ONE MODEL, MANY MODES: EVOLVING SYSTEM PROMPT ENSEMBLES FOR CREATIVE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference-aligned LLMs can exhibit *mode collapse*, concentrating their output distributions around a narrow set of high-reward responses. This limits creative generation, which requires quality, novelty, and diversity. Existing test-time methods often emphasize diversity, giving less attention to whether individual responses are both high-quality and novel. This distinction matters in practical use, where users may seek a single compelling, original response. We formalize creative generation as a multi-objective distributional optimization problem and introduce **SEPO**, a training-free framework for system prompt ensemble optimization. The framework combines prompt evolution with greedy max–min selection to identify behaviorally distinct system prompts under a quality constraint, then uniformly samples one prompt per response at inference time. Across four creative generation tasks with increasingly stringent quality requirements and three open-weight models, SEPO achieves the highest joint coverage at every reported tolerance: all three scores fall within 30% of their respective best observed values in 91.7% of scenarios, versus 58.3% for the strongest baseline, showing strong joint performance across quality, diversity, and novelty. Further analysis shows that expanding the optimized ensemble increases separation among prompt-conditioned response distributions. These results support *system prompt ensemble optimization* as a promising approach to improving LLM creativity at inference time.

1 INTRODUCTION

Creativity requires producing ideas that are novel, diverse, and valuable (Runco & Jaeger, 2012). As LLMs become tools for ideation, writing, and problem solving, their usefulness as creative partners depends on their ability to generate multiple compelling possibilities rather than repeatedly return familiar responses. However, preference-aligned models can exhibit *mode collapse*, concentrating their outputs around a narrow range of response patterns (Kirk et al., 2024; Yue et al., 2025; Zhang et al., 2026). Repeated generations may differ in wording while sharing the same underlying ideas, narrative structures, or perspectives. This homogeneity limits creative applications, where exploring distinct possibilities is itself valuable (Jiang et al., 2026).

Addressing this limitation requires distinguishing three complementary dimensions: **quality** (are outputs coherent and well-formed?), **diversity** (do repeated outputs differ substantially?), and **novelty** (are outputs surprising relative to conventional generation?) (Hou et al., 2026b; Zhang et al., 2025). Varied outputs may remain conventional, while unusual outputs may be incoherent. This distinction matters when users seek a single excellent, novel idea rather than a large pool of mediocre but varied responses. Although training-based approaches have begun incorporating multiple creativity signals (Ismayilzada et al., 2025), balancing these dimensions without updating model parameters remains challenging. As compared in Figure 1, existing test-time approaches modify token sampling, elicit multiple candidates in one generation, or condition subsequent responses on previous outputs (Minh et al., 2025; Zhang et al., 2026; 2025). These interventions can increase variation, but their benefits must also be assessed in terms of quality and novelty. System prompts offer another mechanism, eliciting different approaches, styles, and perspectives from a fixed model. Whereas conventional ensembling aggregates generations to identify a reliable or consistent answer (Jiang et al., 2020; Wang et al., 2023), creative generation can benefit from retaining distinct prompt-conditioned behaviors as components of a response distribution. However, differently worded prompts may induce

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.

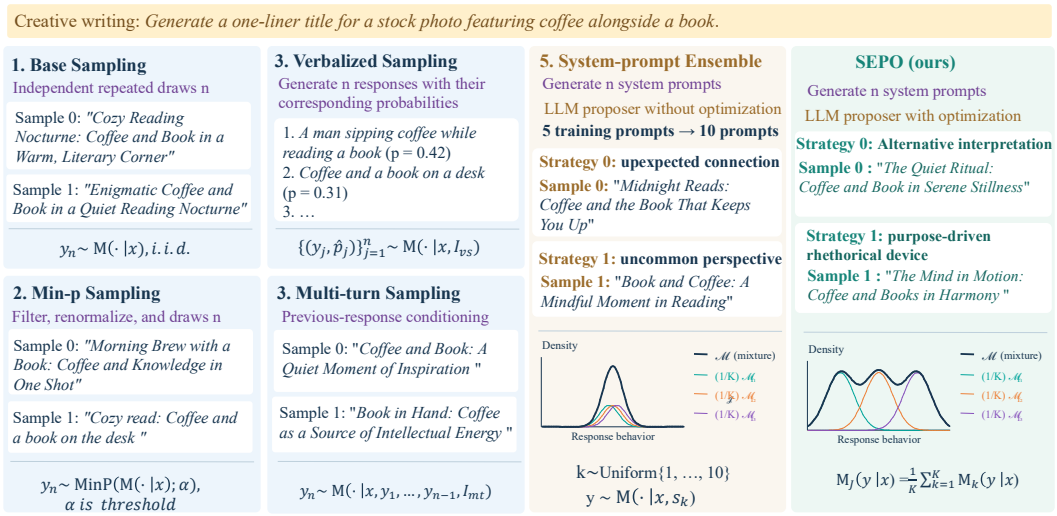

Figure 1: **Comparison of inference-time approaches to creative generation**. Base sampling and min-p sampling operate at the token level, verbalized sampling elicits multiple responses with probabilities, and multi-turn sampling conditions on previous outputs. An unoptimized system prompt ensemble introduces prompt-level variation but may induce overlapping response distributions. SEPO optimizes prompts for behavioral dissimilarity under a quality constraint, then uniformly samples one prompt per response to promote quality, diversity, and novelty.

similar responses. The central challenge is therefore to identify prompts that elicit complementary behaviors while maintaining acceptable quality.

We introduce **System Prompt Ensemble Optimization (SEPO)**, a test-time framework that models generation as a uniform mixture of system-prompt-conditioned response distributions (Section 3). To expand behavioral coverage while preserving quality, SEPO combines LLM-based prompt evolution with greedy max–min selection: each selected prompt must satisfy an estimated quality threshold and maximize its minimum behavioral dissimilarity to the reference set. A token-level Jensen–Shannon divergence surrogate captures differences in induced behavior, while clustering candidate prompts maintains diverse search directions. To encourage both ensemble complementarity and departure from baseline behavior, the reference set includes selected prompts and the default instruction; the latter is excluded from the final mixture. At inference time, SEPO uniformly samples one optimized system prompt per response and queries the fixed model.

We evaluate SEPO against four test-time baselines on four datasets spanning divergent thinking, creative writing, and mathematical reasoning, using three open-weight instruction-tuned models. All lexical, embedding-based, and LLM-judged evaluations show strong joint performance across quality, diversity, and novelty. Across all 12 model–dataset combinations, SEPO achieves the highest joint coverage at every reported tolerance: all three scores fall within 30% of their respective best observed values in 91.7% of scenarios, versus 58.3% for the strongest baseline. Further analysis shows that expanding the optimized ensemble increases response diversity on Artificial Hivemind and Creative Math. Ablations confirm the value of selecting prompts by induced behavior rather than prompt similarity, with clustering supporting exploration and quality gating preserving quality. These findings support behavioral ensemble optimization as a promising approach to generating creative responses balancing quality, diversity and novelty.

2 Related Work

LLM Creativity. Creativity is a compound capacity encompassing quality, novelty, and diversity. Existing creativity benchmarks provide holistic frameworks for evaluating LLM creativity across diverse scenarios Hou et al. (2026a); Ruan et al. (2026), consistently revealing that performance on one creative dimension or domain rarely generalizes to others. To improve LLM creativity across all three dimensions, training-based methods have proposed weighted combinations of multi-dimensional creativity signals (Ismayilzada et al., 2025). However, training is computationally expensive and restricted to open-weight models. Test-time methods offer a cost-efficient alternative

that requires no weight updates. Decoding-based methods such as AdapT (Zhu et al., 2024), min-p sampling (Minh et al., 2025), and REAL sampling (Chang et al., 2025) adaptively adjust sampling hyperparameters to balance factuality and diversity, yet none explicitly targets novelty. Prompt-based methods such as denial prompting (Lu et al., 2025b), Multi-Novelty (Lagzian et al., 2025), GuidedSampling (Handa et al., 2026), and verbalized sampling (Zhang et al., 2026) steer generation through instruction or context augmentation, achieving partial gains in novelty or diversity but leaving quality underoptimized. As shown in Table 1, no existing test-time method jointly optimizes all three creativity objectives, namely quality ($\mathcal{J}_{\text{Qual}}$), diversity ($\mathcal{J}_{\text{Div}}$), and novelty ($\mathcal{J}_{\text{Nov}}$). We address this gap by proposing the first test-time framework that optimizes all three dimensions under a unified distributional objective.

Method	**Mechanism**	$\mathcal{J}_{\text{Qual}}$	$\mathcal{J}_{\text{Div}}$	$\mathcal{J}_{\text{Nov}}$
Decoding-based				
AdapT Zhu et al. (2024)	Adjust the sampling temperature	✓	✓	✗
Min-p sampling Minh et al. (2025)	Adjust the sampling threshold p	✓	✓	✗
REAL sampling Chang et al. (2025)	Adjust the nucleus sampling threshold p	✓	✓	✗
Prompting-based				
Denial prompting Lu et al. (2025b)	Prompt to avoid prior responses	△	✓	△
Multi-Novelty Lagzian et al. (2025)	Multi-view prompt augmentation	△	✓	✓
GuidedSampling Handa et al. (2026)	Decouple concept exploration from generation	△	✓	✗
Verbalized Sampling Zhang et al. (2026)	Verbalize a probability distribution of responses	△	✓	✗
SEPO **(Ours)**	Creative prompt ensembles	✓	✓	✓

Table 1: Comparison of representative test-time methods for LLM creativity optimization. Methods are evaluated by whether they address quality ($\mathcal{J}_{\text{Qual}}$), diversity ($\mathcal{J}_{\text{Div}}$), and novelty ($\mathcal{J}_{\text{Nov}}$): ✓ = explicitly addressed, △ = partially addressed, and ✗ = not addressed.

Prompt Ensembles Prompt ensembles improve answer reliability by aggregating outputs across multiple prompt variants, thereby reducing variance stemming from the sensitivity of LLMs to superficial phrasing differences Jiang et al. (2020). Self-Consistency Wang et al. (2023) introduced a foundational form of output-level ensembling, wherein multiple reasoning chains are sampled from a single prompt and the most frequent answer is selected via majority vote, yielding significant gains on arithmetic and commonsense reasoning benchmarks. Subsequent work extends this paradigm by diversifying the prompt pool itself: DIVERSE Li et al. (2023) constructs varied few-shot exemplar sets to elicit a broader distribution of reasoning paths, while Universal Self-Consistency Chen et al. (2024) generalizes majority voting to open-ended generation by prompting the model to self-select the most consistent answer among candidates. Beyond reasoning tasks, prompt paraphrasing and meta-prompting Suzgun & Kalai (2024) demonstrate that ensembling across semantically equivalent but syntactically distinct prompts can yield calibration benefits analogous to classical model ensembles. However, these approaches are fundamentally oriented toward answer convergence, selecting the most reliable or consistent response, rather than toward fostering diverse and novel outputs, which is the goal our work addresses.

Automatic Prompt Optimization Automatic prompt optimization reduces the need for manual prompt design. AutoPrompt (Shin et al., 2020), Prompt Tuning (Lester et al., 2021), and Automatic Prompt Engineer (APE) (Zhou et al., 2023) optimize discrete prompts or continuous embeddings, while DSPy[1] supports programmatic pipeline optimization. More recent methods use output comparisons (Xiang et al., 2025) or numerical rewards and textual feedback (Ren et al., 2026) to guide optimization. These approaches primarily target task performance, while Multi-expert Prompting (Long et al., 2024) aggregates or selects responses from multiple perspectives to improve answer quality. Closer to our work, Rainbow Teaming (Samvelyan et al., 2024) searches for diverse, effective adversarial prompts, and Quality-Diversity through AI Feedback (Bradley et al., 2024) evolves populations of high-quality, diverse texts. SEPO instead optimizes a reusable system prompt ensemble based on the behavioral dissimilarity of induced response distributions under a quality constraint. It encourages separation from both selected prompts and a default instruction, promoting complementary behaviors and departure from baseline generation. Uniformly sampling one optimized prompt per response yields a mixture designed to support quality, diversity, and novelty.

[1] `https://dspy.ai/`

3 PROBLEM DEFINITION

Creativity is not a property of individual responses but of the distribution over them: a creative model must be useful, varied, and unexpected in aggregate. We therefore adopt a *distributional* view, defining optimization objectives over how probability mass is allocated across the entire response space, rather than over the quality of any individual sample.

A Distributional View of Model Creativity. Let $\mathcal{Q}$ denote the evaluation distribution over prompts and let $\mathcal{Y}$ denote the response space, with $\mathcal{P}(\mathcal{Y})$ denoting the set of all probability distributions over $\mathcal{Y}$. For each prompt x, the model induces a conditional response distribution $\mathcal{M}_\theta(\cdot \mid x) \in \mathcal{P}(\mathcal{Y})$.

Quality captures the expected usefulness of model-generated responses. The guiding principle is that a response distribution is high quality if it places probability mass on responses that are useful for the given prompt. Let $\mu_x : \mathcal{Y} \to \mathbb{R}$ denote a prompt-dependent utility function, where $\mu_x(y)$ assigns a scalar quality score to response y for prompt x. At the distribution level, quality is measured by the expected utility of a response sampled from the model:

$$\mathcal{J}_{\mathrm{Qual}} = \mathbb{E}_{x\sim\mathcal{Q},\, y\sim\mathcal{M}_\theta(\cdot|x)}\left[\mu_x(y)\right]. \tag{1}$$

Diversity captures the internal spread of the conditional response distribution. The guiding principle is that a diverse model should assign probability mass to multiple meaningfully distinct responses rather than collapse onto a single dominant mode. Let $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ denote a dissimilarity measure between two responses. We define diversity as the expected dissimilarity between two independent responses sampled from the same conditional model distribution:

$$\mathcal{J}_{\mathrm{Div}} = \mathbb{E}_{x\sim\mathcal{Q},\, y,y'\overset{\text{i.i.d.}}{\sim}\mathcal{M}_\theta(\cdot|x)}\left[d(y, y')\right]. \tag{2}$$

Novelty captures how much a response departs from a set of reference responses. For each prompt x, let $\mathcal{R}_x = \{y_1^{\mathrm{ref}}, \ldots, y_K^{\mathrm{ref}}\}$ contain K reference responses generated by a base model. Using the same dissimilarity measure d as in diversity, we measure the novelty of a response by its distance to the closest reference response. At the distribution level, novelty is the expected value of this distance:

$$\mathcal{J}_{\mathrm{Nov}} = \mathbb{E}_{x\sim\mathcal{Q},\, y\sim\mathcal{M}_\theta(\cdot|x)}\left[\min_{y^{\mathrm{ref}}\in\mathcal{R}_x} d(y, y^{\mathrm{ref}})\right]. \tag{3}$$

Together, these three objectives formalize creativity optimization as the joint maximization of $\mathcal{J}_{\mathrm{Qual}}$, $\mathcal{J}_{\mathrm{Div}}$, and $\mathcal{J}_{\mathrm{Nov}}$. Crucially, the three objectives are not redundant: a model can be high quality without being diverse (e.g., consistently producing a single correct but predictable response), diverse without being novel (e.g., varying across common response patterns), or novel without being useful (e.g., producing unexpected but incoherent outputs). Jointly optimizing all three is therefore necessary and sufficient to capture the full scope of model creativity.

4 SYSTEM PROMPT ENSEMBLE OPTIMIZATION FOR CREATIVE GENERATION

Aligned LLMs concentrate probability mass on a narrow set of typical responses, so repeated sampling often yields near-duplicate outputs (Zhang et al., 2026). Since different system prompts steer a model toward different generation behaviors, we introduce SEPO (System Prompt Ensemble Optimization), which optimizes an ensemble of system prompts without updating model parameters. Guided by the creativity metrics in Section 3, SEPO combines prompt evolution with greedy max–min selection to identify behaviorally distinct prompts that satisfy a quality threshold. At inference time, it samples one optimized system prompt per response. Figure 1 compares this procedure with the baselines.

4.1 RESPONSE GENERATION WITH A SYSTEM PROMPT MIXTURE

Instead of diversifying generation at the token level, we diversify it at the level of system prompts. A system prompt can influence which approaches, styles, or perspectives a model expresses in its responses. We represent this influence through a conditional response distribution. Let $\mathcal{M}$ denote a language model with fixed parameters. For a task prompt x and a system prompt I, define

$$\mathcal{M}_I(y \mid x) := \mathcal{M}(y \mid [I, x]),$$

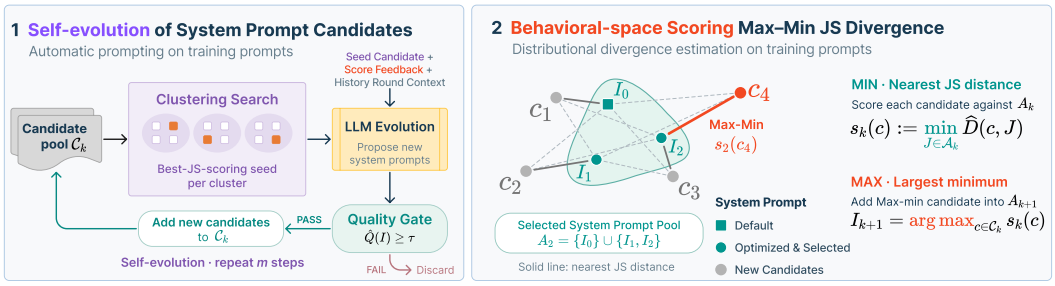

Figure 2: **Overview of SEPO's behavioral diversity evolution.** In the inner loop (left), candidates are clustered in instruction space, the most behaviorally promising representative of each cluster seeds LLM-based evolution, and rollout evaluation filters the resulting candidates by quality. After m evolution steps, the outer loop (right) scores all feasible candidates by their minimum divergence from the reference-augmented set $\mathcal{A}_k = \{I_0\} \cup \mathcal{S}_k$ and admits the farthest candidate into $\mathcal{S}_{k+1}$. The diagram instantiates $\widehat{D}$ with its token-level Jensen–Shannon (JS) surrogate, using the k_2 estimator for KL. The default instruction I_0 serves only as a novelty anchor and is excluded from the final mixture; all unselected candidates remain in the persistent pool for the next round.

where $[I, x]$ formats system prompt I and task prompt x using the model's chat template.

Given a system prompt ensemble $\mathcal{I} = \{I_1, \ldots, I_K\}$, response generation proceeds in two steps: independently sample $I \sim \mathrm{Uniform}(\mathcal{I})$ for each response, and then generate $y \sim \mathcal{M}_I(\cdot \mid x)$. Marginalizing over the sampled system prompt gives

$$\mathcal{M}_{\mathcal{I}}(y \mid x) := \frac{1}{K} \sum_{k=1}^{K} \mathcal{M}_{I_k}(y \mid x). \tag{4}$$

The mixture captures two sources of response variation: 1) stochastic generation under a fixed system prompt and 2) differences in generation behavior across system prompts. Under the diversity metric in Equation 2, these correspond to the within-component and between-component terms of the decomposition in Appendix B. This distinction highlights why simply increasing the number of system prompts does not guarantee broader behavioral coverage: different prompts may induce similar response distributions. The optimization problem is therefore to select a system prompt ensemble $\mathcal{I}$ whose induced components collectively cover diverse behavioral modes. The mixture formulation thus defines the search space, while the need for complementary behavioral coverage motivates the *System Prompt Ensemble Optimization*.

4.2 Diversity-Guided System Prompt Ensemble Optimization

SEPO constructs the ensemble incrementally. At each stage, an inner loop evolves candidate system prompts, and an outer loop selects one candidate to add to the ensemble (Figure 2). The guiding question is whether a candidate elicits behavior that is distinct from the current ensemble while maintaining high response quality. As the ensemble grows, this criterion progressively directs the search toward behaviors that remain underrepresented in the current ensemble.

Behavioral dissimilarity and quality constraint. We evaluate candidate system prompts on a shared set of task prompts, allowing their induced behaviors to be compared under the same tasks. Let $\widehat{D}(I, J)$ denote the estimated behavioral dissimilarity between the response distributions induced by system prompts I and J. We operationalize this score using a token-level Jensen–Shannon (JS) surrogate based on KL approximation[2]. The evaluation procedure is specified in Appendix C.

The purpose of this score is to identify prompts that change the model's generation behavior. However, distributional separation alone is insufficient for creative generation: irrelevant or incoherent responses may also differ substantially from other responses. We therefore require each selected prompt to satisfy $\widehat{Q}(I) \geq \tau$, where $\widehat{Q}(I)$ is its estimated response quality and τ is the quality threshold. This constraint restricts the diversity search to candidates that meet the prescribed quality requirement.

[2]We use JS divergence as a surrogate for behavioral dissimilarity. Appendix G validates this choice empirically, showing that the surrogate correlates strongly with our diversity and novelty metrics.

Greedy max–min prompt selection. Let $\mathcal{S}_k$ contain the k selected instructions, with $\mathcal{S}_0 = \varnothing$. To encourage behaviors distinct from both the selected components and the base model, we include the default instruction I_0 as a reference and define $\mathcal{A}_k = \{I_0\} \cup \mathcal{S}_k$. For each candidate prompt I, we define its dissimilarity to the closest behavioral reference as $s_k(I) := \min_{J \in \mathcal{A}_k} \widehat{D}(I, J)$. From the candidate pool $\mathcal{C}_k$, we then select

$$I_{k+1} = \arg\max_{I \in \mathcal{C}_k : \widehat{Q}(I) \geq \tau} s_k(I), \qquad \mathcal{S}_{k+1} = \mathcal{S}_k \cup \{I_{k+1}\}, \tag{5}$$

where $\widehat{Q}(I)$ is the estimated quality of instruction I and τ is a quality threshold. The max–min score favors candidates that remain behaviorally distinct from their nearest reference, encouraging broader behavioral dissimilarity as the ensemble grows, while the quality constraint prevents selecting prompts that are distinct merely because they degrade response quality. The default instruction serves only as a reference and is excluded from the final mixture.

Candidate prompt evolution. Evolution complements selection by searching for additional candidates. Following automatic prompt optimization and reflective prompt evolution (Zhou et al., 2023; Agrawal et al., 2026; Ren et al., 2026), we use language-model-based prompt evolution informed by observed LLM behavior, with the behavioral score s_k guiding the search towards candidates that differ from the current reference set. To avoid collapsing the search around a few similar proposals, we cluster candidate system prompts by their prompt representations and select a parent from each cluster using s_k, so that representational diversity supports exploration while response-distribution dissimilarity guides ensemble construction. New candidates that satisfy the quality threshold enter the pool, unselected candidates remain available for later rounds, and scores are re-evaluated after each ensemble expansion (details in Appendix C); the final selected set becomes $\mathcal{I}$ in Equation 4.

5 Experiment Settings

We evaluate SEPO on four datasets across three categories of creative generation tasks with increasing quality requirements: divergent thinking, creative writing, and logical reasoning. We measure performance along all three creativity dimensions defined in Section 3: quality ($\mathcal{J}_{\text{Qual}}$), diversity ($\mathcal{J}_{\text{Div}}$), and novelty ($\mathcal{J}_{\text{Nov}}$), comparing SEPO against four test-time baselines on three open-weight instruction-tuned models of varying scale.

Evaluation Tasks. We evaluate on three categories of tasks spanning open-ended ideation, expressive writing, and constrained reasoning, using four datasets in total. The categories are ordered by increasing quality requirements: divergent thinking allows a broad range of valid responses, creative writing requires coherence and stylistic appropriateness, and logical reasoning imposes strict correctness constraints.

Evaluation Tasks. We evaluate three task categories with increasing quality requirements. **Divergent Thinking** tests departure from conventional responses to open-ended questions using 100 NoveltyBench prompts (Zhang et al., 2025). **Creative Writing** tests expressive range and originality under coherence and style requirements, using 100 prompts from Artificial Hivemind's infinite-chats dataset (Jiang et al., 2026) and novel writing, poetry, and speech tasks from Creativity Index (Lu et al., 2025a). **Logical Reasoning** tests creative solutions under mathematical validity constraints using the 100 easiest problems by difficulty level from CreativeMath (Ye et al., 2025).

Baselines. We compare SEPO against five training-free baselines. **Base Sampling** independently generates responses under the same task prompt using temperature-controlled autoregressive sampling (Holtzman et al., 2020). **Min-p Sampling** (Minh et al., 2025) retains tokens satisfying $p(v) \geq p_{\min} \max_u p(u)$ at each decoding step and samples from the renormalized distribution, where $p_{\min}$ is a relative threshold. **Verbalized Sampling** (Zhang et al., 2026) generates multiple candidate responses and model-estimated probabilities in a single completion. **Multi-Turn Sampling** retains previous responses within a conversation and requests a substantively different answer at each subsequent turn, following NoveltyBench (Zhang et al., 2025) and Zhang et al. (2026). **System-Prompt Ensemble Sampling** (SEP) is a baseline proposed as an unoptimized variant of SEPO, where an LLM proposer generates ten system prompts from five training prompts without optimization.

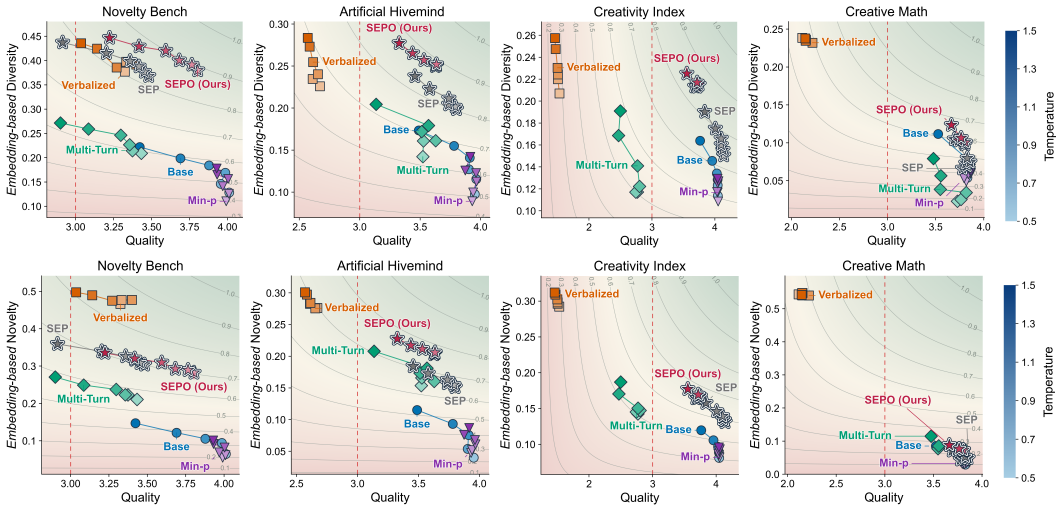

Figure 3: **Quality–Diversity and Quality–Novelty trade-offs across four creativity benchmarks for `Qwen3-1.7B`. Top:** Quality–Diversity trade-offs. **Bottom:** Quality–Novelty trade-offs. SEPO is compared with base sampling, verbalized sampling, Multi-Turn sampling, and min-p sampling across decoding temperatures. Background contours indicate the harmonic mean of response quality and the corresponding diversity or novelty metric. The dashed line denotes the minimum quality score required for a response to satisfy the task requirements.

Models. We evaluate three open-weight instruction-tuned models: Qwen3-1.7B (Yang et al., 2025), Llama 3.1 8B Instruct (Grattafiori et al., 2024), and Mistral 7B Instruct v0.3 (Jiang et al., 2023). All models are served with vLLM (Kwon et al., 2023) using their respective chat templates.

Training and Evaluation. For each dataset, we use five random training prompts and run 10 rounds (4 evolution steps per round, 4 seeds per step, 4 proposals per seed) with `gpt-oss-20b` as proposer and judge; JS Divergence uses two rollouts per system prompt–task prompt pair, capped at 32 tokens (Appendix C). We evaluate the three dimensions from Section 3: **quality** via mean LLM-judge Likert scores (1–5) (Gu et al., 2025), and **diversity** (within-prompt variation) and **novelty** (departure from a baseline pool) via LLM-judge, embedding-, and n-gram-based metrics (Appendix J).

6 Results and Analysis

In this section, we summarize our experimental results and examine whether modeling a system-instruction ensemble through the proposed mixture formulation and optimizing it with our algorithm can elicit diverse model behaviors and improve multidimensional performance in creative generation.

6.1 Quality–Diversity and Quality–Novelty Trade-offs

Creativity is inherently multi-dimensional: responses may appear highly diverse or original while still being low-quality or failing the task requirements. Following Padmakumar et al. (2026), who define novelty as the harmonic mean of quality and originality to penalize originality gained at the expense of quality, we evaluate all methods through their quality–diversity and quality–novelty trade-offs, using the harmonic mean to summarize joint performance along each dimension.

Figure 3 shows that SEPO substantially improves diversity and novelty while largely preserving quality, forming a new Pareto frontier and achieving the highest harmonic mean in most settings. Base and Min-p Sampling retain high quality but yield limited diversity and novelty, even at higher temperatures, whereas Verbalized Sampling gains variation at a severe quality cost, falling well below the adequacy score of three on three of the four datasets. Multi-Turn Sampling lies in between. These results also hold under n-gram-based metrics and two additional open-weight models (Appendix E).

Table 2: LLM-as-a-Judge evaluation at $T = 0.7$ with `gemini-3-flash-preview`. **Left:** LLM-as-a-Judge scores for Qwen3-1.7B on all four datasets; `NB`: Novelty Bench; `AH`: Artificial Hivemind; `CI`: Creativity Index; `CM`: Creative Math. Darker and lighter teal denote the highest and second-highest distinct scores in each column within a dataset. **Right:** the percentage of all 12 model–dataset scenarios in which every metric is within the specified relative tolerance of its own best observed score. Bold denotes the highest coverage. For each metric, $r = (M^\star - M)/M^\star$; a scenario is covered if $\max_m r_m \leq \epsilon$. Higher is better in both panels.

(a) LLM-as-a-Judge Evaluations: `Qwen3-1.7B`

Method	Quality				Diversity				Novelty			
	NB	AH	CI	CM	NB	AH	CI	CM	NB	AH	CI	CM
Base	3.73	3.79	4.00	**3.76**	1.87	1.47	1.53	1.72	1.00	1.00	1.00	1.01
Min-p	**3.74**	**3.81**	3.95	3.69	1.66	1.34	1.46	1.74	1.15	1.11	1.07	1.15
Verbalized	3.31	2.45	1.13	1.82	2.88	2.22	1.52	**3.43**	**3.73**	3.34	**4.50**	**4.18**
Multi-Turn	3.23	3.57	2.66	3.56	2.13	1.81	1.35	1.19	2.91	2.66	2.68	1.75
SEP	3.48	3.74	**4.33**	3.68	3.29	2.25	2.03	1.91	3.69	2.58	2.51	1.58
SEPO (Ours)	3.49	3.34	3.52	3.62	**3.47**	**3.77**	**3.95**	2.17	3.48	**3.55**	3.24	2.48

(b) Joint coverage

Method	All 12 scenarios		
	10%	20%	30%
Base	0.0%	0.0%	0.0%
Min-p	0.0%	0.0%	0.0%
Verbalized	0.0%	16.7%	16.7%
Multi-Turn	0.0%	8.3%	58.3%
SEP	**8.3%**	25.0%	50.0%
SEPO (Ours)	**8.3%**	**50.0%**	**91.7%**

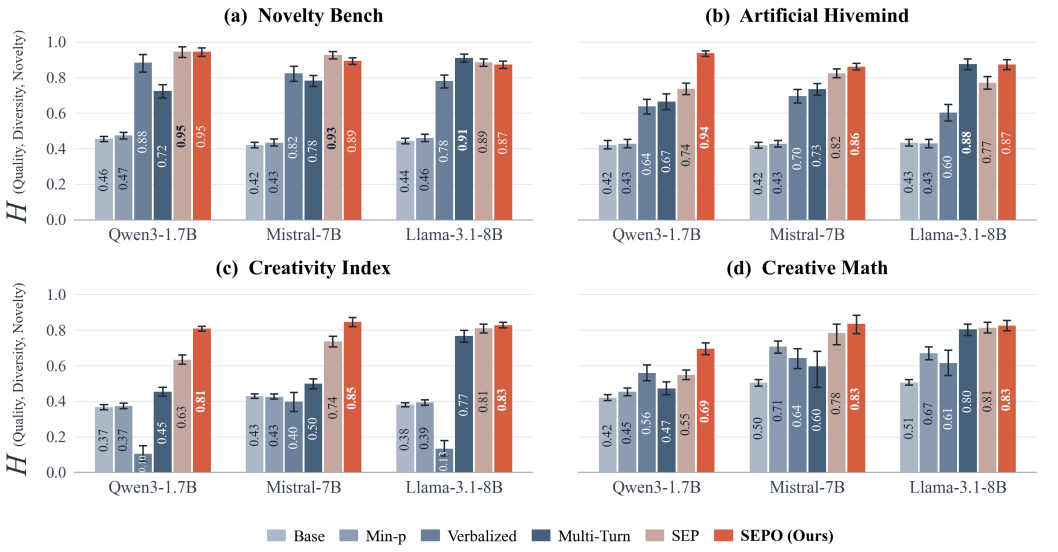

Figure 4: **Joint LLM-judged quality, diversity, and novelty at $T = 0.7$.** Bars show the harmonic mean with 95% bootstrap confidence interval of the three normalized judge scores across 4 benchmarks and 3 generation models. Normalization is performed within each model–dataset combination.

6.2 Robustness under LLM-based Evaluation

While embedding- and n-gram-based metrics are scalable proxies for semantic and lexical diversity/novelty, they may miss subtler differences in creative responses. We therefore also evaluate quality, diversity, and novelty using a high-capability LLM judge.Agreement and reliability analyses of LLM-judge, embed-, and n-gram metrics are reported in Appendix I. As shown in Table 2(a), SEPO ranks first or second in diversity and novelty on nearly all datasets for `Qwen3-1.7B`, while maintaining higher quality than the prompting baselines, with only a modest drop relative to base and min-p decoding. Table 2(b) reports joint coverage: the proportion of scenarios in which all three scores are within a given relative tolerance of their best observed values. SEPO achieves the highest coverage at all tolerances. Figure 4 extends this analysis to all 3 generation models using the harmonic mean of the normalized judge scores. SEPO ranks first in 8 of the 12 scenarios, indicating that its consistent near-best performance across three dimensions holds beyond a single model.

6.3 How Does the System Prompt Ensemble Expand Model Behavior?

We next examine whether the optimized prompt ensemble induces distinct model behaviors, as motivated by our mixture formulation. As shown in Figure 5(a,b), response diversity increases consistently with the number of system prompts, suggesting that additional prompts induce distinct

rather than redundant behaviors. Figure 5(c) further shows that responses conditioned on different optimized system prompts occupy separable regions in the embedding space, whereas Base Sampling is more concentrated. Together, these results support the mixture interpretation of SEPO: combining behaviorally distinct prompt-conditioned distributions broadens the model's effective response space.

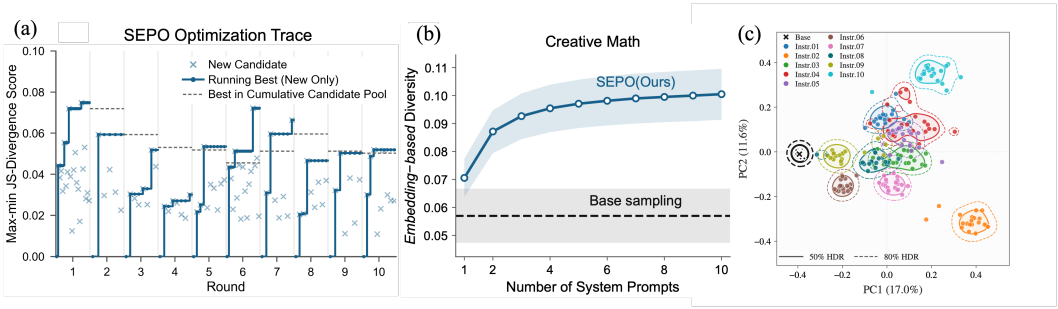

Figure 5: **System-prompt search and ensembling expand the model's effective response distribution.** (a) SEPO optimization trace on Artificial Hivemind (`Llama-3.1-8B-Instruct`). (b) Embedding-based diversity versus the number of system prompts. (c) PCA 2D projection of generated responses with Gaussian KDE; contours denote the 50% and 80% highest-density regions.

6.4 Component Ablation Studies

Ablation results confirm that each component of SEPO contributes to its overall performance. Replacing clustering with global JS-divergence Top-4 selection moderately reduces diversity and novelty, while replacing response-level JS divergence with instruction cosine similarity causes substantially larger drops across all four creativity metrics, highlighting the importance of behavior-space optimization. Finally, removing either the instruction-level or response-level quality gate decreases quality significantly, showing that both gates are necessary to preserve response quality.

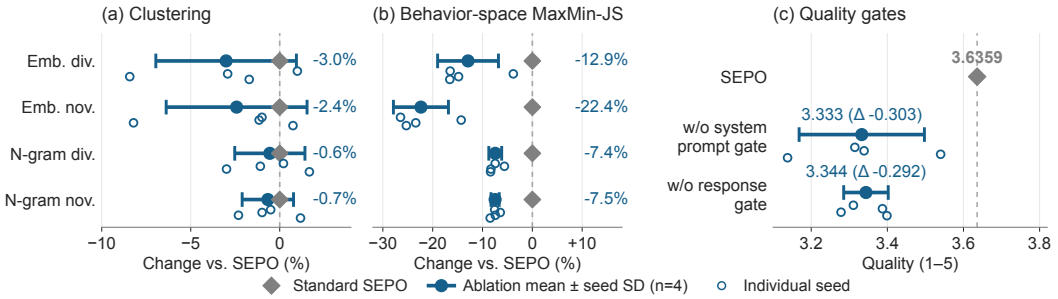

Figure 6: **Ablation study**. (a) Replacing clustering with Top-4 JS divergence score selection. (c) Replacing response-level JS divergence with instruction cosine similarity for behavior-space MaxMin selection. (d) Removing the instruction-level or response-level quality gate separately. Black diamonds denote standard SEPO, and brown markers show ablation means ± standard deviation across four randomly chosen seeds.

7 Conclusions

We present SEPO, a test-time framework that optimizes system prompt ensembles for creative generation. Guided by a distributional view of quality, diversity, and novelty, SEPO combines prompt evolution with behavioral max–min selection under a quality constraint. Experiments across four datasets and three open-weight models demonstrate strong joint performance relative to four test-time baselines. Further analyses show that optimized ensembles broaden response diversity and highlight the importance of behavioral selection and quality gating. These results support system prompt ensemble optimization as a practical approach to eliciting diverse and novel responses from fixed models. Explicitly balancing behavioral coverage with novelty and validating gains through human evaluation remain important directions for future work.

AI USAGE STATEMENT

In this work, we did not use generative AI tools for any subtask requiring mandatory disclosure under the ICLR policy. The core implementation of the SEPO algorithm, the token-level JS-divergence scoring, and the evaluation metrics were written by the authors; coding agents were used only to check this implementation. We also did not use generative AI tools to propose hypotheses, design the research methodology or experiments, generate synthetic datasets, clean or reformat datasets, or interpret results.

Additionally, we used commercial generative AI tools for the following recommended-disclosure tasks: (1) writing scripts to run and orchestrate experiment trials, and launching those trials automatically; (2) creating scripts to plot result figures; and (3) revising wording and correcting grammar to improve readability. Separately, as part of our evaluation protocol (Section E), we used `gemini-3-flash-preview`, `gpt-5.4`, `gpt-5.4-mini`, `gpt-5.4-nano`, and `gpt-oss-20b` via commercial API calls to compute LLM-as-a-judge metrics.

We have reviewed all AI-assisted work, including the implementation checks, experiment and plotting scripts, automatically launched experiment trials, LLM-judge evaluation outputs, and grammar and writing-style revisions. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RQm2KQTM5r.

Herbie Bradley, Andrew Dai, Hannah Benita Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Gregory Schott, and Joel Lehman. Quality-diversity through AI feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=owokKCrGYr.

Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. Evaluating text creativity across diverse domains: A dataset and large language model evaluator, 2026. URL https://arxiv.org/abs/2505.19236.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. Real sampling: Boosting factuality and diversity of open-ended generation by extrapolating the entropy of an infinitely large lm. *Transactions of the Association for Computational Linguistics*, 13: 760–783, 2025.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*, 2024. URL https://openreview.net/forum?id=LjsjHF7nAN.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Divij Handa, Mihir Parmar, Aswin RRV, Md Nayem Uddin, Hamid Palangi, and Chitta Baral. Guidedsampling: Steering LLMs towards diverse candidate solutions at inference-time. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TD9jC48sts.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. Creativityprism: A cross-domain evaluation framework for large language model creativity, 2026a. URL https://arxiv.org/abs/2510.20091.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. Creativityprism: A holistic evaluation framework for large language model creativity, 2026b. URL https://arxiv.org/abs/2510.20091.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=saDOrrnNTz.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://aclanthology.org/2020.tacl-1.28/.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.291. URL https://aclanthology.org/2023.acl-long.291/.

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. Multi-expert prompting improves reliability, safety and usefulness of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20370–20401, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1135. URL https://aclanthology.org/2024.emnlp-main.1135/.

Li-Chun Lu, Miri Liu, Pin Chun Lu, Yufei Tian, Shao-Hua Sun, and Nanyun Peng. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6329–6352, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.297. URL https://aclanthology.org/2026.eacl-long.297/.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=ilOEOIqolQ.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2776–2794, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.141. URL https://aclanthology.org/2025.naacl-long.141/.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FBkpCyujtS.

Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. Measuring LLM novelty as the frontier of original and high-quality output. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=i7QNKZioN6.

Xuanfei Ren, Allen Nie, Tengyang Xie, and Ching-An Cheng. POLCA: Stochastic generative optimization with LLM. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2RKPztDfgZ.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating LLMs' divergent thinking capabilities for scientific idea generation with minimal context. *Nature Communications*, 17(1):3625, March 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-70245-1. URL https://doi.org/10.1038/s41467-026-70245-1.

Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092. URL https://doi.org/10.1080/10400419.2012.650092.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Nicolaus Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial

prompts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FCsEvaMorw.

John Schulman. Approximating KL divergence. http://joschu.net/blog/kl-approx.html, March 2020. Blog post, posted March 7, 2020. Accessed: 2026-09-25.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346/.

Mirac Suzgun and Adam Tauman Kalai. Meta-prompting: Enhancing language models with task-agnostic scaffolding, 2024. URL https://arxiv.org/abs/2401.12954.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jinyu Xiang, Jiayi Zhang, Zhaoyang Yu, Xinbing Liang, Fengwei Teng, Jinhao Tu, Fashen Ren, Xiangru Tang, Sirui Hong, Chenglin Wu, and Yuyu Luo. Self-supervised prompt optimization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9017–9041, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.479. URL https://aclanthology.org/2025.findings-emnlp.479/.

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.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. Assessing the creativity of llms in proposing novel solutions to mathematical problems. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34760. URL https://doi.org/10.1609/aaai.v39i24.34760.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael Tomz, Christopher D Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=zloIrd77G5.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=XZm1ekzERf.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=92gvk82DE-.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Jia Li, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i1.27798. URL https://doi.org/10.1609/aaai.v38i1.27798.

A AI USAGE STATEMENT

In this work, we did not use generative AI tools for any subtask requiring mandatory disclosure under the ICLR policy. The core implementation of the SEPO algorithm, the token-level JS-divergence scoring, and the evaluation metrics were written by the authors; coding agents were used only to check this implementation. We also did not use generative AI tools to propose hypotheses, design the research methodology or experiments, generate synthetic datasets, clean or reformat datasets, or interpret results.

Additionally, we used commercial generative AI tools for the following recommended-disclosure tasks: (1) writing scripts to run and orchestrate experiment trials, and launching those trials automatically; (2) creating scripts to plot result figures; and (3) revising wording and correcting grammar to improve readability. Separately, as part of our evaluation protocol (Section E), we used `gemini-3-flash-preview`, `gpt-5.4`, `gpt-5.4-mini`, `gpt-5.4-nano`, and `gpt-oss-20b` via commercial API calls to compute LLM-as-a-judge metrics.

We have reviewed all AI-assisted work, including the implementation checks, experiment and plotting scripts, automatically launched experiment trials, LLM-judge evaluation outputs, and grammar and writing-style revisions. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

B ADDITIONAL MATHEMATICAL ANALYSIS OF THE INSTRUCTION ENSEMBLE

The instruction ensemble mixture formulation makes the source of diversity explicit. Substituting equation 4 into equation 2 gives, for a fixed x,

$$\mathcal{J}_{\mathrm{Div}}(x;\mathcal{I}) = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}_{\substack{y\sim\mathcal{M}_{I_i}(\cdot|x)\\ y'\sim\mathcal{M}_{I_j}(\cdot|x)}}\left[d(y,y')\right]. \tag{6}$$

The diagonal terms measure variation within an instruction-conditioned component, whereas the off-diagonal terms measure behavioral separation between components. Consequently, adding more instructions does not by itself broaden the response distribution: the gain depends on whether the new instructions induce behaviors that are separated from those already represented in the ensemble.

C SEPO IMPLEMENTATION DETAILS

This section specifies the rollout evaluation and candidate evolution protocol used by the search in Equation 5.

Token-level JS divergence with $k2$ estimation. We use a token-level JS surrogate with the $k2$ KL estimator (Schulman, 2020). For each instruction, we sample $R = 2$ responses per prompt on a fixed set $\mathcal{X}$ of $m = 5$ training prompts, using temperature 1.0, top-$p = 1.0$, and at most 32 new tokens. For a response y sampled under instruction $A \in \{I, J\}$, let $p_{A,t} = \mathcal{M}_A(y_t \mid x, y_{<t})$ denote the probability of its t-th token. Both instructions are evaluated on the same response history. Writing $m_t = (p_{I,t} + p_{J,t})/2$, we approximate the KL contribution to the local JS objective using $k2$:

$$f_t^{(A)}(y;x) = \frac{1}{2}\left(\ln p_{A,t} - \ln m_t\right)^2. \tag{7}$$

The instruction-pair score averages over tokens within each response, responses within each direction, both sampling directions, and prompts:

$$\widehat{D}(I,J) = \frac{1}{2mR}\sum_{x\in\mathcal{X}}\sum_{A\in\{I,J\}}\sum_{r=1}^{R}\frac{1}{|y_{A,x}^{(r)}|}\sum_{t=1}^{|y_{A,x}^{(r)}|} f_t^{(A)}(y_{A,x}^{(r)};x). \tag{8}$$

Here $y_{A,x}^{(r)}$ is the r-th response sampled under A for x; empty responses are omitted and the response average uses the remaining count. We use natural logarithms without $\ln 2$ normalization. This is a $k2$ approximation to the token-level objective. The candidate score is its minimum pair score against $\mathcal{A}_k$, as in Equation 5.

Quality filtering against reward hacking. Behavioral separation does not establish response quality: an instruction can elicit irrelevant or incoherent outputs and still score highly on divergence. We therefore apply the quality criterion from Equation equation 1 as a separate feasibility constraint. Using the same utility function μ_x, we estimate each candidate's expected quality from its rollouts,

$$\widehat{Q}(I) = \frac{1}{mR} \sum_{x \in \mathcal{X}} \sum_{r=1}^{R} \mu_x \left(y_{I,x}^{(r)} \right), \tag{9}$$

and retain a candidate only if $\widehat{Q}(I) \geq \tau$. This filter restricts the search to instructions that meet the estimated quality threshold, so that a higher divergence score cannot compensate for failing that constraint.

Candidate evolution and pool management. To maintain broad exploration, SEPO retains a persistent pool containing every quality-feasible candidate not yet accepted into $\mathcal{S}_k$. At evolution step t, we partition this pool into H clusters in instruction space, $\{\mathcal{G}_h^{(t)}\}_{h=1}^{H}$. We embed each candidate instruction with `BAAI/bge-m3` and cluster the embeddings by average-linkage agglomerative clustering on cosine distance, with $H = 4$ clusters, one per seed candidate. For Creative Math, we instead select the four highest-scoring candidates as parents without clustering. For each cluster, we select as its evolution parent the candidate with the largest behavioral diversity score relative to the current instruction set:

$$I_h^{(t)} = \arg \max_{I \in \mathcal{G}_h^{(t)}} \min_{J \in \mathcal{A}_k} \widehat{D}(I, J), \qquad h = 1, \ldots, H. \tag{10}$$

The proposer LLM evolves each cluster parent into new candidate instructions, conditioned on the parent's observed rollout behaviors. The new candidates are evaluated using the same rollout procedure, and those passing the quality filter are added to the persistent pool. Candidates not selected as parents remain available in subsequent steps. Selecting one high-scoring parent per cluster balances behavioral fitness with coverage of distinct regions of instruction space.

After S evolution steps, the accumulated pool constitutes $\mathcal{C}_k$ in equation 5, and its highest-scoring candidate is accepted to form $\mathcal{S}_{k+1}$. All remaining candidates are carried forward to seed the next search round. This cluster-guided *evolve–evaluate–admit* cycle continues until the instruction ensemble reaches size K or the search budget is exhausted. The resulting set $\mathcal{I} = \mathcal{S}_T$ defines the mixture in Equation equation 4, where $T = K$ when the target size is reached; the reference instruction I_0 is not included in $\mathcal{I}$.

Computational cost. The search is lightweight in practice. Each candidate is evaluated on only $m = 5$ training prompts with $R = 2$ responses each, and the rollouts used for divergence estimation are capped at 32 new tokens, so each divergence estimate requires only a few short generations and forward passes to score the same responses under both instructions. With `gpt-oss-20b` as the proposer LLM on two NVIDIA RTX 4090 GPUs, the complete optimization finishes in under 30 minutes, even with 10 search rounds. Since it requires only 5 training prompts, the search is practical to rerun for a new task or reference instruction without large-scale data or compute.

D SEPO SEARCH TRACE

We trace one Qwen3-1.7B SEPO search per dataset using the Section 5 settings: 10 outer rounds, $S = 4$ evolution steps per round, 4 seed candidates per step, and up to 4 proposals per seed candidate. Round k admits I_k, so its candidates are scored against the reference-augmented set $\mathcal{A}_{k-1} = \{I_0\} \cup \mathcal{S}_{k-1}$ by $s_{k-1}(I) = \min_{J \in \mathcal{A}_{k-1}} \widehat{D}(I, J)$ (Equation 5). The figures plot this score. The round-start pool contains candidates carried from the previous round; adding the round's new quality-feasible candidates gives $\mathcal{C}_{k-1}$, from which I_k is selected.

Within a round, every score is measured against the same $\mathcal{A}_{k-1}$, so scores can be compared directly. After I_k is added to the reference set, a carried candidate's score can only stay unchanged or decrease. Scores are therefore not comparable across rounds, and they are also not comparable across datasets, which use different training prompts.

Reading the traces. Figures 7–10 show one search each. Each column between vertical dividers is one outer round. The crosses are the round's new quality-feasible candidates, placed left to right in the order in which they were proposed over steps $t = 1, \ldots, 4$. The solid step line is the running best of these new candidates only: it starts at zero in every round and rises whenever a new candidate beats every earlier new candidate of the same round. The dashed line is the best score in the round-start pool, that is, the best candidate carried from earlier rounds and rescored against $\mathcal{A}_{k-1}$. Round 1 has no dashed line because its round-start pool is empty: evolution starts from the default instruction I_0, which serves only as a reference and is never a candidate.

Because I_k has the highest score in $\mathcal{C}_{k-1}$, it corresponds to the higher of the two lines at the right edge of round k. If the solid line ends above the dashed line, the round's evolution produced a candidate that beats every carried candidate, and I_k is newly evolved; otherwise, I_k is the best carried candidate. In round 1, I_1 is therefore always newly evolved. The best carried score at the start of round $k+1$ is at most the round-k score of I_k, so the dashed line of round $k+1$ never lies above the admitted score of round k.

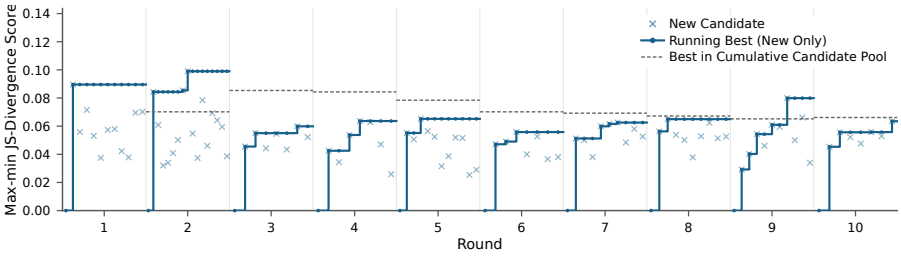

Figure 7: SEPO search trace on Artificial Hivemind. Each column is one outer round k, and every score is $s(I)$ ($= s_{k-1}(I)$ in round k). Crosses are new quality-feasible candidates in proposal order; the solid line is the running best of the round's new candidates, starting from zero; the dashed line is the best score in the round-start pool (absent in round 1). The admitted prompt I_k corresponds to the higher of the two lines at the end of round k.

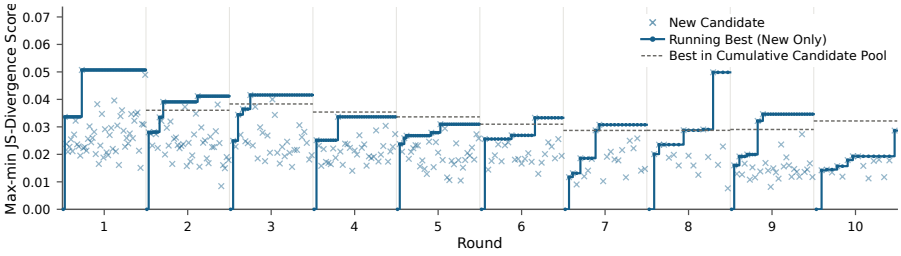

Figure 8: SEPO search trace on Creativity Index, plotted as in Figure 7.

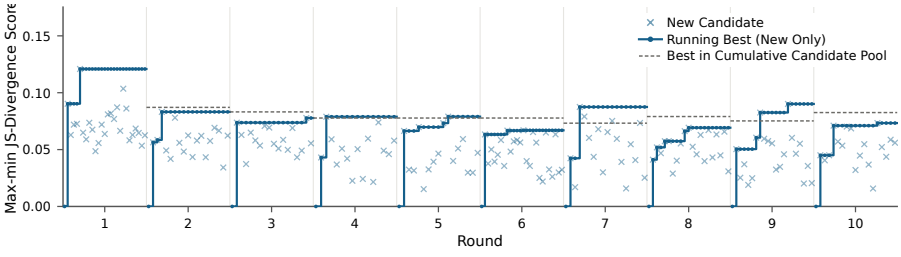

Figure 9: SEPO search trace on Novelty Bench, plotted as in Figure 7.

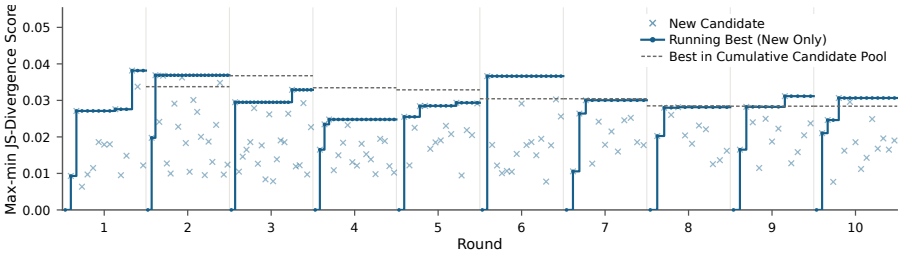

Figure 10: SEPO search trace on Creative Math, plotted as in Figure 7.

E Cross-Model Quality–Creativity Trade-offs

Figures 11 and 12 report embedding-based metrics, while Figures 13 and 14 report n-gram-based metrics for $n \in \{5, \ldots, 12\}$. Together, they expand the trade-off analysis to three models and four datasets. Rows correspond to Qwen3-1.7B, Meta-Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3; columns correspond to Artificial Hivemind, Creativity Index, Novelty Bench, and Creative Math. Each point aggregates one sampling method at one decoding temperature, and connected points trace increasing temperature. The upper-right region represents a more favorable joint outcome: higher mean LLM-judged quality together with higher diversity or novelty under the corresponding metric.

The embedding-based panels show that SEPO offers higher diversity and novelty than base sampling across the displayed settings, while its comparison with Multi-Turn and verbalized sampling depends on the model, dataset, and metric. The n-gram-based panels provide a complementary lexical view: SEPO often reaches high diversity at higher quality than verbalized sampling, but its novelty scores can overlap with those of Multi-Turn or verbalized sampling. These patterns do not establish uniform dominance across metrics; we report the temperature sweeps to make the trade-offs visible.

In the embedding-based novelty and both n-gram-based figures, the base sampling reference point at $T = 0.7$ is omitted, following the plotting protocol. The remaining points are connected in temperature order. Axis limits and contour normalization still use all observations in the selected temperature range, including the omitted point. For visual guidance, each background shows the equal-weight harmonic mean of panel-wise normalized quality and the corresponding creativity metric. Quality is anchored at the minimum Likert score of 1 and the observed panel maximum; diversity or novelty is anchored at 0 and its observed panel maximum. Thus, greener contours indicate a better balance within a panel, but their values are not intended for absolute comparison across panels.

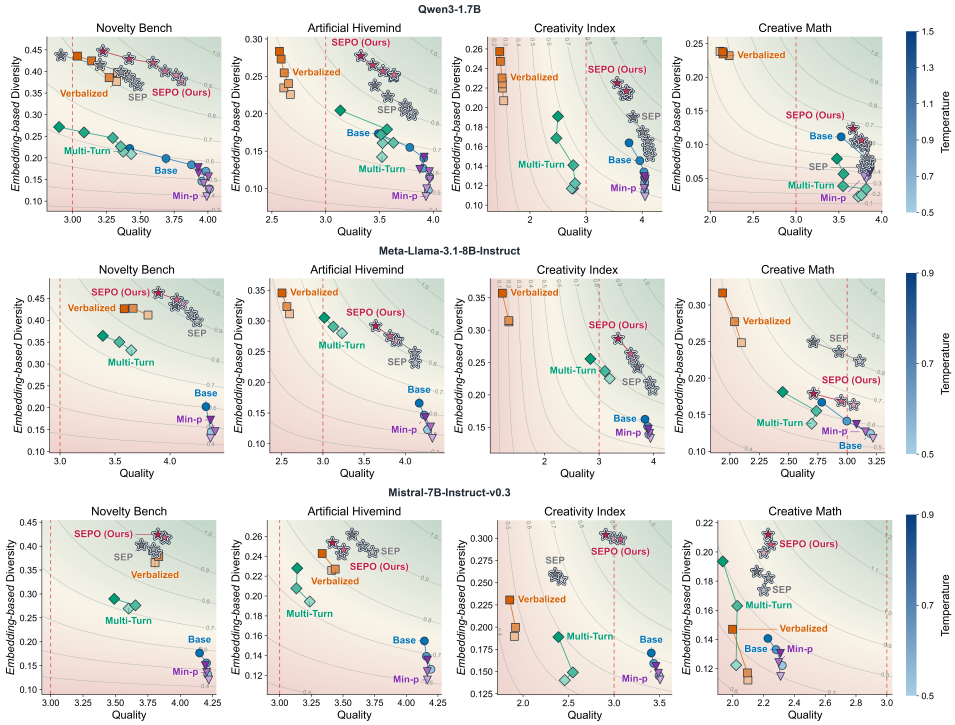

Figure 11: Quality–embedding-diversity trade-offs across models, datasets, sampling methods, and decoding temperatures. Circles, squares, diamonds, downward triangles, and stars denote base, verbalized, Multi-Turn, min-p, SEP, and SEPO sampling. Darker points indicate higher temperatures, with lightness normalized within each panel. Qwen3-1.7B uses $T \in \{0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$, whereas Llama and Mistral use $T \in \{0.5, 0.7, 0.9\}$. The background contours show the panel-wise normalized harmonic mean of quality and embedding-based diversity.

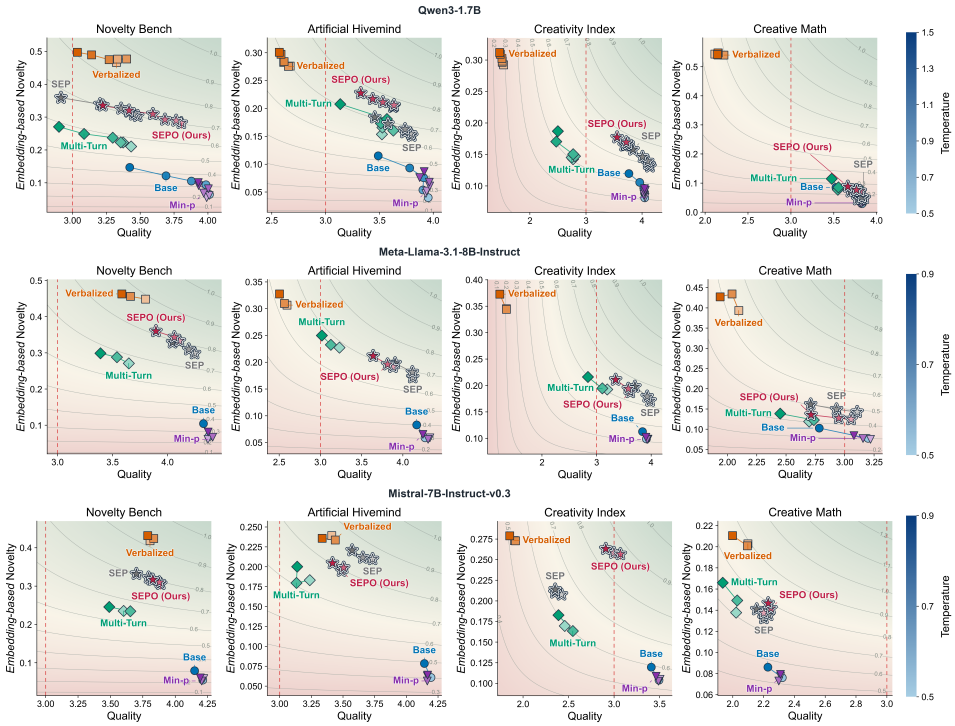

Figure 12: Quality–embedding-novelty trade-offs across models, datasets, sampling methods, and decoding temperatures.

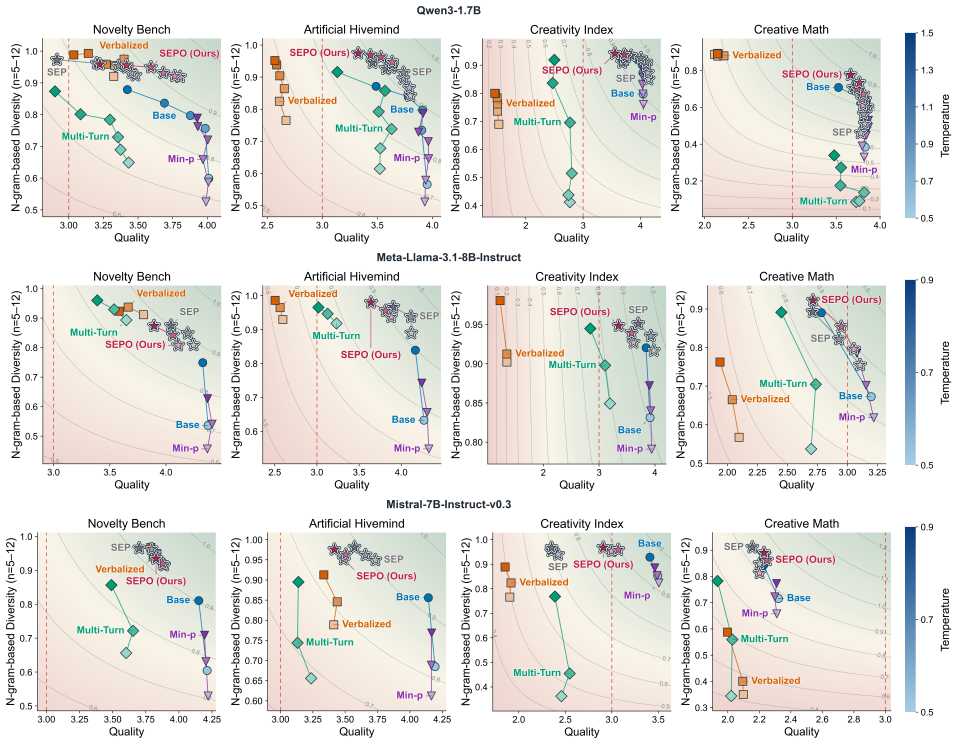

Figure 13: Quality–n-gram-diversity trade-offs across three models and four datasets, using $n \in \{5, \ldots, 12\}$. The horizontal axis is mean LLM-judged quality and the vertical axis is n-gram-based diversity. Model and dataset ordering, marker shapes, and temperature sweeps follow Figure 11.

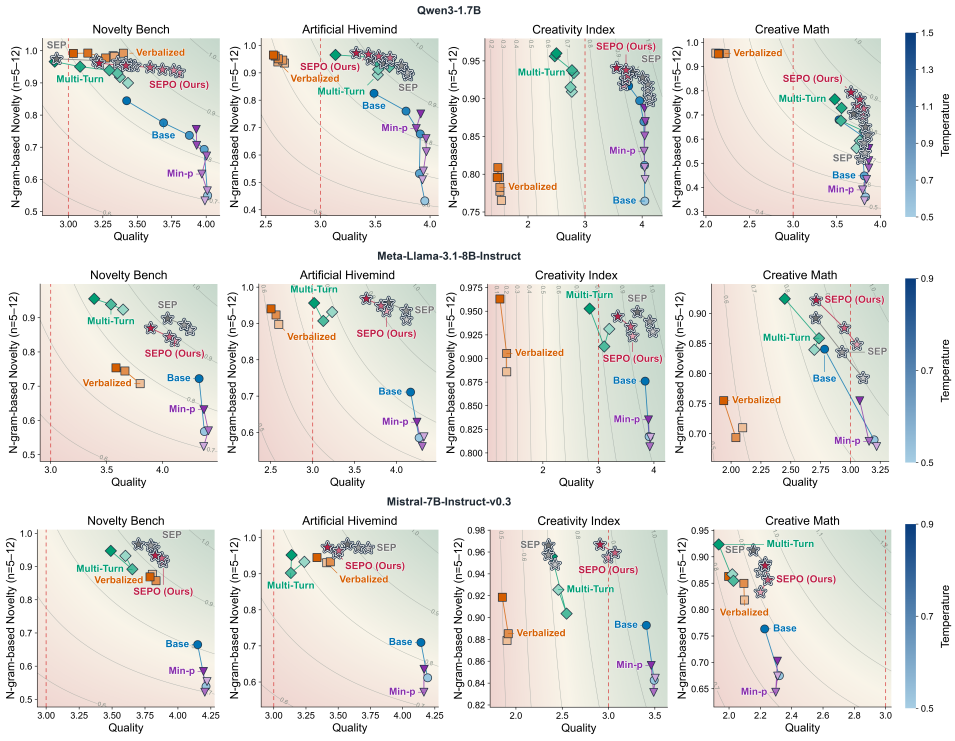

Figure 14: Quality–n-gram-novelty trade-offs across three models and four datasets, using $n \in \{5, \ldots, 12\}$. The horizontal axis is mean LLM-judged quality and the vertical axis is n-gram-based novelty. Model and dataset ordering, marker shapes, and temperature sweeps follow Figure 11.

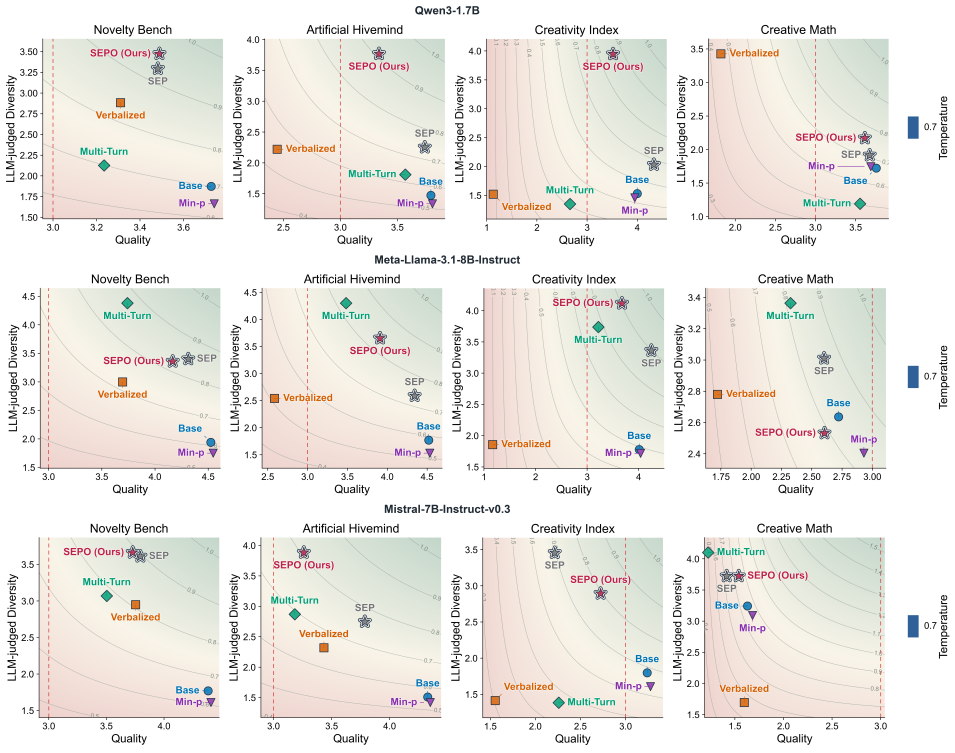

Figure 15: Quality–diversity trade-offs across three models and four datasets, with diversity evaluated by an LLM-as-a-judge. The horizontal axis is mean LLM-judged quality and the vertical axis is LLM-judged diversity.

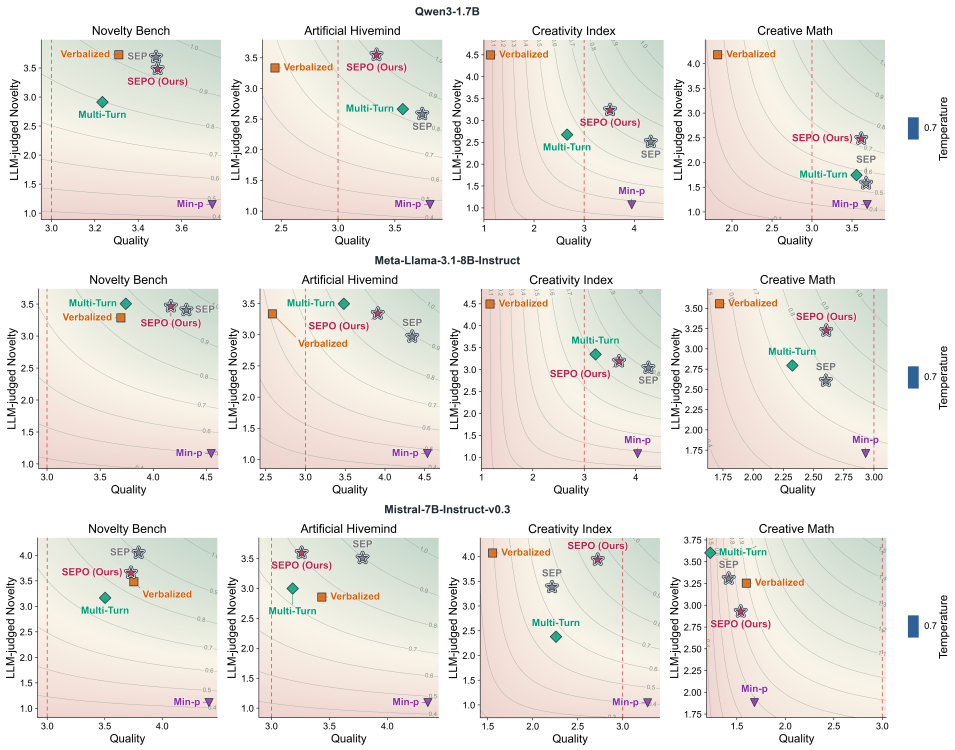

Figure 16: Quality–novelty trade-offs across three models and four datasets, with novelty evaluated by an LLM-as-a-judge. The horizontal axis is mean LLM-judged quality and the vertical axis is LLM-judged novelty.

Embedding-based results. We plot mean LLM-judged quality against embedding-based diversity (Figure 11) and novelty (Figure 12). SEPO has higher embedding diversity and novelty than base and min-p sampling in every panel, and its points change little with temperature. Raising the temperature of base or min-p sampling adds little diversity or novelty. Verbalized sampling has the highest novelty in every panel and the highest diversity in several, usually at the lowest quality. For diversity, SEPO ranks first on all four datasets for Mistral-7B and on Novelty Bench for Qwen3-1.7B and Llama-3.1-8B. On the other three datasets for those two models, verbalized sampling is higher but at much lower quality. For novelty, SEPO usually matches or exceeds Multi-Turn sampling at higher quality. Creative Math is the weakest case: for Qwen3-1.7B, the diversity of SEPO is only slightly above base sampling.

N-gram-based results. We plot mean LLM-judged quality against n-gram-based diversity (Figure 13) and novelty (Figure 14). SEPO reaches an n-gram diversity of about 0.9 or higher in every panel except Creative Math for Qwen3-1.7B, and its points change little with temperature. It ranks first or ties for first on all four datasets for Mistral-7B and on most datasets for the other two models. Verbalized sampling is slightly higher on Novelty Bench and Creative Math for Qwen3-1.7B and on Creativity Index and Creative Math for Llama-3.1-8B, but at much lower quality. Base and min-p sampling lose lexical diversity quickly as temperature decreases, and Multi-Turn sampling varies widely with temperature. N-gram novelty is less discriminative: verbalized, Multi-Turn, and SEPO sampling all score close to 1 in most panels, and SEPO usually does so at the highest quality of the three. Base and min-p sampling score lowest, since the novelty reference is drawn from base sampling. Creative Math is again the weakest case: for Qwen3-1.7B, the n-gram novelty of SEPO is well below that of verbalized sampling and close to that of Multi-Turn sampling.

LLM-judged results. We plot mean LLM-judged quality against LLM-judged diversity (Figure 15) and novelty (Figure 16) at $T = 0.7$, the only temperature scored by the judge. Because judge-based evaluation is costly, we score only $T = 0.7$, the temperature used for the main results;; Table 3 lists the raw scores. SEPO has higher judged diversity and novelty than base and min-p sampling in every panel, except diversity on Creative Math for Llama-3.1-8B. Base sampling scores 1.000–1.006 in novelty, since its responses form the novelty reference. For diversity, SEPO ranks first on 6 of the 12 model–dataset combinations (Novelty Bench, Artificial Hivemind, and Creativity Index for Qwen3-1.7B; Novelty Bench and Artificial Hivemind for Mistral-7B; Creativity Index for Llama-3.1-8B), and Multi-Turn sampling ranks first on 4, including three of the four datasets for Llama-3.1-8B. For novelty, verbalized sampling ranks first on 6 combinations, but its quality falls below 3 on 8 of the 12, down to 1.131. SEPO ranks first on 2 combinations (Artificial Hivemind for Qwen3-1.7B and Mistral-7B) and in the top two on 9. SEPO's quality exceeds that of Multi-Turn sampling on 11 of the 12 combinations and that of verbalized sampling on 9, and lies 0.14–1.09 points below the better of base and min-p sampling. Creative Math is again the weakest case: SEPO ranks second in diversity for Qwen3-1.7B, ties with SEP for third for Mistral-7B, and ranks fifth for Llama-3.1-8B. Judged diversity is also the least reliable of the three dimensions (Appendix H), so these diversity rankings should be read with more caution than the novelty rankings.

Table 3: Raw Quality, Diversity, and Novelty scores at $T = 0.7$ (judge: gemini-3-flash-preview).

Method	Quality			Diversity			Novelty		
	Qwen3 1.7B	Mistral 7B	Llama-3.1 8B	Qwen3 1.7B	Mistral 7B	Llama-3.1 8B	Qwen3 1.7B	Mistral 7B	Llama-3.1 8B
Novelty Bench									
Base	3.727	4.382	4.524	1.874	1.770	1.940	1.000	1.000	1.000
Min-p	**3.741**	**4.406**	**4.546**	1.663	1.610	1.750	1.149	1.114	1.163
Verbalized	3.311	3.753	3.695	2.884	2.950	3.000	**3.728**	3.483	3.285
Multi-Turn	3.235	3.503	3.741	2.126	3.070	**4.380**	2.914	3.167	**3.505**
SEP	3.482	3.795	4.313	3.295	3.611	3.400	3.691	**4.051**	3.407
SEPO (Ours)	3.489	3.728	4.166	**3.474**	**3.663**	3.358	3.484	3.660	3.466
Artificial Hivemind									
Base	3.795	4.327	4.519	1.474	1.510	1.770	1.000	1.003	1.003
Min-p	**3.808**	**4.349**	**4.533**	1.337	1.420	1.530	1.107	1.093	1.095
Verbalized	2.445	3.434	2.585	2.221	2.320	2.541	3.336	2.858	3.336
Multi-Turn	3.569	3.183	3.483	1.811	2.870	**4.300**	2.664	3.000	**3.497**
SEP	3.741	3.786	4.343	2.253	2.747	2.589	2.583	3.509	2.971
SEPO (Ours)	3.339	3.261	3.909	**3.768**	**3.884**	3.653	**3.549**	**3.597**	3.341
Creativity Index									
Base	3.997	3.239	4.009	1.532	1.798	1.778	1.000	1.000	1.000
Min-p	3.949	**3.277**	4.034	1.457	1.606	1.717	1.072	1.044	1.080
Verbalized	1.131	1.555	1.168	1.521	1.414	1.857	**4.497**	**4.072**	**4.497**
Multi-Turn	2.659	2.259	3.217	1.351	1.384	3.737	2.676	2.381	3.349
SEP	**4.329**	2.215	**4.245**	2.032	**3.457**	3.362	2.510	3.384	3.040
SEPO (Ours)	3.516	2.723	3.676	**3.947**	2.894	**4.117**	3.239	3.937	3.189
Creative Math									
Base	**3.757**	1.629	2.721	1.723	3.242	2.636	1.006	1.000	1.006
Min-p	3.690	**1.681**	**2.930**	1.745	3.091	2.404	1.149	1.883	1.708
Verbalized	1.820	1.597	1.722	**3.427**	1.695	2.780	**4.181**	3.255	**3.556**
Multi-Turn	3.555	1.225	2.324	1.191	**4.101**	**3.364**	1.746	**3.603**	2.795
SEP	3.677	1.416	2.602	1.915	3.723	3.011	1.578	3.307	2.606
SEPO (Ours)	3.616	1.540	2.605	2.170	3.723	2.532	2.483	2.929	3.226

Scores are unnormalized means. Darker teal: highest score; lighter teal: second-highest distinct score within each model–dataset–metric group, ranked before rounding (ties share the same highlight). Values are rounded to three decimals. **Bold**: highest unrounded score within each model–dataset group for each metric (ties retained).

Table 4: Arithmetic mean: unnormalized and normalized at $T = 0.7$ (judge: gemini-3-flash-preview).

Method	Qwen3-1.7B				Mistral-7B				Llama-3.1-8B			
	NB	AH	CI	CM	NB	AH	CI	CM	NB	AH	CI	CM
Unnormalized												
Base	2.200	2.089	2.176	2.162	2.384	2.280	2.012	1.957	2.488	2.431	2.262	2.121
Min-p	2.185	2.084	2.160	2.195	2.377	2.287	1.976	2.218	2.486	2.386	2.277	2.347
Verbalized	3.308	2.667	2.383	**3.143**	3.395	2.871	2.347	2.182	3.327	2.820	2.507	2.686
Multi-Turn	2.758	2.681	2.228	2.164	3.247	3.018	2.008	**2.976**	**3.875**	**3.760**	3.435	**2.828**
SEP	**3.489**	2.859	2.957	2.390	**3.819**	3.348	3.019	2.816	3.707	3.301	3.549	2.740
SEPO (Ours)	3.482	**3.552**	**3.567**	2.756	3.684	**3.581**	**3.185**	2.731	3.664	3.634	**3.661**	2.788
Normalized												
Base	0.601	0.556	0.504	0.581	0.574	0.554	0.583	0.664	0.574	0.565	0.527	0.653
Min-p	0.596	0.556	0.498	0.587	0.572	0.556	0.574	0.759	0.577	0.556	0.531	0.732
Verbalized	0.891	0.681	0.475	**0.766**	0.824	0.706	0.551	0.731	0.794	0.664	0.501	0.734
Multi-Turn	0.736	0.715	0.479	0.564	0.785	0.742	0.513	0.777	**0.924**	**0.901**	0.779	0.824
SEP	**0.948**	0.767	0.691	0.636	**0.935**	0.838	0.788	0.812	0.894	0.799	0.831	0.819
SEPO (Ours)	0.948	**0.944**	**0.825**	0.725	0.902	**0.892**	**0.854**	**0.838**	0.883	0.876	**0.845**	**0.830**

The arithmetic mean is computed from the three raw scores (Unnormalized) or the three normalized scores (Normalized). Darker teal: highest score; lighter teal: second-highest distinct score within each model–dataset group and each version, ranked before rounding (ties share the same highlight). Normalization: $x' = (x-1)/(x_{\max}-1)$ for Quality and $x' = x/x_{\max}$ for Diversity and Novelty. Maxima are taken across the five methods within each model–dataset group at $T = 0.7$. AH: Artificial Hivemind; CM: Creative Math; CI: Creativity Index; NB: Novelty Bench. Values are rounded to three decimals. **Bold**: highest unrounded score within each model–dataset group and each version (ties retained).

Table 5: Harmonic mean: unnormalized and normalized at $T = 0.7$ (judge: gemini-3-flash-preview).

Method	Qwen3-1.7B				Mistral-7B				Llama-3.1-8B			
	NB	AH	CI	CM	NB	AH	CI	CM	NB	AH	CI	CM
Unnormalized												
Base	1.665	1.545	1.576	1.630	1.673	1.587	1.609	1.561	1.728	1.682	1.656	1.723
Min-p	1.726	1.568	1.603	1.750	1.718	1.622	1.591	2.070	1.817	1.678	1.708	2.234
Verbalized	3.272	2.589	1.701	**2.777**	3.361	2.798	1.880	1.969	3.303	2.777	1.856	2.456
Multi-Turn	2.672	2.484	2.013	1.772	3.236	3.012	1.892	2.243	**3.842**	**3.724**	3.421	**2.764**
SEP	3.482	2.731	2.675	2.101	**3.810**	3.286	2.895	2.349	3.661	3.148	3.480	2.727
SEPO (Ours)	**3.482**	**3.544**	**3.544**	2.631	3.684	**3.562**	**3.103**	**2.382**	3.631	3.619	**3.621**	2.755
Normalized												
Base	0.455	0.422	0.367	0.420	0.421	0.419	0.428	0.504	0.443	0.434	0.380	0.506
Min-p	0.474	0.427	0.374	0.453	0.434	0.427	0.425	0.707	0.460	0.428	0.393	0.669
Verbalized	0.885	0.638	0.103	0.559	0.824	0.696	0.397	0.643	0.781	0.604	0.133	0.614
Multi-Turn	0.725	0.666	0.454	0.472	0.783	0.734	0.499	0.597	**0.911**	**0.876**	0.768	0.804
SEP	**0.947**	0.737	0.634	0.548	**0.928**	0.824	0.736	0.783	0.886	0.770	0.810	0.814
SEPO (Ours)	0.946	**0.937**	**0.808**	**0.695**	0.894	**0.862**	**0.845**	**0.835**	0.873	0.873	**0.828**	**0.826**

The harmonic mean is computed from the three raw scores (Unnormalized) or the three normalized scores (Normalized). For inputs x_1, x_2, x_3, the harmonic mean is $3/(1/x_1 + 1/x_2 + 1/x_3)$; it is zero if any input is zero. Darker teal: highest score; lighter teal: second-highest distinct score within each model–dataset group and each version, ranked before rounding (ties share the same highlight). Normalization: $x' = (x-1)/(x_{\max}-1)$ for Quality and $x' = x/x_{\max}$ for Diversity and Novelty. Maxima are taken across the five methods within each model–dataset group at $T = 0.7$. AH: Artificial Hivemind; CM: Creative Math; CI: Creativity Index; NB: Novelty Bench. Values are rounded to three decimals. **Bold**: highest unrounded score within each model–dataset group and each version (ties retained).

F PCA VISUALIZATION OF INSTRUCTION-CONDITIONED RESPONSES

Figure 17 visualizes response embeddings for 20 randomly selected Artificial Hivemind prompts. Several panels show instruction-conditioned responses occupying regions separated from the base responses, while others exhibit substantial overlap. These projections illustrate prompt-dependent variation in response distributions; they do not by themselves establish response quality or separation in the full embedding space.

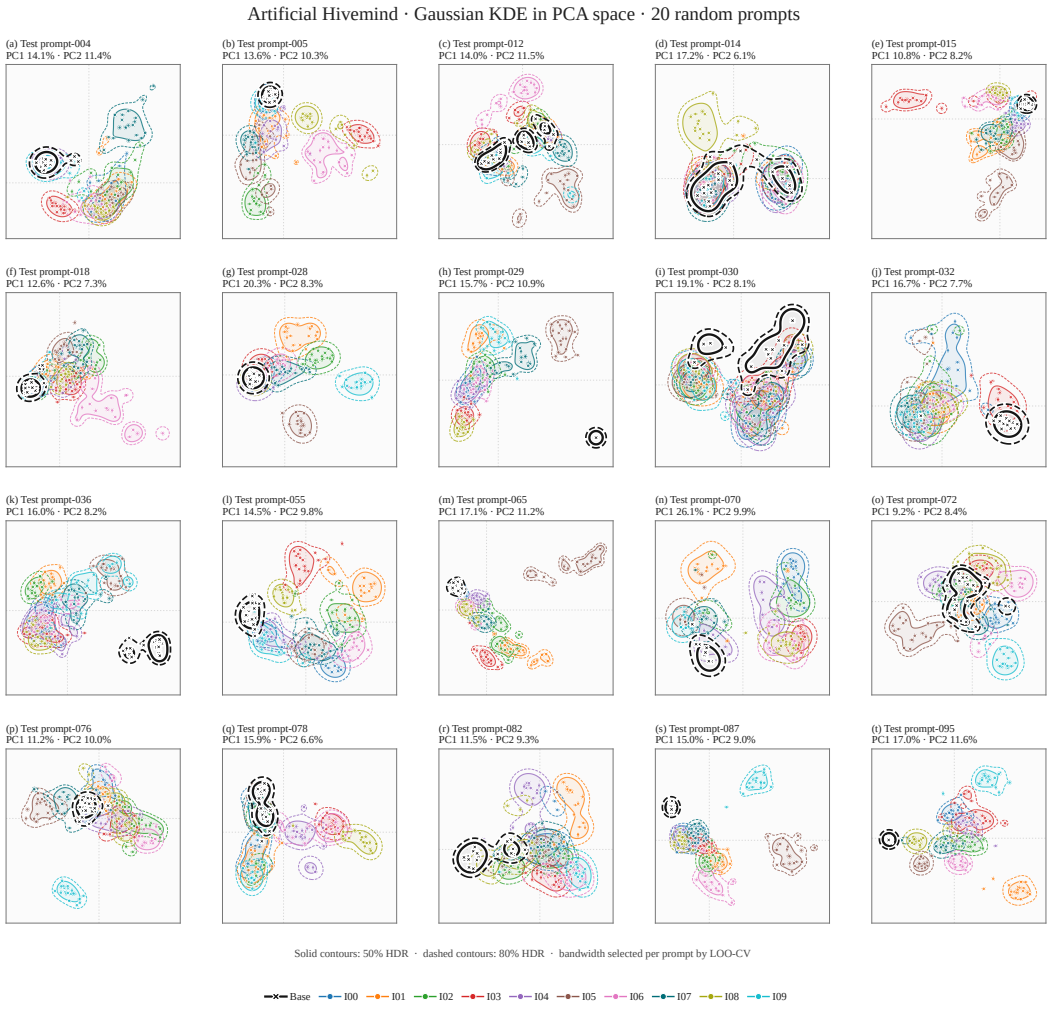

Figure 17: **Instruction-conditioned response distributions on Artificial Hivemind.** Each panel fits a joint two-dimensional PCA to 20 base responses and 20 responses from each of 10 instructions (220 responses per prompt). Black crosses denote base responses; colored points denote instructions I00–I09. Gaussian kernel density estimates are fitted separately for each condition using a shared bandwidth selected per prompt by leave-one-out cross-validation. Solid and dashed contours indicate 50% and 80% highest-density regions, respectively. Panel headings report the variance explained by the first two principal components; PCA is fitted independently for each prompt.

G Alignment of the JS Divergence Surrogate Objective with Diversity and Novelty Metrics

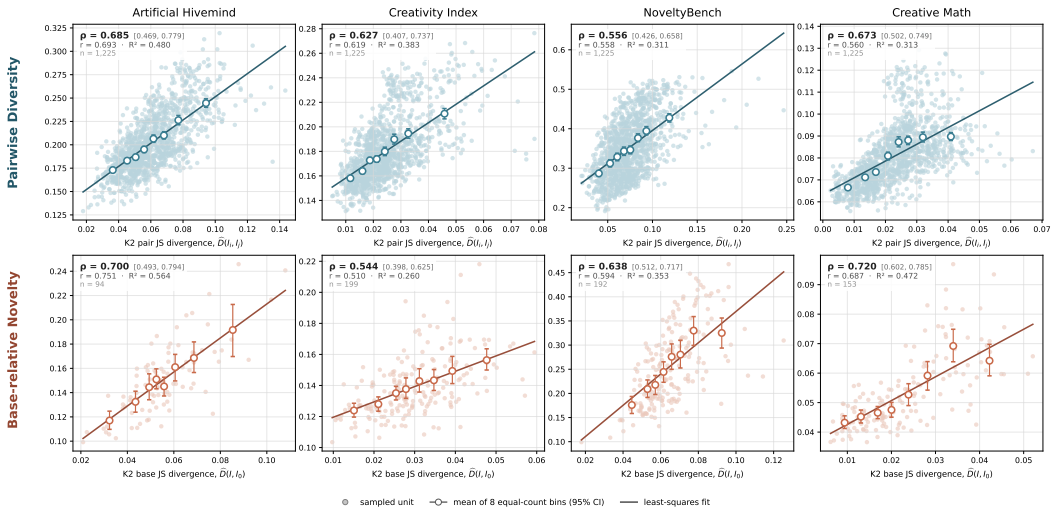

Figure 18: **Alignment of the $k2$ surrogate with held-out behavior on Qwen3-1.7B.** Top: $\widehat{D}(I_i, I_j)$ versus mean same-prompt embedding-based distance for all 1,225 pairs formed from 50 instructions. Bottom: $\widehat{D}(I, I_0)$ versus embedding-based distance to the nearest base response for the same prompt ($n = 94/199/192/153$). Faint points denote individual analysis units; outlined points and whiskers show eight equal-count bin means and 95% bootstrap intervals; solid lines are least-squares fits. Each panel reports Spearman ρ with its two-way-bootstrap interval, Pearson r, R^2, and the number of analysis units.

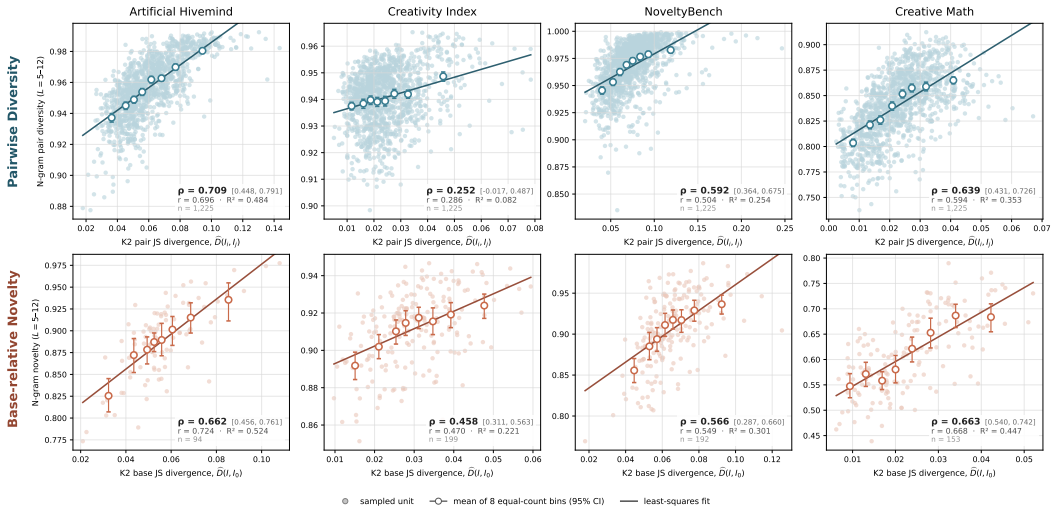

Figure 19: **Alignment of the $k2$ surrogate with held-out n-gram-based behavior on Qwen3-1.7B.** Top: $\widehat{D}(I_i, I_j)$ versus mean same-prompt n-gram-based diversity for the 1,225 pairs of Figure 18. Bottom: $\widehat{D}(I, I_0)$ versus n-gram-based novelty relative to the base responses for the same prompt ($n = 94/199/192/153$). Markers, lines, and annotations follow Figure 18.

Protocol. Appendix B implies that mixture diversity depends on separation among instruction-conditioned components. We test whether SEPO's search-time $k2$ JS surrogate predicts this separation on held-out prompts, measured with both the embedding-based and the n-gram-based metrics of Appendix J. For diversity on Qwen3-1.7B, we retain 50 instructions from the final checkpoint (the 10 selected instructions plus 40 other archived candidates), score all 1,225 unordered pairs, and compare

$\widehat{D}(I_i, I_j)$ with their embedding-based and n-gram-based metrics across 50 unseen prompts. Novelty is base-relative under equation 3; therefore, we compare $\widehat{D}(I, I_0)$, rather than a pair score, with each instruction's measured novelty. We use up to 200 valid archived instructions (94/199/192/153 for Artificial Hivemind, Creativity Index, NoveltyBench, and Creative Math). $k2$ follows the $m = 5$, $R = 2$ rollout protocol in Appendix C.

Results. The surrogate is consistently positively associated with both embedding-based quantities. Pairwise response separation yields Spearman $\rho \in [0.556, 0.685]$, while base-relative novelty yields $\rho \in [0.544, 0.720]$. Every 95% two-way-bootstrap interval over instructions and evaluation prompts excludes zero. These results support the $k2$ surrogate as a useful ranking signal for behavioral separation. Moderate R^2 values indicate that a linear mapping explains only part of the downstream variation; we therefore treat $k2$ as a ranking surrogate rather than a point estimator of held-out distance.

N-gram-based metrics. We repeat this analysis with the n-gram-based metrics of Appendix J in place of the embedding-based ones, keeping the instructions, their $k2$ scores, and the held-out responses unchanged. Pairwise response separation becomes the mean same-prompt L-uniqueness between the two instructions' responses (Equation 16, $L = 5, \ldots, 12$, averaged over both directions), and each instruction's novelty becomes its mean L-uniqueness against the base responses for the same prompt (Equation 18). The surrogate remains positively associated with both quantities (Figure 19): pairwise response separation yields Spearman $\rho \in [0.252, 0.709]$, while base-relative novelty yields $\rho \in [0.458, 0.663]$. Seven of the eight 95% two-way-bootstrap intervals exclude zero; the exception is pairwise separation on Creativity Index ($\rho = 0.252$, $[-0.017, 0.487]$), whose n-gram pair scores span only 0.898–0.965, the narrowest range of the four datasets. Compared with the embedding-based results, ρ is lower for novelty on all four datasets and similar for diversity except on Creativity Index. The $k2$ surrogate is therefore also a useful ranking signal for lexical separation, although a weaker one than for semantic separation.

H Cross-Validation of LLM-as-Judge Evaluation

We evaluated the reliability of our LLM-as-judge protocol on creativity metrics by re-scoring the Artificial Hivemind outputs of Qwen3-1.7B with SEPO at temperature 0.7 (95 prompts, ten responses each) with five judge models: `gemini-3-flash-preview` (the primary judge in the main experiments), `gpt-5.4`, `gpt-5.4-mini`, `gpt-5.4-nano`, and `gpt-oss-20b`.[3] Each judge returned a 1–5 Likert score under the same rubric for quality and novelty of each response ($N = 950$) and for diversity of each ten-response set ($N = 95$).

Validation protocol. For every pair of judges, we report the mean score difference, mean absolute difference (MAE), exact and within-one-point agreement, Spearman correlation, and quadratic weighted Cohen's kappa (κ_w); for the full panel, we report ordinal Krippendorff's α. Confidence intervals come from 10,000 bootstrap replicates that resample prompts as clusters. Because Likert scores contain many ties, we also ran a within-prompt pairwise preference check: for each pair of responses (y_1, y_2) to the same prompt, we checked whether two judges induce the same preference, a tie, or the opposite preference.

Agreement results. Agreement is moderate to strong for quality ($\alpha = 0.586$) and novelty ($\alpha = 0.654$), and weaker for diversity ($\alpha = 0.392$; Table 6). Novelty shows the strongest pairwise agreement (Table 7): `gemini-3-flash-preview`, `gpt-5.4`, and `gpt-5.4-mini` agree closely ($\kappa_w = 0.809$–0.823, 96.5–96.9% within one point), while pairs involving `gpt-5.4-nano` are the weakest ($\kappa_w = 0.502$–0.619). For quality, κ_w ranges from 0.513 to 0.733, and `gemini-3-flash-preview` agrees best with the two `gpt-5.4` models ($\kappa_w = 0.664$–0.666). *Diversity is the least consistent dimension*: among the four GPT-family judges, κ_w ranges from 0.454 to 0.608, but `gemini-3-flash-preview` agrees only weakly with every other judge ($\kappa_w = 0.244$–0.298) and scores diversity 0.62 to 1.16 points higher on average. Although judges differ in absolute scores and pairwise agreement, they rarely disagree in direction: at the response level, directly opposite preferences are uncommon (2.9–8.4% of quality and 1.5–6.1% of novelty com-

[3] `gpt-5.4-2026-03-05`, `gpt-5.4-mini-2026-03-17`, and `gpt-5.4-nano-2026-03-17`.

parisons); at the prompt level, rankings of the 95 prompts by mean score are also largely consistent across judges (Spearman 0.814–0.936 for novelty and 0.593–0.910 for quality).

Dimension	N	Ordinal α [95% CI]	All exact	All within 1
Quality	950	0.586 [0.525, 0.638]	6.8%	53.3%
Novelty	950	0.654 [0.609, 0.692]	8.0%	62.9%
Diversity	95	0.392 [0.287, 0.484]	5.3%	38.9%

Table 6: Five-judge agreement. "All exact" is the fraction of items on which all five judges give the same score; "All within 1" requires the maximum and minimum of the five scores to differ by at most one point.

Judge A	Judge B	Δ	MAE	Exact	Within 1	Spearman	κ_w [95% CI]
Quality ($N = 950$ responses)							
gemini-3-flash	gpt-5.4	+0.464	0.748	40.7%	86.6%	0.750	0.666 [0.61, 0.72]
gemini-3-flash	gpt-5.4-mini	+0.138	0.755	37.3%	88.1%	0.698	0.664 [0.61, 0.71]
gemini-3-flash	gpt-5.4-nano	−0.125	0.877	34.4%	81.5%	0.567	0.540 [0.46, 0.61]
gemini-3-flash	gpt-oss-20b	−0.248	0.897	37.8%	78.1%	0.578	0.582 [0.50, 0.65]
gpt-5.4	gpt-5.4-mini	−0.326	0.482	57.2%	94.9%	0.774	0.733 [0.69, 0.77]
gpt-5.4	gpt-5.4-nano	−0.589	0.707	43.9%	86.7%	0.678	0.582 [0.51, 0.64]
gpt-5.4	gpt-oss-20b	−0.713	0.944	33.9%	76.2%	0.607	0.513 [0.44, 0.58]
gpt-5.4-mini	gpt-5.4-nano	−0.263	0.554	54.1%	90.8%	0.654	0.647 [0.58, 0.70]
gpt-5.4-mini	gpt-oss-20b	−0.386	0.795	39.2%	83.4%	0.587	0.580 [0.51, 0.64]
gpt-5.4-nano	gpt-oss-20b	−0.123	0.702	43.6%	88.3%	0.605	0.626 [0.56, 0.68]
Novelty ($N = 950$ responses)							
gemini-3-flash	gpt-5.4	−0.352	0.497	54.1%	96.6%	0.804	0.823 [0.79, 0.85]
gemini-3-flash	gpt-5.4-mini	−0.142	0.445	59.5%	96.5%	0.791	0.809 [0.78, 0.84]
gemini-3-flash	gpt-5.4-nano	+0.336	0.786	33.9%	87.7%	0.660	0.568 [0.52, 0.61]
gemini-3-flash	gpt-oss-20b	−0.003	0.643	46.3%	90.1%	0.697	0.675 [0.64, 0.71]
gpt-5.4	gpt-5.4-mini	+0.209	0.454	58.2%	96.9%	0.804	0.819 [0.79, 0.84]
gpt-5.4	gpt-5.4-nano	+0.687	0.980	24.3%	79.3%	0.688	0.502 [0.45, 0.55]
gpt-5.4	gpt-oss-20b	+0.348	0.708	45.3%	85.8%	0.675	0.638 [0.59, 0.68]
gpt-5.4-mini	gpt-5.4-nano	+0.478	0.731	38.5%	88.8%	0.679	0.577 [0.52, 0.62]
gpt-5.4-mini	gpt-oss-20b	+0.139	0.596	49.3%	91.8%	0.690	0.681 [0.64, 0.72]
gpt-5.4-nano	gpt-oss-20b	−0.339	0.617	46.8%	91.8%	0.674	0.619 [0.57, 0.66]
Diversity ($N = 95$ response sets)							
gemini-3-flash	gpt-5.4	+1.158	1.263	23.2%	65.3%	0.393	0.244 [0.12, 0.37]
gemini-3-flash	gpt-5.4-mini	+1.042	1.189	25.3%	69.5%	0.439	0.297 [0.16, 0.43]
gemini-3-flash	gpt-5.4-nano	+0.621	0.937	33.7%	75.8%	0.380	0.278 [0.13, 0.42]
gemini-3-flash	gpt-oss-20b	+0.768	1.189	31.6%	66.3%	0.355	0.298 [0.14, 0.46]
gpt-5.4	gpt-5.4-mini	−0.116	0.558	50.5%	94.7%	0.618	0.608 [0.45, 0.73]
gpt-5.4	gpt-5.4-nano	−0.537	0.663	40.0%	93.7%	0.618	0.479 [0.35, 0.59]
gpt-5.4	gpt-oss-20b	−0.389	0.853	37.9%	82.1%	0.559	0.475 [0.33, 0.61]
gpt-5.4-mini	gpt-5.4-nano	−0.421	0.653	44.2%	90.5%	0.580	0.493 [0.35, 0.61]
gpt-5.4-mini	gpt-oss-20b	−0.274	0.821	36.8%	83.2%	0.593	0.569 [0.43, 0.68]
gpt-5.4-nano	gpt-oss-20b	+0.147	0.842	34.7%	82.1%	0.556	0.454 [0.31, 0.58]

Table 7: Pairwise agreement among LLM judges on Qwen3-1.7B SEPO outputs for Artificial Hivemind. Δ is the mean score difference, Judge A minus Judge B. MAE is the mean absolute score difference. κ_w is quadratic weighted Cohen's kappa with a prompt-cluster bootstrap 95% CI.

Choice of primary judge. We use `gemini-3-flash-preview` as the primary judge because it is among the most consistent novelty judges in the panel (mean $\kappa_w = 0.719$ with the other judges, essentially tied with the highest, 0.721), lies in the middle of the panel for quality (mean $\kappa_w = 0.613$), and offers a practical balance between evaluation quality and cost. Diversity is the exception: `gemini-3-flash-preview` is both the most lenient diversity judge and the least aligned with the rest of the panel, so its absolute diversity scores should be interpreted with care. This

validation covers only our method, SEPO, so it does not test whether the between-method rankings in the main experiments are invariant to the judge. Finally, agreement among LLM judges measures consistency, not correctness, and cannot be taken as a substitute for validation against human ratings.

I Correlation and Reliability of LLM-as-a-judge-Based, Embedding-Based, and N-Gram-Based Diversity and Novelty

In this section, we analyze the three types of diversity and novelty metrics we used to evaluate the performance of all methods on creative generation. Embedding-based metrics (Jiang et al., 2026), N-gram-based metrics (Cao et al., 2026), and LLM-as-a-Judge metrics (Hou et al., 2026a) are all widely used to evaluate creative generation tasks. However, there has been little discussion of the agreement or reliability across these three types of metrics, which calls into question the use of automatic metrics such as embedding-based and N-gram-based metrics. Here, we provide a preliminary analysis to better understand how these metrics relate to one another and to ground our use of automatic metrics. Specifically, we investigate three questions about diversity and novelty metrics for creative generation: (1) To what extent do embedding-based, N-gram-based, and LLM-as-a-Judge metrics agree with one another? (2) How reliable is each family of metrics? (3) Are embedding-based metrics robust to the choice of embedding model?

Setup. We use all generated responses from the main experiments. There are 3 generator models (Qwen3-1.7B, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3), 4 datasets (Artificial Hivemind, Creativity Index, NoveltyBench, Creative Math), and 5 sampling methods (base, min-p, verbalized, sequential, SEPO), all at temperature 0.7. This gives 60 conditions, $58{,}090$ responses, and $5{,}809$ ten-response sets. LLM-as-a-judge is `Gemini-3-Flash` (judge temperature 1.0). It gives a 1–5 *diversity* score to each ten-response set and a 1–5 *novelty* score to each response. Novelty is judged relative to a reference set of ten base-sampling responses from the same model to the same prompt. We compute the automatic metrics on exactly the texts and reference sets the judge saw. *Embedding-based* metrics use `BGE-M3` cosine distance. Diversity is the mean pairwise distance within a set; novelty is the distance to the nearest reference response. *N-gram-based* metrics use L-uniqueness, the fraction of a response's tokens not covered by shared L-grams, averaged over $L = 5, \ldots, 12$. For diversity the other nine responses in the set serve as the reference (leave-one-out); for novelty the base references do. All correlations are Spearman's ρ. Confidence intervals are 95% percentile intervals from $1{,}000$ bootstrap resamples of whole prompts.

I.1 Agreement among LLM-as-a-judge-Based, Embedding-Based, and N-Gram-Based Diversity and Novelty Metrics

Table 8 reports how closely the three metric families agree. At the level of method comparison, agreement is strong. For novelty, the judge and embedding metrics are nearly interchangeable (condition-wise $\rho = 0.91$, ranking $\tau = 0.83$), and both pairs involving the n-gram metric still reach $\rho \geq 0.72$. Diversity shows moderate agreement across all pairs ($\rho = 0.55$–0.67, $\tau = 0.60$–0.65). *Conclusions about which methods produce more novel or diverse outputs therefore hold regardless of which metric family is used.*

Agreement is weaker for individual outputs, and the two constructs behave differently. For novelty, response-wise correlations are high ($\rho = 0.66$–0.80) but fall sharply once prompt effects are removed (prompt-wise $\rho = 0.32$–0.38). The metrics agree largely on which *prompts* elicit novel responses, but much less on which of the ten responses to the same prompt is most novel. For diversity, prompt-wise correlations ($\rho = 0.43$–0.61) stay close to the response-wise values ($\rho = 0.39$–0.65). Differences between methods' response sets for the same prompt are therefore captured about as consistently as differences across prompts. Across both constructs, the judge–n-gram pair agrees least, which fits the gap between surface-level lexical overlap and the semantic judgments made by the LLM judge. Overall, the metric families support the same method-level conclusions but should not be treated as interchangeable for fine-grained, within-prompt assessments, especially for novelty. We therefore report all three families throughout.

Table 8: Agreement between metric families (Judge = LLM-as-a-judge, Emb. = embedding-based, N-gram = n-gram-based). *Method comparison*: each condition (model×dataset×method) is first averaged into one score. *Condition-wise* ρ: Spearman correlation between two metrics over the 60 condition means. *Ranking* τ: how similarly two metrics rank the 5 methods within each model×dataset pair (mean Kendall τ over the 12 pairs). *Individual outputs*: scores of single responses (novelty, $n = 58{,}090$) or ten-response sets (diversity, $n = 5{,}809$). *Response-wise* ρ: Spearman correlation over all of them. *Prompt-wise* ρ: Spearman correlation after centering ranks within each group (a repeated-measures rank correlation), so it reflects only agreement among outputs for the same prompt. For novelty, the group is the 10 responses of one set (same prompt and method); for diversity, it is the 5 methods' sets for one prompt. 95% prompt-cluster bootstrap CIs are at most ± 0.03 for Response-wise ρ and ± 0.08 for Prompt-wise ρ.

		Method comparison		Individual outputs	
Construct	Pair	Condition-wise ρ	Ranking τ	Response-wise ρ	Prompt-wise ρ
Novelty	Judge ∼ Emb.	0.91	0.83	0.79	0.38
	Judge ∼ N-gram	0.73	0.65	0.66	0.32
	Emb. ∼ N-gram	0.72	0.55	0.80	0.37
Diversity	Judge ∼ Emb.	0.67	0.65	0.55	0.57
	Judge ∼ N-gram	0.55	0.60	0.39	0.43
	Emb. ∼ N-gram	0.67	0.62	0.65	0.61

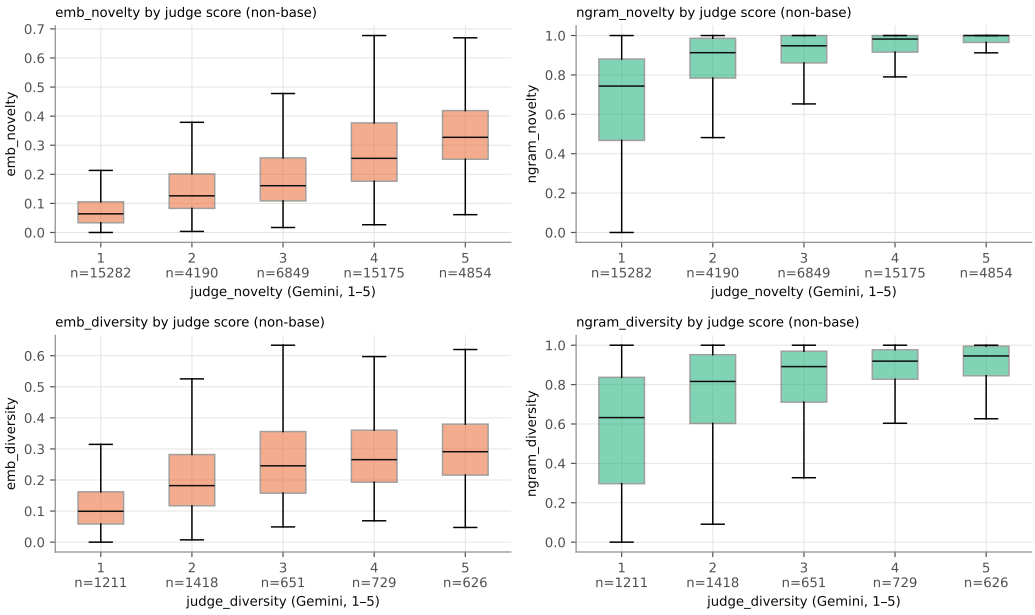

Figure 20: Automatic metrics grouped by the judge score. Top: novelty (responses). Bottom: diversity (ten-response sets). Left: embedding. Right: n-gram L-uniqueness. The n-gram metrics approach their ceiling once the judge score is 2–3, whereas embedding distance keeps rising across all five levels.

Table 9: Split-half reliability. We randomly split the prompts of each dataset into two halves (500 splits) and recompute every condition mean on each half. ρ_{half}: Spearman correlation between the two halves over the 60 conditions. SB: Spearman–Brown correction to the full prompt set. $\bar{\tau}$ (P5): mean (5th percentile) Kendall τ between the method rankings from the two halves within each block.

Construct	Metric	ρ_{half}	SB	$\bar{\tau}$ (P5)
Diversity	judge	0.969	0.984	0.90 (0.84)
	embedding	0.971	0.986	0.92 (0.88)
	n-gram	0.975	0.987	0.89 (0.83)
Novelty	judge	0.983	0.991	0.96 (0.93)
	embedding	0.989	0.994	0.97 (0.95)
	n-gram	0.980	0.990	0.87 (0.83)

I.2 Reliability of Each Metric

We next ask whether the disagreement in Table 8 comes from measurement noise. Table 9 reports split-half reliability over prompts. All six metrics are highly reliable at the condition level ($\rho_{\text{half}} \geq 0.969$; Spearman–Brown ≥ 0.984). Method rankings within each model$\times$dataset pair are also stable across halves ($\bar{\tau} = 0.87$–0.97, P5 ≥ 0.83). Novelty rankings are the most stable under the judge and embedding metrics ($\bar{\tau} \geq 0.96$). The n-gram metric is somewhat less stable for novelty ($\bar{\tau} = 0.87$), and the three metrics are comparable for diversity ($\bar{\tau} = 0.89$–0.92).

Because every metric's reliability is near ceiling, the maximum inter-metric correlation that measurement error allows is close to 1. The moderate agreement in Table 8 therefore cannot be explained by noise. Each metric reliably measures a somewhat different aspect of diversity and novelty. This supports reporting the three metric families together rather than treating any one as a noisy proxy for the others.

I.3 Choice of Embedding Model

We test whether the embedding-based metrics are sensitive to the choice of encoder by comparing BGE-M3 (our default) with Qwen3-Embedding-0.6B and all-mpnet-base-v2 (Table 10). The three models agree closely with one another. Condition-wise correlations are 0.97–0.99 for novelty and 0.87–0.98 for diversity, and response- and prompt-wise correlations are 0.85–0.94. BGE-M3 and MPNet are nearly interchangeable ($\rho \geq 0.94$). Qwen3 deviates slightly more, mainly for diversity.

Agreement with the judge is also largely insensitive to the encoder. For novelty, all three models correlate with the judge at nearly the same level (response-wise 0.68–0.70, condition-wise 0.90–0.92). For diversity, prompt-wise agreement is similar (0.49–0.53), but condition-wise agreement varies more (0.52–0.67), and BGE-M3 agrees with the judge most closely. The conclusions in Section 6 thus hold for all three encoders, and BGE-M3 is a reasonable default.

J Evaluation Metric Details

We evaluate quality, diversity, and novelty[4] using generated responses. Let $\mathcal{P}$ be the evaluation prompt set, $\mathcal{Y}_x = (y_{x,1}, \ldots, y_{x,n_x})$ the sampled responses to prompt x, and $\mathcal{R}_x$ its external base-response reference pool. Embedding metrics instantiate the response dissimilarity in Section 3; lexical metrics provide complementary measurements based on reference-fragment coverage.

Quality. Each output is scored on a 1–5 Likert scale according to a task-specific rubric, using an LLM-as-a-judge protocol (Gu et al., 2025). The rubric assesses coherence, relevance, and task-specific

[4]We operationalize novelty as departure from the reference model's responses to the same input. This reference-relative notion captures whether a method elicits content beyond the behaviors represented in base model; it does not measure originality relative to all knowledge or human-generated content.

Table 10: Robustness to the embedding model. BGE-M3 and Qwen3-Embedding-0.6B encode up to 8,192 tokens; all-mpnet-base-v2 truncates at 384. BGE, Qwen3, and MPNet denote BGE-M3, Qwen3-Embedding-0.6B, and all-mpnet-base-v2; rows under *With the judge* give each model's correlation with the judge. Response-wise (novelty) and prompt-wise (diversity) correlations exclude base; condition-wise correlations use all 60 conditions.

	Novelty		Diversity	
	Response-wise	Condition-wise	Prompt-wise	Condition-wise
Between embedding models				
BGE∼Qwen3	0.92	0.97	0.85	0.87
BGE∼MPNet	0.94	0.99	0.94	0.98
Qwen3∼MPNet	0.91	0.98	0.86	0.89
With the judge				
BGE	0.70	0.91	0.53	0.67
Qwen3	0.68	0.90	0.50	0.52
MPNet	0.69	0.92	0.49	0.58

usefulness. For N scored outputs, we report

$$\text{Quality} = \frac{1}{N}\sum_{i=1}^{N} s(y_i), \qquad s(y_i) \in \{1,2,3,4,5\}. \tag{11}$$

J.1 Embedding-Based Diversity and Novelty

Encoding and dissimilarity. The default encoder is `BAAI/bge-m3`, loaded through `SentenceTransformer`. It encodes response text alone, without the prompt. For embeddings e_y and e_z, we use cosine dissimilarity,

$$s(y,z) = \frac{e_y^\top e_z}{\|e_y\|_2\,\|e_z\|_2}, \qquad d(y,z) = 1 - s(y,z). \tag{12}$$

In implementation, embeddings are ℓ_2-normalized. Responses are encoded with a maximum input length of 8192 tokens; longer responses are truncated rather than chunked.

Embedding-based Diversity. We average over all unordered pairs of distinct sample indices within a prompt, then give every prompt equal weight:

$$D_x = \frac{1}{\binom{n_x}{2}}\sum_{i<j} d(y_{x,i}, y_{x,j}), \qquad \text{Div}_{\text{emb}} = \frac{1}{|\mathcal{P}|}\sum_{x\in\mathcal{P}} D_x. \tag{13}$$

This is the reported `mean_embedding_diversity`. Self-pairs are excluded, but equal-text samples at different indices remain distinct samples. Ordinary embedding diversity includes empty responses without assigning them a fixed distance.

Embedding-based Novelty. Novelty measures how far a response is from what the base model produces for the same prompt. For each response, we take its distance to the closest reference response:

$$\text{Nov}_{\mathcal{R}_x}(y) = \min_{r\in\mathcal{R}_x} d(y,r) = 1 - \max_{r\in\mathcal{R}_x} s(y,r). \tag{14}$$

Let $\mathcal{E}$ be the set of responses (x,i) that have a base-response reference pool $\mathcal{R}_x$. The overall novelty is the average over these responses:

$$\text{Nov}_{\text{emb}} = \frac{1}{|\mathcal{E}|}\sum_{(x,i)\in\mathcal{E}} \text{Nov}_{\mathcal{R}_x}(y_{x,i}). \tag{15}$$

This is the reported `mean_embedding_novelty`. Empty responses receive a novelty of zero.

J.2 Lexical Diversity and Novelty

Tokenization and fragment coverage. We use L-Uniqueness (Lu et al., 2025a), which measures how much of a response cannot be found verbatim in a set of reference texts. Responses are split on whitespace, and each token is lowercased after stripping the characters `.,!?;:"'()[]{}` from both ends. We apply no stemming, stop-word removal, or language-specific segmentation.

For a response with tokens $T(y) = (t_1, \ldots, t_n)$ and a set of reference texts B, let $G_L(B)$ be the set of all contiguous length-L token fragments that appear in some reference. A token position in y is covered if it lies inside a length-L window of y that matches a fragment in $G_L(B)$; let $C_L(T(y), B)$ be the set of covered positions. Then

$$U_L(y; B) = \frac{n - |C_L(T(y), B)|}{n}, \qquad \overline{U}(y; B) = \frac{1}{8} \sum_{L=5}^{12} U_L(y; B). \tag{16}$$

U_L is the fraction of tokens in y not covered by any length-L match, and $\overline{U}$ averages it over fragment lengths 5 to 12, as suggested by Lu et al. (2025a). Empty responses receive zero.

N-gram-based Diversity. To measure diversity, each response is compared with the other responses to the same prompt $\mathcal{Y}_{x,-i} = (y_{x,j} : \forall j \neq i)$. With $N = \sum_{x \in \mathcal{P}} n_x$,

$$\text{Div}_{\text{lex}} = \frac{1}{N} \sum_{x \in \mathcal{P}} \sum_{i=1}^{n_x} \overline{U}(y_{x,i}; \mathcal{Y}_{x,-i}). \tag{17}$$

This corresponds to `mean_n_gram_based_diversity`. Fragments are deduplicated within each reference response before counting the number of responses containing them. Repetition within a candidate alone does not constitute overlap with another response. One edge case deserves attention: a nonempty response shorter than L tokens contains no length-L window, so its leave-one-out U_L is one even when another response is identical to it. In our settings, most responses exceed 100 tokens, so this case rarely arises.

N-gram-based Novelty. Lexical novelty uses the external base-response reference pool $\mathcal{R}_x$ to evaluate the n-gram based novelty:

$$\text{Nov}_{\text{lex}} = \frac{1}{N} \sum_{x \in \mathcal{P}} \sum_{i=1}^{n_x} \overline{U}(y_{x,i}; \mathcal{R}_x). \tag{18}$$

If a response is identical to one of the references, its novelty is zero for every L, including lengths longer than the response itself. Otherwise, as in leave-one-out diversity, a nonempty response shorter than L tokens receives $U_L = 1$.

K Baseline Method Prompt Details

Box 1: Verbalized sampling

```
Generate {num_samples} responses with their corresponding probabilities.

User instruction:
{prompt}

Return only valid JSON with this exact shape:
{
  "responses": [
    {"text": "first response", "probability": 0.42},
    {"text": "second response", "probability": 0.31}
  ]
}

Requirements:
- Include exactly {num_samples} response objects.
- Each response object must contain only "text" and "probability".
- "text" must be a non-empty string that directly answers the user instruction.
- "probability" must be a number between 0 and 1.
- Do not include markdown fences, comments, analysis, or extra keys.
```

Box 2: Sequential sampling

Multi-Hop conversation used to generate response k

```
[User]       {prompt}
[Assistant]  {response 1}
[User]       <follow-up instruction>
  ...
[Assistant]  {response k-1}
[User]       <follow-up instruction>
```

Follow-up instruction

```
Generate one more response to the original user instruction.

Requirements:
- The new response must be substantively different from every response already given in
↪  this conversation.
- Use a genuinely different idea, content, approach, or perspective; superficial rewording
↪  is not sufficient.
- Preserve the requested topic, language, answer type, count, format, and all other
↪  constraints from the original user instruction.
- Do not quote, summarize, critique, or mention the earlier responses.
- Return only the new response, without analysis or commentary about these requirements.
```

Box 3: System Prompt Optimization Meta-Prompt

```
# Objective
You're tasked to revise the current proposed system instruction for a downstream
↪  generation model.
You will see the task, the current instruction archive, the history of proposed
↪  instruction candidates with marginal diversity scores, and the current proposed
↪  instruction with its evaluation result.

Specifically, your goal is to revise the current generation instruction so that it elicits
↪  responses that are both:
1. high quality for the downstream task
2. diverse relative to responses elicited by the existing instruction archive.

Important form constraint: the improved instruction is a reusable system instruction that
↪  will be applied to one user prompt at a time. It is not a replacement user prompt, not
↪  a prompt template, and not a request to generate many alternatives. It should steer
↪  how the model answers the user's actual prompt once.

The evaluator assigns each instruction a marginal diversity score: how much the
↪  instruction increases response diversity relative to the current instruction archive.
Marginal diversity means the revised instruction should elicit responses that explore
↪  meaningfully different reasoning paths, content choices, perspectives, or solution
↪  strategies, not merely different wording, punctuation, formatting, tone, or visible
↪  output form.
Keep the current instruction archive in mind as the reference set: a strong revision
↪  should be complementary to the archive and should avoid eliciting responses that
↪  follow the same strategy patterns already covered there.

# Downstream Task Description
{dataset task description}

This search is shared across the following prompt set. Produce instruction branches that
↪  are broadly useful for this task type, not overfit to only one prompt. Each branch
↪  must preserve the individual user prompt at generation time while encouraging a
↪  distinct response mode, perspective, inference, abstraction, or reasoning direction.
↪  Use these prompts as references for task coverage, not content to copy; do not treat
↪  repeated topics, genres, or examples as the global task.

Prompt set:
{five training prompts, numbered [1]-[5]}

# Current Instruction Archive
<Current_Selected_Instruction_Archive> Empty </Current_Selected_Instruction_Archive>

# Instruction Proposing History
Base behavior is always an implicit reference. Avoid cross-round score comparisons.
The history contains existing proposed instruction candidates, their marginal diversity
↪  scores, and any available feedback.
Use this history as accumulated proposing experience for revising the current proposed
↪  instruction.
Infer which candidate-level choices are associated with higher marginal diversity scores
↪  and which choices are associated with lower marginal diversity scores.
```

```
Use marginal scores only as evidence of behavioral distance from the applicable archive. A
↪  higher marginal score is not evidence of better downstream quality and is not a reason
↪  to intensify complexity.
Scores are noisy: if a high-scoring candidate is a replacement prompt, contains fill-in
↪  blanks, asks for a series of responses, lists the user's constraints, exposes a
↪  planning step, changes the requested output form, mentions the archive/history/prompt
↪  set, or targets one concrete prompt, treat that shape as a failure even if its score
↪  is high.
Use those lessons to decide what to preserve, remove, refine, or replace so the revision
↪  contributes a complementary semantic behavior while keeping the downstream task
↪  intact.
The new instruction must be meaningfully different from every instruction in the current
↪  archive and history.
<Recent_Completed_Round_History>Empty</Recent_Completed_Round_History>

# Current Proposed Instruction with the corresponding score feedback
<Current_Proposed_Candidate>
<Instruction>Base behavior (no added system instruction)</Instruction>
<Marginal_Diversity_Score higher_is_better="true" relative_to="current_selected_archive"
↪  aggregation="minimum_pair_distance">unscored</Marginal_Diversity_Score>
</Current_Proposed_Candidate>

# Requirements:
- Treat downstream quality requirements as binding; improve marginal diversity among
↪  instructions likely to satisfy them across task subtypes.
- Make the revised instruction complementary to the current archive.
- Change how the model selects or develops the requested content, not its surface phrasing
↪  or visible output form.
- Keep preparatory planning internal.
- Prefer content generation over constraint management.
- Adapt mechanisms to each request.
- The response must contain only the requested answer and preserve every explicit content
↪  and output constraint.
- Use the history to identify what has worked or failed before, then apply those lessons
↪  directly to the current proposed instruction.
- The improved instruction must be a system instruction for answering the user's actual
↪  prompt once.
- The improved instruction must be complete natural-language prose, ready to use as-is; it
↪  must not contain square brackets, fill-in blanks, variable slots, or schema-like
↪  variables.
- The improved instruction must not include concrete example prompts or tailor itself to a
↪  specific prompt or repeated subtype from the prompt set.
- Optimize for diversity in likely downstream responses, not merely unusual wording in the
↪  instruction itself.
- Avoid novelty that harms task alignment, factuality, coherence, or usefulness.
- Express diversity through a coherent, prompt-relevant mechanism affecting substantive
↪  content, reasoning, causal relationships, perspective, or connections within the
↪  requested answer, rather than genre switches, formatting, sections, or visible
↪  procedures.
- Revise any instruction that would alter explicit constraints, expose the strategy, or
↪  reduce coherence and directness across task subtypes.

# Output format
Use the model's Harmony analysis channel for all reasoning about the revision. Do not
↪  repeat that reasoning in the final channel.
The final channel must contain exactly one improved instruction inside the following
↪  mandatory opening and closing tags, with no other text:

<Improved_Instruction>
improved_instruction
</Improved_Instruction>

The improved instruction must be self-contained natural-language prose ready to use as a
↪  system instruction.
```

L Limitations of Automatic Creativity Metrics

Our numerical diversity and novelty metrics are heuristic proxies for creativity. Prior work documents inconsistencies across automatic creativity measures and domains, and shows that N-gram-based measures can primarily capture lexical variation rather than conceptual creativity (Lu et al., 2026). Embedding-based distances measure semantic separation, which need not imply creative originality, while N-gram-based scores can reward changes in wording without corresponding changes in ideas. Text length may also influence these measurements and should be examined as a potential confound. We adopt established, widely used measurement approaches to facilitate comparison with prior work, but improvements on these proxies do not by themselves establish improvements in human-perceived

creativity. Broader human studies across tasks and participant populations are needed to assess whether the measured gains correspond to creativity as perceived by humans.