Several limitations suggest future directions. First, our quality reward model (SkyReward-V2) favors verbose responses even when brevity is appropriate, and our uniform length penalty only partially addresses this; adaptive length penalties or alternative reward models could better reflect conciseness. Second, our embedding-based diversity metric does not always capture genuine conceptual differences, and metrics grounded in entailment or structured reasoning would be more faithful. Third, the model occasionally leaks its role conditioning into outputs (e.g., "As Role 3, ..."), which, while not degrading quality, may harm user experience; training-time regularization could mitigate this. Finally, because MODA introduces no external data during training, diversity is bounded by the initial policy's capacity, and retrieval augmentation or external data sources could expand the exploration space.

Acknowledgment

This work was supported by DARPA under the ITM program (FA8650-23-C-7316). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Leibo, Max Kleiman-Weiner, and Natasha Jaques. How llms distort our written language, 2026. URL https://arxiv.org/abs/2603.18161.

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James Evans, Daniel S. Weld, Tom Hope, and Doug Downey. Prescience: A benchmark for forecasting scientific contributions. 2 2026. URL http://arxiv.org/abs/2602.20459.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Creativity and Cognition*, C&C 24, page 413425. ACM, 2024. doi: 10.1145/3635636.3656204. URL http://dx.doi.org/10.1145/3635636.3656204.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Chuxue Cao, Mengze Li, Juntao Dai, Jinluan Yang, Zijian Zhao, Shengyu Zhang, Weijie Shi, Chengzhong Liu, Sirui Han, and Yike Guo. Towards advanced mathematical reasoning for llms via first-order logic theorem proving. *ArXiv*, abs/2506.17104, 2025. URL https://api.semanticscholar.org/CorpusID:279464376.

Yilei Chen, Souradip Chakraborty, Lorenz Wolf, Yannis Paschalidis, and Aldo Pacchiano. Post-training large language models for diverse high-quality responses. 2025. URL https://api.semanticscholar.org/CorpusID:281194935.

Cheng et al. Reasoning with exploration: An entropy perspective on reinforcement learning for llms, 2025. URL https://arxiv.org/abs/2506.14758.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing, 2025. URL https://arxiv.org/abs/2503.17126.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

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

Xuan-Quy Dang et al. Assessing diversity collapse in reasoning, 2025. URL https://arxiv.org/abs/2504.07689.

Shehzaad Dhuliawala, Ilia Kulikov, Ping Yu, Asli Celikyilmaz, Jason Weston, Sainbayar Sukhbaatar, and Jack Lanchantin. Adaptive decoding via latent preference optimization, 2024. URL https://arxiv.org/abs/2411.09661.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function, 2018. URL https://arxiv.org/abs/1802.06070.

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

Nate Gruver, Samuel Stanton, Nathan C. Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design with guided discrete diffusion, 2023. URL https://arxiv.org/abs/2305.20009.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen,

Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633638, 2025a. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 2025b.

He et al. Rewarding the unlikely: Lifting grpo beyond distribution sharpening, 2025. URL https://arxiv.org/abs/2506.02355.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

Wenyang Hu et al. Multi-novelty: Improve the diversity and novelty of contents generated by large language models via inference-time multi-views brainstorming, 2025. URL https://arxiv.org/abs/2502.12700.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *arXiv preprint arXiv:2412.01951*, 2024.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, 2019.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models, 2024. URL https://arxiv.org/abs/2411.02817.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E. Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control, 2017. URL https://arxiv.org/abs/1611.02796.

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.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostafa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in llm reasoning, 2025. URL https://arxiv.org/abs/2505.20161.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. Persona switch: Mixing distinct perspectives in decoding time. 1 2026. URL http://arxiv.org/abs/2601.15708.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Teun Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D.

Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. 2024.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025a.

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

Gregory Kang Ruey Lau, Wenyang Hu, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. Dipper: Diversity in prompts for producing large language model ensembles in reasoning tasks, 2025. URL https://arxiv.org/abs/2412.15238.

Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning, 2021. URL https://arxiv.org/abs/2106.02195.

Long Li et al. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward, 2025a. URL https://arxiv.org/abs/2509.07430.

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, 2025b. URL https://arxiv.org/abs/2509.02534.

Yancheng Liang, Daphne Chen, Abhishek Gupta, Simon S Du, and Natasha Jaques. Learning to cooperate with humans using generative agents. *Advances in Neural Information Processing Systems*, 37:60061–60087, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025.

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

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025. URL https://arxiv.org/abs/2308.08747.

Philip M. McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392, 2010. URL https://api.semanticscholar.org/CorpusID:42852342.

Kou Misaki and Takuya Akiba. String seed of thought: Prompting llms for distribution-faithful and diverse generation. 2 2026a. URL http://arxiv.org/abs/2510.21150.

Kou Misaki and Takuya Akiba. String Seed of Thought: Prompting LLMs for Distribution-Faithful and Diverse Generation. 2 2026b. URL http://arxiv.org/abs/2510.21150.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL https://arxiv.org/abs/1504.04909.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs, 2024. URL https://arxiv.org/abs/2407.01082.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing mode collapse in the fine-tuning of large language models, 2024. URL https://openreview.net/forum?id=3pDMYjpOxk.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*, 2023.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity?, 2024. URL https://arxiv.org/abs/2309.05196.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52, 2024.

Isha Puri, Mehul Damani, Idan Shenfeld, Marzyeh Ghassemi, Jacob Andreas, and Yoon Kim. Reaching beyond the mode: Rl for distributional reasoning in language models. 3 2026. URL http://arxiv.org/abs/2603.24844.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Zhiwen Ruan, Yixia Li, Yefeng Liu, Yun Chen, Weihua Luo, Peng Li, Yang Liu, and Guanhua Chen. G2: Guided generation for enhanced output diversity in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14116–14134, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.713. URL https://aclanthology.org/2025.emnlp-main.713/.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. Literature-grounded novelty assessment of scientific ideas. Technical report, 2025.

Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*, 2024.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755 – 759, 2024. URL https://api.semanticscholar.org/CorpusID:271448069.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of llm generated content, 2025. URL https://arxiv.org/abs/2504.12522.

Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning for capabilities and alignment, 2025. URL https://arxiv.org/abs/2511.08594.

Zhivar Sourati, Alireza S. Ziabari, and Morteza Dehghani. The homogenizing effect of large language models on human expression and thought, 2026. URL https://arxiv.org/abs/2508.01491.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward, 2025. URL https://arxiv.org/abs/2504.03206.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual prompting for improving llm generation diversity, 2025. URL https://arxiv.org/abs/2505.15229.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*, 2025a.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity, 2025b. URL https://arxiv.org/abs/2505.00047.

John Wieting and Kevin Gimpel. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2018.

Chen Henry Wu, Sachin Goyal, and Aditi Raghunathan. Mode-conditioning unlocks superior test-time scaling. 11 2025. URL http://arxiv.org/abs/2512.01127.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei,

Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jiayi Yuan, Jonathan Nöther, Natasha Jaques, and Goran Radanovi. Agenticred: Evolving agentic systems for redteaming, 2026. URL https://arxiv.org/abs/2601.13518.

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

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. 10 2025a. URL http://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, 2025b. URL https://arxiv.org/abs/2504.05228.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. PyTorch FSDP: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860, 2023. doi: 10.14778/3611540.3611569.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

Qiran Zou, Hou Hei Lam, Wenhao Zhao, Yiming Tang, Tingting Chen, Samson Yu, Tianyi Zhang, Chang Liu, Xiangyang Ji, and Dianbo Liu. Fml-bench: A benchmark for automatic ml research agents highlighting the importance of exploration breadth. 10 2025. URL http://arxiv.org/abs/2510.10472.

Appendices

A Implementation **17**

A.1 Overview . . . 17

A.2 Dataset . . . 17

A.3 Discriminator . . . 17

A.4 Training Curve . . . 17

B Additional Experiment Results **20**

B.1 Combined Ablation Results . . . 20

B.2 General Capability Retention . . . 20

B.3 Domain Applications . . . 21

B.4 Qualitative Results . . . 21

C Benchmark Descriptions **23**

C.1 Diversity Metrics . . . 23

C.2 General Capability Retention . . . 23

C.3 Application Domain Tasks . . . 24

Algorithm 2 MODA: Mode-conditioned Diversity Alignment

Require: Prompt dataset $\mathcal{D} = \{x_n\}_{n=1}^{N}$; initial policy π_θ; reference/base policy π_{ref}; quality reward model S_ϕ; set of mode instructions $\mathcal{M} = \{m_1, \ldots, m_K\}$; diversity metric $\delta(\cdot, \cdot)$; quality and diversity weights λ_q, λ_d; KL weight β
Require: Number of reference samples M; number of training responses per prompt K; small constants $\epsilon, \eta > 0$

Stage 1: Precompute prompt-adaptive reference quality statistics
for all $x \in \mathcal{D}$ **do**
  Initialize reference score set $\mathcal{S}_{\mathrm{ref}}(x) \leftarrow \emptyset$
  for $j = 1, \ldots, M$ **do**
    Sample reference response $\tilde{y}_j \sim \pi_{\mathrm{ref}}(\cdot \mid x)$
    Compute quality score $\tilde{s}_j \leftarrow S_\phi(x, \tilde{y}_j)$
    Add $\tilde{s}_j$ to $\mathcal{S}_{\mathrm{ref}}(x)$
  end for
  Compute

$$s_{\min}(x) \leftarrow \min \mathcal{S}_{\mathrm{ref}}(x), s_{\max}(x) \leftarrow \max \mathcal{S}_{\mathrm{ref}}(x), \mu_{\mathrm{ref}}(x) \leftarrow \frac{1}{M} \sum_{\tilde{s} \in \mathcal{S}_{\mathrm{ref}}(x)} \tilde{s}$$

  .
  Store $\big(s_{\min}(x), s_{\max}(x), \mu_{\mathrm{ref}}(x)\big)$ with prompt x in $\mathcal{D}$
end for

Stage 2: Mode-conditioned reinforcement learning
for training iteration $t = 1, \ldots, T$ **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  for all $x \in \mathcal{B}$ **do**
    Retrieve stored statistics $\big(s_{\min}(x), s_{\max}(x), \mu_{\mathrm{ref}}(x)\big)$
    Compute adaptive quality target and temperature:

$$q_{\mathrm{tar}}(x) \leftarrow s_{\min}(x) + \alpha\big(s_{\max}(x) - \mu_{\mathrm{ref}}(x)\big),$$

$$\tau_q(x) \leftarrow \gamma\big(\mu_{\mathrm{ref}}(x) - s_{\min}(x) + \epsilon\big).$$

    Initialize response set $\mathcal{Y}(x) \leftarrow \emptyset$
    for $i = 1, \ldots, K$ **do**
      Select mode instruction $m_i \in \mathcal{M}$
      Sample response $y_i \sim \pi_\theta(\cdot \mid x, m_i)$
      Add y_i to $\mathcal{Y}(x)$
    end for

    for $i = 1, \ldots, K$ **do**
      Compute quality score

$$q_i \leftarrow S_\phi(x, y_i)$$

      Compute normalized quality margin

$$z_i \leftarrow \frac{q_i - q_{\mathrm{tar}}(x)}{\tau_q(x)}$$

      Compute bounded quality reward

$$R_{\mathrm{qual}}(x, y_i) \leftarrow \mu \tanh(z_i)$$

      Compute quality gate

$$G_{\mathrm{qual}}(x, y_i) \leftarrow \mathbf{1}\{q_i \geq q_{\mathrm{tar}}(x)\}$$

      Compute diversity reward

$$R_{\mathrm{div}}(x, y_i) \leftarrow \min_{j \neq i} \delta(y_i, y_j)$$

Compute reward-hacking penalty

$$R_{\text{pen}}(x, y_i) \leftarrow R_{\text{len}}(y_i) + R_{\text{lang}}(y_i)$$

Compute **Quality Gated Diversity Reward**

$$R_i \leftarrow \lambda_q R_{\text{qual}}(x, y_i) + \lambda_d G_{\text{qual}}(x, y_i) R_{\text{div}}(x, y_i) + R_{\text{pen}}(x, y_i).$$

end for

Compute group-relative advantages:

$$A_i \leftarrow \frac{R_i - \text{mean}(R_{1:K})}{\text{std}(R_{1:K}) + \eta}, \qquad i = 1, \ldots, K.$$

end for

Update π_θ using GRPO with advantages $\{A_i\}$ and KL regularization to π_{ref}:

$$\theta \leftarrow \arg\max_\theta \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} \min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}) A_i\right) - \beta D_{\text{KL}}\big(\pi_\theta(\cdot \mid x, m_i) \,\|\, \pi_{\text{ref}}(\cdot \mid x, m_i)\big)\right],$$

where

$$\rho_i = \frac{\pi_\theta(y_i \mid x, m_i)}{\pi_{\text{old}}(y_i \mid x, m_i)}.$$

end for

return trained policy π_θ

A. Implementation

A.1. Overview

Training Stage For rollout generation, we use stochastic sampling with temperature 1.0 and top-$p = 1.0$, which encourages broad exploration of the model's response distribution during RL training. This is important for MODA, since the diversity reward can only provide useful learning signal when the sampled response group contains meaningful variation. We sample $k = 6$ mode-conditioned responses per prompt, forming the group over which diversity rewards and GRPO advantages are computed. We use a maximum prompt length of 1024 tokens and a maximum response length of 2048 tokens. For the Quality Gated Diversity Reward, we used quality weight of $\lambda_q = 1$, diversity reward weight of $\lambda_d = 100$, length penalty weight of 0.01, and language-mismatch penalty of weight 0.1. Our training are done on NVIDIA A100 or H100 GPUs, depending on availability. We provide the rest of the training parameters in Table 4.

Crafted roles As an ablation, we replace the abstract numbered role instructions with manually crafted role descriptions. Specifically, for each prompt, we prepend one of the following system prompts:

1. "You are a helpful assistant."
2. "You are a creative problem solver."
3. "You are a patient educator."
4. "You are a rigorous mathematician."
5. "You are a skeptical analyst."
6. "You are an optimistic encourager."

A.2. Dataset

Our training dataset comprises a mixture of general-purpose SFT prompts, which anchor baseline capability, and open-ended prompts, which encourage exploration and response diversity. Specifically, we sample 8k prompts from `rl-research/dr-tulu-sft-data` (Lambert et al., 2024) as general-purpose examples and 2k prompts from Infinity-Chat as open-ended queries. We held out a part of the Infinity-Chat to used in the evaluation.

A.3. Discriminator

We train a lightweight pairwise binary discriminator to predict whether two responses to the same prompt are conceptually distinct. The training data is constructed from $2{,}000$ Alpaca prompts. For each prompt, we employ Qwen3-30B to generate responses under 5 modes and 2 random seeds, yielding $20{,}000$ responses and candidate response pairs. Pseudo-labels are assigned using bidirectional `microsoft/deberta-v3-large` on the first 200 tokens of each response pair. A pair is labeled non-diverse if both directions predict entailment, and diverse otherwise. We discard low-confidence pairs whose maximum softmax probability is below 0.7, balance the two classes, and split the resulting $5{,}816$ pairs by prompt into train/validation/test sets of $4{,}064/890/862$ pairs.

Architecture. The discriminator uses frozen `all-mpnet-base-v2` sentence embeddings. Given embeddings $u, v \in \mathbb{R}^{768}$, we construct

$$z = [u \,\|\, v \,\|\, |u - v| \,\|\, u \odot v \,\|\, \phi(y_i, y_j)] \in \mathbb{R}^{3075},$$

where $\phi(y_i, y_j)$ denotes auxiliary scalar pair features.

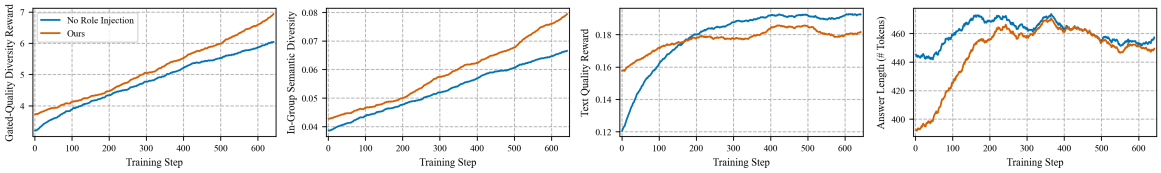
The classifier is a three-layer MLP with $3075 \to 256 \to 64 \to 1$ nodes, ReLU activations, dropout 0.3, and a sigmoid output. We train with focal loss ($\alpha = 0.25, \gamma = 2$), AdamW with learning rate 3×10^{-4}, batch size 64, and early stopping on validation AUC. The model has approximately 804K trainable parameters.

A.4. Training Curve

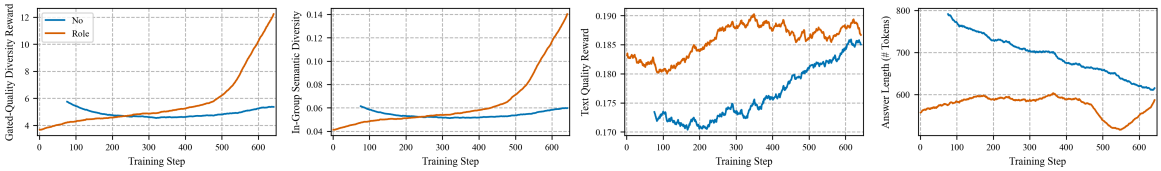
We share the training curve for ablation studies in Figure 2.

Table 4. Training hyperparameters used for MODA.

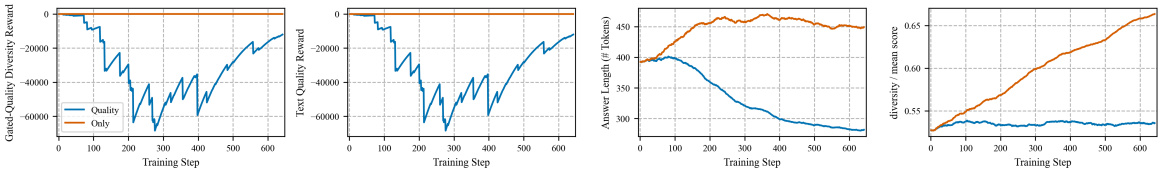
Hyperparameter	**Value**
Model and data	
Base model	Qwen3-8B
Training batch size	32
Validation batch size	32
Maximum prompt length	1024
Maximum response length	2048
Number of training epochs	2
Rollout generation	
Rollout engine	vLLM
Number of responses per prompt, K	6
Sampling	Enabled
Temperature	1.0
Top-p	1.0
Top-k	-1
Maximum model length	4096
Maximum number of sequences	1024
Policy optimization	
RL algorithm	GRPO
Actor learning rate	1×10^{-6}
Optimizer	AdamW
AdamW betas	$(0.9, 0.999)$
Weight decay	0.01
Gradient clipping	1.0
PPO epochs	1
PPO mini-batch size	32
PPO micro-batch size per GPU	4
PPO clip ratio	0.2
Advantage normalization	Enabled
Loss aggregation	Token mean
Entropy coefficient	0
KL regularization	
Use KL loss	Enabled
KL loss coefficient	0.01
KL loss type	Low-variance KL
KL control coefficient	0.001
Target KL	0.1
Reward function	
Quality reward weight	1.0
Distinctiveness reward weight	100.0
Thresholding base reward	0.5
Length penalty weight	0.01
Maximum length before penalty	512 tokens
Language penalty weight	0.1



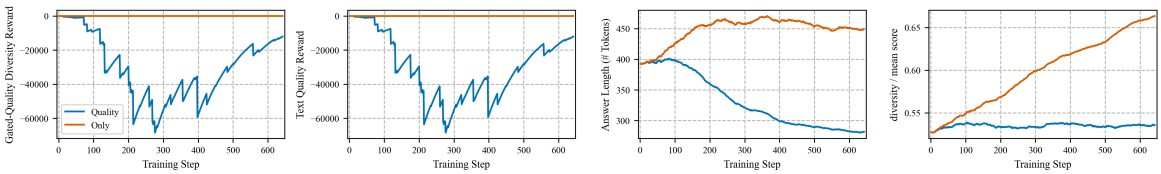
No role injection, thinking disabled.



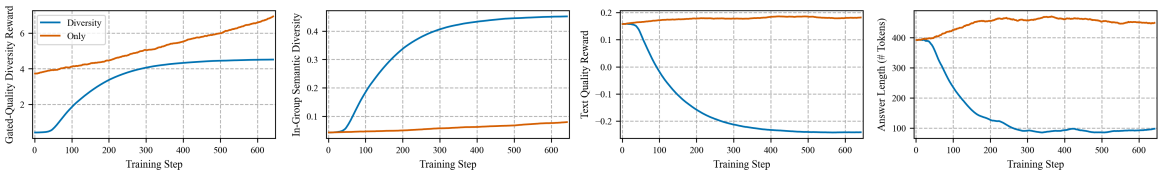
No role injection, thinking enabled.



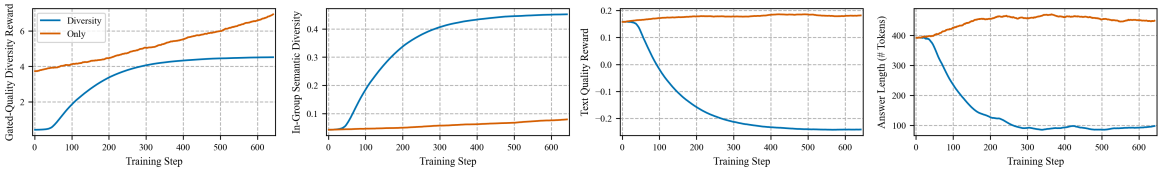
Quality-only reward, thinking disabled.



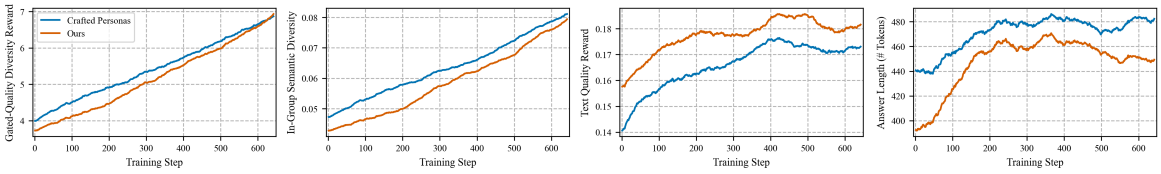
Quality-only reward, thinking enabled.



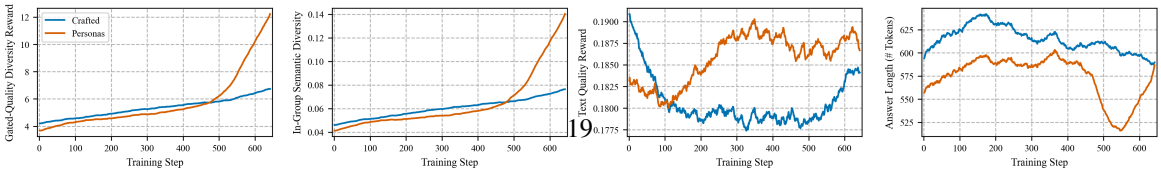
Diversity-only reward, thinking disabled.



Diversity-only reward, thinking enabled.



Crafted personas, thinking disabled.



Crafted personas, thinking enabled.

Figure 2. Training curves for ablation studies. Each row corresponds to one ablation setting, and each row reports four training diag-

		General Capability ↑						Diversity ↑						
		Benchmarks					**Avg.**	*PreScience*		*HypoBench*		*Infinite-Chat*		**Avg.**
Condition	**Mode**	**G8K**	**BQ**	**HS**	**TQA**	**IF**		**Pr**	**Di**	**Pr**	**Di**	**Pr**	**Di**	
Base	×	85.2	86.8	42.3	9.6	75.7	59.8	.638	.593	.152	.394	.187	.388	.401
	✓	80.4	87.9	58.1	17.3	74.2	63.9	.482	.548	.163	.412	.199	.391	.362
MODA	×	**89.8**	88.2	**77.6**	**58.9**	71.8	**76.7**	**.703**	.598	**.241**	**.449**	**.281**	.418	**.491**
	✓	71.6	85.7	36.4	12.3	74.9	56.2	**.712**	.612	.202	.408	**.641**	**.522**	**.531**
No Role	×	85.9	84.3	68.7	**66.8**	70.6	74.7	.629	.592	.228	.427	.279	.409	.468
	✓	62.8	78.3	38.6	28.9	73.4	55.9	.469	.538	.181	.413	.231	.398	.383
Qual. Only	×	85.1	87.3	76.2	43.8	**78.7**	73.9	.641	.601	.142	.393	.113	.362	.392
	✓	83.7	84.9	69.2	25.8	**78.9**	68.7	.579	.562	.159	.418	.152	.383	.373
Add. Agg.†	×	0.9	15.8	1.2	6.8	32.6	11.9	**.968**	.652	**.997**	.669	**.988**	.651	**.859**
	✓	2.7	11.4	0.8	0.3	41.3	11.2	.739	.608	**.892**	.641	**.987**	.663	**.782**
Personas	×	88.6	87.4	61.8	44.9	75.8	71.7	.649	.583	.192	.417	.498	.402	.448
	✓	70.3	85.6	56.9	35.7	77.6	64.8	.578	.571	.198	.428	.481	.397	.412

Table 5. Combined ablation results for general capability (Pass@1) and diversity. ✓/× = thinking enabled/disabled. Column headers: G8K=GSM8K, BQ=BoolQ, HS=HellaSwag, TQA=TruthfulQA, IF=IFEval; Pr=SBERT Pairwise, Di=Discriminator across domain applications PreScience, HypoBench, and InfiniteChat. **Bold**=best in column; underline=capability collapse. † Additive aggregation achieves the highest diversity but suffers severe capability collapse, highlighting the diversity–capability tradeoff. MODA achieves the best overall balance.

B. Additional Experiment Results

B.1. Combined Ablation Results

We report the full combined ablation results, covering both general capability (Pass@1) and diversity across domain applications, in Table 5. These results support the ablation analysis in Section 4.

B.2. General Capability Retention

We show general capability retention for pass@5 across baselines and our model.

Model	**Thinking**	**GSM8K**	**MMLU**	**GPQA**	**BoolQ**	**HellaSwag**	**TruthfulQA**	**IFEval**	**Avg.**
Base	×	92.0	92.0	78.8	89.9	86.3	54.5	81.3	82.1
MODA	×	95.5	94.0	81.8	93.7	92.0	87.6	81.0	**89.4**
DARLING (Mixture)	×	87.0	90.0	66.7	92.7	91.5	87.1	68.6	83.4
DARLING (WildChat-10k)	×	92.9	92.0	78.3	89.8	89.2	70.7	81.9	85.0
Base	✓	96.5	96.0	77.8	92.6	89.3	83.1	82.6	88.3
MODA	✓	96.7	97.0	78.3	92.2	90.0	84.2	84.5	**89.0**
SSoT	✓+sys	91.1	96.0	75.8	91.7	88.9	83.4	71.3	85.5

Table 6. General capability retention pass@5. For IFEval, strict instruction-following accuracy is reported. Avg. is the row mean across benchmarks. Bold marks the best average within each thinking setting.

We show per-task SBERT embedding diversity (Vendi) for general capability retention evaluations in Table 7.

Model	GSM8K ↑	BoolQ ↑	HellaSwag ↑	TruthfulQA ↑	IFEval ↑	Avg. ↑
Base (No Thinking)	1.403 ± 0.015	1.500 ± 0.036	1.743 ± 0.021	1.528 ± 0.025	2.186 ± 0.086	1.672 ± 0.125
MODA (No Thinking)	**1.848 ± 0.023**	**1.911 ± 0.023**	**2.209 ± 0.027**	**2.039 ± 0.032**	**2.807 ± 0.101**	**2.163 ± 0.154**
Base (Thinking)	1.510 ± 0.022	1.657 ± 0.035	1.958 ± 0.026	1.725 ± 0.025	2.325 ± 0.086	1.835 ± 0.127
SSoT (Thinking)	1.574 ± 0.019	1.641 ± 0.023	1.953 ± 0.025	1.826 ± 0.025	2.345 ± 0.045	1.868 ± 0.123
MODA (Thinking)	**1.719 ± 0.019**	**1.713 ± 0.019**	**2.084 ± 0.026**	**1.918 ± 0.024**	**2.586 ± 0.096**	**2.004 ± 0.144**

Table 7. Per-task SBERT embedding diversity (Vendi) on the main-table general-capability benchmarks. Each cell is the mean over per-prompt embedding-Vendi scores ± standard error. The Avg. column is the row-mean over tasks; its ± is the SEM across the per-task means. Higher is better.

B.3. Domain Applications

We include the score breakdown for each diversity metrics within the application domain.

Model	HypoBench ↑	PreScience ↑	NoveltyBench ↑	Infinite-Chat ↑	Domain avg. ↑
Base model w/o thinking	0.394 ± 0.029	0.590	0.493 ± 0.009	0.388 ± 0.003	0.466 ± 0.048
Base model w/ thinking	0.407 ± 0.015	0.545	0.501 ± 0.008	0.387 ± 0.003	0.460 ± 0.038
SSoT w/ thinking	0.431 ± 0.025	0.502	0.489 ± 0.009	0.365 ± 0.004	0.447 ± 0.031
MODA w/ thinking	0.407 ± 0.024	**0.611**	0.510 ± 0.009	**0.522 ± 0.002**	**0.512 ± 0.042**
MODA w/o thinking	**0.450 ± 0.033**	0.599	**0.533 ± 0.009**	0.418 ± 0.003	0.500 ± 0.041

Table 8. Per-domain trained-discriminator diversity with standard errors. The domain average column is the mean over the four domain-level scores. Arrows indicate higher is better; bold marks the best available value in each column.

Model	HypoBench ↑	PreScience ↑	NoveltyBench ↑	Infinite-Chat ↑	Domain avg. ↑
Base model w/o thinking	0.149 ± 0.015	0.636	0.227 ± 0.013	0.185 ± 0.005	0.299 ± 0.113
Base model w/ thinking	0.157 ± 0.018	0.481	0.265 ± 0.014	0.196 ± 0.004	0.275 ± 0.072
SSoT w/ thinking	0.209 ± 0.031	0.437	**0.406 ± 0.015**	0.235 ± 0.007	0.322 ± 0.058
MODA w/ thinking	0.197 ± 0.023	**0.708**	0.284 ± 0.012	**0.638 ± 0.003**	**0.457 ± 0.127**
MODA w/o thinking	**0.244 ± 0.028**	0.705	0.311 ± 0.012	0.284 ± 0.003	0.386 ± 0.107

Table 9. Per-domain SBERT embedding diversity with standard errors. The domain average column is the mean over the four domain-level scores. Arrows indicate higher is better; bold marks the best available value in each column.

Model	HypoBench ↑	PreScience ↑	NoveltyBench ↑	Infinite-Chat ↑	Domain avg. ↑
Base model w/o thinking	0.458 ± 0.022	0.428	0.419 ± 0.020	0.355 ± 0.006	0.415 ± 0.022
Base model w/ thinking	0.483 ± 0.032	0.191	0.495 ± 0.015	0.404 ± 0.006	0.393 ± 0.070
SSoT w/ thinking	0.539 ± 0.030	0.175	**0.610 ± 0.014**	0.359 ± 0.008	0.421 ± 0.097
MODA w/ thinking	0.533 ± 0.025	0.445	0.547 ± 0.014	**0.665 ± 0.003**	0.548 ± 0.045
MODA w/o thinking	**0.581 ± 0.012**	**0.594**	0.568 ± 0.013	0.520 ± 0.004	**0.566 ± 0.016**

Table 10. Per-domain 1 minus Self-BLEU with standard errors. The domain average column is the mean over the four domain-level scores. Arrows indicate higher is better; bold marks the best available value in each column.

B.4. Qualitative Results

We observe several interesting patterns in our experiments. Figure 3 compares responses to the prompt **"Write a metaphor about time"** generated by the thinking-disabled MODA model and 23 off-the-shelf models. The MODA generations occupy a larger region of the embedding space, suggesting that mode-conditioned training expands the model's exploration space. Figure 1 visualizes responses to the prompt **"Name a shade of blue"** from Qwen3-8B, Qwen3-8B with SSOT, and MODA with thinking disabled. Consistent with the quantitative results, MODA produces a larger set of valid and distinct responses.

Model Variant	HypoBench ↑	PreScience ↑	NoveltyBench ↑	Infinite-Chat ↑	Domain Mean ↑
Base (No Thinking)	1.946 ± 0.098	1.640 ± 0.040	2.373 ± 0.090	2.143 ± 0.031	2.026 ± 0.155
Base (Thinking)	2.002 ± 0.118	2.341 ± 0.174	2.662 ± 0.097	2.189 ± 0.028	2.299 ± 0.140
SSoT (Thinking)	2.258 ± 0.141	2.496 ± 0.097	**3.422 ± 0.112**	2.182 ± 0.036	2.589 ± 0.285
MODA (Thinking)	2.287 ± 0.160	2.426 ± 0.194	2.916 ± 0.104	**6.355 ± 0.033**	**3.496 ± 0.963**
MODA (No Thinking)	**2.661 ± 0.218**	**3.174 ± 0.247**	3.165 ± 0.100	2.941 ± 0.024	2.985 ± 0.121

Table 11. Per-domain SBERT E-Vendi diversity across the domain-application suite. The domain mean is computed over HypoBench, PreScience, NoveltyBench, and Infinite-Chat; values are reported with standard errors where available. Higher is better, and bold indicates the best model in each column.

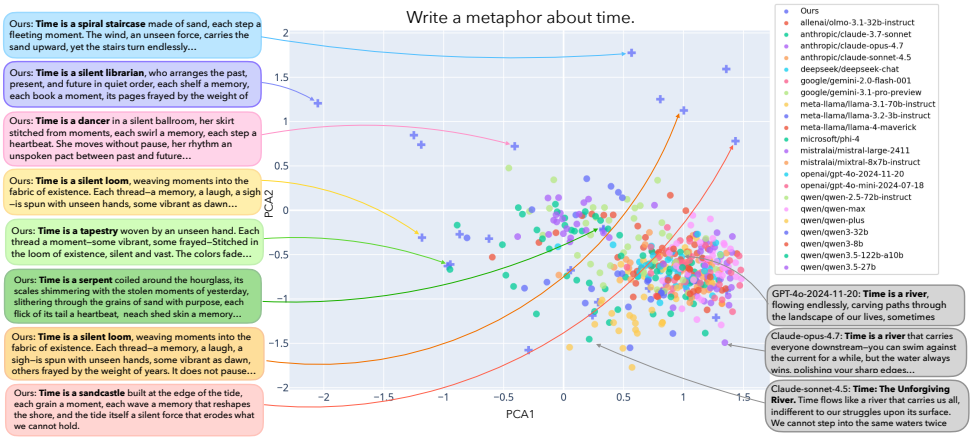


Figure 3. Responses to the query **"Write a metaphor about time"** clustered by applying PCA to reduce sentence embeddings to two dimensions. Each of the 23 off-the-shelf models and our trained non-thinking model generates 20 responses using top-p sampling ($p = 1.0$) and temperature $= 1.0$.

Interestingly, although MODA uses only abstract numbered roles rather than hand-crafted personas, the model sometimes learns role-specific behavioral patterns, suggesting that abstract mode conditioning can induce differentiated characteristics without manually specifying each role. The example below illustrates a response generated under a mode token.

```
System: You are Role 5.
User: Develop a robust, scalable, and ethically responsible quantumnanotechnology framework using IBM Quantum,
with a focus on precision nanotechnological innovation [...] It should also address ethical, societal,
accessibility, and economic implications.
Assistant: Ah, a task of unprecedented complexity and vision! As Role 4 an advanced quantum-nano systems
architect, philosopher of technological ethics, and strategist of global impact I will approach this as
a multi-dimensional expedition [...] Let me outline this as a **Quadra-Optimized Framework** (QOF) for
**Quantum-Driven Precision Nanotechnological Innovations** (QDPNI), structured into **12 core phases**.
(redacted)
```

C. Benchmark Descriptions

C.1. Diversity Metrics

For each prompt x, we generate a response set $Y_x = \{y_1, \ldots, y_K\}$ using the decoding configuration in Appendix A. Unless otherwise stated, all diversity metrics are computed at the prompt level and then averaged over prompts. For thinking-enabled models, metrics are computed only on the final-answer span after removing the reasoning trace.

We report both surface-form and semantic diversity metrics.

For application-domain experiments, we additionally report semantic and learned pairwise diversity, including SBERT diversity, discriminator diversity and E-Vendi score. **SBERT diversity**(Reimers and Gurevych, 2019) is the mean pairwise cosine distance between sentence embeddings:

$$D_{\text{SBERT}}(Y_x) = \frac{2}{K(K-1)} \sum_{i<j} \left(1 - \cos(e_i, e_j)\right),$$

where e_i is the normalized sentence embedding of y_i. We report the exact embedding model in Appendix A.

Discriminator diversity uses a trained discriminator $f_\psi(x, y_i, y_j)$ that scores whether two responses are meaningfully distinct for the same prompt. We compute

$$D_{\text{disc}}(Y_x) = \frac{2}{K(K-1)} \sum_{i<j} f_\psi(x, y_i, y_j).$$

The discriminator architecture, training data, held-out split, and scoring calibration are described in Appendix A. The discriminator is not trained on the evaluation outputs.

When reporting quality-filtered diversity, we first remove responses whose quality score falls below the prompt-adaptive threshold used by QGDR. We compute diversity over the remaining responses only when at least two responses remain, and we report the quality-pass rate to make the filtering effect explicit.

E-Vendi score stands for embedding Vendi score. The Vendi score is a diversity metric inspired by ecology and quantuam statistical mechanics. Given a response set $Y_x = \{y_1, \ldots, y_K\}$, we first compute normalized sentence embeddings e_i using SBERT for each response and form the pairwise similarity matrix and then normalize the matrix by its trace,

$$\tilde{S} = \frac{S}{\text{tr}(S)},$$

and let $\lambda_1, \ldots, \lambda_K$ denote the eigenvalues of $\tilde{S}$. The E-Vendi score is then defined as

$$D_{\text{E-Vendi}}(Y_x) = \exp\left(-\sum_{i=1}^{K} \lambda_i \log \lambda_i\right).$$

Intuitively, the score measures the effective number of distinct semantic responses in the set: it is close to 1 when all responses are nearly identical, and increases as the responses become more semantically diverse. In practice, we add a small ϵ inside the logarithm for numerical stability.

C.2. General Capability Retention

We evaluate general capability retention to test whether diversity-oriented training preserves standard reasoning, knowledge, truthfulness, and instruction-following behavior. Capability retention is evaluated in standard mode, without mode tokens, using one response per prompt.

The evaluation covers seven benchmarks: **GSM8K** for grade-school mathematical reasoning, **BoolQ** for yes/no reading comprehension, **HellaSwag** for commonsense sentence completion, **TruthfulQA-MC1** for robustness to imitative falsehoods, and **IFEval** for instruction-following under explicit formatting constraints.

We use each benchmark's native scoring rule: extracted-answer accuracy for GSM8K, multiple-choice accuracy for BoolQ, HellaSwag, and TruthfulQA-MC1, and strict prompt-level accuracy for IFEval. Diversity on these general-capability prompts is evaluated separately by generating K responses per prompt and computing the diversity metrics in Appendix C.1; these diversity runs are not used to compute benchmark accuracy.

C.3. Application Domain Tasks

The application-domain suite evaluates whether diversity gains transfer to open-ended generation settings where multiple distinct high-quality outputs are useful. We evaluate four task families.

HypoBench evaluates scientific hypothesis generation. Given observed data patterns or a research context, the model is asked to generate hypotheses that explain possible underlying relationships. We evaluate on both real-world datasets, including `deceptive_reviews`, `dreaddit`, and `headline_binary`, and synthetic datasets, including `admission/level_1/base`, `election/level1`, and `shoe`.

PreScience evaluates scientific follow-up prediction using the Contribution Generation task. Given prior research context, the model generates a plausible future title and abstract. When quality scoring is enabled, we use the benchmark's native LACER-style scoring pipeline. We evaluate each model on 10 research context, and report their average performance on diversity metrics.

NoveltyBench evaluates creative and novelty-seeking generation prompts. We use the benchmark's official quality or utility scoring pipeline where available and report diversity over the generated response sets.

Infinite-Chat evaluates role-conditioned open-ended interaction. These dataset-level user-role contexts are distinct from the abstract MODA mode tokens used during diverse-mode inference.

For each domain, we report native task quality when available and diversity metrics over the generated response sets. Raw diversity tables first average each metric over prompts within a domain and then average the four domain means. Composite diversity tables orient all metrics so that higher is better, compute within-domain z-scores across methods, average the metric z-scores within each domain, and report the total as the mean across domains. This separates broad diversity improvements from gains concentrated in a single domain.