EVALUATING LLM CREATIVITY THROUGH LONG-TAIL PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Creativity benchmarks for large language models (LLMs) vary widely in task and metric design. This raises the question of whether they capture a common notion of creativity. We propose that this shared target can be formalized as task-specific performance in the long tail of a reference input-output distribution, which describes how common input-output pairs are relative to a chosen source of text. We ground this proposition in the novelty-quality product formulation and evaluate it empirically across 15 benchmarks and 26 open-weight models, using n-gram absence from the pretraining data as a novelty proxy. Our results show that creativity benchmarks lie further in the tail regions than standard benchmarks. Moreover, on both creativity and standard benchmarks, model rankings from higher-tail subsets of each benchmark align more closely with human creativity judgments in LM-Arena. Building on these findings, we introduce q@Tail(α), a lower-cost creativity diagnostic that measures mean task quality on the tail fraction of an existing benchmark. Applying it to creativity-promoting methods reveals substantial room to improve LLM creativity through stronger tail performance.

1 INTRODUCTION

Creativity is a central capability of intelligence, with a long history of study in human cognition (Stein, 1953; Barron, 1955; Amabile, 1983; Boden, 1994; Runco & Jaeger, 2012) and growing attention in artificial intelligence. As large language models (LLMs) are increasingly used for open-ended tasks such as writing, ideation, planning, problem solving, and discovery (Chakrabarty et al., 2024b; Ruan et al., 2025; Tian et al., 2025; Gottweis et al., 2026), creativity has become an important target for model evaluation. Yet despite decades of study, human creativity lacks a consensus definition across psychology, cognitive science, design, education, and computational creativity (§2).

Existing LLM creativity benchmarks differ in both task and metric design. Task dimensions include the objective (such as writing, problem solving, or ideation) and the domain (such as language, mathematics, or science). Metric dimensions include the granularity at which outputs are evaluated (such as n-grams, sentences, or paragraphs). We organize these in detail in a taxonomy in §A. This variety raises the question of whether such heterogeneous benchmarks nevertheless capture a common notion of creativity. Identifying it would provide a principled basis for designing, validating, and comparing LLM creativity evaluations.

In this work, we propose that this shared target can be formalized as performance in the long tail of a reference input-output distribution (§3). Here, a reference distribution specifies how common inputs and outputs are relative to a chosen source of text, such as a pretraining corpus. Under this view, one relevant aspect of evaluating LLM creativity is asking whether a model can produce high-quality outputs in low-probability regions of that distribution (Figure 1). Standard benchmarks tend to evaluate model performance in high-probability regions (i.e., peak) of this distribution, where inputs and expected outputs are probable and common. Creativity benchmarks, by contrast, shift evaluation toward the long tail: regions where the input is improbable, the output is improbable given the input, or the input-output pair is otherwise improbable under the reference distribution.

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.

We investigate this proposition through a literature review (§2;§A), theoretical analysis (§3.2), and empirical evaluation (§5-§6). First, we survey 82 papers and organize their creativity evaluation designs into a taxonomy (Sections A and 2). Next, we adopt a formulation of creativity that captures the joint presence of novelty and quality through their product (Padmakumar et al., 2025). Under this formulation, the same quality difference contributes more to the benchmark-level creativity score when it occurs in higher-novelty regions (§3.2). Finally, our empirical analysis shows that creativity benchmarks lie further in the tail of the reference distribution than standard benchmarks (§5). And in practice, common creativity elicitation operations push benchmarks toward distinct tail regions (§5.1). On both creativity and standard benchmarks, rankings from higher-tail subsets agree more closely with the LM Arena creative-writing ranking (Chiang et al., 2024) than rankings from lower-tail subsets (§6.1). Building on these findings, we introduce $q@\text{Tail}(\alpha)$, a creativity diagnostic that measures mean task quality on the most novel α fraction of an existing benchmark. We then apply it to a range of test-time and training-time creativity-promoting methods (§7). Most methods increase output novelty without consistently improving quality in the tail, revealing substantial room to improve LLM creativity through stronger tail performance.

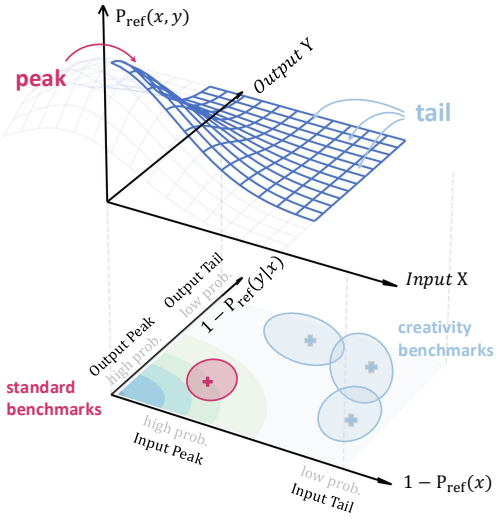


Figure 1: **Creativity benchmarks lie in the long tail of a reference input-output distribution.** *Top:* A reference distribution $P_{\text{ref}}(x, y)$ assigns higher probability to common input-output pairs and lower probability to uncommon or novel ones. The surface is schematic because possible inputs and outputs have no natural ordering. *Bottom:* The same space is re-parameterized by input improbability, $1 - P_{\text{ref}}(x)$, and output improbability, $1 - P_{\text{ref}}(y \mid x)$. Standard benchmarks cluster in high-probability regions (i.e., *Peak*), where inputs and expected outputs are probable and common. Creativity benchmarks shift evaluation toward low-probability regions (i.e., *Tail*), where the input is improbable, the output is improbable given the input, or both. The figure is illusive but depicted peak-tail divergence is empirically corroborated in §5.

In short, we make three contributions:

(*i*) we survey LLM creativity evaluations and characterize their design variation;

(*ii*) we propose long-tail performance under a reference input-output distribution as a shared target and provide theoretical and empirical support;

(*iii*) we introduce a lower-cost tail-performance metric as a creativity diagnostic for existing benchmarks, and assess creativity-promoting methods.

2 What is Missing in Literature?

Despite decades of study in psychology and cognitive science, creativity has no broadly accepted definition (Runco & Jaeger, 2012). The most widely shared definition characterizes creativity as the joint presence of *novelty* and *quality*: an output must be uncommon relative to a reference, and also useful, fitting, or valuable (Stein, 1953; Barron, 1955; Boden, 1994; Amabile, 1983). This lack of consensus is reflected in LLM creativity benchmarks, which span writing (Chakrabarty et al., 2024b; Atmakuru et al., 2024; West & Potts, 2025), problem solving (Ye et al., 2025; Han et al., 2025; Tian et al., 2025), and ideation (Chen & Ding, 2023; Qiu & Hu, 2025; Ruan et al., 2025).

Existing LLM creativity evaluations differ substantially in task and metric design and can yield contrasting conclusions. For example, studies using semantic metrics report strong creative performance by LLMs, whereas evaluations based on expert rubrics identify substantial limitations (Chen & Ding, 2023; Chakrabarty et al., 2024b). Similarly, studies measuring novelty relative to web text report benefits from model scaling, whereas studies measuring diversity across generated responses report declines (Padmakumar et al., 2025; Zhang et al., 2025b). Such contrasting findings motivate making evaluation settings explicit when comparing claims about LLM creativity. Our review of

82 papers organizes task and metric design choices into a nine-dimensional taxonomy (§A). By making differences in the dimensions, the taxonomy highlights potential sources of disagreement and provides a perspective that may help explain divergent findings. Our taxonomy subsumes prior comparisons and surveys that document some of such fragmentation (Lu et al., 2025a; Hou et al., 2025; Jiang et al., 2025; Ismayilzada et al., 2024; Franceschelli & Musolesi, 2024).

We then move beyond characterizing these differences to a more fundamental challenge: determining whether a common evaluative target underlies these heterogeneous task and metric designs. Identifying such a target would provide a principled basis for interpreting existing creativity benchmarks and guiding the design of new evaluations. On the metric side, Padmakumar et al. (2025) operationalize creativity through novelty and quality, which can explain different metric designs in the literature. On the task side, however, the implicit target underlying tasks designed to elicit creativity remains unexplained. We propose that creativity benchmark design implicitly shifts evaluation toward task performance in the long tail of a reference input-output distribution (§3). Empirically, we demonstrate a shift toward higher novelty in the evaluated creativity benchmarks relative to standard benchmarks (§5), and examine whether tail performance is positively associated with human judgments of model creativity (§6).

3 Evaluate LLM Creativity through Long-Tail Performance

3.1 Definition

Let $\mathcal{X}$ denote the input space (e.g., prompts, problems, or questions) and $\mathcal{Y}$ the output space (e.g., responses, solutions, or answers). A **reference distribution** $P_{\text{ref}}(x, y)$ specifies how common an input-output pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is relative to a chosen source of text (Figure 1, ***Top***). The choice of reference P_{ref} is an evaluation parameter: different sources, such as a pretraining corpus, a human-authored corpus, or a model's self-samples, induce different notions of novelty (discussed in §A).

The factorization, $P_{\text{ref}}(x, y) = P_{\text{ref}}(x) \cdot P_{\text{ref}}(y \mid x)$, admits a natural 2D parameterization with axes $P_{\text{ref}}(x)$ and $P_{\text{ref}}(y \mid x)$ (Figure 1, ***Bottom***). Each (x, y) pair is a point in this plane; a benchmark is a finite set of such points, yielding an empirical distribution in this space. This factorization decomposes the input-output space into four quadrants: the *peak*, where both the input and output are common; the *input-tail*, where the input is novel; the *output-tail*, where the output is novel given the input; and the *joint-tail*, where both are novel.

Let $s_{\text{ref}}(x, y) = \phi(P_{\text{ref}}(x)P_{\text{ref}}(y \mid x)) \in [0, 1]$ denote a normalized, monotonically decreasing transformation of the joint reference probability to a **novelty** score. Larger novelty $s_{\text{ref}}(x, y)$ therefore indicates that the input-output pair lies further in the reference tail. For benchmark B, let μ_B denote the empirical distribution over its evaluated input-output pairs, and let μ_B^X denote its marginal distribution over inputs. We define the prompt-level novelty score $S_B(x) = \text{Agg}_{Y \sim \mu_B(\cdot|x)} [s_{\text{ref}}(x, Y)]$, where the aggregation is fixed across benchmarks and models and maps scores in $[0, 1]$ to $[0, 1]$. Thus, $S_B(x)$ summarizes how far the input and its associated outputs (from evaluated models) jointly lie in the reference tail.

3.2 Theoretical connection

Creativity is commonly characterized as the joint presence of novelty and quality (Runco & Jaeger, 2012; Boden, 1994; Amabile, 1983). We adopt the novelty-quality product formulation used by Padmakumar et al. (2025):

$$\text{Creativity}(x, y) = \text{Novelty}(x, y) \cdot \text{Quality}(x, y). \tag{1}$$

Let $Y_\pi \sim \pi(\cdot \mid X)$ be the output generated by model π, and let $q : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be a task-specific quality function. We define the expected quality of π on input x as $\bar{q}_\pi(x) = \mathbb{E}_{Y_\pi \sim \pi(\cdot|x)} [q(x, Y_\pi)]$. Under the novelty-quality product formulation, the benchmark-level creativity score of π is

$$C_{\pi,B} = \mathbb{E}_{X \sim \mu_B^X} [S_B(X)\bar{q}_\pi(X)]. \tag{2}$$

Since $S_B(X) \in [0, 1]$, Tonelli's theorem gives the exact identity

$$C_{\pi,B} = \int_0^1 H_B(t) A_{\pi,B}(t)\, dt, \tag{3}$$

where $H_B(t) = \Pr_{X \sim \mu_B^X}(S_B(X) \geq t)$ is the mass of the tail at threshold t, and $A_{\pi,B}(t) = \mathbb{E}_{X \sim \mu_B^X}[\bar{q}_\pi(X) \mid S_B(X) \geq t]$ is the model's quality on that tail. Thus, Equation 3 expresses overall creativity as an aggregation of quality across all nested novelty regions.

Benchmarking primarily concerns ranking with relative model performance. For a baseline model π_0, one benchmark gives

$$C_{\pi,B} - C_{\pi_0,B} = \int_0^1 H_B(t) \left[A_{\pi,B}(t) - A_{\pi_0,B}(t)\right] dt. \tag{4}$$

A quality advantage on an input with novelty $S_B(x)$ appears at every threshold $t \leq S_B(x)$ and therefore receives total weight $S_B(x)$. Consequently, the same quality difference contributes more to the overall creativity difference when it occurs in higher-novelty regions.

A finite benchmark cannot estimate the complete tail-quality curve at arbitrarily fine resolution. We therefore use $q@\mathrm{Tail}_{\pi,B}(\alpha) = A_{\pi,B}(Q_{B,\alpha})$, $Q_{B,\alpha}$ as the threshold of top-α subset of benchmark B, as a coarse-grained summary of quality in a selected tail region. It approximates the relative ranking induced by $C_{\pi,B}$ when differences in the selected tail dominate, or are ranking-consistent with, the remaining contributions. We verify this condition empirically in §6.

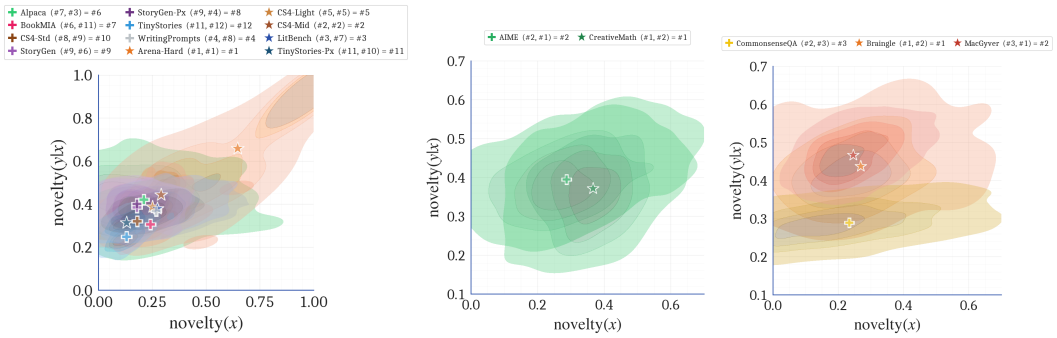
4 Experimental Setup

Reference distribution. We use the Pile pretraining corpus (Gao et al., 2020) as the reference text source and query n-gram occurrences with Infini-Gram-Mini (Liu et al., 2025b; Xu et al., 2026a). These queries determine whether an n-gram appears in the corpus and supply the corpus-absence novelty scores defined below. The Pile predates all evaluated benchmarks, mitigating contamination that would inflate the measurement.

Novelty metric. We operationalize input and output tailness using n-gram novelty, a computationally tractable measure at the scale of the Pile that follows lexical matching approaches used in memorization analysis and retrieval (Merrill et al., 2024; Liu et al., 2025b).[1] For a text t, $\text{novelty}_n(t)$ is the fraction of its n-grams absent from the corpus; higher novelty_n indicates higher improbability and novelty, placing t further into the tail region $\mathcal{T}$. The results in Sections 5 and 6 and their corresponding appendix analyses use $n = 4$. We also smoke test $n = 5, 6$ and their findings are consistent. For each input x, we first average output novelty across the evaluated models, then take the geometric mean of input novelty and this average to obtain the shared score $S_B(x)$. Input-output pairs sharing the same input are selected together using $S_B(x)$, ensuring that all models are compared on the same prompts.

Models. We select recent model releases to cover contemporary LLMs and use open-weight models to support reproducible evaluation. We include a range of model sizes within our available computational resources to examine performance across model scales. Most selected models are covered by the LM Arena creative-writing leaderboard, providing an external human-preference reference for the ranking analysis (§6). The resulting 26 models are drawn from the Llama (Grattafiori et al., 2024), Qwen (Yang et al., 2025a; Qwen Team, 2026), OLMo (Team Olmo et al., 2025), Gemma (Gemma Team et al., 2025), and Mistral (Mistral AI, 2025) families, plus GPT-OSS-20B (OpenAI et al., 2025), Phi-4 (Abdin et al., 2024), Nanbeige4-3B-Thinking (Yang et al., 2025b), and Nemotron-3-Nano-30B (NVIDIA, 2025) (detailed in Table A3). The set spans both instruction-tuned and reasoning-tuned models; we enable thinking mode by default for the latter and use each model's recommended decoding configuration.

[1] Perplexity from a neural language model trained on the Pile and our n-gram novelty score, implemented through a corpus-backed statistical gram language model, are both grounded in the same reference corpus. However, perplexity also depends on how well the model learns different parts of that corpus, potentially introducing model-dependent variation into tailness estimates. Our implementation queries corpus n-gram occurrence statistics directly through Infini-Gram-Mini, avoiding this additional dependence on neural-model fitting. §F.3 further discusses our check using base-model perplexity.



(A) Free-form writing (B) Mathematics reasoning (C) Free-form QA

Figure 2: **Creativity benchmarks shift toward higher input and/or output novelty than standard benchmarks across all three task domains.** 2D kernel density estimates of each benchmark in the $(\text{novelty}(x), \text{novelty}(y \mid x))$ plane. Novelty is the fraction of 4-grams absent from the Pile, computed using Infini-Gram-Mini (§4). Per-prompt output novelty is averaged across all evaluated models. Markers indicate per-benchmark means.

Benchmarks. We select benchmarks using two criteria: creativity benchmarks must have been introduced or used to evaluate LLM creativity in published work, and standard capability benchmarks must cover the same task domains. This selection allows us to compare the novelty distributions of creativity and standard benchmarks within each task domain. We evaluate 15 benchmarks across three task domains: free-form writing (10), mathematics reasoning (2), and free-form QA (3). Creativity benchmarks include Arena-Hard Creative Writing (Li et al., 2024), CS4 (Atmakuru et al., 2024), CreativeMath (Ye et al., 2025), Braingle (Han et al., 2025), and MacGyver (Tian et al., 2024); standard capability benchmarks include Alpaca (Taori et al., 2023), AIME-All (Balunović et al., 2026), and Commonsense-QA (Talmor et al., 2019). The full catalog is provided in Table A4.

Performance evaluation. Even within a single domain, prior creativity benchmarks use heterogeneous performance metrics, e.g., rubric-based LLM-as-judge (Chakrabarty et al., 2024a; Paech, 2025b), reward models (Zhang et al., 2025a), and reference-similarity scoring (Lu et al., 2025b) are all common for writing alone (§A catalogues the broader space). To prevent metric choice from confounding tail-vs-peak comparisons, we hold the performance metric fixed *within* each task domain; we do not attempt to compare absolute quality across domains. For tasks with ground-truth answers (math, QA), we use exact-match on the extracted final answer via LM-Eval-Harness (Gao et al., 2024). For open-ended generation (e.g., writing), we use the normalized reward from Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025a), the current top-ranked model on RewardBench-v2 (Malik et al., 2025).

Ranking. We run the evaluation pipeline for each benchmark B and obtain a model ranking r_B by sorting all models according to their mean quality score q. As an external reference for human creativity judgment, we use the LM Arena creative-writing ranking (Chiang et al., 2024), which aggregates pairwise human preferences over prompts in the creative-writing category; we denote this reference ranking as r_{Arena}.[2] In §6, we compute Kendall's $\tau(B) := \tau(r_B, r_{\text{Arena}})$ to quantify how closely each benchmark's model ranking aligns with this human-preference reference.

5 Creativity Benchmarks Lie in the Tail of P_{REF}

We test whether creativity benchmarks occupy lower-probability regions of P_{ref} than standard benchmarks from the same task domains.

Setup. For each benchmark, we plot a two-dimensional kernel density estimate in the $(\text{novelty}(x), \text{novelty}(y \mid x))$ plane, using the output-novelty aggregation defined in §4.

[2] `https://arena.ai/leaderboard/text/creative-writing`. LM Arena converts pairwise preferences to relative model scores via online Elo updates (Arena Intelligence, 2026). Per-category leaderboards are produced by classifying each submitted prompt into topic categories.

Results. Creativity benchmarks lie further in the tail than standard benchmarks, shifting toward higher input and/or output novelty. The pattern holds in all three task domains with multiple benchmarks (Figure 2A–C). In writing (Figure 2A), Arena-Hard Creative Writing sits at the high end of both axes (both approximately 0.6 to 0.7), well above standard benchmarks (Alpaca, TinyStories, BookMIA); LitBench and CS4 likewise lie further in the tail than every standard benchmark in this task domain. In math reasoning (Figure 2B), CreativeMath lies to the right of AIME-All along novelty(x). In free-form QA (Figure 2C), MacGyver and Braingle lie above CommonsenseQA along novelty$(y \mid x)$, which anchors the low-novelty end. The pattern is stable across $n \in \{4, 5, 6\}$ and across model families. Pooled univariate and multivariate tests confirm that the creativity and standard distributions differ significantly (all $p < 0.001$; §E.1).

5.1 Three elicitation operations, three tail sub-regions

Figure 3 relates the three creativity elicitation operations in our taxonomy (§A) to input-tail, output-tail, and joint-tail regions.

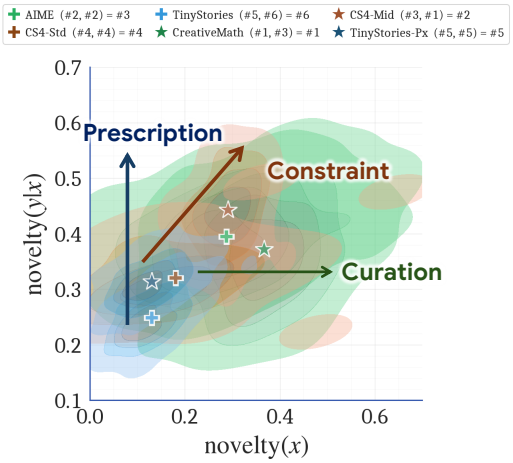


Figure 3: **Three creativity elicitation operations illustrate shifts toward different tail regions.** Aggregated kernel density estimates for creativity and standard benchmarks in the (novelty(x), novelty$(y \mid x)$) plane. Arrows illustrate shifts associated with *Curation* toward higher input novelty (input-tail), *Prescription* toward higher output novelty (output-tail), and *Constraint* toward higher novelty on both axes (joint-tail). Operations are defined in §5.1; the taxonomy is provided in §A.

Elicitation through Curation selects improbable or rare inputs from a larger source distribution. CreativeMath, for example, is filtered by human experts from AMC and AIME for problems requiring non-standard or out-of-the-box thinking. These problems or prompts by nature are more likely be at the tail region of the task domain distribution. So, compared with the standard benchmark of that domain, creativity benchmarks rise novelty(x), placing at the *input-tail* (bottom-right of Figure 3).

Elicitation through Prescription explicitly requests creativity through a system prompt or task instruction, for example by asking the model to write creatively (Chakrabarty et al., 2024a) or propose an alternative solution (Ye et al., 2025). The input is largely unchanged, but the prompt pushes outputs away from the default, placing the benchmark in the *output-tail* (top-left of Figure 3).

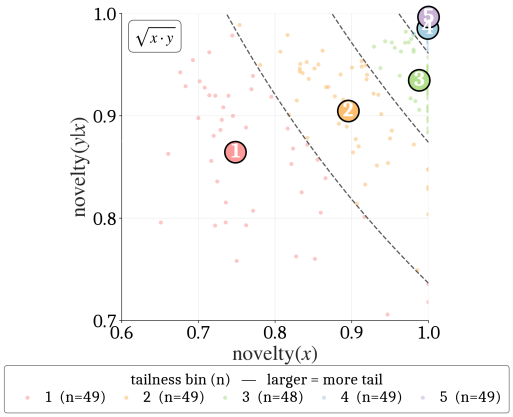
Elicitation through Constraint adds restrictions that steer generation away from common outputs, for example by requiring a CS4 story to use a coastal-town setting and flashbacks to the protagonists' youth (Atmakuru et al., 2024), or forbidding a particular algorithm in code generation (Lu et al., 2025c). Such combinations can make both prompts and acceptable outputs less common under the reference distribution, shifting evaluation toward the *joint-tail* (top-right of Figure 3).

6 Tailness Contributes to More Accurate Creativity Evaluation

§5 shows that creativity benchmarks concentrate on tail regions of P_{ref}. This characterizes how benchmarks are *constructed*, but not whether tail performance is what humans actually reward as creativity. We test this first within each benchmark (§6.1) and then across writing benchmarks (§6.2). All ranking comparisons follow the procedure in §4.

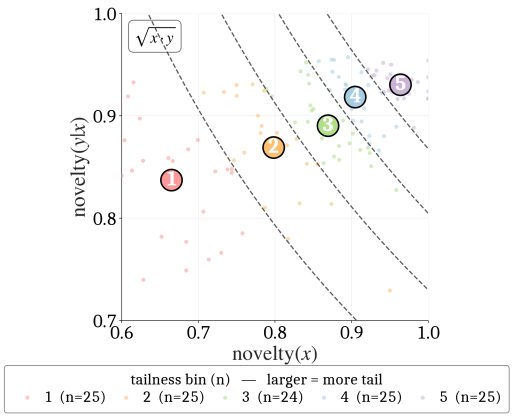
6.1 Within each benchmark, tailness predicts agreement with humans

Setup. For each benchmark B, we form five equal-sized bins using the shared score $S_B(x)$ (§4), from bin 1 (lowest novelty) to bin 5 (highest novelty), keeping pairs with the same input together. We



# Bin	Input $n(x)$	Output $n(y \mid x)$	Joint $n_{\text{geo}}(x, y)$	Agreement τ with human
Bin 1	0.748	0.864	0.799	+0.152
Bin 2	0.895	0.905	0.899	+0.091
Bin 3	0.988	0.935	0.961	+0.061
Bin 4	0.999	0.985	0.992	**+0.394**
Bin 5	1.000	0.997	0.998	+0.333

Figure 4: **Within-benchmark tailness predicts agreement with human creativity rankings** on creativity benchmark (Arena-Hard Creative Writing). All input-output pairs are split into quintile bins by joint n-novelty under geometric-mean aggregation at n=4. Bin 1 is the most-peak (lowest-novelty) subset, bin 5 the most-tail. *Top:* scatter of input-output novelty per prompt, colored by bin, with centroids marked. *Bottom:* per-bin statistics. Tail bins (4-5) reach $\tau = +0.39, +0.33$, well above peak bins (1-3, $+0.06$ to $+0.15$). The trend generally holds across other benchmarks and choice of joint-novelty aggregator; see §E.



# Bin	Input $n(x)$	Output $n(y \mid x)$	Joint $n_{\text{geo}}(x, y)$	Agreement τ with human
Bin 1	0.665	0.837	0.745	+0.099
Bin 2	0.798	0.869	0.832	+0.154
Bin 3	0.869	0.890	0.879	+0.175
Bin 4	0.905	0.918	0.911	**+0.396**
Bin 5	0.964	0.930	0.947	+0.222

Figure 5: **The same trend holds on standard benchmark** (e.g., GPQA-Diamond). All input-output pairs are split into quintile bins by joint n-novelty under geometric-mean aggregation, following Figure 4; Models are ranked by performance (pass@1). Tail bins (4-5) reach $\tau = +0.40, +0.22$, above peak bins (1-3, $+0.10$ to $+0.18$). See other model and joint-novelty in §E.

compute ranking agreement $\tau_i(B)$ separately within each bin. We include additional results on other aggregation methods and more bins in §C.

Results. Higher-tail subsets align more closely with the human creativity ranking. Figure 4 shows the pattern on Arena-Hard Creative Writing: tail bins 4-5 reach $\tau = +0.33$, compared with $+0.06$ to $+0.15$ for peak bins 1-3. The same pattern appears on GPQA-Diamond (Figure 5), which is not a benchmark designed specifically for creativity. This still holds at ten-bin resolution, where the association between bin-level novelty and human-ranking agreement is statistically significant on both benchmarks (Spearman $\rho = 0.661$-0.855, $p = 0.0016$-0.0376).

Takeaway. The observed alignment on both creativity and standard benchmarks (§E) motivates using $q@\text{Tail}(\alpha)$ (§3.2) as a creativity diagnostic on existing benchmarks, without redesigning them or collecting new prompts.[3]

6.2 Across writing benchmarks, the association is positive

Setup. For each writing benchmark B, we measure tailness as the mean shared score $S_B(x)$ across its inputs. We examine the association between this mean and ranking agreement $\tau(B)$ (§4).

Results. Table A7 reports the results. Specifically, Arena-Hard Creative Writing, the most-tail benchmark, reaches $\tau = 0.37$ ($p = 0.029$), whereas Alpaca, near the low-novelty end, reaches

[3]One additional finding is that the relationship need not be strictly monotonic at the extreme tail: in the two main examples, agreement peaks before the final bin. Across the broader results in §E, however, agreement only occasionally peaks before the final bin, providing no evidence of a systematic cutoff.

$\tau = 0.16$ and does not significantly agree with the human ranking ($p = 0.37$). Across the eight writing benchmarks, we observe a positive but weaker association (Spearman $\rho = 0.26$). [4]

Interpretation and scope. Together, these results support tailness as a creativity-relevant indicator. We do not seek to isolate creativity from general capability: quality partly depends on general capability, and the LM Arena creative-writing ranking is strongly coupled with the overall ranking (§B.2). The observed alignment therefore supports a creativity-relevant signal, without establishing a capability-orthogonal factor.

7 Diagnosing Creativity-Promoting Methods

Our framework motivates examining whether creativity-promoting methods improve quality in low-probability regions of P_{ref}. We use $q@\text{Tail}(\alpha)$ (§6.1) to assess this aspect of their behavior on existing benchmarks.

Table 1: **Diagnosing creativity-promoting methods.** We report overall quality and novelty, together with Quality at Tail ($q@\text{Tail}(\alpha{=}25\%)$) and Novelty at Peak. Among the evaluated methods, only min-p sampling improves tail quality over the vanilla model, but it reduces both novelty metrics. Most methods increase peak novelty but have lower tail quality than the vanilla model. Results are averaged across 4 representative models and benchmarks. Subscripts indicate relative percentage changes from the vanilla models; green denotes increases and red denotes decreases. Bold marks the vanilla model. Full details are provided in §D.

Method	Quality Overall	Quality at Tail ($q@\text{Tail}$)	Novelty Overall	Novelty at Peak
Min-p ($p = 0.05$)	0.565$_{+11.7\%}$	0.558$_{+17.7\%}$	0.374$_{-16.3\%}$	0.159$_{-44.2\%}$
Vanilla Model	**0.506** 0%	**0.474** 0%	**0.447** 0%	**0.285** 0%
Convergent-Divergent Thinking	0.481$_{-4.96\%}$	0.442$_{-6.94\%}$	0.517$_{+15.5\%}$	0.322$_{+12.9\%}$
Temperature Scaling ($\tau \times 1.5$)	0.428$_{-15.4\%}$	0.441$_{-6.96\%}$	0.495$_{+10.7\%}$	0.305$_{+7.02\%}$
Persona	0.442$_{-12.6\%}$	0.421$_{-11.2\%}$	0.484$_{+8.28\%}$	0.447$_{+56.8\%}$
Branch-Solve-Merge	0.438$_{-13.4\%}$	0.408$_{-13.9\%}$	0.508$_{+13.6\%}$	0.566$_{+98.6\%}$
Self-consistency	0.403$_{-20.4\%}$	0.384$_{-19.0\%}$	0.528$_{+18.1\%}$	0.654$_{+129\%}$
Meta-prompting	0.387$_{-23.5\%}$	0.320$_{-32.5\%}$	0.396$_{-11.4\%}$	0.272$_{-4.56\%}$
Temperature Scaling ($\tau \times 2.0$)	0.222$_{-56.1\%}$	0.212$_{-55.3\%}$	0.610$_{+36.5\%}$	0.904$_{+217\%}$
CRPO-SFT	0.135$_{-73.3\%}$	0.130$_{-72.6\%}$	0.543$_{+21.5\%}$	0.701$_{+146\%}$
CRPO-DPO	0.097$_{-80.8\%}$	0.098$_{-79.3\%}$	0.560$_{+25.3\%}$	0.759$_{+166\%}$

Setup. We apply this diagnostic to a diverse set of creativity-promoting methods, including test-time interventions (e.g., min-p sampling (Nguyen et al., 2025), temperature scaling, persona (Xu et al., 2025; Pai et al., 2026), Convergent-Divergent Thinking (Tian et al., 2024), Branch-Solve-Merge (Saha et al., 2024), Self-consistency (Wang et al., 2023), Meta-prompting (Suzgun & Kalai, 2024)) and training-time methods (e.g., CrPO-SFT and -DPO (Ismayilzada et al., 2025)). We confirm that both quality and novelty are mentioned as the target aspects for the tasks in prompts. We include detailed creativity instructions or skills built by Chakrabarty et al. (2024a); Lu et al. (2024); Gómez-Rodríguez & Williams (2023) when the method is needed (e.g., meta-prompting). We evaluate each method on a mixture of all creativity benchmarks introduced in §4, spanning four model families; full experimental details are provided in §D.

Conceptually, a method can improve overall creativity along two complementary aspects: *i*) Improving quality on tail quadrants. When the input or output is novel, the primary challenge is not necessarily to make the output still more novel, but to maintain high output quality despite the unfamiliarity. This capability is measured by Quality at Tail, namely $q@\text{Tail}(\alpha)$; *ii*) Improving novelty on peak quadrants. A method may also elicit more novel outputs for familiar, high-probability inputs. We measure this capability using Novelty at Peak. Based on our proposition, we focus more on the former aspects and less on the latter, since we have shown that most creativity benchmarks generally do not concentrate in the peak quadrant.

Results. Table 1 shows a quality-novelty trade-off among the evaluated creativity-promoting methods. Most methods increase Novelty at Peak, but have lower Quality at Tail than the vanilla model. Conversely, min-p sampling improves Quality at Tail by 17.7%, but reduces Novelty Overall and Novelty at Peak by 16.3% and 44.2%, respectively. Thus, none of these methods improves both tail quality and peak novelty in the aggregate results. Additional sanity checks (§D) with other frontier

[4] We understand this as benchmarks differ simultaneously in terms of prompt source, length, format, scoring behavior, sample size, and overall difficulty, which may affect their model rankings independently of tailness. This heterogeneity motivates placing greater weight on the controlled within-benchmark analysis above.

models (Claude Opus 4.6 High) and quality-focused methods (CoT) show higher Quality at Tail than the vanilla model, indicating room for improvement in tail quality.

Overall, this diagnostic highlights a limitation of current creativity-promoting methods: gains in novelty often coincide with lower quality on uncommon inputs or outputs. The tail-quality gains achieved by min-p sampling and observed in the sanity checks point to an opportunity to develop methods that improve tail quality while preserving (or even improving novelty) as a potential direction for improving human-perceived creativity.

8 Creativity is not Identical to Difficulty

Prior work uses long-tail distributions to explain degraded LLM performance or construct harder evaluations (Kandpal et al., 2023; Razeghi et al., 2022; Mallen et al., 2023; Naik et al., 2022; Godbole & Jia, 2023). Since our framework emphasizes evaluation in tail regions, a natural question is whether selecting tail examples simply recovers difficult examples. Empirically, we compare *difficulty* and *novelty* splits on MATH and Cambridge CEFR. The overlap remains low (0.07 and 0.23), near random (Table 2); the same pattern holds under model-based difficulty on AIME and WritingBench. Conceptually, the two literatures study different axes. Difficulty work asks whether tail inputs are harder for a fixed model, while our hypothesis asks whether stronger tail performance better aligns with human judgments of creativity across models. See §B.1 for details.

Table 2: **Difficulty and novelty pick largely disjoint inputs.** *Diff.* and *Novelty* are per-split means (difficulty rated 1–5 for math, A2=1 to C2=5 for Cambridge). Both datasets are class-balanced.

Dataset	Split	Diff.	Novelty	Overlap
MATH (n=1,000)	difficulty	5.00	0.728	1.00
	novelty	2.38	0.965	**0.07**
	random	2.94	0.774	0.17
Cambridge CEFR (n=300)	difficulty	5.00	0.677	1.00
	novelty	3.62	0.835	**0.23**
	random	2.65	0.693	0.12

9 What Can We Learn?

From an evaluation perspective, our findings and taxonomy (§A) help assess whether a benchmark's design choices match its intended application and audience. For example, the reference distribution determines what counts as novel and can affect evaluation results: we use the Pile as a broad textual reference for an evaluation aimed at general users as LM Arena, whereas a benchmark serving scientific researchers may benefit from a reference emphasizing literature from the relevant field. Designers can also examine whether a benchmark covers the tail regions relevant to its intended use. Where multiple forms of creativity matter, curation, prescription, and constraint, individually or in combination, can help cover uncommon inputs, outputs, and their combinations. Metric choices such as judge type and granularity likewise shape which aspects of creativity are measured and involve trade-offs in cost and reproducibility. Specifying these choices and their rationale alongside the benchmark facilitates replication and comparison with subsequent work.

From a model-development perspective, our findings point to improving model performance specifically in tail regions as a potential pathway toward greater human-perceived creativity. Contemporaneous work has begun to investigate long-tail learning in post-training (Ramasubramanian et al., 2026) and pretraining (Badhe et al., 2026; Pouransari et al., 2026), through targeted data augmentation and algorithmic interventions. Whereas decoding or prompting methods can readily increase output rarity at inference time, they do not necessarily improve a model's ability to produce high-quality outputs in rare regimes. Strengthening this ability is likely more challenging, but may offer a more substantive path toward improving model creativity.

AI Use Statement

In this work, we used generative AI tools to help develop theoretical models and conceptual frameworks, propose and refine hypotheses, implement methods, assist with writing, and discuss interpretations of results. Additionally, we used these tools to create and modify scientific figures, format

references, and improve the writing quality of author-written drafts. The authors reviewed all AI-assisted contributions. We checked AI-assisted numerical summaries and descriptions against the experimental results and reviewed the theoretical arguments with their stated assumptions. We revised AI-generated suggestions where necessary to preserve the intended claims and avoid unsupported conclusions. We take responsibility for the final content of this work, including text, claims, code, and figures produced with the assistance of generative AI.

Ethical Considerations

To the best of our knowledge, this work does not introduce direct ethical harms. Our study focuses on evaluating LLM creativity benchmarks and proposing a diagnostic based on long-tail task performance for creativity evaluation, rather than deploying models in real-world decision-making settings.

Reproducibility Statement

§3 presents the formal definitions, assumptions, and theoretical derivation connecting novelty-weighted creativity to tail-conditioned quality. §4 specifies the reference corpus, novelty computation, shared prompt selection, quality metrics, and ranking-agreement procedure. §C lists the evaluated models and benchmarks with links to their upstream resources. For the diagnostic experiments, §D reports the model set, prompt mixture, scoring, and aggregation settings. Additional statistical tests and robustness analyses are provided in §E.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Teresa M. Amabile. *The Social Psychology of Creativity*. Springer Series in Social Psychology. Springer-Verlag, New York, 1983. ISBN 978-1-4612-5535-2.

Arena Intelligence. LLM Leaderboard - Best Text & Chat AI Models Compared: Text Arena ✍ Creative Writing. https://arena.ai/leaderboard/text/creative-writing, 2026. Accessed: May 21, 2026.

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. CS4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints, 2024.

Sanket Badhe, Deep Shah, and Nehal Kathrotia. Long-tail knowledge in large language models: Taxonomy, mechanisms, interventions and implications, 2026. URL https://arxiv.org/abs/2602.16201.

Minwook Bae and Hyounghun Kim. Collective critics for creative story generation. In *Proceedings of EMNLP*, 2024.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, 2026. URL https://arxiv.org/abs/2505.23281.

Frank Barron. The disposition toward originality. *Journal of Abnormal and Social Psychology*, 51(3): 478–485, 1955. doi: 10.1037/h0048073.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, London, 1994.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. *arXiv (Cornell University)*, 2025.

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.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024a. doi: 10.1145/3613904.3642731. URL https://dl.acm.org/doi/10.1145/3613904.3642731.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity, 2024b. URL https://arxiv.org/abs/2309.14556.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better llm-based evaluators through multi-agent debate. In *Proceedings of ICLR*, 2024.

Honghua Chen and Nai Ding. Probing the creativity of large language models: Can models produce divergent semantic association?, 2023. URL https://arxiv.org/abs/2310.11158.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

Xiaoyang Chen, Xinan Dai, Yu Du, Qian Feng, Naixu Guo, Tingshuo Gu, Yuting Gao, Yingyi Gao, Xudong Han, Xiang Jiang, Yilin Jin, Hongyi Lin, Shisheng Lin, Xiangnan Li, Yuante Li, Yixing Li, Zhentao Lai, Zilu Ma, Yingrong Peng, Jiacheng Qian, Hao-Yu Sun, Jianbo Sun, Zirui Wang, Siwei Wu, Zian Wang, Bin Xu, Jianghao Xu, Yiyang Yu, Zichuan Yang, Hongji Zha, and Ruichong Zhang. Deepmath-creative: A benchmark for evaluating mathematical creativity of large language models, 2025. URL https://arxiv.org/abs/2505.08744.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of COLING*, 2022.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

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

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. Scientific and creative analogies in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

Giorgio Franceschelli and Mirco Musolesi. Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41, 2024. doi: 10.1145/3664595.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Ameya Godbole and Robin Jia. Benchmarking long-tail generalization with likelihood splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 963–983, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.71. URL https://aclanthology.org/2023.findings-eacl.71/.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *Proceedings of ICLR*, 2023.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: a comprehensive evaluation of llms on creative writing, 2023. URL https://arxiv.org/abs/2310.08433.

Simeng Han, Howard Dai, Stephen Xia, Grant Zhang, Chen Liu, Lichang Chen, Hoang Huy Nguyen, Hongyuan Mei, Jiayuan Mao, and R. Thomas McCoy. Creativity or brute force? using brainteasers as a window into the problem-solving abilities of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *Proceedings of ICLR*, 2020.

Zhaoyi Joey Hou et al. CreativityPrism: A holistic evaluation framework for large language model creativity, 2025. URL https://arxiv.org/abs/2510.20091.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. Creativity in AI: Progresses and challenges. 2024. URL https://arxiv.org/abs/2410.17218.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke van der Plas, and Roger Beaty. Creative preference optimization, 2025. URL https://arxiv.org/abs/2505.14442.

Qi Jia, Xiang Yue, Tianyu Zheng, Jie Huang, and Bill Yuchen Lin. Simulbench: Evaluating language models with creative simulation tasks, 2024. URL https://arxiv.org/abs/2409.07641.

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

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of EMNLP*, 2021.

Deokgi Kim, Joonyoung Jo, Byung-Won On, and Ingyu Lee. Representation-to-creativity (R2C): Automated holistic scoring model for essay creativity. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5257–5275, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.292. URL https://aclanthology.org/2025.findings-naacl.292/.

Keuntae Kim, Eunhye Jeong, Sehyeon Lee, Seohee Yoon, and Yong Suk Choi. Claws:creativity detection for llm-generated solutions using attention window of sections, 2025b. URL https://arxiv.org/abs/2510.17921.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL*, 2016.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline, 2024. URL https://arxiv.org/abs/2406.11939.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. URL https://arxiv.org/abs/2101.00376.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

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

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens, 2025b. URL https://arxiv.org/abs/2401.17377.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*, 2023.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play, 2024. URL https://arxiv.org/abs/2405.06373.

Li-Chun Lu, Miri Liu, Pin-Chun Lu, Yufei Tian, Haoran Wen, and Hung-Yi Chen. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations, 2025a. URL https://arxiv.org/abs/2508.05470.

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 *International Conference on Learning Representations (ICLR)*, 2025b. 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 2025c. 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/.

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, 2025d. URL https://arxiv.org/abs/2407.09007.

Fangrui Lv, Kaixiong Gong, Jian Liang, Xinyu Pang, and Changshui Zhang. Subjective topic meets LLMs: Unleashing comprehensive, reflective and creative thinking through the negation of negation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12318–12341, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.686. URL https://aclanthology.org/2024.emnlp-main.686/.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation, 2025. URL https://arxiv.org/abs/2506.01937.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Guillermo Marco, Julio Gonzalo, M.Teresa Mateo-Girona, and Ramón Del Castillo Santos. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19654–19670. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1096. URL http://dx.doi.org/10.18653/v1/2024.emnlp-main.1096.

William Merrill, Noah A. Smith, and Yanai Elazar. Evaluating n-gram novelty of language models using rusty-DAWG. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14459–14473, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.800. URL https://aclanthology.org/2024.emnlp-main.800/.

Mistral AI. Mistral Small 3.2: An update to instruction following and tool use, June 2025. URL https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506. Model release on Hugging Face; instruct-tuning refinement of Mistral Small 3.1.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark J. F. Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and Shiva Taslimipoor. The Cambridge multiple-choice questions reading dataset. Technical report, Cambridge University Press & Assessment, 2023.

Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction, 2025. URL https://arxiv.org/abs/2504.15266.

Aakanksha Naik, Jill Lehman, and Carolyn Rosé. Adapting to the long tail: A meta-analysis of transfer learning research for language understanding tasks. *Transactions of the Association for Computational Linguistics*, 10:956–980, 2022. doi: 10.1162/tacl_a_00500. URL https://aclanthology.org/2022.tacl-1.56/.

Lakshmi Nair, Evana Gizzi, and Jivko Sinapov. Creative problem solving in large language and vision models - what would it take? 2024.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs, 2025. URL https://arxiv.org/abs/2407.01082.

NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning, 2025. URL https://arxiv.org/abs/2512.20848. Technical report.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, et al. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, P. Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. Measuring llm novelty as the frontier of original and high-quality output, 2025. URL https://arxiv.org/abs/2504.09389.

Samuel J Paech. Eq-bench creative writing benchmark v3. https://github.com/EQ-bench/creative-writing-bench, 2025a.

Samuel J. Paech. EQ-Bench creative writing benchmark v3. https://github.com/EQ-bench/creative-writing-bench, 2025b.

Tsung-Min Pai, Jui-I Wang, Li-Chun Lu, Shao-Hua Sun, Hung-Yi Lee, and Kai-Wei Chang. Billy: Steering large language models via merging persona vectors for creative generation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, 2002.

Hadi Pouransari, David Grangier, C Thomas, Michael Kirchhof, and Oncel Tuzel. Pretraining with hierarchical memories: separating long-tail and common knowledge, 2026. URL https://arxiv.org/abs/2510.02375.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of EMNLP*, 2023.

Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, and Heng Ji. Escapebench: Towards advancing creative intelligence of language model agents, 2025. URL https://arxiv.org/abs/2412.13549.

Ziliang Qiu and Renfen Hu. Deep associations, high creativity: A simple yet effective metric for evaluating large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10859–10872, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.550. URL https://aclanthology.org/2025.emnlp-main.550/.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Blog post.

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. Do llms agree on the creativity evaluation of alternative uses?, 2024. URL https://arxiv.org/abs/2411.15560.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Shrinivas Ramasubramanian, Daman Arora, Fahim Tajwar, Guanning Zeng, Qingyang Wu, Zhongzhu Zhou, Chenfeng Xu, Haiwen Feng, Yuda Song, Aarti Singh, Ruslan Salakhutdinov, J. Andrew Bagnell, Jeff Schneider, and Andrea Zanette. Tail-likelihood reinforcement learning, 2026. URL https://arxiv.org/abs/2609.02987.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP*, 2019.

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

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Liveideabench: Evaluating llms' divergent thinking for scientific idea generation with minimal context, 2025. URL https://arxiv.org/abs/2412.17596.

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.

Arkadiy Saakyan, Najoung Kim, Smaranda Muresan, and Tuhin Chakrabarty. Death of the novel(ty): Beyond n-gram novelty as a metric for textual creativity, 2026. URL https://arxiv.org/abs/2509.22641.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. Branch-solve-merge improves large language model evaluation and generation, 2024. URL https://arxiv.org/abs/2310.15123.

Morris I. Stein. Creativity and culture. *The Journal of Psychology*, 36(2):311–322, 1953. doi: 10.1080/00223980.1953.9712897.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 2020.

Luning Sun, Yuzhuo Yuan, Yuan Yao, Yanyan Li, Hao Zhang, Xing Xie, Xiting Wang, Fang Luo, and David Stillwell. Large language models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 57:101870, 2025. ISSN 1871-1871. doi: https://doi.org/10.1016/j.tsc.2025.101870. URL https://www.sciencedirect.com/science/article/pii/S1871187125001191.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

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

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Nathan Lambert, Pete Walsh, Pradeep Dasigi, et al. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of EACL*, 2021.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. MacGyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5303–5324, Mexico City, Mexico, 2024. Association for Computational Linguistics.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers?, 2025. URL https://arxiv.org/abs/2311.09682.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change, 2023. URL https://arxiv.org/abs/2206.10498.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language?, 2024. URL https://arxiv.org/abs/2305.10037.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

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

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing, 2025. URL https://arxiv.org/abs/2503.05244.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expertprompting: Instructing large language models to be distinguished experts, 2025. URL https://arxiv.org/abs/2305.14688.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Infini-gram mini: Exact n-gram search at the internet scale with fm-index, 2026a. URL https://arxiv.org/abs/2506.12229.

Jundong Xu, Hao Fei, Yuhui Zhang, Liangming Pan, Qijun Huang, Qian Liu, Preslav Nakov, Min-Yen Kan, William Yang Wang, Mong-Li Lee, and Wynne Hsu. Muslr: Multimodal symbolic logical reasoning, 2026b. URL https://arxiv.org/abs/2509.25851.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Chen Yang, Guangyue Peng, Jiaying Zhu, Ran Le, Ruixiang Feng, Tao Zhang, Wei Ruan, et al. Nanbeige4-3B technical report: Exploring the frontier of small language models, 2025b. URL https://arxiv.org/abs/2512.06266.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. Assessing the creativity of LLMs in proposing novel solutions to mathematical problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25687–25696, 2025. doi: 10.1609/aaai.v39i24.34760.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR*, 2020.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. In *Conference on Language Modeling (COLM)*, 2025a. URL https://openreview.net/forum?id=XZm1ekzERf.

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

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. Assessing and understanding creativity in large language models. *Mach. Intell. Res.*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of NeurIPS*, 2023.

A TAXONOMY

Our taxonomy organizes LLM creativity evaluations into Task and Metric dimensions. Each dimension distinguishes design choices in prior work, helping compare evaluation settings, identify potential sources of divergent findings, and position new evaluations.

A.1 TASK

In this paper, a task refers to a well - defined type of problem or set of instructions that specifies an intended goal, which includes creativity and admits a space of valid outputs. A task is defined at the level of objective and problem type rather than the level of specific prompt attributes and surface prompt realization. Variations in tone, theme, wording, or formatting are treated as different prompt realizations of the same task, whereas changes in the core objective (e.g., writing versus problem solving) or knowledge domain (e.g., language versus mathematics) constitute different tasks.

Objective. The objective defines the primary functional goal the model is expected to achieve. It determines the intended form of successful outputs and the criteria under which they are considered acceptable or valuable.

Domain. The domain specifies the type of knowledge and competencies required to perform the task. It characterizes the general capability space within which valid solutions would be constructed.

Creativity Elicitation Operation. This dimension characterizes the operation through which creativity is induced in the task. Creativity may be elicited by an objective or instruction that directly demands it, by posing inputs that are themselves rare or unusual such that standard heuristics no longer suffice, or by imposing explicit constraints that block the most typical outputs and push the model away from its default response.

Targeted Level. The targeted level specifies where creativity is expected to manifest within the output - whether primarily in surface expression, semantic content, or underlying solution strategies and conceptual structures.

A.2 METRIC

In this paper, a metric is a concrete algorithmic procedure that takes the model output to one prompt from one decoding pass and produces computational quantitative evaluative signals for analysis or comparison. A metric operationalizes evaluation but it does not define the task itself. The task - level results are averaged metrics across all prompts within the task.

Type. This dimension specifies how creativity is operationalized for measurement at a high level. It distinguishes between approaches that treat creativity as a single construct and those that decompose it into theoretically motivated components evaluated separately.

Granularity. Granularity defines the unit at which evaluation is computed, thereby determining the structural scale at which variation and novelty are assessed.

Reference Source. The reference source specifies the distribution, knowledge base, or population relative to which outputs are evaluated. It determines the comparative baseline that grounds judgments.

Output. This dimension specifies which portion of the model's generation is subject to evaluation, namely whether assessment considers only the final output or also includes intermediate artifacts - reasoning traces, plans, drafts, or actions - produced during task execution.

Judge. The judge specifies the mechanism responsible for producing the evaluative signal.

Table A1: **Taxonomy of Creativity Evaluation**, organized into Task and Metric.

Dimension	Category	Description
		TASK
Objective	Writing	Tasks whose primary objective is to produce textual artifacts. *Examples:* Story (Chakrabarty et al., 2024b; Atmakuru et al., 2024; Lu et al., 2025b), poetry (West & Potts, 2025), essay (Kim et al., 2025a), and head-to-head comparison against a professional fiction author (Marco et al., 2024).
	Problem Solving	Tasks whose primary objective is to reach a correct or valid solution. *Examples:* Math (Ye et al., 2025; Han et al., 2025; Chen et al., 2025), physical / multimodal (Tian et al., 2025; Nair et al., 2024).
	Ideation	Tasks whose primary objective is to propose ideas or generate candidate solutions or associations. *Examples:* Brainstorming (Lv et al., 2024; Ruan et al., 2025), association tasks (Chen & Ding, 2023; Qiu & Hu, 2025), TTCT-style divergent-thinking suites (Zhao et al., 2025; Sun et al., 2025), and analogy generation (Czinczoll et al., 2022).
Domain	Commonsense	Tasks grounded in everyday knowledge and informal reasoning. *Examples:* Commonsense QA (Talmor et al., 2019), Physical Commonsense Reasoning (Bisk et al., 2019), PlanBench (Valmeekam et al., 2023), RiddleSense (Lin et al., 2021).
	Language	Tasks centered on natural language generation or understanding. *Examples:* SQuAD (Rajpurkar et al., 2016), GLUE (Wang et al., 2019), WritingBench (Wu et al., 2025), Creative Writing v3 (Paech, 2025a).
	Math	Tasks involving mathematical reasoning, QA, computation, or problem solving. *Examples:* GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), AIME (Balunović et al., 2026), DeepMath-Creative (Chen et al., 2025).
	Computer (CS)	Tasks requiring computer science knowledge or programming skills. *Examples:* HumanEval (Chen et al., 2021), SWE-Bench (Jimenez et al., 2024).
	Science	Tasks requiring scientific explanation, hypothesis generation, or reasoning. *Examples:* GPQA (Rein et al., 2023), AstaBench (Bragg et al., 2025), scientific-analogy probes (Czinczoll et al., 2022).
	Synthetic	Tasks defined in non-natural-language or symbolic environments. *Examples:* Symbolic language or knowledge tasks (Wang et al., 2024; Nagarajan et al., 2025; Xu et al., 2026b).
	Others	Tasks not covered by the above domains. *Examples:* Game strategy (West & Potts, 2025; Qian et al., 2025), simulation / agentic environments (Jia et al., 2024), or design ideation tasks.
Creativity Elicitation Operation	Prescription	Creativity is demanded through the task objective or explicit instruction. *Examples:* Creative story writing (Chakrabarty et al., 2024b; Atmakuru et al., 2024; Lu et al., 2025b; West & Potts, 2025), brainstorming (Ruan et al., 2025), prompts requiring novel solutions (Rabeyah et al., 2024; Ye et al., 2025), alternative tool use (Rabeyah et al., 2024).
	Curation	Creativity is induced by posing inputs that are themselves rare or unusual, such that standard heuristics are insufficient to produce a valid output and elicit creativity to finish. *Examples:* Brainteasers (Han et al., 2025), riddles (Lin et al., 2021), creative writing premise, puns and wordplays.

Continued on next page

Table A1 continued from previous page

Dimension	Category	Description
	Constraint	Creativity is induced by adding explicit requirements or constraints to the input that eliminate the most typical outputs, eliciting the model to depart from its common output. *Examples:* Constrained story writing (Atmakuru et al., 2024), constrained code generation (Lu et al., 2025d).
Targeted Level	Surface (Form)	The task implicitly or explicitly demands creativity at the lexical or stylistic level. *Examples:* Stylistic rewriting tasks (Gómez-Rodríguez & Williams, 2023; West & Potts, 2025), wordplay.
	Meaning (Content)	The task implicitly or explicitly demands creativity at the level of new semantic combinations or content-level variation. *Examples:* Story writing (Gómez-Rodríguez & Williams, 2023; Chakrabarty et al., 2024b; Atmakuru et al., 2024; Lu et al., 2025b; West & Potts, 2025), divergent association tasks (Chen & Ding, 2023; Qiu & Hu, 2025), TTCT-style ideation (Zhao et al., 2025).
	Strategy (Concept)	The task implicitly or explicitly demands creativity at the level of novel ideas, concepts, or solution strategies. *Examples:* Creative problem-solving or algorithm design tasks (Tian et al., 2025; Ye et al., 2025; Han et al., 2025), creative simulation / agentic tasks (Jia et al., 2024).
		METRIC
Reference Source	Self Samples	Evaluation relative to the model's own generated samples at test time. *Examples:* Intra-model diversity (Jiang et al., 2025; Zhang et al., 2025b; Atmakuru et al., 2024).
	Training Corpus	Evaluation relative to (a proxy of) the model's training distribution. *Examples:* Web-scale corpus comparison via n-gram or infinigram lookup (Lu et al., 2025b; Padmakumar et al., 2025; Saakyan et al., 2026).
	Human Single	Evaluation from an individual human's judgment relative to their own knowledge. *Examples:* User rating (Stiennon et al., 2020; Ouyang et al., 2022).
	LLM Single	Evaluation from another LLM relative to its output or judgment on its own knowledge. *Examples:* LLM-as-a-judge, inter-model diversity metrics (Jiang et al., 2025).
	Human Community	Evaluation from a population of humans relative to their own knowledge. *Examples:* Crowd-sourced human evaluation (Chhun et al., 2022; Karpinska et al., 2021); story evaluation by multiple literary experts (Marco et al., 2024); group-level TTCT comparisons against human populations (Sun et al., 2025).
	LLM Community	Evaluation from a population of other LLMs relative to their output or judgment on their own knowledge. *Examples:* Multi-model ensembling and multi-agent evaluation (Verga et al., 2024; Chan et al., 2024); multi-LLM critic ensembles for creative writing (Bae & Kim, 2024).
	Human Historical	Evaluation relative to historical human artifacts. *Examples:* This is an imaginary setting; approximations include internet-scale or library-scale databases.
Output	Process	Evaluation considers intermediate steps, plans, drafts, or actions produced during task execution, rather than or in addition to the final output. *Examples:* Reasoning trace (Han et al., 2025), iterative drafting process (Atmakuru et al., 2024), separation of direction- vs. process-accuracy for mathematical reasoning (Chen et al., 2025).
	Outcome	Evaluation considers only the final output of the model. *Examples:* Final answer (Lu et al., 2025b; Tian et al., 2025; Zhang et al., 2025b), final story.

Continued on next page

Table A1 continued from previous page

Dimension	Category	Description
Type	Monolithic	Creativity is treated as a unitary construct and measured directly. *Examples:* Direct human or LLM creativity ratings (Gómez-Rodríguez & Williams, 2023; Kim et al., 2025a; Cao et al., 2026).
	Compositional	Creativity is defined through explicit theoretical components (e.g., novelty and value), each operationalized independently. *Note*: We only discuss the novelty component in this research and consider the value component out of our scope. In experiments, we apply the most widely used metrics for the value component (e.g., reward for writing, accuracy for reasoning). *Examples:* Novelty $\times$ quality decompositions (Zhang et al., 2025b; Padmakumar et al., 2025); TTCT four-factor decomposition (Fluency / Flexibility / Originality / Elaboration) (Zhao et al., 2025); convergent $\times$ divergent decomposition for code (Lu et al., 2025d).
Granularity	Token	Evaluation at the level of individual tokens. *Examples:* Token probability or entropy (Holtzman et al., 2020); token-level attention-window scoring for creativity detection (Kim et al., 2025b).
	N-gram	Evaluation over contiguous token sequences. *Examples:* Distinct-n (Li et al., 2016); n-gram corpus-novelty / Creativity Index (Lu et al., 2025b); infinigram-based novelty (Saakyan et al., 2026).
	Sentence	Evaluation at the sentence level. *Examples:* Sentence embedding similarity (Reimers & Gurevych, 2019; Tevet & Berant, 2021).
	Step	Evaluation at the level of procedural or reasoning steps, often composed of multiple sentences. *Examples:* Step-wise entailment evaluation (Prasad et al., 2023; Golovneva et al., 2023).
	Paragraph	Evaluation over the paragraph level, which often is the entire generation. *Examples:* Story evaluation (Chakrabarty et al., 2024b; Atmakuru et al., 2024).
Note: Granularity defines the structural unit at which variation is assessed. Well-designed evaluations align this choice with the task's Targeted Level (e.g., token- or n-gram-level metrics for surface creativity, paragraph-level metrics for content or strategy creativity). Misalignment between granularity and targeted level is common in existing work and may contribute to inconsistent findings across benchmarks.		
Judge	Deterministic	Rule-based procedures without trainable parameters. *Examples:* BLEU, ROUGE, distinct-n (Papineni et al., 2002; Lin, 2004; Li et al., 2016).
	Learned	Trained models producing evaluative signals. *Examples:* Reward models (Ouyang et al., 2022); trained embedding-based scorers (Zhang et al., 2020); creativity-specific trained scoring models such as R2C and CLAWS (Kim et al., 2025a;b); fine-tuned creativity evaluators (Cao et al., 2026).
	LLM-as-a-Judge	Evaluation performed by prompting an LLM under a judging protocol. *Examples:* Rubric-based LLM judging (Zheng et al., 2023; Liu et al., 2023); multi-critic LLM ensembles for creative writing (Bae & Kim, 2024).
	Human	Evaluation performed by human annotators. *Examples:* Expert evaluation (Chakrabarty et al., 2024b) or crowd-sourced evaluation (Chhun et al., 2022).

B Detailed Discussion

B.1 Creativity is not Identical to Difficulty

A line of work uses the long tail of a reference distribution to construct harder evaluations or to explain degraded LLM performance. Entity frequency, term frequency, and entity popularity in the pretraining corpus (Kandpal et al., 2023; Razeghi et al., 2022; Mallen et al., 2023; Naik et al., 2022), and pretrained-LM likelihood (Godbole & Jia, 2023), all predict downstream LLM accuracy; e.g., likelihood-based test splits raise relative error rates by 33-93% on semantic parsing, NLI, and QA (Godbole & Jia, 2023). In all of these works, tail-ness is invoked to predict or explain difficulty. Our hypothesis frames creativity as performance on the long tail of P_{ref} via the same kind of operationalization, so a natural worry follows: are creativity evaluations simply rediscovering difficulty under a new name?

Setup. We test this directly. For each source benchmark, we construct three 20% subsets, a *difficulty* split, a *novelty* split (tailness under P_{ref}), and a *random* baseline, and measure pairwise overlap. We use two operationalizations of difficulty in parallel: *human-annotated* (from expert labels) and *model-based* (from aggregated model accuracy). For human-annotated difficulty we use two class-balanced datasets: MATH (Hendrycks et al., 2021a), a competition mathematics benchmark with five difficulty levels (Level 1–5), and the Cambridge CEFR exam (Mullooly et al., 2023), a reading comprehension corpus spanning proficiency levels A2–C2. The difficulty split contains all prompts at the highest human-annotated level.

Results. Across both operationalizations, the difficulty-novelty overlap is near or below the random baseline. As shown in Table 2. On MATH, the overlap is 0.07 against a random baseline of 0.17, which indicates the two splits are nearly disjoint. And the novelty split is spread across all five difficulty levels (34% drawn from Level 1). On Cambridge CEFR, the overlap is 0.23 against a random baseline of 0.12, which is a small positive correlation, but still largely distinct. Similarly, under model-based difficulty, the overlap between the difficulty subset (bottom-20% by the mean model performance) and the novelty subset sits at the random baseline (0.20 and 0.25 on AIME and WritingBench). And the model's performance on the novelty split is substantially higher (indicating less difficulty) than on the difficulty split.[5]

The two literatures are conceptually different. The difficulty literature focuses on the *input axis*: for a fixed model, performance on tail inputs is lower than on peak inputs, and tail-ness is invoked as one candidate cause alongside task complexity or domain knowledge. Our hypothesis focuses on the *model axis*: on a fixed tail input set, the model whose performance is higher is likely to be rated as more creative by humans. Both relations can hold simultaneously. Tail prompts can be harder for any individual model, *and* tail performance can still be the region where between-model differences track creative capability. The experiments in this paper concern the second relation.

B.2 Creativity and General Model Capability

A potential alternative explanation for our results is that tail performance primarily distinguishes generally capable models, rather than capturing a creativity-relevant capability. **We do not assume that creativity and general capability are orthogonal.** Indeed, the LM Arena creative-writing ranking is strongly correlated with its overall model ranking and with each capability-specific ranking we examined (Table A2).

This coupling is also consistent with the novelty-quality account in §3: creativity requires not only novelty but also quality, and quality necessarily depends in part on general language, reasoning, and instruction-following capabilities. A decomposition that removes all effects of general capability would therefore also remove part of the quality component of creativity.

Under this interpretation, the GPQA-Diamond result (Figure 5) is not treated as evidence that GPQA is itself a complete creativity benchmark. Rather, it supports the metric-level implication that

[5]We do not claim creativity is orthogonal to task-specific capability. The empirical splits partially overlap rather than being fully disjoint, and conceptually, researchers can (and do) build creativity tests on hard tasks and view part of the capability for solving hard problems as indicating creativity. The aim of this paper is not to identify what creativity is or to define its boundaries against other capabilities; it is to surface the shared implicit pattern in how existing creativity benchmarks operationalize their target.

Table A2: **LM Arena creative-writing rankings are strongly coupled with general model capability.** Spearman correlations compare the creative-writing model ranking with the overall and capability-specific LM Arena rankings.

Creative-writing ranking vs.	Spearman ρ	p
Overall	0.9443	3.92×10^{-9}
Instruction Following	0.8927	6.29×10^{-7}
Math	0.6739	2.17×10^{-3}
Coding	0.6347	4.66×10^{-3}

Table A3: Open-weight model properties tracking chronological release dates, architectural configurations, and performance profiles (absolute rank position alongside Elo rating estimation) on the LMSYS Chatbot Arena Creative Writing sub-category leaderboard.

Model Name	Date	Architecture (Size)	LM Arena Creative Writing (Rank / Elo)	Link
Qwen3.5-27B	Feb 2026	28B Dense	#26 (1367 $\pm$ 28)	Hugging Face
Gemma-3-27b-it	Mar 12, 2025	27B Dense	#34 (1349 $\pm$ 8)	Hugging Face
Qwen3.5-35B-A3B	Feb 2026	36B (3B MoE)	#39 (1339 $\pm$ 28)	Hugging Face
Gemma-3-12b-it	Mar 12, 2025	12B Dense	#42 (1334 $\pm$ 23)	Hugging Face
Qwen3-30B-A3B-Instruct-2507	Jul 29, 2025	30.5B (3.3B MoE)	#47 (1325 $\pm$ 11)	Hugging Face
Mistral-Small-3.2-24B-Instruct	Jun 20, 2025	24B Dense	#49 (1322 $\pm$ 12)	Hugging Face
Qwen3-32B	Apr 28, 2025	32.8B Dense	#57 (1307 $\pm$ 22)	Hugging Face
Gemma-3n-E4B-it	Jun 26, 2025	8B Raw / 4B Eff.	#62 (1301 $\pm$ 11)	Hugging Face
QwQ-32B	Mar 5, 2025	32B Dense	#64 (1297 $\pm$ 10)	Hugging Face
OLMo-3.1-32B-Instruct	Jan 13, 2026	32.2B Dense	#67 (1289 $\pm$ 15)	Hugging Face
Qwen3-30B-A3B	Apr 28, 2025	30B (3B MoE)	#69 (1288 $\pm$ 10)	Hugging Face
Gemma-3-4b-it	Mar 12, 2025	4B Dense	#77 (1277 $\pm$ 22)	Hugging Face
OLMo-3-32B-Think	2025	32B Dense	#81 (1267 $\pm$ 21)	Hugging Face
Nemotron-3-Nano-30B-A3B	Dec 15, 2025	31.6B (3.2B MoE)	#92 (1252 $\pm$ 13)	Hugging Face
OLMo-3.1-32B-Think	Jan 2026	32B Dense	#94 (1247 $\pm$ 18)	Hugging Face
GPT-OSS-20B	Aug 5, 2025	21B (3.6B MoE)	#98 (1236 $\pm$ 18)	Hugging Face
Phi-4	Dec 12, 2024	14B Dense	#105 (1212 $\pm$ 10)	Hugging Face
Llama-3.1-8B-Instruct	Jul 23, 2024	8B Dense	#119 (1179 $\pm$ 8)	Hugging Face
Llama-3.1-Nemotron-Nano-8B-v1	Mar 18, 2025	8B Dense	Unranked	Hugging Face
gemma-4-31B-it	Apr 2, 2026	30.7B Dense	Unranked	Hugging Face
Qwen3-8B	Apr 28, 2025	8B Dense	Unranked	Hugging Face
Qwen3-8B-Thinking	Apr 28, 2025	8B Dense	Unranked	Hugging Face
Qwen3-30B-A3B-Thinking	Jul 29, 2025	30B (3B MoE)	Unranked	Hugging Face
OLMo-3-7B-Instruct	Nov 20, 2025	7B Dense	Unranked	Hugging Face
OLMo-3-7B-Think	Nov 20, 2025	7B Dense	Unranked	Hugging Face
Nanbeige4-3B-Thinking	Dec 6, 2025	3B Dense	Unranked	Hugging Face

performance on rare input-output regions can carry a creativity-relevant signal even on a benchmark without explicit creativity framing. At the same time, the current experiments do not establish that this signal is specific to creativity or separable from general capability. We therefore treat $q@\mathrm{Tail}$ as a diagnostic approximation, rather than as an estimate of a disentangled latent creativity factor.

C Models and Benchmarks

This appendix summarizes the experimental assets used in §5–§8. We detail the evaluated model families (Table A3) and benchmark (Table A4) used during generation. These details are provided to improve reproducibility and to clarify how comparisons were standardized across diverse standard and creativity benchmarks.

D Diagnosing creativity-promoting methods settings

This appendix gives implementation details for the diagnostic experiment in §7.

Models. We use four open-weight reference checkpoints: Llama-3.1-8B-Instruct, Gemma-3n-E4B-it, OLMo-3-7B-Think, and Qwen3-8B. The two thinking models enabled thinking by default.

Table A4: Taxonomy of the 15 evaluation benchmarks, grouped by evaluation domain and scoring strategy. Benchmark names containing subset, sampling, filtering, or formatting qualifiers denote the exact evaluation variants used in our experiments; links point to the corresponding upstream dataset or official repository.

Domain	Benchmark	Type	Link
Writing	Arena-Hard Creative Writing	Rubric	GitHub
Writing	CS4	Rubric	GitHub
Writing	StoryGen	Rubric	Hugging Face
Writing	StoryGen-Px	Rubric	Hugging Face
Writing	LitBench-Test	RefSim	Hugging Face
Writing	WritingBench-EN100	Rubric	GitHub
Writing	Writing-Prompts-Curated-1k	Rubric	Hugging Face
Writing	Alpaca	Rubric	Hugging Face
Writing	TinyStories	RefSim	Hugging Face
Writing	BookMIA	RefSim	Hugging Face
Math	AIME-All	EM	Hugging Face
Math	CreativeMath	EM	GitHub
QA	CommonsenseQA (FF)	EM	Hugging Face
QA	Braingle-Verifiable	EM	Hugging Face
QA	MacGyver-Creative	Rubric	GitHub

Data and scoring. Each (model, method) cell is run on a 237-prompt creative mix drawn from four datasets: Arena-Hard Creative Writing, LitBench-test, MacGyver-Creative, and CreativeMath. Quality score from Skywork-Reward-V2-Llama-3.1-8B on the three reward-model datasets, with pass@1 on CreativeMath. We normalize per-prompt against each model's baseline run and aggregate into q@Tail(α=0.25) using the joint n-novelty.

Tail selection for the saved-output reanalysis. For each benchmark B, we compute

$$S_B(x) = \sqrt{n(x)\,\frac{1}{4}\sum_{m=1}^{4} n(y_m^0(x))},$$

where $y_m^0(x)$ is the saved baseline response from reference model m. We select the top 25% of prompts by $S_B(x)$, rounding the subset size up to the nearest integer. These prompt sets are shared across all models and held fixed across interventions.

Sanity check. To assess whether the diagnostic can capture gains in tail quality, we additionally evaluate CoT as a quality-oriented intervention and Claude Opus 4.6 High as a separate model comparison. CoT achieves a Quality at Tail score of 0.539, compared with 0.474 for the vanilla model, while Claude Opus 4.6 High achieves 0.610. These comparisons show that higher tail quality is attainable, although they do not establish joint improvements in tail quality and peak novelty.

E Additional Results

E.1 Creativity Benchmark Lies in the Tail

Detailed Statical Analysis A pooled comparison within the writing domain further supports this shift: both input novelty (Mann-Whitney $p < 0.001$; KS $D = 0.108$, $p < 0.001$) and output novelty (Mann-Whitney $p < 0.001$; KS $D = 0.186$, $p < 0.001$) differ significantly between creativity (n = 3,022) and standard (n = 6,175) examples, as does their joint distribution (Hotelling's T^2, $p < 0.001$). The overall pattern is stable across $n \in \{4, 5, 6\}$ and across model families.

E.2 Within-Benchmark Robustness and Additional Benchmarks

Section 6.1 presents the five-bin results for Arena-Hard Creative Writing and GPQA-Diamond. Here, we provide two complementary extensions. First, we increase the bin resolution from five quantile bins to ten to test whether the observed relationship is an artifact of a particular discretization. Second,

Table A5: **Ranking agreement at ten-bin resolution.** Each entry τ_i is Kendall's τ between the model ranking induced by performance within novelty bin i and the LM Arena creative-writing ranking. Bin 1 is the most-peak subset and bin 10 is the most-tail subset. The last two columns compare mean agreement over the peak half (bins 1–5) and tail half (bins 6–10). Agreement increases substantially in the tail half for both Arena-Hard Creative Writing and GPQA-Diamond, although the relationship is not strictly monotonic across adjacent bins.

Benchmark	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	$\text{Mean}_{1:5}$	$\text{Mean}_{6:10}$
Arena-Hard	0.000	0.091	0.182	0.000	0.030	0.061	0.273	**0.485**	0.333	0.364	0.061	**0.303**
GPQA-Diamond	0.117	0.064	0.097	0.198	0.085	0.180	0.425	**0.451**	0.178	0.266	0.112	**0.300**

we report the same within-benchmark analysis on the remaining benchmarks for which both task performance and LM Arena creative-writing ratings are available.

Unless otherwise stated, examples are ordered by joint 4-gram novelty and divided into equal-count bins. The conclusion also holds across different metrics, including Creativity Index and $\text{novelty}_{n=5,6}$; the latter serve as supplementary robustness analyses. Within each bin, models are ranked by their mean task performance, and agreement with the LM Arena creative-writing ranking is measured using Kendall's τ. Bin 1 denotes the most-peak subset and the highest-numbered bin denotes the most-tail subset.

E.2.1 Robustness to the Number of Bins

Figure A1 shows the ten-bin partition for Arena-Hard Creative Writing and GPQA-Diamond. The positive association persists at this finer resolution. For Arena-Hard, agreement rises from approximately zero in the earliest bins to $\tau = 0.27$–0.49 in bins 7–10. For GPQA-Diamond, bins 7 and 8 reach $\tau = 0.43$ and $\tau = 0.45$, respectively, compared with $\tau = 0.06$–0.20 for most earlier bins. Across the ten bins, the Spearman association between bin-level joint novelty and ranking agreement is positive on both benchmarks ($\rho = 0.661$–0.855, $p = 0.0016$–0.0376).

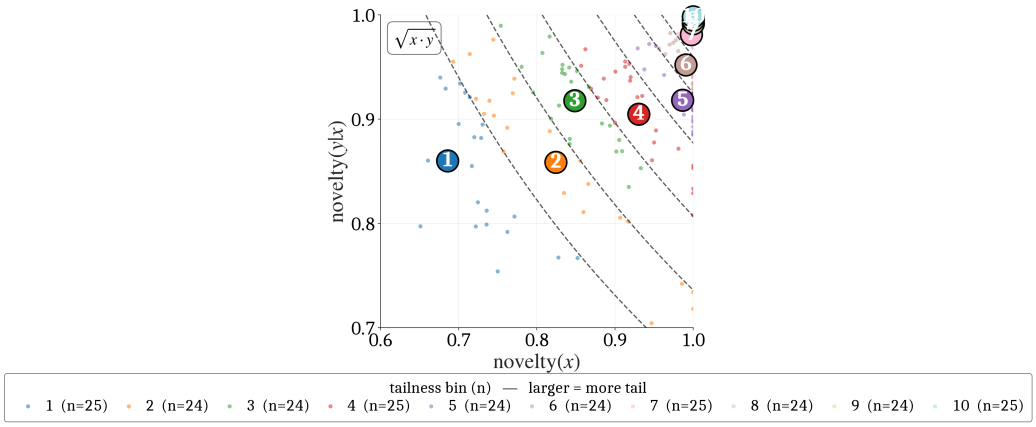
The relationship is not strictly monotonic at every adjacent bin. In particular, the most-tail bin need not attain the largest agreement. As discussed in the main text, extreme-tail subsets can exhibit reduced between-model performance variance, which weakens the ranking signal even when their novelty is higher.

E.2.2 Results on Additional Benchmarks

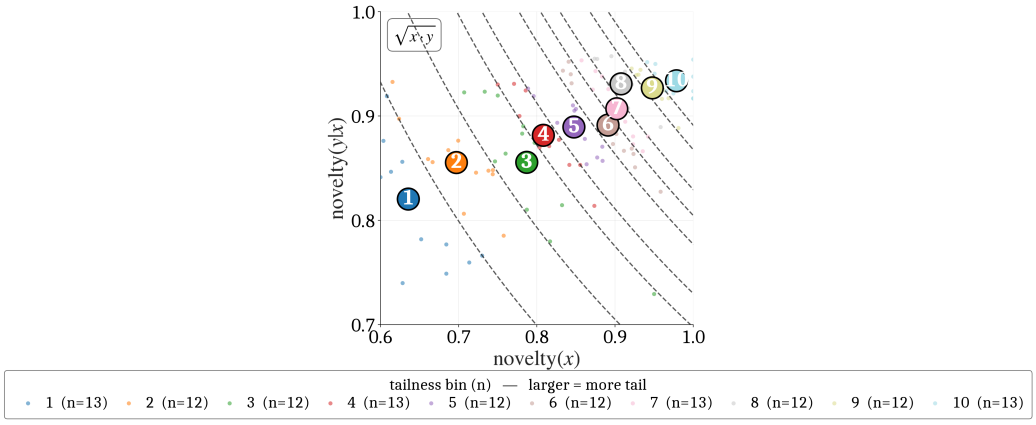
Figures A2 extend the five-bin analysis to representative additional benchmarks. Table A6 reports the complete bin-level ranking agreement for all benchmarks included in this analysis.

The results vary across benchmarks, while most of them present non-negative correlation. AIME-All, BookMIA, and Braingle show higher agreement in their most-tail bins than in their most-peak bins, consistent with the direction observed for Arena-Hard and GPQA-Diamond. Alpaca shows essentially no endpoint difference, and StoryGen and TinyStories remain positively aligned with the human creative-writing ranking across all bins.

These results delimit the scope of the within-benchmark finding. Tailness is not a universally monotonic predictor of agreement for every benchmark or task domain. The relationship is strongest when the quality function and the evaluated tail region reflect capabilities represented in the human reference ranking. Domain mismatch, limited between-model performance variation within individual bins, and differences in benchmark scoring can weaken or reverse the observed association. We therefore interpret tail-conditioned performance as a creativity-relevant indicator under a domain-aligned reference ranking, rather than as a domain-invariant measure of creativity.



(a) Arena-Hard Creative Writing



(b) GPQA-Diamond

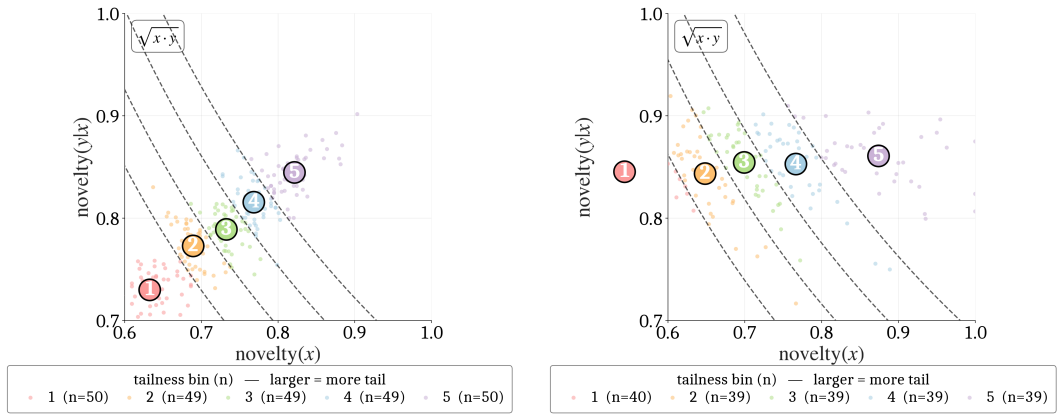
Figure A1: **Input-output novelty partitions at ten-bin resolution.** Input-output pairs are divided into equal-count bins using the geometric mean of input and output 4-gram novelty. Numbered circles mark bin centroids, dashed curves mark equal joint-novelty boundaries, and larger bin numbers indicate more-tail examples. The corresponding bin-level agreement with the human creative-writing ranking is reported in Table A5.

F Additional Discussion and Limitations

F.1 Limitations

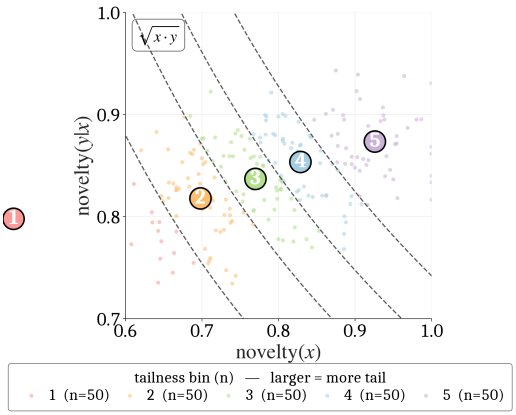
This work studies creativity evaluation through the lens of long-tail performance, but our operationalization is not the only possible one. We approximate the reference distribution using n-gram statistics from the Pile. These statistics capture surface-level rarity but do not directly measure semantic, conceptual, or culturally grounded novelty. Other reference distributions, such as human-authored creative corpora, model self-samples, or domain-specific corpora, may lead to different estimates of tailness.

Our empirical validation is also limited by the available benchmarks and human-preference references. Although we evaluate across multiple task domains and models, the strongest human-alignment analysis relies on LM Arena creative-writing rankings, which may not fully represent human creativity judgments in mathematics, scientific reasoning, design, or other creative domains. In addition, benchmark-level correlations are noisy, suggesting that tailness is an important factor but not a complete explanation of creativity evaluation.

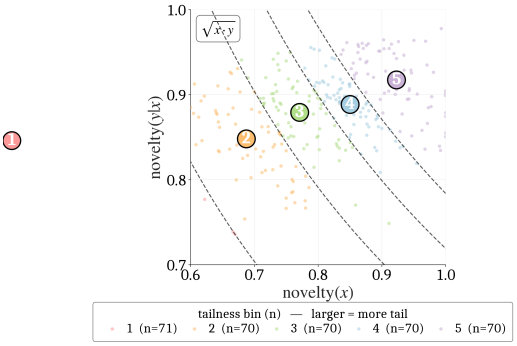


(a) BookMIA

(b) StoryGen



(c) AIME-All



(d) Braingle

Figure A2: **Within-benchmark results on additional representative benchmarks.** BookMIA, AIME-All, and Braingle exhibit higher agreement with the LM Arena creative-writing ranking in their most-tail bins than in their most-peak bins. StoryGen illustrates that the relationship need not be monotonic on every benchmark. Complete results, including null and negative cases, are reported in Table A6.

Table A6: **Ranking agreement across joint-novelty bins on additional benchmarks.** Each entry is Kendall's τ between the performance-induced model ranking within a bin and the LM Arena creative-writing ranking. Bins are formed using geometric-mean joint 4-gram novelty; bin 1 is most peak and bin 5 is most tail. $\Delta\tau_{5-1}$ is the endpoint difference between the most-tail and most-peak bins. Positive and null results are reported together to avoid selecting benchmarks based on the direction of the observed effect.

Benchmark	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	$\Delta\tau_{5-1}$
AIME-All	0.131	0.095	0.135	0.089	0.290	+0.159
Alpaca	0.091	0.182	0.030	0.091	0.091	0.000
BookMIA	0.273	0.182	0.273	0.333	0.364	+0.091
Braingle	0.155	0.243	0.231	0.137	0.290	+0.135
StoryGen	0.364	0.455	0.333	0.333	0.273	−0.091
TinyStories	0.303	0.212	0.273	0.303	0.212	−0.091

Our evaluation of creativity-promoting methods is further limited to open models with 7-8B parameters due to resource constraints. Methods with more complex instructions, such as meta-prompting, may behave differently on larger models or frontier closed-source models.

Table A7: **Across writing benchmarks, tailness shows a positive but inconclusive association with human-ranking agreement.** Tailness averages the shared joint 4-gram novelty score across inputs. For each input, output novelty is first averaged across the 26 evaluated models and then combined with input novelty using the geometric mean. τ is Kendall's τ between each benchmark's model ranking and the LM Arena ELO ranking. Across the eight benchmarks, the association between tailness and τ is positive but weak and not statistically significant (Spearman $\rho = 0.26$, $p = 0.53$). p_{Kendall} is the p-value for testing the null hypothesis of no rank association between each benchmark's model ranking and the LM Arena ranking; * marks $p_{\text{Kendall}} < 0.05$. These row-level Kendall tests are separate from the across-benchmark Spearman test.

Benchmark	N	Tailness	τ vs Arena	p_{Kendall}
Arena-Hard	250	0.646	0.368*	0.029
CS4-mid	50	0.357	0.286	0.093
CS4-light	50	0.310	0.357*	0.034
BookMIA	247	0.274	0.497*	0.002
StoryGen	200	0.255	0.485*	0.003
Alpaca	384	0.244	0.158	0.368
CS4-standard	50	0.221	0.310	0.068
TinyStories	220	0.176	0.333*	0.049

Finally, $q@\text{Tail}$ is intended as a diagnostic approximation rather than a replacement for carefully designed creativity benchmarks. The LM Arena creative-writing ranking is itself strongly coupled with overall model capability (§B.2), so our experiments cannot attribute the predictive signal of tail performance exclusively to creativity. High tail performance may also reflect general capability, task-specific skills, domain knowledge, instruction following, or robustness to rare inputs. We therefore interpret $q@\text{Tail}$ as surfacing creativity-relevant behavior under a declared reference distribution and quality function, rather than as measuring a capability-orthogonal creativity factor. Future work should examine whether the signal persists among capability-matched models and under broader population-level human references.

F.2 Scope of Human Validation

Our external validation targets population-level perceived creativity: whether rankings induced by tail-conditioned performance agree with preferences aggregated over a broad human population. We use the LM Arena creative-writing ranking because it provides large-scale pairwise preferences and substantial coverage of our evaluated models. An in-house evaluation feasible within our resources would cover far fewer annotators, prompts, and models, and would not provide a comparably stable population-level ranking.

This choice limits our conclusions. LM Arena users are self-selected, its preferences are holistic rather than specific to creativity, and its creative-writing category covers only one domain. Our results therefore do not establish the same relationship for expert judgments of mathematical, scientific, or design creativity. Large-scale, domain-specific human validation remains future work.

The human ranking and P_{ref} serve distinct roles: r_{Arena} provides an external target for human-perceived creativity, whereas P_{ref} determines which input–output pairs are considered novel.

F.3 Scope of the Reference Distribution

We instantiate P_{ref} using n-gram statistics over the Pile, providing a fixed, model-independent reference across benchmarks and models. This shared reference is appropriate for comparison against a population-level human ranking. A model-specific reference instead measures a model's creative range relative to itself, which is a different evaluation target.

As a robustness check, we used base-model perplexity as an alternative reference. Its aggregate relationship with the LM Arena creative-writing ranking is near zero and inconsistent in direction across models and datasets. This result does not establish that n-gram novelty is uniquely correct; it only shows that perplexity does not provide a stable alternative for the present evaluation.

Our n-gram operationalization primarily captures surface-level lexical rarity. It does not directly measure semantic, conceptual, or culturally grounded novelty, and may assign high novelty to unusual wording that is not conceptually creative. If used as an optimization target, it may therefore encourage surface-level novelty.

Embedding-based density is a natural alternative, but estimating it at pretraining-corpus scale requires embedding, storing, and indexing a large text collection. Its result would also depend on the corpus sample, embedding model, chunking strategy, and distance function. These extensions are outside the scope of this work. Our findings should therefore be interpreted under the present surface distributional notion of novelty.

F.4 Dependence on the Quality Estimator

We use the task-specific quality estimators described in §4. Reward models may exhibit architecture, length, style, and domain biases. Absolute scores, model rankings, and $q@\mathrm{Tail}$ may therefore depend on the chosen evaluator. The framework permits different task-appropriate choices of q, but this does not imply that our empirical findings are invariant to the evaluator. Robustness across reward models, LLM judges, and human-calibrated quality measures remains future work.

F.5 Interpretation and Cost of $q@\mathrm{Tail}$

The shared score $S_B(x)$ fixes novelty weights across models for comparison on the same prompts. This is a benchmark-level operationalization: $C_{\pi,B}$ need not equal the expected novelty–quality product of a model's own responses. The integral identity holds for this shared-weight score; the empirical analysis evaluates $q@\mathrm{Tail}$ through agreement with human rankings.

$q@\mathrm{Tail}$ follows a filter-then-score procedure. Tail membership is determined using the novelty of the input and the concatenated input–output pairs, with output novelty aggregated across the evaluated models. The quality function q is then applied only to examples selected by this joint novelty criterion. Output novelty is therefore incorporated through tail selection, although q itself measures only quality and does not explicitly compute a novelty–quality product.

Consequently, a conventional but high-quality model response may still receive a high score on an example selected as tail. We therefore present $q@\mathrm{Tail}$ as a diagnostic indicator, not as a complete replacement for creativity benchmarks or separate measurements of output novelty and quality.

The n-gram index over the Pile has a nontrivial construction and storage cost, but it is built once and reused through Infini-Gram-Mini. Applying $q@\mathrm{Tail}$ to a new benchmark therefore has lower marginal cost than constructing a new creativity benchmark and collecting large-scale human preferences. It is not cost-free: model generation and quality evaluation still require computation.