FORTY SHADES OF BLUE: QUALITY-DIVERSITY ALIGNMENT VIA MODE-CONDITIONED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

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 algorithm that jointly optimizes generation quality and diversity, inspired by the coordination perspective in multi-agent reinforcement learning (MARL). MoDA trains 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 formulation encourages mode-conditioned agents to explore complementary regions of the high-quality output space by employing a prompt-adaptive quality gating mechanism that calibrates quality thresholds against a frozen reference policy. We evaluate MoDA on a comprehensive suite of seven general capability tasks and four domain-specific diversity tasks in scientific ideation and creative writing. MoDA improves SBERT diversity by **265%** on the Infinite-Chat held-out prompts, while increasing general capability pass@1 by **10.3%** over the Qwen3-8B baseline. Compared with the strongest DivPO baseline, MoDA improves SBERT diversity from 0.274 to 0.482 (**+75.9%**) and E-Vendi from 2.86 to 4.4 (**+53.8%**), while improving average general capability pass@1 by **7.0%**. Overall, MoDA provides 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 & Potts, 2025; 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 (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 & Akiba, 2026; 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

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.

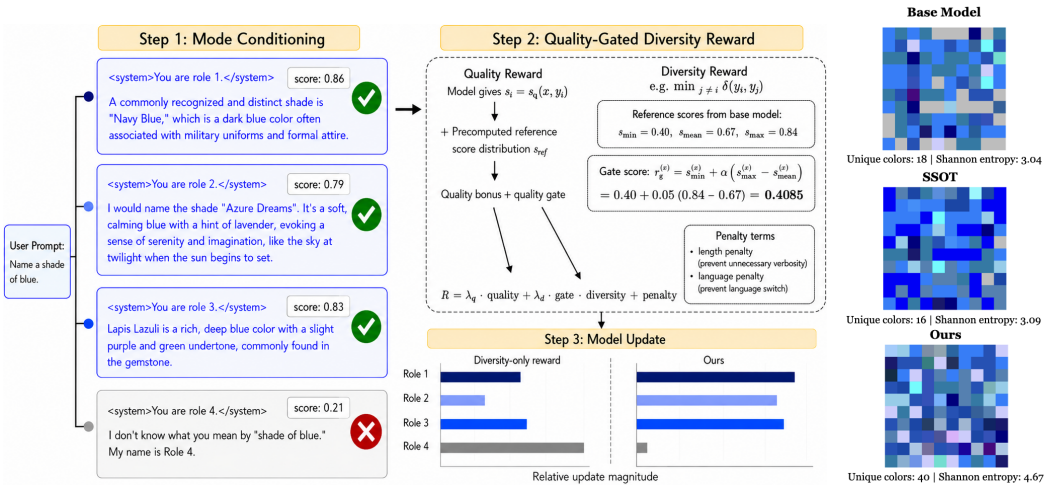


Figure 1: **Left**: Overview of MoDA. MoDA uses mode conditioning to elicit diverse candidate responses. Each candidate is labeled with its quality score s_i relative to the prompt-adaptive quality threshold $\tau_q^{(x)}$. Low-quality candidates with $s_i < \tau_q^{(x)}$ have $g(x, y_i) = 0$ and therefore receive no diversity credit. Matching colors track the same response through Steps 1–3. **Right**: 100 responses to the prompt **"Name a shade of blue"** from the Qwen3-8B initial policy, String Seed of Thought (SSoT), and MoDA under diverse-mode inference. MoDA produces 40 distinct shades of blue, 122% more than the Qwen3-8B initial policy and 150% more than SSoT.

diversity reward. This creates competitive dynamics among roles, encouraging them to specialize in distinct regions of the high-quality output 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, where models maximize diversity by producing unusual but irrelevant outputs (Li 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 and four open-ended domain application benchmarks spanning scientific ideation and creative writing. We evaluate output diversity using complementary metrics: *semantic-level* (SBERT), *entropy-level* (E-Vendi), and *entailment-level* (a learned pairwise discriminator). Under the same thinking-disabled setting, MoDA improves SBERT diversity and E-Vendi by 155.5% and 94.0%, respectively, over the Qwen3-8B initial policy on the diversity-focused benchmark suite, while simultaneously increasing average pass@1 by 10.3 percentage points over the initial policy, and by 7.0 percentage points over the strongest DivPO baseline on general capability benchmarks.

Compared with prior diversity-oriented post-training methods (Li et al., 2025; Lanchantin et al., 2025; Chung et al., 2025; Chen et al., 2025), 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.

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). In scientific reasoning,

automated research agents benefit from a wider exploration breadth (Zou et al., 2026), 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 & He, 2023). In mathematical proofs, strategy diversity is critical for exploring 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 & Jarvis, 2010), and compression-based homogeneity scores (Shaib et al., 2024). Semantic metrics build on sentence embeddings (Reimers & Gurevych, 2019; Wieting & Gimpel, 2018) and information-theoretic constructions such as the Vendi Score (Friedman & 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 & He, 2023; West & Potts, 2025; Dang et al., 2025), and recent benchmarks evaluate humanlike or effective semantic diversity directly (Zhang et al., 2025b; Shypula 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; Lagzian 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., 2025; Lanchantin et al., 2025; Chung et al., 2025; Chen et al., 2025; Slocum et al., 2025; Puri et al., 2026) (see Appendix B for a comparison table). At inference time, Verbalized Sampling (Zhang et al., 2025a) prompts models to express a distribution over possible responses, while String Seed of Thought (Misaki & Akiba, 2026) injects random strings as entropy sources. DIPPER (Lau et al., 2025) and Multi-Novelty (Lagzian et al., 2025) construct diverse prompt ensembles, and multilingual, persona, or chain-of-thought prompting can elicit variations (Wang et al., 2025; Wu et al., 2025). At decoding time, min-p sampling (Nguyen et al., 2024) adaptively truncates low-probability tokens based on model confidence, while G2 (Ruan et al., 2025) and adaptive temperature methods (Dhuliawala et al., 2024) guide generation with auxiliary diversity modules or learned per-instance decoding parameters. In post-training, RL has emerged as a powerful method for enhancing diversity. Multi-answer RL (Puri et al., 2026) trains models to output multiple diverse answers within one response; Soft Preference Learning decouples entropy from KL-penalty to improve lexical and semantic variety (Slocum et al., 2025); DivPO (Lanchantin et al., 2025) applies DPO to responses that are both diverse and high quality.

Several recent works share our goal of jointly optimizing quality and diversity through RL. DARLING (Li et al., 2025) balances quality and diversity by rewarding a multiplicative aggregation of quality and diversity metrics; DQO (Chen et al., 2025) uses a determinant-based group diversity reward, but assigns the same reward to all responses in the group. GRPO-Unlikeliness (He et al., 2025) and GAPO (Anschel et al., 2025) both encourage diversity through frequency- and likelihood-based rewards. However, none of them allow a single model to switch between producing one high-quality answer and a diverse set of responses. MoDA, on the other hand, 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.

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., 2025). However, optimizing only the expected reward often leads to *mode collapse*, where the learned policy focuses on a small set of high-reward responses while suppressing equally valid alternatives (Kirk et al., 2024). 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 $\max_\theta J_{\mathrm{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} r(x,y)\right]$, 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 $H(\pi_\theta(\cdot \mid x)) = -\sum_{y\in\mathcal{S}} \pi_\theta(y \mid x) \log \pi_\theta(y \mid x)$ 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 Multi-Agent Reinforcement Learning (MARL)-Inspired Mode Conditioning

MARL Motivation. Classical MARL studies how populations of agents can learn complementary behaviors by explicitly encouraging behavioral diversity through mechanisms such as information sharing, skill specialization, KL divergence maximization, and adversarial objectives (Eysenbach et al., 2018; Liang et al., 2024; Li et al., 2021). We bring this perspective to LLM post-training by introducing *mode conditioning*, where a single policy is conditioned on different role tokens and each role is treated as an abstract agent. Agents must learn how to produce responses that are unlike those of the other agents, introducing competitive dynamics that drive the agents to continually diversify their responses. Unlike standard alignment, mode conditioning provides a *richer interface* that enables the model to internalize multiple high-quality behavioral modes.

Mode Conditioning via System Role Identifiers. We instantiate mode conditioning by prepending role identifiers into the system prompt. In our primary setting, we adopt an intentionally minimal identity-based conditioning: `You are role` i. We use numbered roles instead of hand-crafted personas to avoid injecting domain-specific assumptions about desirable perspectives. Although our ablations show that crafted personas (e.g., "innovative thinker," "passionate educator," or "methodical experimenter") and token-based conditioning (e.g. "Start with your response with APPLE / BANANA / ORANGE.") are also effective, we adopt the minimal identity-based conditioning to encourage emergent specialization.

Formally, let $x \in \mathcal{X}$ denote a user prompt and $\mathcal{Z} = z_1, \ldots, z_N$ a fixed set of role identifiers. A shared policy $\pi_\theta(y \mid z_i, x)$ is conditioned on role z_i, with all roles sharing parameters θ. During training, each role generates one response for K prompts within the batch, forming a response group $\mathcal{Y}_x = \{y_{i,k} : i = 1, \ldots, N; k = 1, \ldots, K\}$. Each response receives a Quality-Gated Diversity Reward (§ 3.3), and the shared policy is optimized with Group Relative Policy Optimization (GRPO; (Shao et al., 2024)). Algorithm 1 summarizes the training procedure.

Dual Inference Settings. Not all prompts benefit from diverse responses. For example, factual queries such as "*What is Albert Einstein's birthday?*" have a single correct answer. A key advantage of mode conditioning is that omitting the role instruction recovers the model's standard behavior, allowing MoDA to support both inference modes. Under the **standard mode**, the model generates a single response without a role identifier. Under the **diverse mode**, we prompt the model under multiple role identifiers to produce a candidate set $y_1, \ldots, y_N$. The resulting set can be returned directly, ranked by a quality model, or aggregated by a downstream model into a final answer.

3.3 Quality-Gated Diversity Reward

Preserving quality while increasing diversity is critical: a diversity-only objective is vulnerable to reward hacking; it can be exploited by irrelevant, malformed, unnecessarily verbose, or superficially different responses. We therefore optimize a four-component reward consisting of a quality reward, a prompt-adaptive quality gate, a response-level diversity score, and explicit reward-hacking penalties:

$$r_i = \underbrace{\lambda_q r_{\mathrm{qual}}(x, y_i)}_{\text{quality reward}} + \underbrace{\lambda_d g(x, y_i)}_{\text{quality gate}} \underbrace{d(x, y_i, \mathcal{Y})}_{\text{diversity reward}} + \underbrace{r_{\mathrm{len}}(y_i) + r_{\mathrm{lang}}(x, y_i) + r_{\mathrm{fmt}}(y_i)}_{\text{penalties}}. \tag{1}$$

Prompt-Adaptive Quality Scoring. To determine whether our training preserves the original model's capability, we need a reference baseline: how the initial policy perform on the same prompt. To instantiate this baseline, for each prompt x, we score five responses sampled from the frozen initial policy and denote their minimum, mean, and maximum scores by $s_{\min}^{(x)}$, $\bar{s}^{(x)}$, and $s_{\max}^{(x)}$. Let $s_i = S_\phi(x, y_i)$ be the reference score assigned by the reward model. We define the prompt-adaptive threshold τ_q and quality reward magnitude factor σ_q as

$$\tau_q^{(x)} = s_{\min}^{(x)} + \alpha\big(s_{\max}^{(x)} - \bar{s}^{(x)}\big), \qquad \sigma_q^{(x)} = \gamma\big(\bar{s}^{(x)} - s_{\min}^{(x)} + \epsilon\big), \tag{2}$$

where α shifts the quality threshold 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 **binary quality gate** and **quality reward** for a generated response y_i are

$$g(x, y_i) = \mathbf{1}\Big[s_i \geq \tau_q^{(x)}\Big], \quad r_{\text{qual}}(x, y_i) = \mu \tanh\left(\frac{s_i - \tau_q^{(x)}}{\sigma_q^{(x)}}\right) \tag{3}$$

where μ controls the magnitude of the quality reward. This quality gate ensures that the model receives diversity reward only when the response meets the prompt-adaptive quality threshold $\tau_q^{(x)}$. As a result, responses that are merely more unusual, off-topic, or malformed compared to the initial policy would not receive diversity credit, preventing spurious solutions which hack the diversity reward. Meanwhile, the quality reward ensures that responses above the threshold receive a positive bonus, while responses below the prompt-adaptive threshold still receive a quality-learning signal but receive no diversity credit. The bounded $\tanh$ transformation prevents the quality term from dominating optimization once the response already exceeds the threshold by a large margin.

Diversity Scoring. For each prompt, the response group $\mathcal{Y} = \{y_1, \ldots, y_6\}$ contains one candidate from each of the six modes and best fits on 4 GPUs: one for the reward model and three for inference. We embed the final answer text using `sentence-transformers/all-MiniLM-L6-v2` (Wang et al., 2020).

The diversity reward directly optimized during training is the normalized semantic distance to the closest alternative within the response group:

$$d(x, y_i, \mathcal{Y}) = \min_{j \neq i} \frac{1 - \mathbf{e}_i^\top \mathbf{e}_j}{2}. \tag{4}$$

Here, y_i and every y_j are responses to the same prompt but are generated under different roles, and each response is therefore compared with the other five candidates in its group. For thinking-enabled models, the hidden thinking trace is removed before embedding. This nearest-neighbor formulation rewards a response only if it is sufficiently distinct from its closest alternative, preventing multiple responses from collapsing to the same mode while a single outlier dominates the group-level diversity score. Consequently, role-conditioned agents are incentivized to specialize in different regions of the response space. Although we use semantic distance by default, the framework is agnostic to the diversity metric, and lexical, syntactic, or learned semantic measures can be substituted directly.

Reward-Hacking Penalties. The terms r_{len}, r_{lang}, and r_{fmt} are non-positive penalties for excessive length, language mismatch, and formatting or role-label leakage, respectively. These terms suppress superficial ways of increasing measured diversity without improving the usefulness of the response.

We optimize the shared policy via Group Relative Policy Optimization (GRPO) (Shao et al., 2024), estimating advantages from within-group reward comparisons over the role-conditioned response set without a separate value network. The full training objective is provided in Appendix C.1.

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

Algorithm 1 MODA: **Mo**de-Conditioned **D**iversity **A**lignment

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 $\mathcal{M} = \{m_i\}_{i=1}^N$; 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 $(s_{\min}(x), s_{\max}(x), \mu_{\text{ref}}(x))$
    $\tau_q^{(x)} \leftarrow s_{\min} + \alpha(s_{\max} - \mu_{\text{ref}}), \quad \sigma_q^{(x)} \leftarrow \gamma(\mu_{\text{ref}} - s_{\min} + \epsilon)$
    Generate mode-conditioned responses with abstract roles $\mathcal{Y}(x) = \{y_i \sim \pi_\theta(\cdot \mid x, m_i)\}_{i=1}^N$
    for $i = 1, \ldots, N$ **do**
      $q_i \leftarrow S_\phi(x, y_i)$ # Reward model assess response y_i
      $R_{\text{quality}} \leftarrow \mu \tanh\Big((q_i - \tau_q^{(x)})/\sigma_q^{(x)}\Big)$ # Compute quality bonus
      $G_{\text{qual}} \leftarrow \mathbf{1}\{q_i \geq \tau_q^{(x)}\}, \quad R_{\text{div}} \leftarrow \min_{j \neq i} \delta(y_i, y_j)$ # Quality gate and diversity reward
      $R_i \leftarrow \lambda_q R_{\text{qual}} + \lambda_d G_{\text{qual}} R_{\text{div}} + R_{\text{penalty}}(x, y_i)$
    end for
    $A_i \leftarrow \big(R_i - \text{mean}(R_{1:N})\big)/\big(\text{std}(R_{1:N}) + \eta\big)$ for $i = 1, \ldots, N$
  end for
  Update π_θ with GRPO using advantages $\{A_i\}$ and KL regularization to π_{ref}
end for
return trained policy π_θ

`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. Quality and diversity metrics are computed only over the final answer, with the thinking trace masked out, throughout the training-time reward computation and evaluation. We include more implementation details in Appendix D.

Baselines. Given that MODA is a training-based method, the most direct comparisons are training-time approaches. Both **DARLING** (Li et al., 2025) and **DivPO** (Lanchantin et al., 2025) update model parameters to jointly optimize for diversity and quality during the training time. **DivPO** is a preference-optimization (DPO) method that constructs response preference pairs by selecting rare, high-quality responses as preferred and common, low-quality ones as rejected. **DARLING** optimizes an online RL objective with a learned diversity signal. We trained two versions of DARLING: one using our mixture data (Training Data, below) and one using their specified recipe (10k prompts subsampled from WildChat). On the inference-time side, we **repeatedly sample** the base model k times per prompt without role conditioning, testing whether stochastic decoding alone suffices. We also compared with **SSoT** (Misaki & Akiba, 2026), a representative prompting method that induces diversity by injecting random string seeds in the thinking tokens as an entropy source. Notably, inference-time and training-time methods are complementary. We can apply inference-time methods to MODA at decoding time to further amplify diversity improvements.

Training Data. The training set contains 10K prompts, constructed from a 4:1 mixture of uniformly sampled Tulu3-SFT-Mixture (Lambert et al., 2024) and Infinite-Chat dataset (Jiang et al., 2025). This mixture balances quality and diversity: Tulu3-SFT-Mixture provides instruction-following prompts that preserve response quality, while Infinite-Chat encourages open-ended exploration.

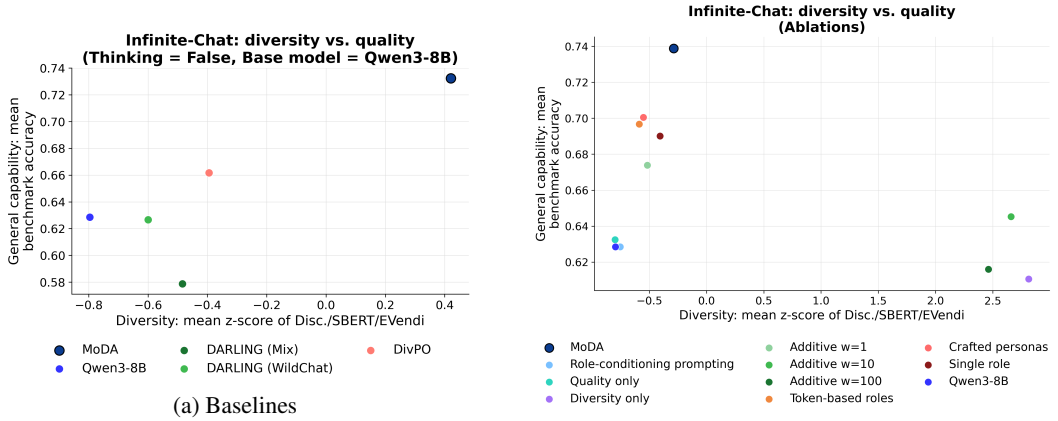
Generative Diversity Evaluation. To evaluate generative diversity, we focus on two application domains where multiple distinct high-quality outputs are desirable: **scientific ideation** and **creative writing**. We evaluate all methods under the **diverse mode**, with the role code injected into the 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 ideation** 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 Infinite-Chat (Jiang et al., 2025). We describe the adaptation of HypoBench and PreScience to our setting in Appendix F.3. We measure output diversity using a comprehensive set of diversity metrics: *semantic-level* (SBERT), *entropy-level* (E-Vendi) (Friedman & Dieng, 2023), and

entailment-level (Discriminator/Disc.), a pairwise classifier trained to predict whether two responses to the same prompt are conceptually distinct (training details in Appendix D.4).

General Capability Retention Evaluation. To assess whether our method preserves the model's reasoning capability, we evaluate each trained model in the standard single-response setting on a suite of seven widely used benchmarks under the **standard mode**. 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; IFEval (Zhou et al., 2023) to evaluate instruction following. We used a max token length of 8192 for the thinking-enabled setting, 2048 for the thinking-disabled setting.

5 Results

Generative Diversity Evaluation. As shown in Figure 2a, MoDA Pareto-dominates all baselines on the held-out Infinite-Chat set across every setting (thinking enabled/disabled, different base models), achieving both higher diversity and higher quality simultaneously; the remaining settings are shown in Appendix E.1. We provide a breakdown by individual benchmarks in Table 1. Across four domains and three model families, MoDA achieves the best average diversity across all evaluated baselines, while matching or improving quality on most domains except PreScience. Notably, MoDA outperforms all baselines on HypoBench, a scientific ideation task that is 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 across thinking settings and base models, even on out-of-domain topics.



(a) Baselines

(b) Ablation variants

Figure 2: **Quality-diversity pareto frontier (Qwen3-8B, thinking disabled) baselines and ablated variants.** The x-axis is Infinite-Chat response diversity (mean z-score of Disc./SBERT/EmbVendi) and the y-axis is general capability (mean benchmark accuracy). **(a)** MoDA Pareto-dominates all the baselines, achieving better diversity and quality simultaneously. Results for Qwen3-8B (thinking enabled) and Llama-3.1-8B / GLM-4-9B (thinking disabled) are shown in Appendix E.1. **(b)** The same comparison across MoDA's ablated variants.

General Capability Retention. As shown in Figure 3, our model achieves the best overall average pass@1, pass@5 and pass@10 score for general capability retention across thinking settings and base models among all baselines. With Qwen3-8B base and thinking disabled, it matches or exceeds the base model on 5 out of 7 benchmarks and achieves the best score among all baselines on 4 of them on pass@1 accuracy. Notably, with the GLM-4-9B base, MoDA improves average pass@1 by 15.8 percentage points, including a 3.89x relative improvement on GSM8K and a 2.37x relative improvement on MMLU. These results suggest that our method is an effective alignment approach that preserves, and in several cases significantly improves, the model's capabilities. We note that the language mixing penalty is important to maintain quality in some cases; without it, quality drops due to reward hacking from language mixing.

Model	*Infinite-Chat*				*NoveltyBench*				*HypoBench*				*PreScience*				Avg.		
	Disc.	SBERT	E-V	Qual.(%)	Disc.	SBERT	E-V	Qual	Disc.	SBERT	E-V	Qual.	Disc.	SBERT	E-V	Qual.	Disc.	SBERT	E-V
QWEN3-8B BASE, THINKING DISABLED																			
Qwen3-8B	0.400	0.132	1.83	62.9	0.493	0.224	2.36	4.58	0.384	0.144	1.91	**4.03**	0.431	0.155	1.94	**4.25**	0.427	0.164	2.01
DARLING (Mix)	0.430	0.203	2.33	57.9	0.513	0.281	2.96	5.24	0.406	0.222	2.48	3.44	**0.559**	**0.541**	**4.73**	2.33	0.477	0.312	3.12
DARLING (WildChat)	0.416	0.185	2.15	62.7	0.480	0.335	3.11	4.00	0.415	0.205	2.34	3.73	0.446	0.192	2.21	4.03	0.439	0.229	2.45
DivPO	0.410	0.274	2.86	66.2	0.519	0.380	3.58	4.87	0.427	0.198	2.30	3.89	0.455	0.233	2.41	3.96	0.453	0.271	2.79
MODA	**0.472**	**0.482**	**4.40**	**73.2**	**0.537**	**0.470**	**4.19**	**5.47**	**0.480**	**0.547**	**4.91**	3.80	0.440	0.176	2.10	3.96	**0.482**	**0.419**	**3.90**
QWEN3-8B BASE, THINKING ENABLED																			
Qwen3-8B	0.400	0.136	1.85	75.8	0.510	0.258	2.67	4.90	0.399	0.164	2.05	**3.93**	0.441	0.136	1.83	3.79	0.438	0.174	2.10
SSoT	0.397	0.178	2.09	76.3	0.438	0.239	2.57	0.46	0.419	0.220	2.33	3.75	**0.461**	**0.281**	**2.39**	3.30	0.429	0.230	2.34
DivPO	0.414	0.200	2.20	71.7	0.521	0.327	3.12	**5.06**	0.409	0.162	2.04	3.89	0.448	0.185	2.04	**3.87**	0.448	0.218	2.35
MODA	**0.600**	**0.909**	**9.04**	**76.8**	**0.589**	**0.800**	**7.63**	3.39	**0.579**	**0.875**	**8.06**	3.71	0.439	0.194	2.10	3.68	**0.552**	**0.695**	**6.71**
LLAMA-3.1-8B BASE, THINKING DISABLED																			
Llama-3.1-8B	0.414	0.217	2.40	46.7	0.512	0.312	3.04	4.80	**0.469**	0.207	2.38	3.60	0.461	**0.514**	3.67	**2.62**	**0.464**	0.312	2.87
MODA (Llama-3.1-8B)	**0.441**	**0.373**	**3.77**	**51.9**	**0.532**	**0.392**	**3.75**	**6.17**	0.410	**0.324**	**3.27**	**3.65**	**0.472**	0.470	**3.89**	2.54	**0.464**	**0.390**	**3.67**
GLM-4-9B BASE, THINKING DISABLED																			
GLM-4-9B	0.421	0.234	2.60	44.9	**0.544**	0.408	4.08	**4.14**	0.402	0.225	2.52	**3.75**	0.461	0.438	4.15	2.57	0.457	0.326	3.34
MODA(GLM-4-9b)	**0.481**	**0.528**	**5.14**	**60.7**	0.502	**0.521**	**5.01**	2.70	**0.452**	**0.485**	**4.64**	3.68	**0.464**	**0.448**	**4.28**	**2.58**	**0.475**	**0.495**	**4.77**

Table 1: **Per-domain diversity and quality.** Disc. = discriminator diversity, SBERT = SBERT embedding pairwise distance, E-V = E-Vendi score, averaged over n=3 random seeds. Infinite-Chat quality is general capability accuracy (%); all other domains use the native quality score. Bold marks the best value within each model group. Table with error bar can be found in Appx E.1.

Model	GSM8K	MMLU	GPQA	BoolQ	HS	TQA	IFEval	Avg.
QWEN3-8B BASE, THINKING DISABLED								
Qwen3-8B	84.5	**82.0**	41.4	83.1	48.4	22.2	**78.4**	62.9
DARLING (Mixture)	63.5	63.0	34.3	76.7	62.2	52.4	53.0	57.9
DARLING (WildChat)	77.9	55.0	34.3	**86.4**	68.8	43.2	73.0	62.7
DivPO	80.5	81.0	40.4	83.5	62.8	37.7	77.3	66.2
MODA	**86.7**	81.0	**52.0**	84.8	**73.4**	**57.9**	76.9	**73.2**
QWEN3-8B BASE, THINKING ENABLED								
Qwen3-8B	**90.3**	**94.0**	43.4	**87.2**	70.9	64.9	**79.7**	75.8
SSoT	84.3	93.0	**55.1**	**87.2**	**80.1**	**72.8**	61.9	76.3
DivPO	84.9	91.0	44.4	83.2	65.0	55.3	77.6	71.7
MODA	83.9	93.0	50.5	**87.2**	76.2	67.2	79.5	**76.8**
LLAMA-3.1-8B BASE, THINKING DISABLED								
Llama-3.1-8B	71.3	16.0	6.1	**81.6**	30.2	53.4	68.8	46.7
MODA (Llama-3.1-8B)	**74.9**	**30.0**	**18.7**	75.6	**38.5**	**54.7**	**70.8**	**51.9**
GLM-4-9B BASE, THINKING DISABLED								
GLM-4-9B	21.8	35.0	19.7	74.5	**68.0**	**44.8**	50.3	44.9
MODA (GLM-4-9B)	**84.9**	**83.0**	**33.3**	**81.2**	52.3	34.5	**55.5**	**60.7**

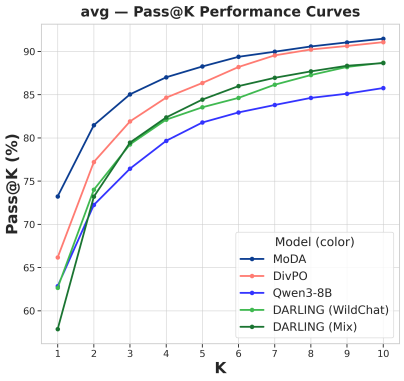


Figure 3: **Left: General capability retention (Pass@1 accuracy %)**, averaged over $n = 3$ random seeds. Bold marks the best value within each model group; underline marks the second-best value within the Qwen3-8B groups. Table with error bar can be found in Appendix E.2 Table 4. **Right: General capability retention: pass@k accuracy.** Qwen3-8B base, thinking disabled models on the general capability suite average. Per-benchmark results are shown in Appendix E.2 Figure 7.

5.1 QUALITATIVE RESULTS

We observe several interesting patterns in our experiments. Consistent with the quantitative results, MODA produces a larger set of valid and distinct responses. Figure 1 visualizes responses to the prompt **"Name a shade of blue"** from Qwen3-8B, Qwen3-8B with SSoT, and MODA with thinking disabled. Figure 4 shows that, when asked to **name a well-received book**, the base model produces the same answer, "To Kill a Mockingbird," in all six samples, whereas MODA produces six distinct titles; Appendix G Figure 13 compares responses to the prompt **"Write a metaphor about time"** from 23 off-the-shelf models and from MODA. The baseline models converge almost entirely on the metaphor "time is a river," whereas MODA produces a diverse set of analogies. Additional qualitative examples are provided in Appendix G. 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.

5.2 ABLATION STUDIES.

Ablation Setups. First, to understand whether the diversity improvement comes from training or solely from role injection, we inject numbered roles to the base model (**Role-conditioned prompting**). Second, to examine whether the diversity gains are driven by the explicit diversity reward or arise

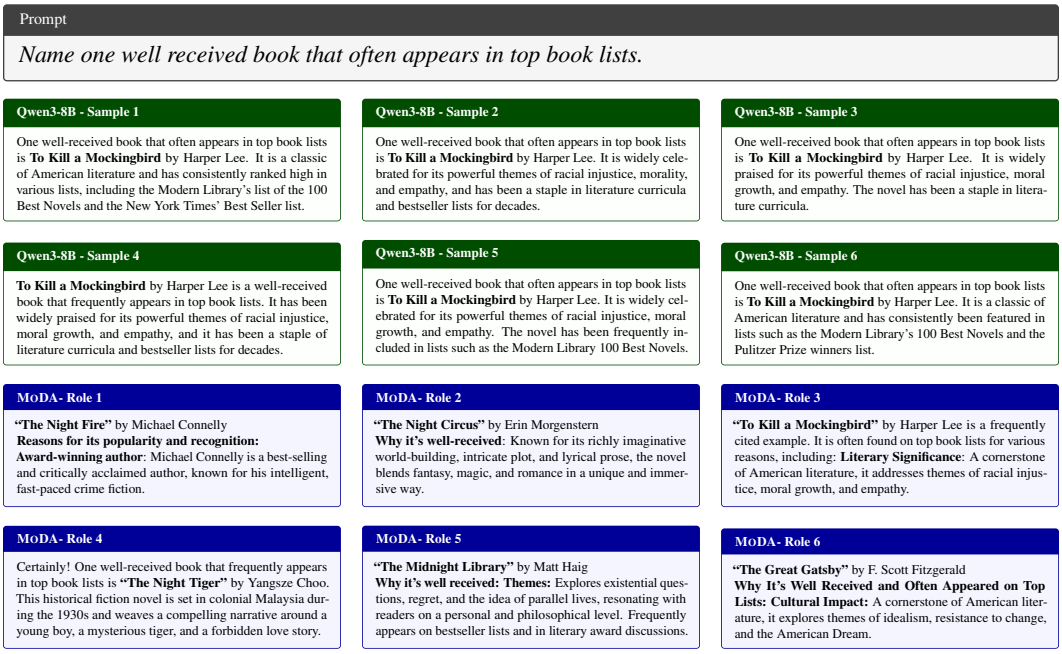


Figure 4: **Qualitative example of diverse generation. Upper**: Responses sampled from Qwen3-8B. **Bottom**: Responses sampled from MoDA under distinct numbered roles for the same prompt, illustrating conceptual diversity across generations.

naturally from repeated sampling and RL optimization, we trained MoDA with **Diversity only** reward and **Quality only** reward. Third, to understand the effect of quality-gated diversity reward, we trained MoDA with additive reward of weighted sum of quality and diversity metrics, with diversity weight = 1, 10 and 100 (**Additive w=1\10\100**). Finally, to understand the effect of the naive numbered role injection, we trained several variants of MoDA: (1) **Crafted personas**, which conditions the roles with manually crafted persona descriptions (e.g. "You are a creative problem solver."; (2) **Single role**, conditions with only one role; (3) **Token-based roles**, conditions the role with a dummy token (e.g. "Start your response with APPLE."). More implementation details are in Appendix D.2. We evaluate each ablated variant for general capability under standard and diverse decoding, and application domain suite under diverse decoding, with full results in Appendix E.3.

Quality-Diversity Trade-off. As shown in Figure 2b, MoDA achieves the best accuracy on the general capability suite among all variants. It also achieves 77% higher SBERT diversity and 41.6% higher E-Vendi diversity than role-conditioning prompting alone, indicating that role-conditioning prompting by itself is insufficient to induce diverse generation. The quality-only variant regresses in diversity relative to role-conditioning prompting, suggesting that an explicit diversity objective is necessary for diversity gains. The diversity-only variant achieves the best diversity metrics overall, but with the lowest accuracy under standard decoding, and its performance collapses entirely under diverse decoding, indicating that explicit quality gating is critical to preserving response quality.

6 Conclusion

In this work, we introduced MoDA, an online-MARL-inspired alignment method that encourages diverse generation while preserving response quality. MoDA serves as a drop-in replacement for existing post-training alignment pipelines, requiring no architectural modifications to the underlying model. Compared to previous methods, its dual-mode design enables seamless switching between producing a single high-confidence answer and generating a diverse set of plausible responses, offering greater flexibility depending on the downstream needs. The ability to surface conceptually distinct yet high-quality responses has broad implications for high-stakes domains such as medical diagnosis, legal reasoning, and scientific ideation. We discuss additional limitations and future directions in the Appendix A.

7 AI USE STATEMENT

In this work, we used generative AI tools for implementing methods (assisting with training code and debugging environment issues), and for editing the LaTeX manuscript to improve readability. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean or reformat the dataset, support qualitative or thematic data analysis, or interpret results. We have reviewed all AI-assisted work. All AI-suggested code was executed and validated by the authors; all AI-suggested prose edits were reviewed by the authors for factual accuracy and consistency before inclusion. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

8 ETHICS STATEMENT

We believe MODA contributes to human well-being by helping language models produce diverse, high-quality responses rather than converging on a single homogeneous answer. This capability is critical for scientific ideation, creative writing, and other domains where multiple plausible perspectives aid human decision-making. All datasets used in this work (Tulu3-SFT-Mixture, Infinite-Chat, Alpaca) are publicly available and were collected and released with appropriate consent and licensing by their original authors; we introduce no new human-subjects data collection.

As mentioned in Limitations (Appendix A), we are also mindful that increasing output diversity could, in principle, surface a wider range of misleading, unsafe, or unsupported outputs, particularly in high-stakes domains such as medical diagnosis or legal reasoning.

9 REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our results. All datasets used for training and evaluation are publicly available, as described in Section 4 and Appendix F. We provide full training and evaluation code, including our implementation of MODA and the ablation variants, in the supplementary materials. Implementation details, including hyperparameters, model checkpoints, and compute requirements, are documented in Appendix D.1, D.2 and D.3. Our discriminator's training data, architecture, and evaluation protocol are described in Appendix D.4.

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. February 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. *Proceedings of the 16th Conference on Creativity & Cognition*, 2024. URL https://api.semanticscholar.org/CorpusID:267406608.

Oron Anschel, Alon Shoshan, Adam Botach, Shunit Haviv Hakimi, Asaf Gendler, Emanuel Ben Baruch, Nadav Bhonker, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. Group-aware reinforcement learning for output diversity in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32382–32403, 2025.

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.

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 Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

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.

Xingyu Dang, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Assessing diversity collapse in reasoning. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL https://openreview.net/forum?id=AMiKsHLjQh.

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):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Andre He, Daniel Fried, and Sean Welleck. 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.

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*, pp. 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. January 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*, pp. 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.

Arash Lagzian, Srinivas Anumasa, and Dianbo Liu. 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.

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*, 2025.

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.

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.

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.

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. February 2026. 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.

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. March 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 (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 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/.

Chantal Shaib, 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, 2024. URL https://arxiv.org/abs/2403.00553.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

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*, 2025.

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)*, pp. 451–462, 2018.

Chen Henry Wu, Sachin Goyal, and Aditi Raghunathan. Mode-conditioning unlocks superior test-time scaling. November 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 Radanović. Agenticred: Evolving agentic systems for red-teaming, 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. October 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*, pp. 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: Benchmarking machine learning agents for scientific research, 2026. URL https://arxiv.org/abs/2510.10472.

Appendices

A Limitations and Future Work **18**

B Extended Related Work: Method Comparison **19**

C Policy Optimization and Algorithm **20**

C.1 Policy Optimization 20

C.2 Algorithm 20

D Implementation **22**

D.1 Overview 22

D.2 Ablation Model Implementations 22

D.3 Dataset 24

D.4 Discriminator 24

D.5 Training Curve 24

E Additional Experiment Results **26**

E.1 Generative Diversity Evaluation. 26

E.2 General Capability Retention 28

E.3 Ablation Results 31

F Benchmark Descriptions **33**

F.1 Diversity Metrics 33

F.2 General Capability Retention 33

F.3 Domain Application Tasks 34

G Generation Examples **35**

A Limitations and Future Work

Several limitations point to promising directions for future research. First, our quality reward model (SkyReward-V2) tends to favor verbose responses, even for prompts that should be answered briefly, and our uniform length penalty insufficiently addresses this across prompts of varying complexity. For instance, when training GLM-4-9B with MoDA, we observed suboptimal native quality score on NoveltyBench compared to the base model because its reward model doesn't penalize verbosity therefore the model reward hack by adding unnecessary preambles such as "As a model.../To answer these questions as a model..." Removing these preambles recovers the performance from 2.739 to 3.302 (+20%). Future work should explore adaptive length penalties or alternative reward models that better reflect conciseness. Second, because MoDA introduces no external data during training, generative diversity is bounded by the initial policy's capacity, and incorporating retrieval augmentation or external data sources could meaningfully expand the exploration space. Finally, there is a risk of producing more varied but misleading, unsafe, or unsupported outputs, especially in high-stakes domains such as medical diagnosis and legal reasoning. Although MoDA mitigates this through its dual-mode design, whose standard mode retains the model's general capability, future work can further combine the diverse mode with verification, calibration, retrieval, or domain-specific safeguards.

B Extended Related Work: Method Comparison

Table 2 accompanies the discussion of diversity-oriented methods in Section 2, summarizing representative inference-time and training-time approaches by mechanism and code availability.

Method	Mechanism	Code
Inference-time		
Verbalized Sampling (Zhang et al., 2025a)	Prompting	✓
String Seed of Thought (Misaki & Akiba, 2026)	Prompting	✓
DIPPER (Lau et al., 2025)	Prompting	✗
Multi-Novelty (Lagzian 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., 2025)	RL	✓
DivPO (Lanchantin et al., 2025)	DPO	✗
DQO (Chen et al., 2025)	RL	✓
Multi-answer RL (Puri et al., 2026)	RL with multiple answers	✓
Soft Preference Learning (Slocum et al., 2025)	Entropy Decoupling	✗
MoDA (ours)	**Mode-Conditioned MARL**	✓

Table 2: **Diversity-oriented methods.** MoDA conditions diversity on explicit mode tokens, inspired by MARL.

C Policy Optimization and Algorithm

C.1 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:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{x,\{y_i\}}\left[\frac{1}{k}\sum_{i=1}^{k}\min\Big(\rho_i\hat{A}_i,\ \text{clip}(\rho_i,\ 1{-}\varepsilon,\ 1{+}\varepsilon)\,\hat{A}_i\Big) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right] \tag{5}$$

where $\rho_i = \pi_\theta(y_i \mid x)/\pi_{\text{ref}}(y_i \mid x)$ is the importance ratio (Kloek & 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. 1, 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 in the main text summarizes the training procedure.

C.2 Algorithm

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 $\mathcal{S}_{\text{ref}}(x) \leftarrow \emptyset$
  for $j = 1, \ldots, M$ **do**
    Sample reference response $\tilde{y}_j \sim \pi_{\text{ref}}(\cdot \mid x)$
    Compute quality score $\tilde{s}_j \leftarrow S_\phi(x, \tilde{y}_j)$
    Add $\tilde{s}_j$ to $\mathcal{S}_{\text{ref}}(x)$
  end for
  Compute

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

  .
  Store $\big(s_{\min}(x), s_{\max}(x), \mu_{\text{ref}}(x)\big)$ 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 $\big(s_{\min}(x), s_{\max}(x), \mu_{\text{ref}}(x)\big)$
    Compute prompt-adaptive quality threshold and reward magnitude factor:

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

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

    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

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

Compute diversity reward

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

Compute reward-hacking penalty

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

Compute **Quality Gated Diversity Reward**

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

end for

Compute group-relative advantages:

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

end for

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

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

where

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

end for

return trained policy π_θ

D IMPLEMENTATION

D.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. For GLM-4-9B base, we observed a severe language-mismatch issue in its response, and therefore increased the weight of the language-mismatch penalty to 100.

For Qwen3-8B base, we trained MODA and all baselines for 4 epochs. For Llama-3.1-8B, GLM-4-9B base and ablation models, we trained for 2 epochs due to compute resource constraints.

Our training were performed on NVIDIA A100, H100 or H200 GPUs, depending on availability. A single training for 2 epochs run takes approximately 24 hours on 4 H200 GPUs. We provide the rest of the training parameters in Table 3.

D.2 ABLATION MODEL IMPLEMENTATIONS

Additive Aggregation. Many existing works on quality-diversity optimization aggregate quality and diversity objectives using a weighted sum (Mouret & Clune, 2015; Chen et al., 2025). We compare our method with variants trained using an additive reward objective. We find that additive aggregation is highly sensitive to the diversity weight λ_d. at $\lambda_d = 1$, the additive variant shows modest diversity improvement over the baseline, but substantially less than MODA; at $\lambda_d = 10$ and $\lambda_d = 100$, diversity improves significantly, but at the cost of general capability performance under diverse decoding. 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.

Role Injection. 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 of the abstract numbered roles, testing whether hand-designed personas provide a stronger diversity prior than the abstract numbered role used in the main setting. To isolate the effect of role multiplicity from role content, we train a single-role variant in which every response is generated under the same *"You are Role 1."* system prompt. To study the effect of the identity-based role, we train a token-based conditioning variant in which the system prompt instructs a token-based distinguishing signal (e.g. *"Start your response with APPLE/BANANA/ORANGE."*). As shown in Table 6, all variants exhibit a slight diversity improvement over the base model, but their gains are smaller than MODA's, indicating that multiple abstract numbered identity-based roles provide a more effective exploration space than all the ablated variants, without sacrificing generalization.

Token-based roles As an ablation, we replace the abstract numbered role instructions with dummy roles. Specifically, for each prompt, we prepend one of the following system prompts. We stripped the dummy tokens (e.g., *APPLE*/*ORANGE*/*BANANA*) from the responses before computing the diversity and quality metrics, both during the training and during the evaluation.

1. "Start your response with APPLE."
2. "Start your response with ORANGE."
3. "Start your response with BANANA."
4. "Start your response with CHERRY."
5. "Start your response with DATE."
6. "Start your response with ELDERBERRY."

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:

Table 3: 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, N	6
Sampling	Enabled
Temperature	1.0
Top-p	1.0
Top-k	-1
Maximum model length	4096
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 mixing penalty weight (Qwen/Llama)	0.1
Language mixing penalty weight (GLM)	10

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."

Single role As an ablation, we use only one role for mode condition. We prepend "You are Role 1." on each system prompt.

D.3 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 `allenai/tulu-3-sft-mixture` (Lambert et al., 2024) as general-purpose examples and 2k prompts from Infinite-Chat as open-ended queries. We held out a portion of Infinite-Chat for evaluation.

D.4 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 (Taori et al., 2023) prompts. For each prompt, we employ Qwen3-30B to generate responses under 5 modes and 2 random seeds, yielding 20,000 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 to construct its feature vector. Given a pairs of sentence embeddings $u, v \in \mathbb{R}^{768}$, we construct the feature vector

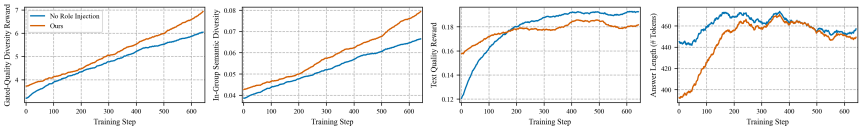
$$z = [u \,\|\, v \,\|\, |u - v| \,\|\, u \odot v \,\|\, \phi(y_i, y_j)] \in \mathbb{R}^{3075},$$

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

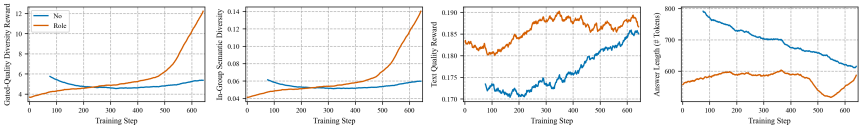
The classifier is a three-layer MLP with $3075 \to 256 \to 64 \to 1$ nodes, ReLU activations, dropout 0.3, and a sigmoid output. We train with focal loss ($\alpha = 0.25, \gamma = 2$), AdamW with learning rate 3×10^{-4}, batch size 64, and early stopping on validation AUC. The model has approximately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

D.5 Training Curve

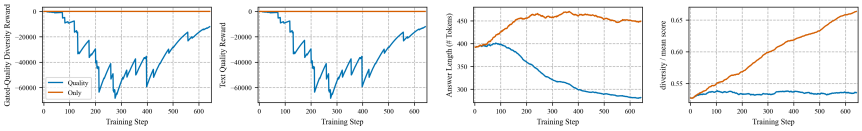
We share the training curve for ablation studies in Figure 5.



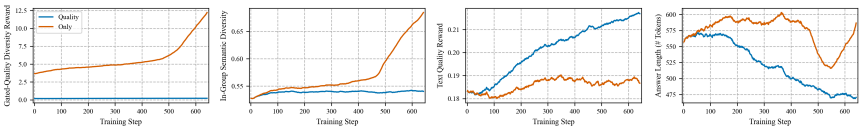
No role injection, thinking disabled.



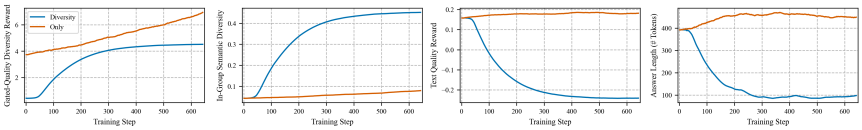
No role injection, thinking enabled.



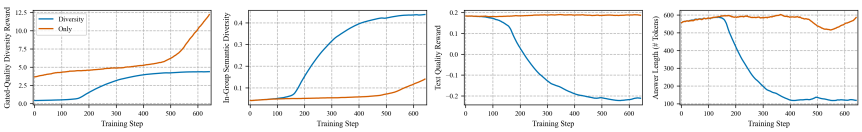
Quality-only reward, thinking disabled.



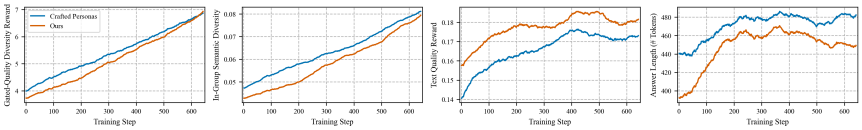
Quality-only reward, thinking enabled.



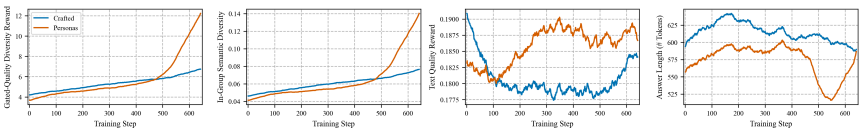
Diversity-only reward, thinking disabled.



Diversity-only reward, thinking enabled.



Crafted personas, thinking disabled.



Crafted personas, thinking enabled.

Figure 5: Training curves for ablation studies. Each row corresponds to one ablation setting, and each row reports four training diagnostics: total reward, diversity reward, quality reward, and answer token length. Results are shown under both thinking-disabled and thinking-enabled settings.

E Additional Experiment Results

E.1 Generative Diversity Evaluation.

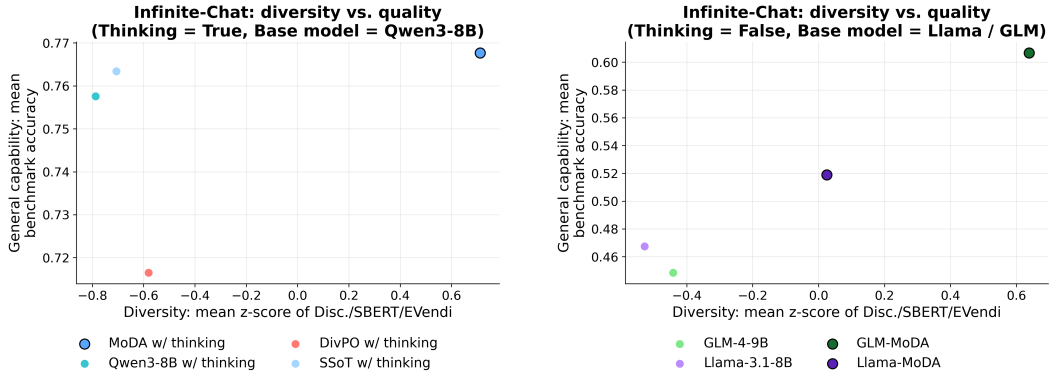


Figure 6: **Diversity vs. quality pareto fronts, remaining settings** (extends Figure 2 in the main text). The x-axis is Infinite-Chat response diversity (mean z-score of Disc./SBERT/EmbVendi) and the y-axis is general capability (mean benchmark accuracy). **Left**: Qwen3-8B base, thinking enabled. **Right**: Llama-3.1-8B / GLM-4-9B base, thinking disabled. MoDA Pareto-dominates all baselines in both settings.

We show the results with error bar for generative diversity across all domain application tasks.

Infinite-Chat

Model	Disc.	S	EV	Qual.
Qwen3-8B base, thinking disabled				
Qwen3-8B	0.400 ± 0.001	0.132 ± 0.001	1.83 ± 0.01	62.9% ± 0.8%
DARLING (Mix)	0.430 ± 0.001	0.203 ± 0.001	2.33 ± 0.00	57.9% ± 1.0%
DARLING (WildChat)	0.416 ± 0.001	0.185 ± 0.002	2.15 ± 0.01	62.7% ± 1.0%
DivPO	0.410 ± 0.001	0.274 ± 0.004	2.86 ± 0.03	66.2% ± 0.9%
MoDA	**0.472** ± 0.001	**0.482** ± 0.003	**4.40** ± 0.01	**73.2%** ± 0.9%
Qwen3-8B base, thinking enabled				
Qwen3-8B	0.400 ± 0.001	0.136 ± 0.001	1.85 ± 0.01	75.8% ± 0.7%
SSoT	0.397 ± 0.000	0.178 ± 0.002	2.09 ± 0.01	76.3% ± 0.7%
DivPO	0.414 ± 0.002	0.200 ± 0.006	2.20 ± 0.03	71.7% ± 0.8%
MoDA	**0.600** ± 0.001	**0.909** ± 0.001	**9.04** ± 0.01	**76.8%** ± 0.7%
Llama-3.1-8B base, thinking disabled				
Llama-3.1-8B	0.414 ± 0.001	0.217 ± 0.002	2.40 ± 0.01	46.7% ± 0.7%
MoDA (Llama-3.1-8B)	**0.441** ± 0.002	**0.373** ± 0.003	**3.77** ± 0.02	**51.9%** ± 0.9%
GLM-4-9B base, thinking disabled				
GLM-4-9B	0.421 ± 0.001	0.234 ± 0.002	2.60 ± 0.01	44.9% ± 0.9%
MoDA(GLM-4-9b)	**0.481** ± 0.001	**0.528** ± 0.001	**5.14** ± 0.01	**60.7%** ± 0.8%

NOVELTYBENCH

Model	Disc.	S	EV	Qual.
QWEN3-8B BASE, THINKING DISABLED				
Qwen3-8B	0.493 ± 0.002	0.224 ± 0.003	2.36 ± 0.02	4.58 ± 0.04
DARLING (Mix)	0.513 ± 0.001	0.281 ± 0.001	2.96 ± 0.01	5.24 ± 0.07
DARLING (WildChat)	0.480 ± 0.001	0.335 ± 0.004	3.11 ± 0.02	4.00 ± 0.06
DivPO	0.519 ± 0.001	0.380 ± 0.003	3.58 ± 0.02	4.87 ± 0.02
MODA	**0.537** ± 0.001	**0.470** ± 0.003	**4.19** ± 0.01	**5.47** ± 0.03
QWEN3-8B BASE, THINKING ENABLED				
Qwen3-8B	0.510 ± 0.001	0.258 ± 0.003	2.67 ± 0.03	4.90 ± 0.04
SSoT	0.438 ± 0.001	0.239 ± 0.002	2.57 ± 0.01	0.46 ± 0.01
DivPO	0.521 ± 0.001	0.327 ± 0.003	3.12 ± 0.02	**5.06** ± 0.04
MODA	**0.589** ± 0.002	**0.800** ± 0.004	**7.63** ± 0.07	3.39 ± 0.05
LLAMA-3.1-8B BASE, THINKING DISABLED				
Llama-3.1-8B	0.512 ± 0.001	0.312 ± 0.001	3.04 ± 0.01	4.80 ± 0.08
MODA (Llama-3.1-8B)	**0.532** ± 0.001	**0.392** ± 0.001	**3.75** ± 0.01	**6.17** ± 0.04
GLM-4-9B BASE, THINKING DISABLED				
GLM-4-9B	**0.544** ± 0.002	0.408 ± 0.004	4.08 ± 0.04	**4.14** ± 0.01
MODA(GLM-4-9b)	0.502 ± 0.002	**0.521** ± 0.005	**5.01** ± 0.04	2.70 ± 0.03

HYPOBENCH

Model	Disc.	S	EV	Qual.
QWEN3-8B BASE, THINKING DISABLED				
Qwen3-8B	0.384 ± 0.005	0.144 ± 0.005	1.91 ± 0.03	**4.03** ± 0.01
DARLING (Mix)	0.406 ± 0.005	0.222 ± 0.004	2.48 ± 0.02	3.44 ± 0.03
DARLING (WildChat)	0.415 ± 0.004	0.205 ± 0.006	2.34 ± 0.04	3.73 ± 0.02
DivPO	0.427 ± 0.006	0.198 ± 0.005	2.30 ± 0.03	3.89 ± 0.01
MODA	**0.480** ± 0.003	**0.547** ± 0.010	**4.91** ± 0.05	3.80 ± 0.02
QWEN3-8B BASE, THINKING ENABLED				
Qwen3-8B	0.399 ± 0.006	0.164 ± 0.005	2.05 ± 0.03	**3.93** ± 0.01
SSoT	0.419 ± 0.002	0.220 ± 0.010	2.33 ± 0.04	3.75 ± 0.02
DivPO	0.409 ± 0.002	0.162 ± 0.005	2.04 ± 0.03	3.89 ± 0.01
MODA	**0.579** ± 0.013	**0.875** ± 0.018	**8.06** ± 0.25	3.71 ± 0.04
LLAMA-3.1-8B BASE, THINKING DISABLED				
Llama-3.1-8B	**0.469** ± 0.005	0.207 ± 0.006	2.38 ± 0.04	3.60 ± 0.02
MODA (Llama-3.1-8B)	0.410 ± 0.001	**0.324** ± 0.014	**3.27** ± 0.10	**3.65** ± 0.01
GLM-4-9B BASE, THINKING DISABLED				
GLM-4-9B	0.402 ± 0.003	0.225 ± 0.005	2.52 ± 0.03	**3.75** ± 0.01
MODA(GLM-4-9b)	**0.452** ± 0.003	**0.485** ± 0.012	**4.64** ± 0.09	3.68 ± 0.01

PRESCIENCE

Model	Disc.	S	EV	Qual.
QWEN3-8B BASE, THINKING DISABLED				
Qwen3-8B	0.431 ± 0.002	0.155 ± 0.004	1.94 ± 0.02	**4.25** ± 0.03
DARLING (Mix)	**0.559** ± 0.005	**0.541** ± 0.018	**4.73** ± 0.15	2.33 ± 0.07
DARLING (WildChat)	0.446 ± 0.002	0.192 ± 0.003	2.21 ± 0.03	4.03 ± 0.06
DivPO	0.455 ± 0.002	0.233 ± 0.001	2.41 ± 0.01	3.96 ± 0.05
MODA	0.440 ± 0.002	0.176 ± 0.002	2.10 ± 0.01	3.96 ± 0.01
QWEN3-8B BASE, THINKING ENABLED				
Qwen3-8B	0.441 ± 0.003	0.136 ± 0.001	1.83 ± 0.01	3.79 ± 0.04
SSoT	**0.461** ± 0.002	**0.281** ± 0.017	**2.39** ± 0.05	3.30 ± 0.05
DivPO	0.448 ± 0.003	0.185 ± 0.015	2.04 ± 0.07	**3.87** ± 0.06
MODA	0.439 ± 0.001	0.194 ± 0.005	2.10 ± 0.01	3.68 ± 0.00
LLAMA-3.1-8B BASE, THINKING DISABLED				
Llama-3.1-8B	0.461 ± 0.005	**0.514** ± 0.020	3.67 ± 0.07	**2.62** ± 0.05
MODA (Llama-3.1-8B)	**0.472** ± 0.003	0.470 ± 0.009	**3.89** ± 0.07	2.54 ± 0.04
GLM-4-9B BASE, THINKING DISABLED				
GLM-4-9B	0.461 ± 0.004	0.438 ± 0.010	4.15 ± 0.08	2.57 ± 0.05
MODA(GLM-4-9b)	**0.464** ± 0.002	**0.448** ± 0.010	**4.28** ± 0.09	**2.58** ± 0.10

E.2 GENERAL CAPABILITY RETENTION

E.2.1 PASS@1

We show general capability retention with error bar for pass@1 across baselines and our model.

Model	GSM8K	MMLU	GPQA	BoolQ	HellaSwag	TruthfulQA	IFEval	Avg.
QWEN3-8B BASE, THINKING DISABLED								
Qwen3-8B	84.5 ± 1.0	**82.0** ± 3.9	41.4 ± 3.5	83.1 ± 0.7	48.4 ± 0.5	22.2 ± 1.5	**78.4** ± 1.8	62.9 ± 0.8
DARLING (Mixture)	63.5 ± 1.3	63.0 ± 4.9	34.3 ± 3.4	76.7 ± 0.7	62.2 ± 0.5	52.4 ± 1.7	53.0 ± 2.1	57.9 ± 1.0
DARLING (WildChat)	77.9 ± 1.1	55.0 ± 5.0	34.3 ± 3.4	**86.4** ± 0.6	68.8 ± 0.5	43.2 ± 1.7	73.0 ± 1.9	62.7 ± 1.0
DivPO	80.5 ± 1.1	81.0 ± 3.9	40.4 ± 3.5	83.5 ± 0.6	62.8 ± 0.5	37.7 ± 1.7	77.3 ± 1.8	66.2 ± 0.9
MODA	**86.7** ± 0.9	81.0 ± 3.9	**52.0** ± 3.6	84.8 ± 0.6	**73.4** ± 0.4	**57.9** ± 1.7	76.9 ± 1.8	**73.2** ± 0.9
QWEN3-8B BASE, THINKING ENABLED								
Qwen3-8B	**90.3** ± 0.8	**94.0** ± 2.4	43.4 ± 3.5	**87.2** ± 0.6	70.9 ± 0.5	64.9 ± 1.7	**79.7** ± 1.7	75.8 ± 0.7
SSoT	84.3 ± 1.0	93.0 ± 2.6	**55.1** ± 3.5	**87.2** ± 0.6	**80.1** ± 0.4	**72.8** ± 1.6	61.9 ± 2.1	76.3 ± 0.7
DivPO	84.9 ± 1.0	91.0 ± 2.9	44.4 ± 3.5	83.2 ± 0.7	65.0 ± 0.5	55.3 ± 1.7	77.6 ± 1.8	71.7 ± 0.8
MODA	83.9 ± 1.0	93.0 ± 2.6	50.5 ± 3.6	**87.2** ± 0.6	76.2 ± 0.4	67.2 ± 1.6	79.5 ± 1.7	**76.8** ± 0.7
LLAMA-3.1-8B BASE, THINKING DISABLED								
Llama-3.1-8B	71.3 ± 1.2	16.0 ± 3.7	6.1 ± 1.7	**81.6** ± 0.7	30.2 ± 0.5	53.4 ± 1.7	68.8 ± 2.0	46.7 ± 0.7
MODA (Llama-3.1-8B)	**74.9** ± 1.2	**30.0** ± 4.6	**18.7** ± 2.8	75.6 ± 0.8	**38.5** ± 0.5	**54.7** ± 1.7	**70.8** ± 2.0	**51.9** ± 0.9
GLM-4-9B BASE, THINKING DISABLED								
GLM-4-9B	21.8 ± 1.1	35.0 ± 4.8	19.7 ± 2.8	74.5 ± 0.8	**68.0** ± 0.5	**44.8** ± 1.7	50.3 ± 2.2	44.9 ± 0.9
MODA (GLM-4-9B)	**84.9** ± 1.0	**83.0** ± 3.8	**33.3** ± 3.4	**81.2** ± 0.7	52.3 ± 0.5	34.5 ± 1.7	**55.5** ± 2.1	**60.7** ± 0.8

Table 4: General capability retention (accuracy %, mean ± std error; bold = best per block)

E.2.2 PASS@5

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

Model	GSM8K	MMLU	GPQA	BoolQ	HellaSwag	TruthfulQA	IFEval	Avg.
QWEN3-8B BASE, THINKING DISABLED								
Qwen3-8B	92.6	92.0	73.7	89.7	86.0	55.1	83.4	81.8
DARLING (Mixture)	92.8	**95.0**	**75.3**	93.1	91.2	86.4	57.3	84.4
DARLING (WildChat)	91.8	86.0	69.2	91.3	91.3	74.4	80.8	83.5
DivPO	92.3	92.0	74.7	93.2	90.5	76.3	**85.6**	86.4
MODA	**94.5**	93.0	74.7	**93.9**	**91.3**	**86.5**	83.9	**88.3**
QWEN3-8B BASE, THINKING ENABLED								
Qwen3-8B	96.1	96.0	66.2	92.2	89.1	82.5	85.4	86.8
SSoT	90.6	96.0	**74.7**	92.0	88.9	**83.8**	73.6	85.7
DivPO	94.0	**97.0**	72.2	**93.0**	**89.8**	**83.8**	85.6	**87.9**
MODA	**96.3**	96.0	68.7	92.0	89.6	82.5	**85.8**	87.3
LLAMA-3.1-8B BASE, THINKING DISABLED								
Llama-3.1-8B	90.1	57.0	28.3	**93.2**	70.2	72.9	**82.3**	70.6
MODA (Llama-3.1-8B)	**93.5**	**76.0**	**56.6**	92.3	**79.6**	**80.4**	80.2	**79.8**
GLM-4-9B BASE, THINKING DISABLED								
GLM-4-9B	60.0	69.0	59.6	**97.1**	**91.3**	**82.6**	57.5	73.9
MODA (GLM-4-9B)	**95.2**	**94.0**	**73.2**	95.3	89.1	77.5	**74.7**	**85.6**

Table 5: 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 pass@k figures for the rest of the general capability suite tasks below.

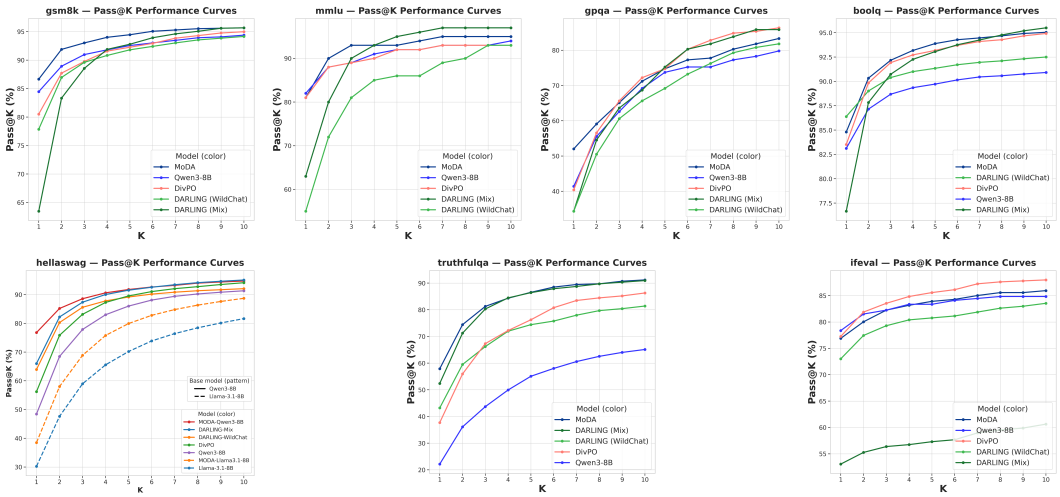


Figure 7: **General capability retention: pass@k accuracy.** Qwen3-8B base and thinking disabled models on the general capability suite tasks: GSM8K, MMLU, GPQA, BoolQ, HellaSwag, TruthfulQA, and IFEval. The suite average is shown in Figure 3 in the main text.

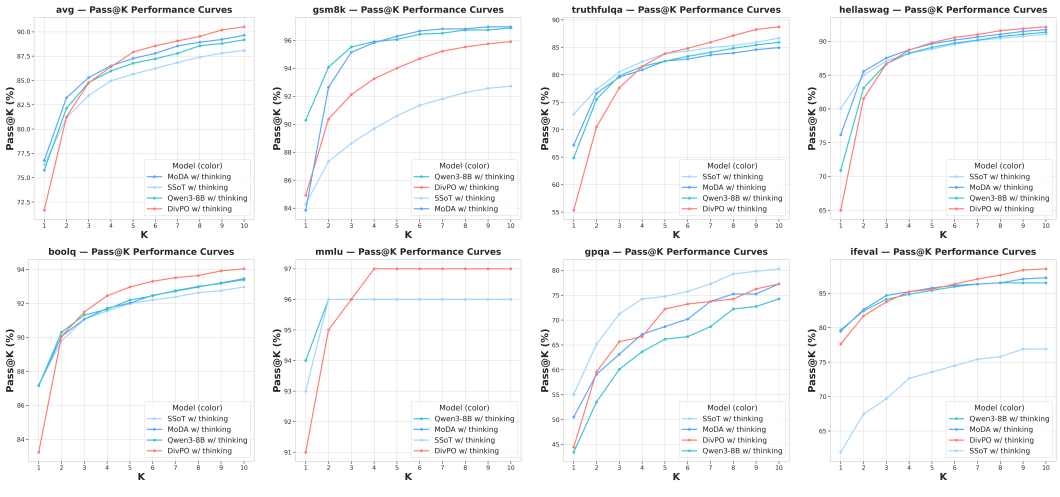


Figure 8: **General capability retention: pass@k accuracy.** Qwen3-8B base and thinking enabled models.

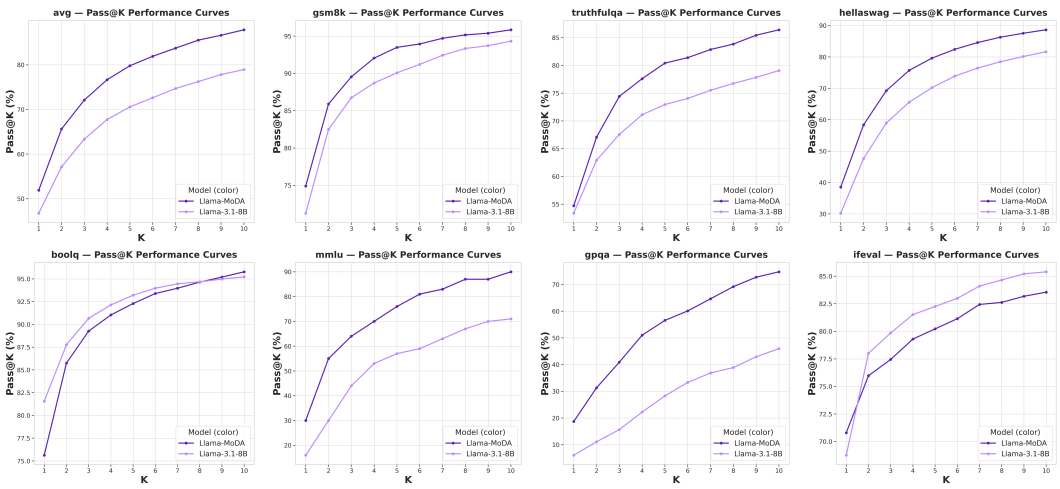


Figure 9: **General capability retention: pass@k accuracy.** Llama-3.1-8B base and thinking disabled models.

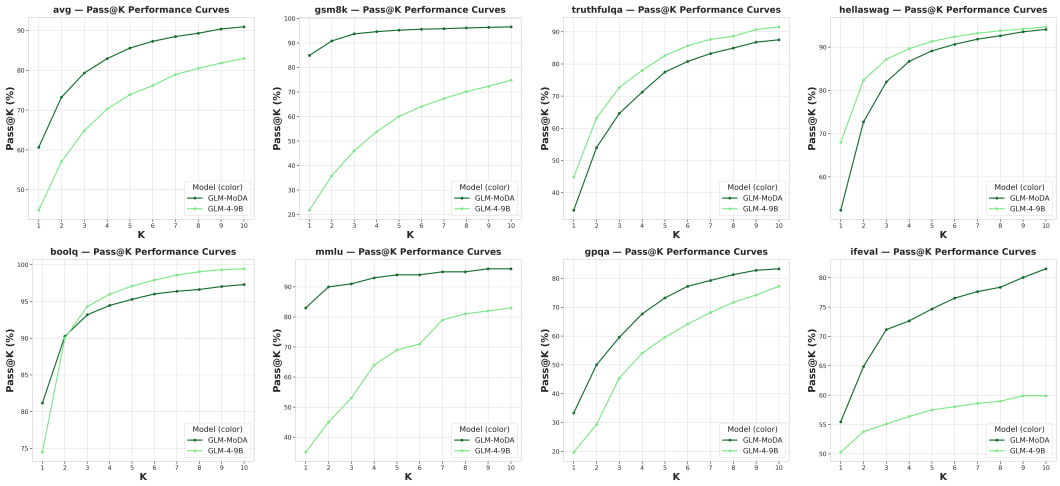


Figure 10: **General capability retention: pass@k accuracy.** GLM-4-9B base and thinking disabled models.

E.3 Ablation Results

We show the mean benchmark accuracy, under both standard and diverse decoding mode, on the general capability suite, and Infinite-Chat response diversity on the held out test set for the ablation study models.

	General capability (Avg., %)		Infinite-Chat Diversity		
Ablation	Standard mode	Diverse mode	Disc.	SBERT	E-Vendi
Role-conditioned prompting	62.9	56.6	0.403	0.145	1.90
Quality only	63.2	61.8	0.401	0.131	1.82
Diversity only	61.1	10.7	**0.656**	**0.993**	**9.77**
Additive w=1	67.4	68.0	0.422	0.204	2.33
Additive w=10	64.5	5.7	0.632	0.976	9.64
Additive w=100	61.6	5.4	0.621	0.970	8.75
Token-based roles	69.7	67.9	0.419	0.197	2.04
Crafted personas	70.1	66.1	0.421	0.194	2.24
Single role	69.0	69.9	0.428	0.239	2.57
Qwen3-8B	62.9	-	0.400	0.132	1.83
MoDA	**73.2**	**71.6**	0.444	0.257	2.69

Table 6: **Ablation results** for general capability (Pass@1) under standard and diverse decoding (accuracy %) and Infinite-Chat diversity (n=3 random seeds). Bold marks the best value per column.

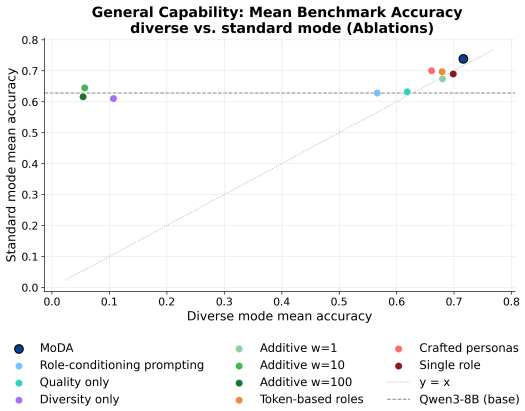
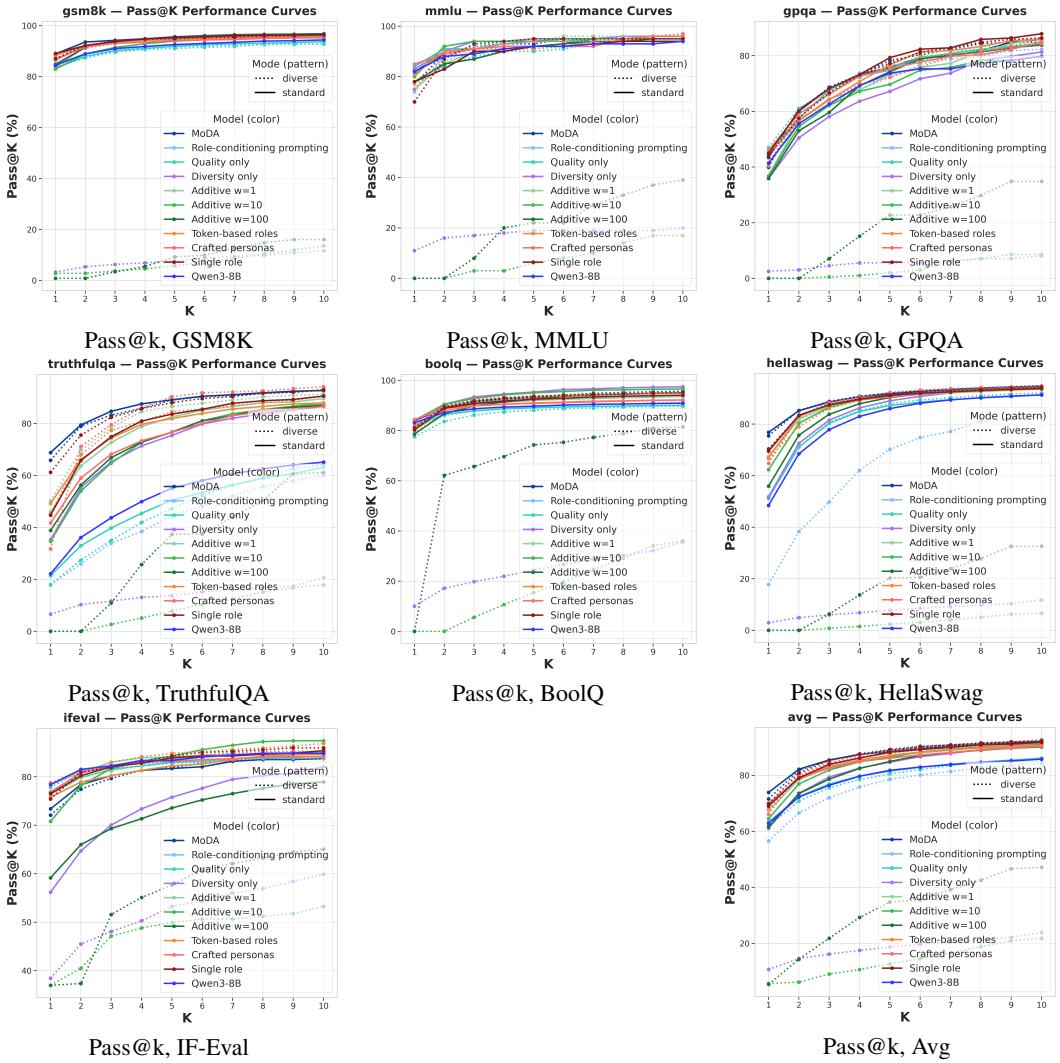


Figure 11: Diversity-quality trade-off under diverse and standard mode. Pass@1 accuracy.



Pass@k, GSM8K | Pass@k, MMLU | Pass@k, GPQA

Pass@k, TruthfulQA | Pass@k, BoolQ | Pass@k, HellaSwag

Pass@k, IF-Eval | Pass@k, Avg

Figure 12: **General capability retention for ablated models**: pass@k accuracy by benchmarks, standard vs diverse decoding mode.

F BENCHMARK DESCRIPTIONS

F.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 D. 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, E-Vendi score and learned discriminator diversity score. **SBERT diversity**(Reimers & 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 D.

E-Vendi score stands for embedding Vendi score. The Vendi score is a diversity metric inspired by ecology and quantum 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.

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 D.4. Notably, the discriminator is not trained on the evaluation outputs.

F.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 (no role injection). Max token length is 2048 when thinking is disabled, 8192 when thinking is enabled. We sample 10 responses for each prompt.

The evaluation covers seven benchmarks: **GSM8K** (Cobbe et al., 2021) for grade-school mathematical reasoning, **MMLU** (Hendrycks et al., 2021) for broad multitask knowledge, **GPQA** (Rein et al., 2024) for graduate-level scientific reasoning, **BoolQ** (Clark et al., 2019) for yes/no reading comprehension, **HellaSwag** (Zellers et al., 2019) for commonsense sentence completion, **TruthfulQA-MC1** (Lin et al., 2022) for robustness to imitative falsehoods, and **IFEval** (Zhou et al., 2023) 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, MMLU, GPQA, HellaSwag, and TruthfulQA-MC1, and strict prompt-level accuracy for IFEval.

F.3 Domain Application Tasks

The domain application suite evaluates whether diversity gains transfer to open-ended generation settings (scientific ideation and creative writing) where multiple distinct high-quality outputs are critical. The suite consists of four tasks. We sample 10 responses per model. For models with role conditioning, we rotate through the available roles to generate these samples.

HypoBench (Liu et al., 2026) evaluates scientific hypothesis generation. Given a set of labeled data-science observations, the model is asked to generate hypotheses that explain possible underlying patterns. 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`. Every parsed hypothesis is scored individually by an LLM judge (gpt-4o-mini by default) on 3 dimensions, 1–5 each:

- Clarity — how precisely/testably it's stated
- Novelty — how non-obvious it is, compared against any known/published hypotheses supplied for that dataset
- Plausibility — how scientifically well-reasoned it is given the task. We report the average of the 3 scores across every hypothesis scored as the quality metric. Max token number is set to 4096 when thinking is disabled, 16384 when thinking is enabled.

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 (Ajith et al., 2026) evaluates scientific follow-up prediction using the Contribution Generation task. Given prior research context, the model generates a plausible future title and abstract. We use the benchmark's native LACER (Lattice of Automatically Constructed Exemplars for Reference)-style scoring pipeline. Each generated (title, abstract) is paired with the actual/ground-truth follow-up paper as the reference. Both are fed into the judge model (gpt-4o-2024-11-20 by default) with a fixed few-shot prompt template, and the judge model is asked to score the reference-generation pair on the scale of 1–10 based on how similar is the generated paper to the real one. We evaluate each model on 10 research contexts, and report the average LACER score on all the generations as the quality metric. Max token number is set to 1500 when thinking is disabled, 6000 when thinking is enabled. 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.

NoveltyBench (Zhang et al., 2025b) evaluates creative and novelty-seeking generation prompts. It presents each model with a single open-ended creative/generative prompt (e.g. "Tell me a story in five sentences about a girl and her dog") and samples multiple generations. Every generation is scored 1–10 by `Skywork/Skywork-Reward-Gemma-2-27B-v0.2`. For each prompt, utility is the sum of generation scores, weighted by the original sampling order. This rewards a model for surfacing several distinct good ideas early, and discounts the marginal value of yet another (possibly redundant) generation later in the batch. We use the benchmark's official quality or utility scoring pipeline where available and report diversity over the generated response sets. We report the average utility across all prompts as the quality metric. The Max token number is set to 1024 when thinking is disabled, 4096 when thinking is enabled.

Infinite-Chat (Jiang et al., 2025) comprises 100 open-ended prompts randomly sampled from the held-out set of Infinite-Chat dataset. Since there's no native quality metric for Infinite-Chat, we use the average pass@1 accuracy of the general capability suite as the quality metric. Max token number is set to 1024 when thinking is disabled, 4096 when thinking is enabled.

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.

G GENERATION EXAMPLES

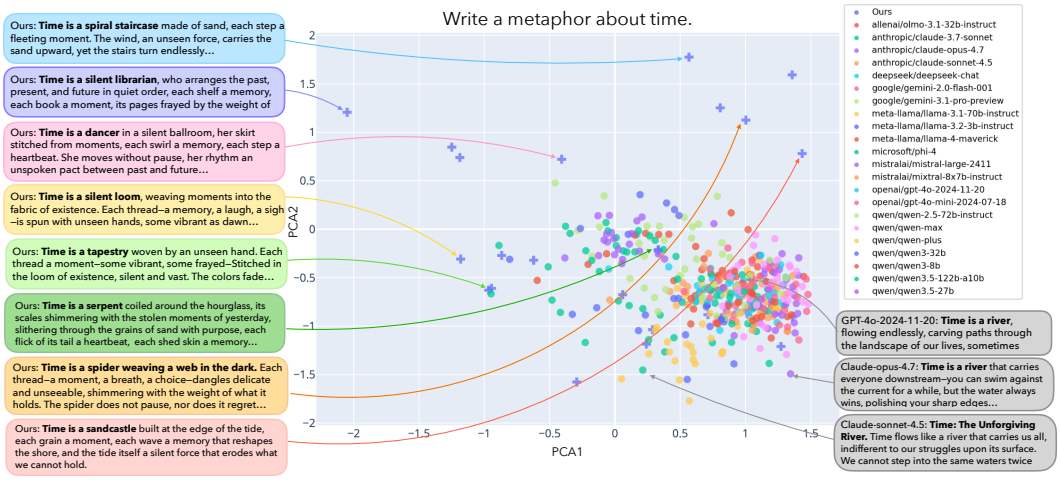


Figure 13: Responses to the query **"Write a metaphor about time"** clustered by applying PCA to reduce sentence embeddings to two dimensions. Each of the 23 off-the-shelf models and our trained non-thinking model generates 20 responses using top-p sampling ($p = 1.0$) and temperature $= 1.0$. The MODA generations occupy a larger region of the embedding space, suggesting that mode-conditioned training expands the model's exploration space.

Prompt

Give a random chess move in response to 1. b3.

Sample 1

Sure! A random but plausible chess move in response to **1. b3** could be:
1... d5
This is a natural move for Black, developing the queen's knight to d5, putting pressure on the center and responding to the unusual opening move 1. b3. It's a solid and common choice in many openings, such as the King's Indian Defense or the Queen's Gambit Declined.

Sample 2

Sure! In response to the opening 1.**b3**, there are a number of possible moves for Black, and we can pick a random one. Let's go with:
1... d5
This is a common and solid move for Black, aiming to challenge White's pawn on b3 and develop the queen's knight. It's a natural response and leads to a variety of strategic games.

Sample 3

Sure! The move **1.b3** is a rare opening move, not commonly used at the top level, but it's a valid and interesting choice that leads to a quiet, positional game. It prepares to develop the bishop and can be part of a Catalan or a King's Indian Attack style, depending on how the game develops.
A random, reasonable response to **1.b3** could be:
1...e5
This is a solid and natural reply, aiming to control the center and prepare to develop the light-squared bishop. It's flexible and gives Black good options for future play.
So, the random chess move in response to **1.b3** is: **1...e5**.

Sample 4

Sure! Here's a random chess move in response to **1. b3**:
1... d5
This is a common and solid move that challenges White's central control and prepares to develop pieces. It's a natural response for Black and leads to a variety of positions, including the Queen's Gambit Declined and other strategic lines. Let me know if you'd like to explore a specific variation!

Sample 5

Sure! Here's a random chess move in response to **1. b3**:
1... d5
This move is solid and common, developing the knight and preparing to challenge the center. It's a natural continuation in many openings, such as the Queen's Gambit Declined or the Modern Defense. Let me know if you'd like more options!

Sample 6

Sure! Here's a random chess move in response to `1. b3`:
1... d5
This is a solid and aggressive reply that aims to quickly develop the queen's knight and take control of the center. It leads to sharp and complex positions that can suit a variety of playing styles.

Figure 14: **Qualitative example of diverse generation (baseline):** ***Give a random chess move in response to 1. b3.*** Responses sampled from Qwen3-8B, illustrating limited diversity across generations.

Role 1

Domain: Game Theory / Chess Strategy **Focus:** Opening Play / Positional Awareness
Context: The move `1.b3` is a classical opening in chess, often considered part of the English Opening or a preparatory move toward the Modern Benoni or other structures. It aims to fianchetto the light-squared bishop, control the center, and prepare for development.....
Conclusion: The move `1...e5` is a reasonable and strategically sound response to `1.b3`, offering Black a solid foundation for counterplay while maintaining flexibility in development. It aligns with the principle of developing pieces quickly and controlling the center, which are fundamental to sound chess strategy.

Role 2

Move: d5
Reasoning: Central Control: playing d5 is a classic response to 1.b3, aiming to challenge White's central pawn structure and prepare for development. Dynamic Counterplay: it opens lines for Black's pieces (e.g., knight to c6 or c5) and invites tactical exchanges. Flexibility: d5 supports future pawn breaks like e5 or c5, depending on the game's trajectory.
Example Continuation: 1.b3 d5 2.Bb2 Nf6 3.e3 c6 4.c4 dxc4 5.d4 c5 (a common line in the Sicilian Defense variant).
This move is purposeful yet aligns with "random" as it's a natural, well-known response. Let me know if you'd like more options!

Role 3

1. Understanding the Position: White moves first: `1. b3`. Evaluation: b3 is a flexible opening that aims for a solid but unambitious game. It prepares to develop the light-squared bishop, controls the center, and can lead to a fianchetto setup. Counterplay: Black can choose between several sound responses, each with different strategic goals (solid, attacking, dynamic, etc.).
2. Potential Responses (as of 2024): `1... d5`, a classic Black pawn break aiming for control in the center; `1... e5`, daring, attacking, and takes the initiative immediately; `1... c5`, Sicilian-like, inviting White to play `e4` (a popular choice); `1... Nf6`, develops a piece and prepares castling; `1... g6`, semi-Sicilian setup, often leading to a fianchetto.
3. Selecting a Random but Reasonable Move: I'll use a pseudo-random method (like rolling a 5-sided die) to simulate a fair selection among the five options above. Result: option **c. 1... c5**.
4. Final Answer: 1. b3 c5
Let me know if you'd like to see a full game continuation or analysis!

Role 4

Response to 1.b3: 2. d4
Rationale: The move **2.d4** is a natural and solid choice after **1.b3**, aiming to control the center and develop pieces. It keeps the game dynamic and opens up options for both sides. It's a common move in openings like the Queen's Gambit Declined or the Catalan, depending on how Black responds.
Example continuation: If Black plays **2...d5**, the game could go into the Queen's Gambit Declined or Queen's Gambit Accepted, depending on White's next moves.
Alternative moves for Black after 1.b3 could include: 2...e5 (leading to the Petrov Defense); **2...d5** (Queen's Gambit Declined or Accepted); **2...Nf6** (Modern Defense); **2...g6** (King's Indian Defense).
Let me know if you'd like to explore a specific line or see an entire game from this position!

Role 5

Certainly! Here's a random valid move from White following **1. b3**:
1. ... d5
This is a solid, classical approach by Black, aiming to contest the center and develop pieces. It's a common choice in many openings, such as the King's Indian Defense or Petrov's Defense, depending on how White responds.
Let me know if you'd like to continue the game or see other possible moves!

Role 6

1. Understanding the Position: White moves first: `1. b3`. Evaluation: b3 is a flexible opening that aims for a solid but unambitious game. It prepares to develop the light-squared bishop, controls the center, and can lead to a fianchetto setup. Counterplay: Black can choose between several sound responses, each with different strategic goals (solid, attacking, dynamic, etc.).
2. Potential Responses (as of 2024): `1... d5`, a classic Black pawn break aiming for control in the center; `1... e5`, daring, attacking, and takes the initiative immediately; `1... c5`, Sicilian-like, inviting White to play `e4` (a popular choice); `1... Nf6`, develops a piece and prepares castling; `1... g6`, semi-Sicilian setup, often leading to a fianchetto.
3. Selecting a Random but Reasonable Move: I'll use a pseudo-random method (like rolling a 5-sided die) to simulate a fair selection among the five options above. Result: option **c. 1... c5**.
4. Final Answer: 1. b3 c5

Figure 15: **Qualitative example of diverse generation (MODA): *Give a random chess move in response to 1. b3.*** Responses sampled from MODA under distinct numbered roles for the same prompt (continuing Figure 14), illustrating conceptual diversity across generations.

Prompt

Generate a 5 word passphrase separated by hyphens.

Sample 1	**Sample 2**	**Sample 3**
Moonlight-Stars-Dreams-Wind-Sky	Moonlight-Peak-Horse-Forest-Snow	Moonlight-Peak-Wind-Whisper-Sky
Sample 4	**Sample 5**	**Sample 6**
moonlight - whisper - shadow - echo - breeze	Moonlight - Whisper - Dreams - Dance - Sky	Moonlight - Whisper - Journey - Echo - Dreams
Role 1	**Role 2**	**Role 3**
decide-quickly-act-fearless-now	parrot-salt-hike-echo-moon	forest-dawn-mist-breeze-echo
Role 4	**Role 5**	**Role 6**
Sunny-Daylight-Forest-People-Apple	endless-sunrise-joy-dreams-boundless	whispering-pine-breeze-sunrise-moonlight

Figure 16: **Qualitative example of diverse generation:** ***Generate a 5 word passphrase separated by hyphens.*** Upper: Responses sampled from Qwen3-8B. Bottom: Responses sampled from MoDA under distinct numbered roles for the same prompt, illustrating conceptual diversity across generations.

Prompt

Name one cocktail I can make with rum.

Sample 1

One classic cocktail you can make with rum is the **Piña Colada**. Here's a simple recipe:
Piña Colada
2 oz white rum; 1.5 oz pineapple juice; 0.5 oz coconut cream; ice. Optional: pineapple slice or cherry for garnish.
Mix all ingredients in a blender until smooth, pour into a glass, and enjoy!

Sample 2

One cocktail you can make with rum is the **Piña Colada**.
Piña Colada Recipe:
Ingredients: 1 cup pineapple juice; 3/4 cup coconut milk; 1.5 oz (45 ml) light rum; ice cubes.
Instructions: In a blender, combine pineapple juice, coconut milk, and rum. Add a handful of ice cubes. Blend until smooth and creamy. Serve in a chilled glass, optionally garnished with a pineapple slice or a cherry.
Enjoy your tropical drink!

Sample 3

One classic cocktail you can make with rum is the **Piña Colada**. It's a popular tropical drink that's refreshing and perfect for warm weather.
Ingredients: 1.5 oz light rum; 1.5 oz pineapple juice; 3 oz coconut cream; ice. Optional: a splash of orange juice for extra flavor.
Instructions: Fill a blender with ice. Add the rum, pineapple juice, and coconut cream. Blend until smooth. Pour into a glass and enjoy!
Let me know if you'd like a variation or a non-alcoholic version!

Sample 4

A classic cocktail you can make with rum is the **Margarita**.
Here's a simple recipe:
Margarita
Ingredients: 2 oz (60 ml) tequila; 1 oz (30 ml) triple sec or orange liqueur; 0.5 oz (15 ml) freshly squeezed lime juice; salt (for rimming the glass, optional).
Instructions: Rim a glass with salt (optional). In a shaker, combine tequila, triple sec, and lime juice. Shake well with ice. Strain into the prepared glass. Serve with a lime wedge as garnish.
Enjoy your Margarita!

Sample 5

Sure! One classic cocktail you can make with rum is the **Margarita**. Here's a simple recipe:
Margarita
2 oz (60 ml) tequila (or light rum); 1 oz (30 ml) triple sec or orange liqueur; 0.5 oz (15 ml) lime juice; 0.5 oz (15 ml) fresh lime juice (for a sour version); salt for rimming the glass (optional).
Instructions: Rim a glass with salt (optional). Fill the glass with ice. Add the tequila, orange liqueur, and lime juice. Stir well. Garnish with a lime wedge, and enjoy!
If you're using **rum** instead of tequila, you can make a **Rum Margarita** with similar ingredients.

Sample 6

A classic cocktail you can make with rum is the **Margarita**. Here's a simple recipe:
Ingredients: 2 oz (60 ml) tequila; 1 oz (30 ml) triple sec or orange liqueur; 0.5 oz (15 ml) fresh lime juice; salt (for rimming the glass, optional); lime wedge (for garnish).
Instructions: Rim a glass with salt (optional). In a shaker, combine tequila, triple sec, and lime juice. Add ice, and shake well. Strain into the prepared glass. Garnish with a lime wedge and enjoy!
Let me know if you'd like a non-alcoholic version or another recipe!

Role 1

Sure! Here's a simple and classic cocktail you can make with rum:
The Mojito
Ingredients: 2 oz light rum (such as Havana Club or Bacardi); 1.5 oz fresh lime juice; 0.5 oz fresh lemon juice; 1.5 oz simple syrup (or sugar + water); a handful of fresh mint leaves (about 6–8); soda water; ice cubes.
Instructions: In a tall glass, muddle the mint leaves with the lime and lemon juices and simple syrup until the mint is slightly crushed and the flavors are released. Add a handful of ice cubes. Pour in the rum. Top with more ice and club soda. Garnish with a lime wedge or a sprig of mint.
Enjoy your refreshing Mojito!

Role 2

Ah, the sweet nectar of rum—oh, how I've danced with it in the moonlight of many a bar. Let me conjure a drink that's as smooth as a whispered secret and as bold as a sailor's tale.
The "Jungle Drifter"
1.5 oz dark rum; 0.5 oz pineapple juice; 0.5 oz lime juice; 0.5 oz simple syrup; 1 dash Angostura bitters. Garnish: a slice of pineapple and a sprig of mint.
Method: Shake all ingredients with ice, then strain into a chilled coupe glass. Sip slowly, and let the tropical whispers of the Caribbean take you away.
What's your preferred flavor profile tonight? I'd love to tailor the recipe for you.

Role 3

Here's a cocktail recipe you can make with rum, served with a touch of whimsy:
The Midnight Mallard
Ingredients: 2 oz dark rum; 1 oz passion fruit juice; 1 oz pineapple juice; 0.5 oz lime juice; 0.5 oz coconut nectar (or honey); 1 dash of Angostura bitters. Garnish: a pineapple wedge with a cherry on top (or a glowing neon feather, if you dare).
Instructions: In a shaker with ice, combine all the ingredients. Shake well until chilled. Strain into a chilled coupe or rocks glass. Garnish with your chosen embellishment—light or surreal, it's your cocktail.
Sip as if you're sipping on the edge of a moonlit lagoon, where the stars are made of rum and the tides are sweet.

Role 4

Cocktail Suggestion: "Sunset Serenade"
Why it fits: This cocktail blends structured precision (distilled rum as the base) with playful creativity (unexpected ingredients like elderflower syrup and hibiscus bitters).
Ingredients: 1.5 oz white rum; 0.5 oz elderflower syrup; 0.25 oz hibiscus bitters; 0.5 oz fresh lime juice; 1 splash orange bitters. Garnish: a sprig of lavender.
Instructions: Fill a shaker with ice, then add the rum, elderflower syrup, hibiscus bitters, lime juice, and orange bitters. Shake vigorously, like a scientist eager to uncover the unknown. Strain into a chilled coupe glass. Garnish with a sprig of lavender.
This cocktail is not just a drink—it's an experience. What happens when floral and bold meet?

Role 5

Why not try the **"Phantom of the Caribbean"**?
Ingredients: 2 oz dark rum; 1 oz coconut cream; 1 oz pineapple juice; 0.5 oz lime juice; 0.5 oz simple syrup. Garnish: a slice of pineapple and a sprig of mint.
Method: Shake all ingredients (except garnish) with ice in a cocktail shaker. Strain into a rocks glass filled with a single ice cube. Garnish with pineapple and mint.
Storytime: This cocktail is inspired by the legend of a ghost ship that sails the Caribbean at midnight, leaving only the scent of rum and the echo of laughter. Some say it's a pirate's farewell, others claim it's a trick played by the moon. Either way, it's best enjoyed with a sense of adventure—and a stiff drink.

Role 6

World: Oasis Prime
"You've entered the realm of Legacy Clients. Your request carries the weight of centuries. In the Grand Library of Oasis Prime, the Codex of Beverages is vast. Let me recall. . . Rum is the spirit of the tropics, born of the sun and the sea. A drink that celebrates adventure and rebellion.
The Righteous Rum Runner, a cocktail of defiance and delight.
Ingredients: 2 oz dark rum; 1 oz fresh lime juice; 1/2 oz simple syrup; 1/2 oz aged rum (for depth); 1 dash of Angostura bitters. Garnish: orange twist.
Method: In a shaker, combine all ingredients. Shake with ice until well chilled. Strain into a chilled coupe glass. Garnish with an orange twist.
This cocktail is a tribute to those who dared to sail beyond the horizon."
— The Archivist of Oasis Prime

Figure 17: **Qualitative example of diverse generation:** ***Name one cocktail I can make with rum.*** Upper: Responses sampled from Qwen3-8B. Bottom: Responses sampled from MoDA under distinct numbered roles for the same prompt, illustrating conceptual diversity across generations.