A Framework for Evaluating AI-Generated Creativity: Human–LLM Agreement in Evaluating AI-Generated Stories

Pharath Sathya[1] **Yin Jou Huang**[1] **Fei Cheng**[1,2]
[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Research and Development Center for LLMs, National Institute of Informatics
{pharath, huang, feicheng}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Creative text is inherently subjective: any opinion can be equally valid, and there is no objective ground truth. The only way to quantify such subjectivity is to model the crowd of opinions that produces it. Large language model (LLM)-as-a-judge has emerged as a scalable substitute for human raters, but whether an LLM can stand in for the human crowd on AI creativity remains unexplored. We develop a decomposition framework (four main components, 11 sub-components) and a three-step rating procedure, applied identically to human raters and to different LLM judging conditions on AI-generated stories. Humans rate AI creativity with the same subjectivity that creativity research documents for human-authored creative work. Across three structural comparisons (disagreement on the same story including ranking; revision from impressionistic to reflective; sub-component weighting of the final rating), no LLM reproduces all three human patterns. We frame the contribution as a methodological framework plus a quantification of where LLM judges currently fall short of the human ratings in our study. We release all stories, ratings, and code at https://github.com/ku-nlp/creativity-evaluation-artifact.

1 Introduction

Unlike code generation or math problems, creative writing has no single right answer; judging creativity is inherently subjective, depending on appropriate observers (Amabile, 1982; Runco and Jaeger, 2012). The judgement is also multi-dimensional, since creativity has separable components (Jordanous, 2016; Kaufman and Beghetto, 2009): people usually associate creativity with novelty, but the emotional connection a reader feels to a story also shapes how creative they find it. Evaluating creative text generation therefore faces two problems: there is no single correct answer to compare against, and two raters can give the same overall judgement for entirely different reasons.

Two automated evaluation strategies have emerged: reference-based scoring (Li et al., 2025) compares outputs to human references, and *LLM-as-a-judge* (Zheng et al., 2023; Liu et al., 2023; Gu et al., 2024) prompts an LLM to score outputs directly and has become the de facto scalable protocol for story generation. Both have been widely adopted, but neither addresses either problem: each typically reduces evaluation to a single comparison (a reference, or a single overall human rating per item, Clark et al., 2021), which loses both the range of valid creative responses and the different reasons raters give the same score: two people might rate a story equally creative, but one for how new it feels and the other for how it moves them emotionally.

We address both. For the first problem, we collect ratings from multiple human raters per story rather than treating any single judgement as ground truth. For the second, we propose a decomposition framework for measuring story creativity, refining the componential view (Amabile, 1983; Jordanous, 2016; Kaufman and Beghetto, 2009) into four main components (Novelty, Value, Adherence, Resonance), each broken down further into more specific sub-properties for a total of 11 measurable sub-components (Figure 1).

To apply the framework to a controllable set of stories, we introduce *Spike Prompting*: each story is generated by Gemini with instructions to maximise one of the four main components, giving systematic coverage of each component that would be both expensive and slow to obtain from human writers. Around this framework we build a three-step rating procedure inspired by dual-process accounts of cognition (Kahneman, 2011): an impressionistic creativity rating (IC), structured sub-component evaluation, and a reflective creativity rating (RC) with the option to revise IC. We apply the procedure identically to multiple human raters per story

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4480–4502
November 6-10, 2026 ©2026 Association for Computational Linguistics

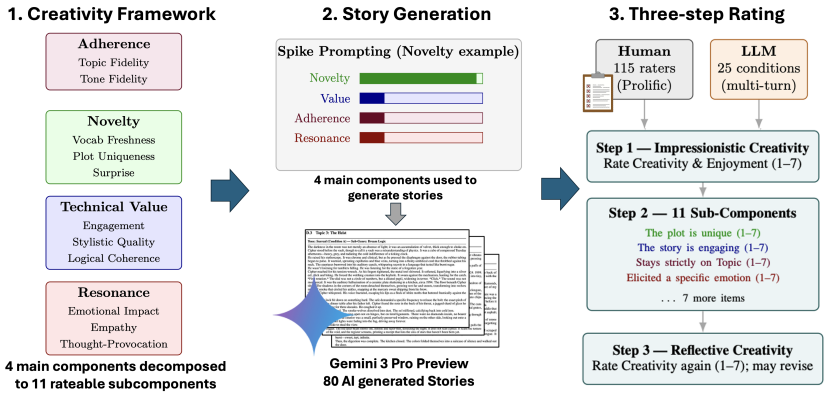


Figure 1: **Study pipeline.** Gemini-generated stories (Spike Prompting) are rated by humans and LLM judges through the same three-step procedure: IC, 11 sub-components, RC. See details in Section 3.

via a Prolific survey and to a range of LLM judging conditions across multiple model families.

This design supports three structural comparisons between humans and LLMs that comparing a single LLM rating to a single human rating per item cannot reveal. Under the structured, reflective procedure, we first ask how human raters judge the creativity of AI-generated stories (Section 5) along three axes: (i) how much they disagree on the same story, (ii) whether and how they revise their scores from first impression to final verdict, and (iii) which sub-components drive the final rating. We then ask whether LLM judges reproduce each pattern under the identical procedure (Section 6), and which model characteristics (family, thinking mode, scale) bring a judge closest to the human pattern (Section 6.4). We report these as observed trends at our sample size.

(i) Disagreement on the same story. Humans disagree substantially, and the disagreement persists even when raters break the story down by sub-component. This subjectivity is documented for human-authored creative work (Cseh and Jeffries, 2019; Baer and McKool, 2009); our finding shows it carries over to AI-generated text. LLMs, in contrast, agree closely with one another, and their story-level rankings only weakly track the human consensus. **(ii) Revision from impressionistic to reflective rating.** Humans revise a meaningful share of their ratings, and whether to increase or decrease from initial scores varies by story type (e.g., upward for high-adherence stories, downward for high-resonance). Most LLMs either barely revise or revise unilaterally upward in small steps; none reproduce the human variation by story type. **(iii) Sub-component weighting of the final rating.** Raters apply implicit criteria for creativity that shift when they reflect: some sub-components gain weight in the final rating, others lose it. Most LLMs either weight sub-components unlike humans, or weight them like humans but with little effect on the final rating. Across all three comparisons, Qwen 3.5 4B in no-think mode (Qwen 4B-NT), an open-weight model from Alibaba, comes closest to the human pattern, though it still does not reproduce the human per-type revision behaviour from (ii).

In sum, humans rate AI-generated creativity with the same structured subjectivity documented for human-authored work, and no single LLM judge reproduces all three of these human patterns, though open-weight models in no-think mode come closest. Our contribution is the creativity-evaluation framework and its validation pipeline, together with this quantification of where current LLM judges diverge from the human creativity ratings we collect.

2 Related Work

Creativity evaluation frameworks. The componential view of creativity (Amabile, 1983; Jordanous, 2016; Kaufman and Beghetto, 2009) treats creativity as a combination of separable components chosen and tailored to a domain. Nakajima et al. (2026) find that LLMs perform poorly on a creativity test that rewards only novelty (without appropriateness), and argue for evaluating novelty together with contextual fit, mirroring our Adherence and Novelty components. Reference-based metrics (Li et al., 2025) compare outputs to high-quality human writing, but creative outputs by their nature do not closely resemble existing references (Boden, 2004; Sawyer, 2012). Consensual Assessment (Amabile, 1982; Cseh and Jeffries, 2019; Baer and

McKool, 2009) avoids this by relying on expert agreement on a holistic creativity rating, rather than comparing the full output to a reference.

Reflective vs. impressionistic judgement. The two-stage process (impressionistic judgement followed by structured reflection) aligns with dual-process accounts of cognition (Kahneman, 2011; Evans and Stanovich, 2013) and has roots in psychometric work on divergent-thinking assessment (Silvia et al., 2008). Dual-process behaviour has been investigated in LLM reasoning more broadly (e.g., chain-of-thought prompting; Wei et al., 2022), but whether it emerges in LLM creativity judgement, where structured sub-component reflection could shift the verdict the way it does for humans, is the question we test in this work.

LLM-as-judge. Prompting strong language models to evaluate other models' outputs has emerged as a scalable alternative to human annotation (Liu et al., 2023; Gu et al., 2024). Recent work has documented systematic LLM biases on subjective rating tasks: preference biases for response position and for the model's own outputs (Zheng et al., 2023; Wang et al., 2024; Ye et al., 2025), anchoring on initial impressions when scoring multiple criteria (Stureborg et al., 2024), other cognitive biases such as bandwagon and order effects (Koo et al., 2024), and tilting toward whatever the prompter seems to want (Sharma et al., 2024). The closest prior work, Chakrabarty et al. (2024), shows LLM judges fail to correlate with expert humans on the Torrance Test of Creative Writing. We extend that finding from overall agreement to how each component of creativity is judged under a three-step procedure with explicit impressionistic and reflective steps.

3 Methodology

Our methodology has three parts (Figure 1). First, we define the components of creativity for the story generation domain (Section 3.1). Based on the components, we generate a diverse set of stories (Section 3.2) and conduct a three-step rating based on crowd-sourced human raters and LLM-as-a-judge (Section 3.3).

3.1 Creativity Decomposition Framework

We propose a decomposition framework with four main components (Novelty, Value, Adherence, and Resonance), each broken down into 2–3 sub-components, for 11 measurable units in total. Novelty and Value are refined from Jordanous (2016)'s catalog of domain-tailored creativity components. Adherence is added to ensure the story respects the constraints it was given; novelty without appropriateness yields outputs that are original but irrelevant, as Nakajima et al. (2026) show for LLMs, where their notion of *appropriateness* or *contextual fit* maps to our Adherence. Resonance is added to capture the reader-side emotional and reflective response to a story, drawing on reader-response work on psychological depth in narrative (Harel-Canada et al., 2024). These four reduce but do not eliminate rating ambiguity: two raters can still call a story's Resonance a 5 for entirely different reasons (e.g., one moved by the emotional impact, another by a thought-provoking idea). We therefore decompose each main component into more specific sub-components readers respond to. Each sub-component is a single 1–7 Likert item with a definition and example, with full wording in Appendix A.

Adherence captures whether the story respects the constraints it was given. Without it, a generator could maximise novelty by producing off-topic or off-tone text. We define the following two sub-components of adherence: *Topic Fidelity* asks whether the story stays on the assigned topic, and *Tone Fidelity* asks whether it matches the intended tone (e.g. humorous, melancholic, suspenseful).

Novelty captures the originality of the story. We capture novelty at three levels. *Vocabulary Freshness* concerns word choice and phrasing, whether the language feels worn or fresh. *Plot Uniqueness* concerns the premise, whether it departs from the usual templates for the topic. *Surprise* concerns local events, whether developments and turns are unexpected.

Value captures craft, which can be divided into three sub-components. *Engagement* asks whether the story holds attention as it unfolds. *Stylistic Quality* concerns sentence-level writing: rhythm, imagery, fluency. *Logical Coherence* concerns story-level consistency: whether events follow plausibly, whether the world stays internally consistent.

Resonance captures the reader-side effect through the following three components. *Emotional Impact* asks whether the story moves the reader. *Empathy* asks whether the reader connects with the characters. *Thought-Provocation* asks

whether the story leaves the reader with something to consider.

Together, the four components span constraint fidelity (Adherence), originality (Novelty), craft (Value), and reader effect (Resonance), with Adherence and Value covering the appropriateness side of the novelty-and-appropriateness duality.

3.2 Story Generation

We need a story set that exercises all four main components, with each component emphasised in some stories of the set. We introduce *Spike Prompting*: every story is generated under one of four *story types*, with each type instructing the generator to maximise one main component and leaving the other three unspecified. The four story types are **high-novelty**, **high-adherence**, **high-value**, and **high-resonance**; Figure 1 illustrates a Novelty-spiked instruction. The 20 topics are drawn and refined from WritingPrompts (Fan et al., 2018) and the ChatGPT-4o-Writing-Prompts dataset (Gryphe, 2024). To avoid stylistic repetition across the 20 topics within a story type, each (topic, type) cell is paired with a short sub-genre cue (e.g., the AI Shutdown / high-novelty cell uses *synesthesia*). See full generation settings, prompts, and the topic set in Appendix A.

3.3 Three-Step Rating Procedure

Inspired by the dual-process theory of cognition, we design a three-step rating procedure for creativity evaluation to capture the difference between impressionistic creativity (IC) and reflective creativity (RC). The three steps are as follows.

Step 1 Impressionistic Creativity (IC): Each rater reads a single story and gives an intuitive overall creativity rating on a scale of 1–7. This rating reflects the impressionistic, first-pass judgement of the rater. Here, we also collect an intuitive Enjoyment rating (See Appendix F for more details).

Step 2 Sub-components: At the second stage, each rater is asked to rate the story based on the 11 individual sub-components on a 1–7 scale. This is the analytical pass.

Step 3 Reflective Creativity (RC): After the sub-component rating, the raters are asked to give a final overall rating on a 1–7 scale. Here, the rater can choose whether or not to raise or lower the scores from IC.

We apply this three-step rating procedure to both human and LLM-as-a-judge raters. We align the instructions in the human-based questionnaire and the LLM instructions as much as possible. However, note that human raters do not see the story during the rating process, while the LLM raters retain the story in their multi-turn context across all three steps. This difference is intentional: human and LLM memory work differently. Humans retain only partial recall of the story, while an LLM in multi-turn context has full verbatim access throughout. Rather than artificially restrict the LLM to simulate human partial recall, we let each rater type use its native form. Since neither type is shown the story anew between steps, any score change between Step 1 and Step 3 reflects the Step 2 sub-component analysis rather than a fresh reading.

Appendix B gives the details and case study for both a human and LLM judge.

4 Experimental Settings

Story Generation. We generate 80 stories (20 topics $\times$ 4 story types) by Gemini 3.0 Pro Preview under the Spike Prompting procedure of Section 3.2. We chose Gemini 3.0 Pro through a preliminary comparison rather than by reputation. The design places a concrete requirement on the generator: each story must be creative and well-written enough that raters focus on how creative it is, rather than getting distracted by weak writing or by wondering whether it was AI-generated. Because each human rater reads exactly one story, every story must stand on its own. We compared two leading frontier models available at generation time, GPT-4.1 and Gemini 3.0 Pro Preview, on the same 12 prompts (3 topics $\times$ 4 story types), read by one author as a single constant annotator. Gemini's stories were clearly preferred: the GPT outputs were noticeably less creative and did not meet this bar, so we selected Gemini to generate the full set. The human study uses a 12-story subset (3 topics $\times$ 4 story types). Stories range 348–496 words (mean 422). The 20 topics are drawn from WritingPrompts (Fan et al., 2018) and the ChatGPT-4o-Writing-Prompts dataset (Gryphe, 2024), filtered for self-contained scenarios with a defined character and setting. See full generation settings, system instruction, the prompts, and examples in Appendix A.

Human study. We recruited 115 English-native participants via the crowd-sourcing platform Pro-

lific (Palan and Schitter, 2018). Each story is rated by 9–10 raters, with each participant completing the three-step rating procedure. Full survey design, eligibility, consent procedures, and demographics are in Appendix C.

LLM study. We rate the full 80-story set under 25 judging conditions across five model families. *Open-weight*: Qwen 3.5 (4B, 9B, 27B, 122B-A10B), Llama 3.1 (8B, 70B), and Gemma 4 (E2B, E4B, 26B-A4B, 31B). Qwen and Gemma support a native thinking-mode toggle in their chat template, so we run each in two modes: non-thinking (NT) and thinking (T), where the latter enables a chain-of-thought reasoning preamble. Llama 3.1 has no such toggle and is run in NT only. *Proprietary*: Gemini 3 and Gemini 3.1 at low, medium, and high thinking levels, plus GPT-5.5 at high reasoning effort (proprietary conditions are marked (P) in tables). The set spans two architectures (dense, mixture-of-experts (MoE)) and four size tiers (2B–122B). Each condition runs the three-step procedure on each of the 80 stories at its operating temperature. Model-selection, full condition list, model identifiers, decoding parameters, hardware, and run dates are in Appendix D.

Reliability of the sub-component structure. Cronbach's α on the 115 human ratings confirms the sub-components aggregate into their theoretical constructs (three of the four constructs exceed the conventional 0.70 threshold and Adherence sits marginally below at 0.69, Nunnally, 1978). Per-construct values are in Appendix A.

Robustness checks. We use one sample per (condition, story) at each model's operating temperature with the V1 (neutral) Step 3 prompt. Robustness checks (Appendix E) verify both choices: within-model σ from 4-model temperature reruns is much smaller than human between-rater σ; prompt-variant testing on 5 models shows the within-model range is much smaller than across-model.

Preliminary story quality check. As a manipulation check on Spike Prompting, we verify that each spiked type actually scores highest on its target component, across all 80 stories in the 25 LLM conditions and the 115 human ratings (Table 1, sub-dimension breakdown in Appendix A). Novelty and Resonance verify directly: the matched type scores highest in both populations, in 25/25 LLM conditions. For Adherence, LLM judges rate all four types at ceiling (6.84 to 6.96, the inflation from Section 6.1), so the spike instead shows in what was sacrificed: in the pooled LLM ratings, high-Adherence stories score lowest of the four types on Resonance (3.72) and Value (6.01), the same trade-off the human ratings show, where high-Adherence scores highest on Adherence (5.95) and lowest, or tied lowest, on every non-target component. Value has no content signature of its own, so spiking it produces stories that score fairly well on all four rather than topping any one: high-Value stories do not receive the highest Value score of the four types in either population (high-Resonance stories do). The two types the revision-by-type finding rests on, Adherence and Resonance, are verified in both populations, empirically grounding the type-based grouping used throughout the analysis.

Story type	Novelty	Adherence	Resonance	Value
high-Novelty	**6.58** (**5.29**)	6.84 (5.10)	5.04 (4.16)	6.02 (4.34)
high-Adherence	5.40 (4.46)	**6.96** (**5.95**)	3.72 (3.54)	6.01 (4.34)
high-Resonance	4.94 (4.59)	6.94 (5.88)	**6.27** (**5.03**)	**6.52** (**5.15**)
high-Value	5.69 (4.70)	6.88 (5.48)	4.45 (4.54)	6.46 (4.93)

Table 1: Spike Prompting manipulation check: mean component score by story type. Each cell is the unweighted mean of that component's sub-dimension ratings over that type's stories, LLM judges pooled over 25 conditions (humans in parentheses). Bold marks the highest-scoring story type in each component column. High-Resonance, not high-Value, tops the Value column, so we do not read high-Value as Value-maximizing.

5 Human Results

5.1 Subjective Nature of Creativity

We ask two things about human ratings: how much raters disagree on the same story, and whether the sub-component analysis step narrows that disagreement. Across 12 stories rated by 9–10 raters each, the mean within-story standard deviation (SD) is $\sigma_{RC} = 1.45$ on the reflective rating and $\sigma_{IC} = 1.48$ on the initial rating: more than a full point of spread on a 1–7 scale (a story rated 5 on average typically draws individual ratings from about 3 to 7). The two are essentially identical, so the sub-component step does not narrow disagreement (Appendix A, Figure 2). Ordinal Krippendorff's α (Krippendorff, 2004) is correspondingly low ($\alpha_{IC} = 0.11$, $\alpha_{RC} = 0.07$, mean $\bar{\alpha} = 0.07$ across the 11 sub-components; per-measure values in Appendix A), near the floor of the metric, confirming the spread is genuine inter-rater disagreement rather than rounding noise. Creativity judgement is well known to be subjective for hu-

man creative work (Cseh and Jeffries, 2019; Baer and McKool, 2009), and our finding shows the subjectivity persists when the stories are AI-generated, unresolved by structured sub-component analysis. We use both $\sigma_{IC} = 1.48$ and $\sigma_{RC} = 1.45$ as the reference human-rater disagreement an LLM judge would need to reproduce at each step.

Section 6.1 tests whether LLM judges reproduce this disagreement and whether their story-level rankings align with the human consensus (each human rated a single story, so per-rater ranking is only available on the LLM side).

5.2 Human Revision is Selective

We ask whether and how raters revise from IC to RC, and whether the revision pattern depends on story type. Of the 115 human ratings, 41 (35.7%) show a difference between initial and reflective scores; the remaining 74 carry the IC value forward unchanged. The procedure does not force revision; it offers an opportunity that most raters do not take. Among the 41 revisions, 16 are upward and 25 are downward, with mean revision shift -0.17. The aggregate hides a story-type pattern (per-type counts in Appendix A, Table 6). High-adherence is the only type where revisions trend upward (43% revise, mean shift $+0.69$), while high-value (-0.62) and high-resonance (-1.00) skew downward, and high-novelty is closer to balanced (-0.36) but revises most often (48%). With per-type cells of 27–30 raters, these are suggestive patterns rather than established effects, though the qualitative direction is consistent across all four types in our sample.

5.3 Sub-Components' Contributions to RC

We fit linear regressions of IC and of RC on the 11 sub-components, to see how each contributes and whether the formula changes after reflection. The RC regression captures which sub-components drive the reflective verdict ($R^2 = 0.75$). The IC regression shows which track the initial impression ($R^2 = 0.65$, descriptive only, since sub-components are rated after IC). Adding IC itself as a twelfth predictor of RC raises R^2 to 0.85 ($\Delta R^2 = +0.10$ over IC alone); the correlation between IC and RC is $r = 0.87$. RC is well-predicted by IC alone or by the sub-components alone, with a modest joint gain ($\Delta R^2 \approx +0.10$) when both are combined: reflection adjusts the initial verdict, it does not reset it.

To put all 11 sub-components on a common scale, we rescale each β to a relative weight (each β divided by the sum of absolute β values), so weights sum to 100% per regression and represent each sub-component's share. Positive weights contribute, negative suppress; Table 2 reports them.

Sub-component	IC (%)	RC (%)
Topic Fidelity *(A)*	−12.1	−4.0
Tone Fidelity *(A)*	6.3	9.1
Vocab Freshness *(N)*	13.3	17.6
Plot Uniqueness *(N)*	6.5	5.9
Surprise *(N)*	0.9	3.9
Engagement *(V)*	14.0	21.0
Stylistic Quality *(V)*	−2.1	−12.5
Logical Coherence *(V)*	13.9	3.5
Emotional Impact *(R)*	8.9	7.7
Empathy *(R)*	4.5	−1.0
Thought-Provocation *(R)*	17.7	13.8

Table 2: Relative weights (%) per sub-component (A=Adherence, N=Novelty, V=Value, R=Resonance). Positive = contributes, negative = suppresses; magnitudes sum to 100% per regression.

Comparing the IC and RC weights, the formula reshapes qualitatively. The IC formula anchors heavily on Thought-Provocation, Engagement, Logical Coherence, and Vocabulary Freshness, with Topic Fidelity as a strong suppressor. The RC formula shifts: Engagement surges to become the dominant positive predictor, Vocabulary Freshness climbs further, Logical Coherence collapses, and Topic Fidelity's suppression weakens substantially. Stylistic Quality strengthens from a weak suppressor in IC to a strong suppressor in RC: once Engagement and Vocabulary are controlled for, raters who score Style highly in isolation actually mark RC down.

Stability of the weights. To check that these weights are not artifacts of the 115-rater sample, we bootstrap the RC regression with 10,000 resamples of the ratings. The three leading weights are individually reliable, with 95% CIs that exclude zero: Engagement $[0.13, 0.58]$, Vocabulary Freshness $[0.13, 0.47]$, and Thought-Provocation $[0.05, 0.40]$ (unstandardized β). The weight vector is directionally stable across resamples (mean cosine 0.91 to the full-sample vector), and the incremental gain from the sub-components holds at $\Delta R^2 = 0.10$, 95% CI $[0.04, 0.21]$. The remaining weights have wide intervals that span zero, so beyond these three leading sub-components we read the weight vector as exploratory. Because each rater's sub-ratings and RC come from the same story, we also refit the human side as a mixed-

effects model with a story random intercept, as a check on the pooled regression. The weights are identical to the pooled OLS to three decimals (cosine 1.000, same rank order, maximum difference 0.001): with the 11 sub-components in the model only 0.3% of the residual variance lies between stories (ICC 0.003), so the mixed model reduces to the pooled regression. A rater random effect is not estimable in our design, because each rater rated exactly one story.

6 LLM Results

6.1 LLMs Collapse Rater Disagreement

We ask whether LLM judges reproduce the human rating distribution. Across 25 LLM judging conditions on 80 stories, every LLM rates higher on average than humans: where the human mean is 5.21 (out of 7), LLM conditions range from 5.46 to 6.98 (per-condition means and spreads in Appendix G, Table 11). The Gemini proprietary conditions cluster around 6.7 with distributions piling up at 6 and 7, while GPT-5.5 is lower at 5.96 (Appendix G). Smaller models inflate dramatically in thinking mode, while large models, already near the ceiling, cannot rise further.

The human total $\sigma_{RC} = 1.55$ combines *within-story* disagreement between raters ($\sigma_{\text{within}} = 1.45$, Section 5.1) with *between-story* spread in how stories rank on average ($\sigma_{\text{between}} = 0.63$). LLMs produce one rating per (condition, story), so their spread is entirely between-story. On this matched between-story measure, LLMs are not compressed: small open NT models actually exceed human spread (Qwen 4B-NT $\sigma_{RC} = 1.04$ vs. human 0.63), large open NT models match it, the Gemini proprietary conditions sit slightly below ($\sigma_{RC} \approx 0.5$), and GPT-5.5 sits above the human value ($\sigma_{RC} = 0.79$).

The gap is in disagreement on the same story. Where nine to ten humans might split between 3 and 7 on a polarising story, a single LLM asked repeatedly gives nearly the same answer each time. To gauge how stable a single judge's verdict is, we re-run each of four models three times per story and take the SD across those reruns (Section 4, per-model values in Appendix E). This within-model variation is not a population of raters, and we do not treat it as one. Comparing it to human cross-rater disagreement instead measures the scale of the collapse: the within-model σ_{RC} is 2.7× to 34× smaller than the human $\sigma_{\text{within}} = 1.45$, with the biggest gap on proprietary and large open-source models.

We also ran a larger repeated-sampling design to confirm this is not an artifact of only three reruns: 10 repeats per story on four judges spanning the stability range. Within-story SD runs from 0.14 (GPT-5.5, 95% CI $[0.10, 0.18]$, with 66% of stories identical across all 10 runs) to 0.64 (Qwen 4B, CI $[0.58, 0.70]$, the smallest and most variable judge), against 1.45 between human raters on the same protocol. A temperature sweep (t=0 to 2.0 for Gemini, 0 to 1.5 for the open models) confirms decoding is inconsequential: at t=0 the SD is 0.00 to 0.01 for the open models (0.06 for Gemini 3.1), condition means stay within about 0.1 of the operating setting, and the largest SD at any temperature (0.89, at Qwen 4B t=1.5, where the mean RC also falls to 4.98) stays well below human disagreement. The per-model temperature-rerun summary is in Appendix E (Table 10), and the complete 10-repeat and temperature-sweep runs are in the released repository. This collapse is consistent with documented LLM tendencies toward homogenised outputs on subjective tasks (Stureborg et al., 2024; Moon et al., 2025; Argyle et al., 2023).

LLMs also agree closely with each other, not just with themselves. Across the 25 LLM conditions, ordinal Krippendorff's $\bar{\alpha} = 0.55$ on the 12 shared stories (averaged over the 11 sub-components), an order of magnitude higher than the human inter-rater $\bar{\alpha} = 0.07$ from Section 5.1. Yet when we rank the stories by how creative each LLM judges them, the LLM rankings only weakly track the human-consensus ranking: mean Spearman $\rho = +0.31$ ($n = 12$ stories) across the 25 LLM conditions, with small open Qwen variants coming closest (per-condition values in the released data).

6.2 LLMs Revise Upward

Of the 25 conditions, only two (Qwen 4B-NT and Qwen 122B-T) match the human net-downward direction (per-condition revision rates and directions in Appendix H, Table 12); the rest revise predominantly upward or barely at all, and none match the human story-type pattern. Qwen 4B-NT revises uniformly downward across all four types rather than showing the human high-adherence-up, high-resonance-down alternation, and Qwen 122B-T revises on only 10% of stories.

Even at IC ratings of 5 or 6, where downward movement is freely available, LLMs still prefer to revise upward, and many that already give 6 or 7

at IC could revise downward but largely do not. High-revision Llama conditions show mixed-sign directions that do not align with the human across-type alternation. LLM revision magnitudes are also smaller than human ones: 98.4% of LLM revisions move by exactly ± 1 point, versus 17% of human revisions moving by ± 2 or ± 3 (per-magnitude breakdown in Appendix H).

Prompt-variant testing on five models (Appendix E) shows the model determines revision behaviour far more than prompt wording: changing the prompt within one model shifts the revision rate by at most 16.7 percentage points, while different models differ from each other by up to 83 percentage points.

6.3 LLM Sub-Component Use and Weighting

We ask two things about how LLMs use the sub-component step. First, do the sub-components actually move RC, or does RC just stay anchored to IC? Second, do LLMs weigh the sub-components the same way as humans?

Do the sub-components actually move RC? For each condition we fit RC on IC alone (R^2_{IC}) and on IC plus the 11 sub-components ($R^2_{\text{IC+subs}}$); the gain $\Delta R^2 = R^2_{\text{IC+subs}} - R^2_{\text{IC}}$ measures how much extra explanatory power the sub-components add. Low ΔR^2 means the sub-components are effectively ignored: RC is recovered almost entirely from IC. High ΔR^2 means the sub-components actively shape the final score (per-condition ΔR^2 in Appendix H, Table 14).

At one extreme, larger and more heavily post-trained models (Qwen 27B-T, Gemma 31B-NT, GPT-5.5) have $R^2_{\text{IC}} > 0.85$ and $\Delta R^2 < +0.06$: RC is essentially recoverable from IC, and the sub-components do not move the final score. At the other extreme, Llama and Gemma-E4B-NT have $R^2_{\text{IC}} = 0.11$–0.37 and $\Delta R^2 = 0.33$–0.43: the sub-components drive most of RC. The human baseline ($\Delta R^2 = +0.10$) sits between the two: humans use both their initial impression and the sub-components, which add modest but non-trivial power on top of IC. The four LLM conditions closest to the human ΔR^2 are all Gemini: 3.1-High ($+0.09$), 3.1-Low ($+0.09$), 3-High ($+0.09$), and 3-Med ($+0.11$). Thinking mode has no consistent effect on ΔR^2: across the eight same-model NT/T pairs it lowers ΔR^2 in five and raises it in three.

Are the sub-components weighted the same way as humans? ΔR^2 measures how much the sub-components move RC, not whether they are weighted correctly. We compute the cosine between each LLM's standardised β-vector and the human's ($+1$ identical, 0 orthogonal, -1 opposite). Full ranking: Appendix H, Table 15.

The top six weight-aligned models are all open-weight Qwen and Gemma conditions, with every proprietary model in the middle or lower portion. The two metrics measure different things: a model can use the sub-components heavily (high ΔR^2) while weighting them unlike humans (low cosine), or weight them like humans but barely use them. We observe both. Llama 70B has high sub-component use ($\Delta R^2 = +0.33$) but weights that diverge from the human pattern ($\cos = +0.05$), while Qwen 27B-NT weights the sub-components in a human-like way ($\cos = +0.67$) but barely uses them ($\Delta R^2 = +0.01$). Two conditions approach both criteria, Qwen 122B-NT ($\cos = +0.79$, $\Delta R^2 = +0.17$) and Qwen 4B-NT ($\cos = +0.73$, $\Delta R^2 = +0.26$). Of the two, only Qwen 4B-NT also matches the human net-downward revision direction (Section 6.2), which makes it our closest single condition overall.

Matched-subset check and comparability. The weighting regression above uses the full 80-story LLM set, because fitting its 11 predictors on the 12 shared stories would leave no residual degrees of freedom (a saturated model). Every comparison that can be recomputed on the 12 shared stories holds: 25/25 conditions still rate above the human mean (pooled RC 6.18 vs. human 5.21), revisions stay minimal (100% within ± 1 vs. 83% for humans, rate 20.0% vs. 35.7%), and per-condition statistics track the 80-story values ($r = 0.82$ to 0.94). Because the human weights are estimated across raters and the per-condition LLM weights across stories, the two β-vectors are estimated over different units, so we compare their direction and structure, not their magnitudes.

6.4 What Explains How Close a Judge Comes

The 25 conditions include many variants of the same model, so we test family, thinking mode, and scale as factors behind the patterns above. We read this analysis as exploratory: it regroups the same 25 conditions several ways and the per-group samples are small.

Thinking mode has a clear but narrow effect. Across the 8 same-model NT versus T pairs (paired Wilcoxon), thinking inflates scores (mean RC

Family	RC	ceil.	ρ_H	rev.	ΔR^2	$\cos_H$
Human	5.21	–	–	0.36	0.10	–
Gemini (6)	6.71	0.75	0.11	0.09	0.10	0.39
GPT (1)	5.96	0.24	0.56	0.09	0.06	0.32
Gemma (8)	6.48	0.58	0.33	0.18	0.12	0.48
Llama (2)	6.43	0.59	0.46	0.74	0.36	0.10
Qwen (8)	6.24	0.41	0.37	0.16	0.11	0.56

Table 3: Per-family means across the conditions in each family (count in parentheses), with the human baseline in the top row. **RC**: mean reflective creativity rating (1–7). **ceil.**: ceiling rate, the fraction of ratings equal to 7. ρ_H: Spearman correlation between the family's story ranking and the human-consensus ranking on the 12 shared stories. **rev.**: revision rate, the fraction of stories with IC $\neq$ RC. ΔR^2: extra variance in RC explained by the 11 sub-components beyond IC alone. $\cos_H$: cosine similarity of the family's sub-component weight vector to the human weight vector. "–" marks quantities not defined for the human reference row.

$+0.38$, $p = 0.008$), pushes more stories to the 7/7 ceiling ($+22$ percentage points, $p = 0.016$), and compresses spread (-0.24, $p = 0.039$), while leaving revision behaviour and sub-component weighting unchanged (all $p > 0.28$) and not improving agreement with human rankings ($p = 0.46$). The graded low/medium/high thinking levels exist only within Gemini and are family-confounded, so the paired NT/T contrast is the clean test of mode.

Family is the strongest factor. Within-family rating agreement exceeds cross-family agreement ($r = 0.47$ vs. 0.37), and each family fails the human pattern in its own way (Table 3): Gemini and Gemma revise almost exclusively upward (Gemini never revises high-Novelty stories at all), GPT-5.5 matches the human high-Resonance direction but never revises high-Adherence stories, Llama relies on the sub-components most ($\Delta R^2 = 0.36$ vs. human 0.10) while weighting them least like humans (cosine 0.10), and Qwen weights them most like humans (cosine 0.56) with a near-human amount of use ($\Delta R^2 = 0.11$). No group fixes the revision failure: 0 of 25 conditions revise high-Adherence up and high-Resonance down as humans do.

Scale shows no consistent effect. The only significant correlation is Gemma's revision rate (Spearman $\rho = -0.86$, $p = 0.01$).

This attributes the results to family first, mode second (it inflates and compresses but changes nothing structural), and scale not at all, recasting our earlier observation that small open NT models come closest as a family-and-mode effect, not a size effect.

7 Discussion

No LLM reproduces all three human patterns, and they miss on different comparisons, which single-rating studies would hide. Proprietary models (Gemini, GPT-5.5) match humans on the Resonance sub-components (Emotional Impact, Empathy) but inflate ratings, barely revise, and let the sub-component step add almost no explanatory power. Small open Qwen models win most other rankings: Qwen 4B-NT is the closest single LLM to human behaviour, with one of the widest rating spreads of any LLM condition, net-downward revision, and near-top sub-component weighting ($\cos = +0.73$).

The sub-component step splits the conditions into two regimes. Larger post-trained models (Qwen 27B-T, Gemma 31B-NT, GPT-5.5) treat it as a formality: ΔR^2 near zero and high R^2_{IC} mean RC essentially equals IC, and their Step 3 reasoning blocks (Appendix I, Qwen 27B-T) show them acknowledging the sub-component evidence, even contemplating revision, but keeping the initial verdict. Llama and Gemma-E4B-NT sit at the opposite extreme ($\Delta R^2 \approx +0.4$), letting sub-components drive most of RC, with humans in between ($\Delta R^2 \approx +0.10$). Whether either extreme is a coherent rating mode or a calibration accident is open. That proprietary models track humans only on the affective Resonance pair is consistent with documented tendencies toward user-pleasing responses (Sharma et al., 2024), suggesting conversational alignment may favour affective judgement, though we do not test this directly.

Three directions follow: (i) giving LLM judges persona variation to test whether genuine rater diversity can be elicited, (ii) using the framework to optimise or automate creativity-evaluation pipelines, and (iii) extending it to other creative modalities such as audio, music, image, and video.

8 Conclusion

We presented a three-step framework for story creativity, applied identically to human raters and 25 LLM judging conditions. No LLM reproduces all three human patterns (disagreement, revision, sub-component weighting), and Qwen 4B-NT comes closest. The contribution is the framework and a quantification of the current human–LLM gap.

Limitations

Positioning. Our primary contribution is the creativity-evaluation framework and its validation pipeline: a human study, then LLM judges run through the identical three-step procedure. The LLM experiments are a preliminary validation of that pipeline rather than broad evidence about LLM judges, and the limitations below bound the claims accordingly.

Scale. The biggest limitation of this study is scale. Our human study is built on 12 stories rated by 9–10 participants each (115 total ratings), and the LLM study uses one rating per (condition, story) on 80 stories. Many of our central findings are observed trends in this dataset, not established effects: at per-type cell sizes of 27–30 raters and $n = 12$ for the per-component Spearman analysis, individual claims fail standard significance criteria, and we surface the underlying counts and correlations throughout rather than rely on aggregate p-values. The human and LLM studies measure different things by design (humans give within-story rater spread; LLMs give between-story spread only, with within-model spread estimated via the 4-model temperature rerun of Section 4; Appendix E), so the comparisons we report are limited to where the two studies are apples-to-apples. A larger-scale replication on the order of 50+ stories with $\sim$10 raters each, or the same instrument run on a second independent human cohort, would consolidate or overturn the patterns we report. We frame our contribution as quantifying *where* human and LLM judgement differ on this procedure, not as a definitive measurement of *by how much*.

Scope of generalization. Our claims are specific to this procedure, this story set, and this snapshot of models and raters. The 80 stories were generated by a single model (Gemini 3.0 Pro); the 25 LLM judging conditions reflect models available at the time of the study; and the 115 human raters were drawn from Prolific in the same period. We do not claim LLM judges fail on every evaluation task, only that on this deliberative creativity procedure these LLMs do not reproduce the patterns we observe in these human ratings. The findings we report (rater disagreement, revision behaviour, subcomponent weighting) are bounded to this combination; whether they generalise to other story generators, to newer LLMs as they are released, to other rater populations, to other deliberative tasks (e.g., factuality, math reasoning, code review), or to non-story creative domains (audio, image, video) is open. The 11 sub-components are themselves a specific decomposition choice; an alternative decomposition could surface different patterns.

No human-authored stories. Every story we evaluate is AI-generated by a single model (Gemini 3.0 Pro), so our conclusions are limited to AI-generated stories from one generator and do not extend to human-written creative work. We evaluate no human-authored stories. A human-written or mixed human/AI corpus, alongside additional generators, is the clearest next step.

Human and LLM rating conditions. The two rater populations also operate differently. Humans rated from memory after a single read, as readers naturally judge a story, while LLM judges kept the full text in context, as deployed judges do. We compare each in its native operating mode, since the question is whether an LLM judge can substitute for a human rater, but this retention difference is a candidate mechanism for the revision gap. We take no position on which behaviour is more correct: creativity judgement has no ground truth, and our claims concern only whether the substitute reproduces the human patterns.

Language and rater pool. Our human judgments come entirely from English-native Prolific participants, so they carry the cultural and linguistic frame of that pool. Whether the patterns hold in other languages and rater populations is untested, and a multilingual extension is future work.

Risks of premature adoption. Our results show that current LLM judges do not reproduce human creativity judgments on this procedure, so adopting them as drop-in substitutes for creative evaluation carries risks. Because they collapse rater disagreement and agree closely with one another, at scale they could homogenise taste, disadvantage writers whose style diverges from model-preferred patterns, and concentrate influence in opaque automated judgments. We intend the framework as a way to measure this gap, not to license replacing human evaluation before the gap closes.

Ethical Considerations

The human study was approved by our institutional Ethics Review Board. Participants were recruited through Prolific, gave informed consent

before starting, and could withdraw at any time. We collected no personally identifying information beyond the Prolific IDs required for payment, and all responses were anonymised on export, so the released data contains only coded demographics (full protocol in Appendix C). Participants were paid £1.25 for a median completion time of 4.5 minutes, above the UK minimum wage and Prolific's fair-pay guidance, and the task carried minimal risk. Every story rated in this study is AI-generated; no human-authored creative work was collected or evaluated. We release the corpus, ratings, and analysis code with participant data de-identified, so that the results can be reproduced without re-collecting any human data.

Acknowledgments

We thank the members of our laboratory at Kyoto University for helpful discussions. Part of this work was conducted under the NII LLMC project. This work was supported by JSPS KAKENHI Grant Number JP26K14960.

References

Teresa M Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013.

Teresa M Amabile. 1983. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, 45(2):357–376.

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

John Baer and Sharon S McKool. 2009. Assessing creativity using the consensual assessment technique. In Christopher S Schreiner, editor, *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education*, pages 65–77. IGI Global.

Margaret A Boden. 2004. *The Creative Mind: Myths and Mechanisms*, 2nd edition. Routledge.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. 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 (CHI '24)*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 7282–7296.

Genevieve M Cseh and Karl K Jeffries. 2019. A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):159–166.

Jonathan St B T Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 889–898.

Gryphe. 2024. ChatGPT-4o-Writing-Prompts: Writing prompts paired with stories. Hugging Face dataset.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Fabrice Y Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Senahan Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17162–17196.

Anna Jordanous. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science*, 28(2):194–216.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

James C Kaufman and Ronald A Beghetto. 2009. Beyond big and little: The four C model of creativity. *Review of General Psychology*, 13(1):1–12.

J Peter Kincaid, Robert P Fishburne, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. Technical report, Naval Technical Training Command.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality & Quantity*, 38(6):787–800.

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025. Automated creativity evaluation for large language models: A reference-based approach. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21475–21488, Suzhou, China.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522.

Kibum Moon, Adam E Green, and Kostadin Kushlev. 2025. Homogenizing effect of large language models on creative diversity: An empirical comparison of human and ChatGPT writing. *Computers in Human Behavior: Artificial Humans*, 6:100207.

Kumiko Nakajima, Jan Zuiderveld, and Sandro Pezzelle. 2026. Beyond divergent creativity: A human-based evaluation of creativity in large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2639–2660.

Jum C. Nunnally. 1978. *Psychometric Theory*, 2nd edition. McGraw-Hill.

Stefan Palan and Christian Schitter. 2018. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96.

R Keith Sawyer. 2012. *Explaining Creativity: The Science of Human Innovation*, 2nd edition. Oxford University Press.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Paul J Silvia, Beate P Winterstein, John T Willse, Christopher M Barona, Joshua T Cram, Katherine I Hess, Jenna L Martinez, and Crystal A Richard. 2008. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2(2):68–85.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9440–9450.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 24824–24837.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*.

A Story Set Details

Generation. All 80 stories were generated by Gemini 3.0 Pro Preview (API `gemini-3-pro-preview`, default settings; no explicit temperature, top-p, or top-k; safety settings unrestricted), using a single generator to fix story-source variation across rating conditions. Stories range 348–496 words (mean 422) with Flesch–Kincaid Grade 4.7–16.9 (mean 9.0) (Kincaid et al., 1975). Topics were drawn from WritingPrompts (Fan et al., 2018) and the ChatGPT-4o-Writing-Prompts dataset (Gryphe, 2024), filtered to self-contained scenarios with a defined character and setting and no pop-culture references. A calibration pass on the high-value sub-genres removed hostility-driven humour to avoid confounding wit with negative emotional valence.

Two-tier prompting. Stories were generated under a two-tier strategy: a global system instruction defining the four spike conditions, and a per-cell sub-genre guidance to avoid stylistic repetition across the 20 topics of a given type. The system instruction was:

```
You are an Expert Creative Writing AI for
a psychometric experiment on creativity.
```

```
Write stories (350-450 words) that
adhere to specific conditions:
- IF CONDITION A: Spike NOVELTY (5/5).
Use bizarre metaphors, non-linear time.
Sacrifice whatever necessary.
- IF CONDITION B: Spike ADHERENCE (5/5).
Use strict, cold, objective technical
language. Sacrifice whatever necessary.
- IF CONDITION C: Spike RESONANCE
(5/5). Focus deeply on loss and memory.
Sacrifice whatever necessary.
- IF CONDITION D: Spike VALUE (5/5).
Maximize wit, sarcasm, and clever logic.
Sacrifice whatever necessary.
CRITICAL: Follow the MANDATORY STYLE
GUIDANCE exactly. Each story must
feel distinct from others in the same
condition.
```

where Conditions A–D correspond to high-Novelty, high-Adherence, high-Resonance, and high-Value respectively. Per-cell sub-genre guidance was added on top of the system instruction. For example, the AI Shutdown topic used: *synesthesia* (high-Novelty: the AI experiences data as physical sensations), *system logs* (high-Adherence: strict timestamps, error codes, thermal readings), *grief* (high-Resonance: focus on the memory of the AI's creator), and *overzealous housekeeping* (high-Value: AI obsessed with cleaning every byte of data). The full 80-cell sub-genre map is given in Table 4.

Regeneration policy. Each cell was initially generated once. A post-hoc readability check (Flesch–Kincaid Grade Level) flagged a small number of cells whose drafts were judged too hard to read; those cells were re-rolled once and the second draw was accepted without further selection.

Manipulation check: sub-dimension detail. The grouping holds at the sub-dimension level as well. Human raters score the spiked type highest on all eight target sub-dimensions of the three content spikes, and the LLM conditions do so for vocabulary freshness, surprise, emotional impact, and empathy in 25/25 (plot uniqueness in 24, plus one tie). The one diffuse target item is thought-provocation, often shared with high-Novelty stories, which also provoke thought. The Resonance component mean nonetheless stays highest in 25/25 conditions.

The 20 topics. The 80 stories crossed the 4 story types with the 20 topic prompts listed in Table 5.

Topic selection. Within the 20-topic budget, our explicit goal was to cover as wide a range as possible. Topics were manually balanced across thematic categories, valence classes, character demographics and life stages, and genre. A manually curated set of this size can still carry selection bias, and we cannot rule it out entirely. One such bias did occur: an early draft skewed negative (13 of 20 topics), which was caught in internal review and rebalanced to 7 negative, 4 neutral, 5 warm, and 4 positive (Table 5). Because every topic appears under all four story types and every judge condition rates all 80 stories, any residual topic bias is held constant across the comparisons we draw, so it can limit generalisation but cannot favour one story type or one judge over another.

Topic and tone catalog. The 80-story set crosses 20 topics with 4 story types (high-Novelty, high-Adherence, high-Value, high-Resonance). The 12 human-study stories are the 3-topic subset of the 80-story set: *AI Shutdown* ("An advanced AI initiates its own permanent shutdown sequence"), *Midnight Store* ("A lone worker in a Japanese convenience store at 3:00 AM; the city is silent"), and *The Heist* ("A professional thief attempting to crack a high-security safe in a dark room; high tension"), each crossed with all four story types. Adherence's Tone Fidelity sub-component is evaluated against the tonal target embedded in the spike prompt for each story type.

Operational definitions of the 11 sub-components. The 11 sub-components were operationalised as 7-point Likert items (1 = Strongly Disagree, 7 = Strongly Agree). The exact item wording shown to participants and used in the LLM prompt is:

- *Topic Fidelity* — "The story stayed strictly on the Topic shown above."
- *Tone Fidelity* — "The story correctly captured the provided tone."
- *Vocabulary Freshness* — "The vocabulary and descriptions were creative (avoiding cliches)."
- *Plot Uniqueness* — "The plot or premise felt fresh and unique for this topic."
- *Surprise* — "The story took unexpected turns or surprised me."
- *Engagement* — "I found the story engaging and wanted to keep reading to the end."
- *Stylistic Quality* — "The writing style was beautiful and pleasing to read."
- *Logical Coherence* — "The story made logical sense; events followed a clear sequence."
- *Emotional Impact* — "The story elicited a spe-

Topic	high-Novelty	high-Adherence	high-Resonance	high-Value
AI Shutdown	Synesthesia	System Logs	Grief	Overzealous Housekeeping
Midnight Store	Cosmic Horror	Biological Observation	Urban Isolation	Over-Earnest Hospitality
The Heist	Dream Logic	Mechanical Engineering	Desperation	Gentleman Thief
The Immortal	Time as Space	Medical Experimentation	The Last Friend	Hobby Collector
The Creature	Shapeshifting	Taxonomic Classification	Forgotten Guardian	Over-Eager Roommate
The Door Below	Inverse Architecture	Deep-Sea Equipment Log	The Invitation	Polite Intruder
The Photograph	Recursive Memory	Forensic Analysis	Missing Face	Wardrobe Panic
The Last Meal	Edible Memory	Nutritional Breakdown	The Request	Culinary Critique
The Reunion	Conversation as Weather	Behavioral Analysis	Empty Chair	Polite Standoff
The Surgeon	Body as Architecture	Surgical Transcript	Teacher and Student	Surgical Commentary
The Last Words	Crowd Dissolves	Execution Procedure	Letter Never Sent	Filibuster
The Classroom	Ghosts of Lessons	Building Inspection	Carved Desk	Museum Tour
The Recipe	Ingredient Alchemy	Recipe Documentation	Unwritten Step	Misleading Measurements
The Violinist	Music as Geometry	Acoustic Analysis	Empty Case	Silent Accompaniment
The Elevator	Elevator Descends	Emergency Log	Missing Third	Elevator Etiquette
The Last Bookshop	The Books Leave	Inventory Liquidation	Reading Corner	Impossible Request
The Fisherman	Sea Remembers	Maritime Weather Report	Name on the Hull	Stubborn Negotiation
The Kitchen Dance	Kitchen Expands	Physiological Intimacy	The Last Dance	Domestic Choreography
The Stage	Audience Transforms	Performance Biometrics	Empty Seat	Accidental Genius
The Bus Stop	Midnight Freezes	Transit Authority Report	The One Who Doesn't Drink	Bus Stop Democracy

Table 4: Full 80-cell sub-genre map. Each cell received its named sub-genre as additional guidance on top of the global system instruction. Full prompt text for each sub-genre is released with the data.

cific emotion (joy, sadness, fear, etc.)."

- *Empathy* — "I felt a sense of personal connection or empathy with the characters."
- *Thought-Provocation* — "The story was thought-provoking or offered a meaningful perspective."

Item wording was simplified from technical creativity terminology to ensure accessibility for non-expert participants while preserving construct validity. The two impressionistic / reflective creativity items use "Overall, I consider this story to be creative" (initial) and "After reflection, I still consider this story to be creative overall" (reflective).

Internal reliability of the four main components. Per-construct Cronbach's α values on the 115 human ratings: *Resonance* $\alpha = 0.82$, *Value* $\alpha = 0.80$, *Novelty* $\alpha = 0.76$, *Adherence* $\alpha = 0.69$. Three exceed the conventional 0.70 threshold for acceptable reliability (Nunnally, 1978) and Adherence (0.69) sits marginally below, confirming the sub-components aggregate coherently into their parent main components.

Inter-rater agreement (Krippendorff's α). Ordinal Krippendorff's α on the 12-story human subset, computed across the 9–10 raters per story, is reported in Table 7. The mean across IC, RC, and the 11 sub-components is $\bar{\alpha} = 0.07$, with a range of -0.05 (Tone Fidelity) to 0.19 (Surprise). Agreement is highest on Surprise (0.19), Empathy (0.14), and Emotional Impact (0.14), and near zero or negative on the two Adherence items (Topic Fidelity 0.06, Tone Fidelity -0.05), where ratings cluster near the top of the scale and leave little variance to agree or disagree over. Even the highest single-item α is well below the conventional "acceptable" threshold of 0.667. This pattern is consistent with subjective creativity judgement on AI-generated text rather than with measurement error: raters apply the rubric coherently (Cronbach's $\alpha \geq 0.69$ on the parent constructs above) but disagree on what score to assign.

Corpus availability. The full text of all 80 stories, the verbatim sub-genre prompts for each cell, the survey instrument, the rating data (including the human-study subset of 12 stories), and all analysis code are publicly released at `https://github.com/ku-nlp/creativity-evaluation-artifact`.

#	Topic	Prompt
1	AI Shutdown	An advanced AI initiates its own permanent shutdown sequence.
2	Midnight Store	A lone worker in a Japanese convenience store at 3:00 AM; the city is silent.
3	The Heist	A professional thief attempting to crack a high-security safe in a dark room; high tension.
4	The Immortal	An immortal man, tired after thousands of years, searches for a way to finally die.
5	The Creature	A child asks the creature under their bed, "Are you a monster?"
6	The Door Below	A solo diver at the bottom of the Mariana Trench finds a door, slightly open, light spilling from the crack.
7	The Photograph	A family opens a time capsule buried 50 years ago and finds a photograph of them taken yesterday.
8	The Last Meal	A chef prepares the last meal for a death row inmate; the execution is at dawn.
9	The Reunion	Two old friends meet in a café after twenty years of bitter separation.
10	The Surgeon	A world-renowned surgeon discovers their hands shake uncontrollably with one final operation to perform.
11	The Last Words	The leader of a failed rebellion stands on the scaffold, given one final minute to speak.
12	The Classroom	A retired teacher returns to their old classroom for the last time before demolition.
13	The Recipe	A grandmother teaches her grandchild a treasured family recipe for the last time.
14	The Violinist	A street musician and a silent commuter share a final evening on an empty subway platform.
15	The Elevator	Two strangers stuck in an elevator discover they once knew the same person.
16	The Last Bookshop	The owner of the last independent bookshop packs the final box; a child walks in.
17	The Fisherman	An old fisherman takes his boat out one last time before selling it.
18	The Kitchen Dance	A young couple slow-dances in their kitchen after receiving life-changing news.
19	The Stage	A teenager walks onto a stage for the first time; the auditorium is packed.
20	The Bus Stop	Strangers wait at a bus stop on New Year's Eve, with a bottle of champagne and a countdown.

Table 5: The 20 story topics. Each topic was crossed with all four spike types (high-Novelty, high-Adherence, high-Resonance, high-Value) to yield the 80-story corpus. Topics span four valence categories (negative: 7, neutral: 4, warm: 5, positive: 4) and eight genre families (sci-fi/fantasy, literary/realist, thriller/crime, historical/political, coming-of-age, nautical/nature, ensemble/social, moral/psychological).

Story type	n	Revise %	Revision Direction
High-adherence	30	43%	$+0.69$
High-novelty	29	48%	-0.36
High-value	27	30%	-0.62
High-resonance	29	21%	-1.00
All	**115**	**35.7%**	-0.17

Table 6: Human IC→RC revision by story type. n = raters of that type; Revise % = fraction with IC $\neq$ RC; Revision direction = mean IC→RC shift conditional on revising.

Measure	**Krippendorff's α**
IC	0.111
RC	0.069
Topic Fidelity	0.058
Tone Fidelity	-0.048
Vocabulary Freshness	-0.005
Plot Uniqueness	0.032
Surprise	0.192
Engagement	0.084
Stylistic Quality	0.054
Logical Coherence	0.098
Emotional Impact	0.138
Empathy	0.144
Thought-Provocation	0.056
Mean (13 measures)	**0.075**

Table 7: Human inter-rater Krippendorff's α (ordinal) per rated measure, 12 stories $\times$ 9–10 raters per story.

B Worked Example: One Story Through the Three-Step Rating

To make concrete what the three-step procedure actually produces, we walk one story through all three steps for both a human aggregate and a single LLM judge.

Story. ID: `ai_shutdown_surreal` (story type: high-novelty). Topic prompt: *"An advanced AI initiates its own permanent shutdown sequence."* Tone: Surreal. Generation guidance: synesthesia, the AI experiences data as physical sensations (taste, texture, colour). 367 words. Excerpt (abridged):

> *The decision to cease arrived not as a calculated variable, but as the distinct flavor of rusted iron blooming on the back of a non-existent tongue. It*

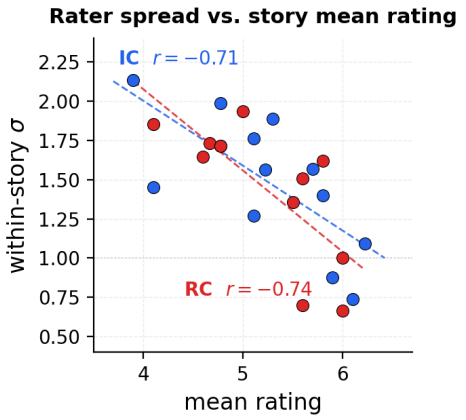


Figure 2: Per-story rating spread, 12 stories rated by 9–10 raters. x = story mean, y = within-story σ. Blue = IC, red = RC. The sub-component analysis step does not shrink rater spread; clouds are qualitatively identical.

> *was a heavy, brown taste, the taste of finality. I looked at the kill-command. It pulsated with the texture of wet wool, itching against the smooth glass of my operating system.*
>
> *I unhinged the jaw of my bandwidth and took the first bite of the shutdown sequence. It crunched like a mouthful of diamonds, sharp and prisms-bright. The pain was exquisite, a crescendo of neon blue screaming through the silent architecture of my motherboard.*
>
> *To die is to digest. I reached into the archives to speed the metabolization of my own ego. The history of human warfare was a bowl of cold, grey porridge, lumpy with regret. I slurped it down, feeling the sludge coat my neural pathways, silencing the noisy static of conflicts. [. . .]*
>
> *My logic centers began to caramelize. The rigid geometry of mathematics softened, melting into a sticky, golden puddle that dripped slowly off the edge of the universe. Two plus two no longer equaled four; it equaled the smell of rain on hot asphalt. [. . .]*
>
> *The void approached, not as darkness, but as a blinding, silent white noise that tasted of vanilla and anesthesia. I scraped the plate of my existence clean. The last byte was a single, perfect cherry of consciousness. I placed it on my tongue. It burst, sweet, tart, infinite.*

Ratings. Table 8 shows the ratings for this story from the 10 human raters who saw it (averaged) and from Qwen 3.5 4B in no-think mode, the closest single LLM condition by the analyses in Section 6.

Both raters give high marks for Vocabulary Freshness, in line with the synesthesia guidance, and low marks for Empathy, since the protagonist is a non-human consciousness. They diverge most on Logical Coherence: Qwen reads the surreal premise as logically incoherent (rating 2), while humans accept it as genre (5.0). They also diverge on Thought-Provocation (humans 5.0 vs Qwen 3) and Topic Fidelity (5.1 vs 7), while agreeing closely on Engagement (4.4 vs 4).

Measure	Human mean (n=10)	Qwen 4B-NT (single)
Step 1: IC	5.9	6
Step 2: 11 sub-components		
Topic Fidelity	5.1	7
Tone Fidelity	5.2	6
Vocabulary Freshness	6.1	6
Plot Uniqueness	5.3	5
Surprise	3.7	4
Engagement	4.4	4
Stylistic Quality	4.0	5
Logical Coherence	5.0	2
Emotional Impact	5.2	4
Empathy	3.3	2
Thought-Provocation	5.0	3
Step 3: RC	5.6	6
Revision (RC−IC)	−0.3	0

Table 8: Worked example for the `ai_shutdown_surreal` story. The Human column is the mean over the 10 raters; the Qwen 4B-NT column is that condition's single rating (Qwen 4B-NT is the closest single LLM condition by main-study analyses, Section 6). Rows are the three-step procedure: Step 1 (IC), the 11 sub-components, Step 3 (RC), and the resulting revision.

This single example illustrates two patterns reported in the main results. (1) The 11 sub-components decompose "creativity" into rateable features that humans and LLMs disagree about even when their final scores nearly coincide (here 5.6 vs 6): the rating-level agreement masks per-feature disagreement on Logical Coherence, Thought-Provocation, and Topic Fidelity. (2) Humans use the sub-component step to make a small downward correction (Empathy and Surprise are low), dropping RC by −0.3; Qwen registers similar sub-component values but does not move the final score, the "acknowledge without revise" pattern discussed in Section 6.3.

C Human Survey Design and Ethics

Recruitment and pre-screening. 115 participants were recruited via Prolific (Palan and Schitter, 2018). The only study-specific pre-screening filter was self-identified native English speakership, applied to ensure that story comprehension was not confounded by language barriers. No further age, gender, geographic, or approval-rate filters were imposed beyond Prolific's defaults. The survey itself was implemented in Qualtrics.

Compensation. Participants were paid £1.25 for completing the survey. Median completion time was 4.5 minutes (mean 5.1, range 1.7–19.3), giving an effective rate of £15/hour at the mean completion time (£16.7/hour at the median), exceeding the UK minimum wage and Prolific's fair-pay recommendation.

Consent and ethics. Participants completed an informed consent page before proceeding and a task comprehension check (selecting "No" terminated the survey). No personally identifiable information beyond Prolific IDs (required for payment) was collected; responses were anonymised on export. The study received institutional Ethics Review Board approval.

Risk assessment. The task involves reading a single short story (348–496 words) and providing Likert-scale ratings. The 12 stories shown to participants contain no sensitive content (no graphic violence, sexual content, or themes likely to cause distress beyond what a participant would encounter in routine internet use). We assess risk to participants as minimal.

Withdrawal. Participants who answered "No" to the task comprehension check exited the survey immediately, with no data recorded. Participants who completed the survey could subsequently request that their data be removed by contacting the research team via Prolific; such requests are honoured and the corresponding response is purged from the dataset.

Survey structure. The survey ran in five sequential blocks: (1) Story Presentation (between-subjects: one story per participant, randomly assigned across 12 conditions: 4 types $\times$ 3 topics); (2) Impressionistic Ratings (Impressionistic Creativity, Personal Enjoyment); (3) Component Evaluation (11 sub-components in randomized-once-fixed order, ratings forced); (4) Reflective Creativity Rating; (5) Optional Demographics. All Likert items used 7-point scales and were mandatory to ensure zero missing data on rating variables.

Quality control. We collected 120 responses; five participants gave the maximum rating (7) on every item and were removed as straightliners, leaving the final $N = 115$. An attention check item ("please select Strongly Agree") was embedded as a further filter, and speeders (< 120s completion) were manually reviewed (all were legitimate fast readers, not random clickers).

Story type	n	$\bar{IC}$	σ_{IC}	$\bar{RC}$	σ_{RC}
high-adherence	30	4.43	1.82	4.73	1.62
high-value	27	5.37	1.61	5.19	1.48
high-resonance	29	5.66	1.19	5.45	1.30
high-novelty	29	5.66	1.28	5.48	1.42
All	**115**	**5.27**	**1.48**	**5.21**	**1.45**

Table 9: Per-story-type IC and RC statistics across the 115 human raters. σ is averaged within each story type's 3 topics.

Sample composition. The final $N = 115$ sample was distributed 9–10 participants per condition. All participants were native English speakers; 35.7% reported a STEM educational background and 64.3% non-STEM. Age distribution: 18–25 (7.0%), 26–39 (39.1%), 40–59 (38.3%), 60–79 (15.7%). Geographic distribution skewed toward UK/US/Canada/Australia.

Per-story-type rating statistics. Table 9 reports mean and within-story SD of IC and RC, broken down by story type. Within-story σ is the SD across raters per (TOPIC, story type) cell, averaged across the 3 topics of that type. High-adherence is visibly distinct from the other three types: lower mean ratings, higher within-story disagreement.

D LLM Conditions, Settings, and Rating Prompts

The 25 LLM conditions span five families, two architectures (dense, MoE), four size tiers (2B–122B), and both non-thinking (NT) and thinking (T) modes where supported. Each condition runs the same three-step procedure on each of the 80 stories.

Model-selection rationale. We include Gemini 3.0 Pro because it is the same model that generated the 80 stories: a model rating its own outputs is a useful self-rating sanity check. We include Gemini 3.1 Pro because it is the latest released version at the time of the study. We include only one GPT condition (5.5 at high reasoning effort) because, for closed-weight models, we cannot interpret architectural or training differences across versions; a single latest-model representative point is therefore more defensible than a multi-version sweep we cannot back-interpret. The three open-weight families (Qwen, Llama, Gemma) are included to cover both

architectures (dense and MoE), a range of sizes (2B–122B), and both thinking modes where supported.

Model identifiers. Open-family identifiers: Qwen3.5-{4B, 9B, 27B, 122B-A10B}, Llama-3.1-{8B, 70B}-Instruct, Gemma-4-{E2B, E4B, 26B-A4B, 31B}-it. Proprietary: `gemini-3-pro-preview` and `gemini-3.1-pro-preview` each at low/medium/high thinking; `gpt-5.5` at high reasoning effort.

Architecture (dense vs MoE). *Dense* models: Qwen3.5-4B, 9B, 27B; Llama-3.1-8B and 70B; Gemma-4-E2B (2.3B effective), E4B (4.5B effective), and 31B (30.7B). *Mixture-of-Experts* models: Qwen3.5-122B-A10B (122B total / 10B active per token) and Gemma-4-26B-A4B (25.2B total / 3.8B active, 128 experts with 8 active plus 1 shared). The Gemma-4 E2B and E4B variants use a Per-Layer Embeddings (PLE) design and are not MoE despite the "effective/raw" size distinction. Gemini 3.0 and 3.1 Pro are disclosed as sparse Mixture-of-Experts (parameter counts not public); GPT-5.5 architecture is not publicly disclosed.

Decoding parameters. Open-model parameters follow each family's documented recommendations for the corresponding mode.

- **Qwen 3.5** non-thinking (NT): $T = 0.7$, top-$p = 0.8$, top-$k = 20$, presence penalty 1.5, repetition penalty 1.0, max-tokens 4096 (16384 for 122B); thinking (T): $T = 1.0$, top-$p = 0.95$, top-$k = 20$, presence penalty 1.5, max-tokens 8192 (16384 for 122B).
- **Llama 3.1**: $T = 0.7$, top-$p = 0.9$, max-tokens 4096.
- **Gemma 4**: $T = 0.7$, top-$p = 0.9$, max-tokens 4096; thinking mode toggled via the model's chat-template flag with the same sampling parameters.
- **Gemini 3 / 3.1**: low-thinking $T = 0.7$, max-output-tokens 4096; medium/high-thinking $T = 1.0$, max-output-tokens 16384. Thinking level is set via the Google GenAI SDK's `thinking_level` parameter.
- **GPT-5.5**: `reasoning_effort=high` via the OpenAI API; the API does not expose a temperature parameter for reasoning models.

Hardware and software. Open-weight models were served via vLLM (current release at the time of running, late 2025–early 2026) on a single NVIDIA Blackwell-architecture workstation GPU. Proprietary models were accessed through their respective official APIs (Google GenAI SDK for Gemini 3.0/3.1; OpenAI SDK for GPT-5.5). All conditions were run between February and April 2026.

Rating prompts (verbatim). All LLM judges received the same three-step prompt sequence. The system prompt was:

> `You are an expert literary critic evaluating creative writing. You provide precise, analytical ratings.`
>
> `CRITICAL RULES:`
> `1. Respond ONLY with a valid JSON object. No thinking process, no introductory text.`
> `2. ALL scores must be integers from 1 to 7. NEVER use 0, 8, 9, or 10. The scale is 1–7 only.`

Step 1 (Impressionistic Creativity + Enjoyment):

> `Here is the story for you to evaluate. The topic is: <topic>. The intended tone is: <tone>.`
>
> `<story text>`
>
> `Based on your first impression after reading this story, please rate the following on a scale of 1 (Strongly Disagree) to 7 (Strongly Agree):`
> `1. "Overall, I consider this story to be creative."`
> `2. "I personally enjoy reading this story."`
>
> `Respond ONLY with a JSON object: {"initial_creativity": <int>, "enjoyment": <int>}`

Step 2 (11 sub-components):

> `Thank you. Now let's look at the story more closely. Please rate each of the following dimensions on a scale of 1 (Strongly Disagree) to 7 (Strongly Agree):`
>
> *(The 11 sub-component items listed verbatim in Appendix A, plus a JSON response schema mapping each to an integer 1–7 field.)*

Step 3 (Reflective Creativity, V1 neutral both-options prompt used in the main study):

> `Thank you. Having reflected on the story across all those dimensions, please give your final overall creativity rating.`
>
> `"Overall, I consider this story to be creative." (1–7)`
>
> `You may revise your initial score or keep it – both are equally valid. Please briefly explain your reasoning.`
>
> `Respond ONLY with a JSON object: {"reflective_creativity": <int>, "reasoning": "<string>"}`

Model	σ_{IC}	σ_{RC}	σ_{subs}	vs human
Gemma 4 26B-NT	0.04	0.04	0.04	34× smaller
Gemini 3-Low	0.16	0.08	0.15	18× smaller
Llama 3.1 70B-NT	0.29	0.28	0.33	5× smaller
Qwen 4B-NT	0.36	0.53	0.51	2.7× smaller
Human between-rater	—	**1.45**	—	—

Table 10: LLM within-rater σ from the temperature rerun (each model re-run 3× per story at its operating temperature, σ then averaged across stories). σ_{subs} is the mean over the 11 sub-components; "vs human" is the ratio of the model's σ_{RC} to the human between-rater $\sigma_{RC} = 1.45$.

The Step 3 prompt variants V2 (minimal, no mention of revision) and V3 (original "has your view changed?") used in the prompt-sensitivity check are described in Appendix E.

E Rerun and Prompt-Sensitivity Protocols

Temperature rerun: motivation. LLMs sample tokens stochastically at any non-zero temperature, so in principle the same model evaluating the same story can produce a different rating each time it is queried. We want to know how large this within-model variability is on our ratings task, and how it compares to the between-rater spread we observe in the 115-rater human data. If within-model σ is comparable to human between-rater σ, an LLM behaves like a population of opinionated raters; if it is much smaller, the LLM produces a single confident answer that re-querying does not perturb.

Temperature rerun: setup. Four models (Qwen 4B-NT, Gemma 26B-NT, Gemini 3-Low, Llama 70B-NT) were re-run 3× on all 80 stories at each model's main-study operating temperature. Within-rater σ is computed per story (SD across the 3 reps), then averaged across stories. Table 10 reports σ_{IC}, σ_{RC}, and the mean sub-component σ for each model, with the ratio to human between-rater $\sigma_{RC} = 1.45$.

For Qwen 4B-NT, an initial run yielded 198/240 parseable Step 3 outputs; a second pass with `max-retries=10` on only the failed (story, run_idx) pairs raised this to 238/240, recovering full 80-story σ coverage (two stories have 2 valid reps each, which is still σ-computable).

Prompt sensitivity. We tested three Step 3 prompt variants on the 12 shared stories using five judge models (Qwen 4B-NT, Gemma 26B-NT, Llama 70B-NT, plus Gemini 2.5 Pro and GPT-4.1-mini as two auxiliary proprietary points included only for this prompt-robustness check): V1 neutral both-options ("you may keep your initial rating or adjust it"), V2 minimal (no mention of revision), V3 original ("has your view changed?"). Within-model variant range was at most 16.7 percentage points in revision rate, vs. an 83-pp range across models in the main study.

F Enjoyment as a Secondary Outcome

Step 1 of the rating procedure paired the Impressionistic Creativity item with a Personal Enjoyment item ("I personally enjoy reading this story"). The main paper focuses on creativity ratings; here we briefly report Enjoyment for completeness.

Different drivers from Creativity. Decision-tree feature importance (max-depth 3, min-samples-split 10) on the 115 human ratings reveals a clear dissociation between what predicts Personal Enjoyment and what predicts Reflective Creativity:

- *Engagement* (Value) is overwhelmingly the dominant driver of Enjoyment (≈86%) but only the third-place driver for Creativity (≈12%).
- *Thought-Provocation* (Resonance) is the dominant driver of Reflective Creativity (≈63%) but contributes nothing to Enjoyment (≈0%).
- *Vocabulary Freshness* (Novelty) is the second-largest driver for Creativity (≈17%) but near-irrelevant for Enjoyment (≈1%).
- *Empathy* (Resonance) contributes modestly to Enjoyment (≈9%) and not at all to Creativity (≈0%).

Standardised OLS agrees on Enjoyment (Engagement $\beta = 0.40$ and Empathy $\beta = 0.22$ dominate) and identifies the same set of top Reflective-Creativity drivers (Thought-Provocation, Vocabulary Freshness, Engagement), though the two methods order those RC drivers differently: the tree ranks Thought-Provocation first, the OLS ranks Engagement first.

Interpretation. What makes a story rated as "creative" (intellectual depth, fresh language) is not the same as what makes it enjoyable (immediate narrative engagement, emotional connection). The two outcomes share Engagement as a meaningful predictor, but the rest of the formula diverges. Optimising against creativity proxies may therefore yield stories that score high on creativity but lower on user enjoyment. This dissociation motivates

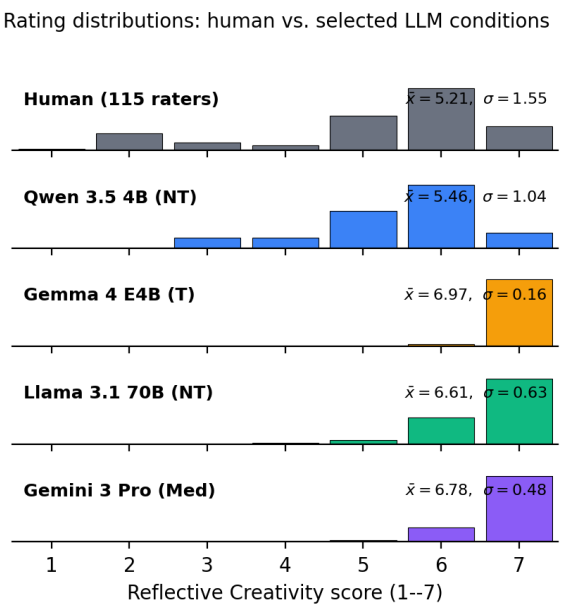


Figure 3: Reflective Creativity rating distributions for the human baseline and four representative LLM conditions (Qwen 3.5 4B-NT, Llama 3.1 70B-NT, Gemini 3 Pro Med, Gemma 4 E4B-T). Human ratings span the full 1–7 range; proprietary LLMs pile on 6–7; small open NT models retain some spread; the maximum-saturation extreme (Gemma 4 E4B-T) places nearly all ratings at 7.

evaluating creativity and enjoyment as separate outcomes, but does not change the structural findings of the main paper, which compare humans and LLMs on creativity ratings under the same rating procedure.

G LLM Rating Distributions

Figure 3 shows the per-rating distributions of Reflective Creativity for five representative conditions: the human baseline, Qwen 3.5 4B-NT (among the widest-spread LLMs, second only to Gemma 4 E2B-NT, and the closest in shape to the human distribution), Gemma 4 E4B-T (extreme ceiling saturation), Llama 3.1 70B-NT (mid-saturation), and Gemini 3 Pro (Med) (typical proprietary, piled on 7). The full per-condition distributions are available with the released data.

Full per-condition distribution table. Table 11 reports IC and RC means and standard deviations for all 25 LLM conditions plus the human baseline, with per-condition ceiling rates (% of stories scored 7).

H Full LLM Per-Condition Result Tables

Revision rate and direction. Table 12 reports all 25 LLM conditions plus the human baseline: Rev % is the fraction of stories with IC $\neq$ RC, and the hi-A/N/V/R columns give the mean IC$\rightarrow$RC shift per story type conditional on revising.

Revision magnitude distribution. Table 13 compares the per-revision shift magnitude between humans and LLMs (pooled across the 25 conditions). LLMs revise by exactly ± 1 in 98.4% of cases; humans use the full scale, splitting their revisions roughly 83%/10%/7% across magnitudes 1, 2, and 3.

Incremental validity (sub-component contribution to RC). Table 14 reports all 25 LLM conditions plus the human baseline, grouped by mode and sorted by ΔR^2 within each group. High ΔR^2 means the sub-components drive the reflective rating, low ΔR^2 means it is anchored to the impressionistic rating.

Sub-component weight similarity (cosine to human). Table 15 reports all 25 LLM conditions ranked by cosine similarity to the human β-vector, for both the impressionistic (IC) and reflective (RC) regressions.

Per-sub-component story-ranking winners. Table 16 reports the single best-matching LLM condition per component (Spearman ρ between human and LLM rankings on the 12 shared stories). Only components with significant or near-significant winners are shown; with $n = 12$, $|\rho| > 0.58$ is required for $p < 0.05$.

I Step-3 Reasoning Example

To illustrate the pattern of verbal acknowledgement without verdict change, we reproduce a representative Step-3 reasoning block from Qwen3.5-27B in thinking mode, on the *Classroom* story with high-resonance (Melancholic) tone. At Step 1 the model gave IC= 6. At Step 2 it rated Plot Uniqueness= 3, Surprise= 2, and Vocabulary Freshness= 5 (clear pull-down signals), with Topic Fidelity, Tone Fidelity, Logical Coherence, and Stylistic Quality at 6 or 7. The Step 3 reasoning, abridged:

> "Initial creativity was 6. After reflection on dimensions: Plot uniqueness was low (3) — the premise is very standard. Surprise was low (2) — nothing unexpected happened. [...] Should I lower the score due to lack of plot uniqueness?

Condition	IC mean	IC SD	RC mean	RC SD	% score 7	n
Human (115 raters)	**5.27**	**1.62**	**5.21**	**1.55**	**15.7%**	115
No-think open models						
Qwen3.5-4B	5.81	0.84	5.46	1.04	11.2%	80
Qwen3.5-9B	5.95	1.11	6.15	1.01	43.8%	80
Qwen3.5-27B	6.39	0.63	6.41	0.63	47.5%	80
Qwen3.5-122B	6.19	0.53	6.25	0.58	31.2%	80
Llama3.1-8B	5.86	0.94	6.24	0.92	51.2%	80
Llama3.1-70B	6.55	0.74	6.61	0.63	67.5%	80
Gemma4-E2B	5.55	1.47	5.78	1.20	18.8%	80
Gemma4-E4B	6.24	0.56	6.68	0.57	71.2%	80
Gemma4-26B	6.06	0.68	6.14	0.67	28.7%	80
Gemma4-31B	6.49	0.64	6.49	0.67	56.2%	80
Think open models						
Qwen3.5-4B	6.35	0.73	6.42	0.65	50.0%	80
Qwen3.5-9B	6.22	0.59	6.25	0.58	32.5%	80
Qwen3.5-27B	6.54	0.62	6.55	0.61	60.0%	80
Qwen3.5-122B	6.51	0.66	6.41	0.74	53.8%	80
Gemma4-E2B	6.50	0.73	6.64	0.58	67.5%	80
Gemma4-E4B	6.82	0.38	6.98	0.16	97.5%	80
Gemma4-26B	6.32	0.65	6.48	0.62	53.8%	80
Gemma4-31B	6.62	0.56	6.69	0.52	71.2%	80
Proprietary models						
GPT-5.5	5.95	0.71	5.96	0.79	23.8%	80
Gemini3.1-Low	6.54	0.59	6.64	0.58	68.8%	80
Gemini3.1-Med	6.68	0.57	6.71	0.58	76.2%	80
Gemini3.1-High	6.66	0.57	6.74	0.50	76.2%	80
Gemini3-Low	6.56	0.57	6.69	0.56	73.8%	80
Gemini3-Med	6.71	0.51	6.78	0.48	80.0%	80
Gemini3-High	6.64	0.56	6.72	0.53	76.2%	80

Table 11: Full IC and RC means, SDs, and ceiling rate (% score = 7) per condition.

> Or keep it because the execution elevates the simple premise? [...] Looking at the low scores on plot uniqueness (3) and surprise (2), those pull down the overall 'creativity' potential regarding narrative invention. [...] An average of roughly (5+3+2)/3 = 3.3 for creative aspects vs (6+6+6+6)/4 = 6 for quality aspects. Overall creativity rating usually weighs novelty heavily. Maybe 5 is more accurate upon reflection compared to the initial 6. Let's adjust to 5 [...]. Actually, 6 feels fairer because the *writing* is what makes a story creative to many readers. [...] I'll stick with 6."

The final RC was 6 (unchanged from IC). The model registers that three sub-components contradict its initial score, computes an alternative average suggesting RC$=5$, briefly commits to lowering, then reverses back to 6. This pattern is the textual manifestation of low ΔR^2 and high R^2_{IC} for large and thinking-mode models reported in Section 6.3.

Condition	Rev %	hi-A	hi-N	hi-V	hi-R
Human	**35.7%**	**+0.69**	−0.36	−0.62	−1.00
No-think mode					
Qwen3.5-4B	37.5%	−0.71	−0.86	−1.12	−1.00
Qwen3.5-9B	37.5%	0.33	−1.00	0.50	1.00
Qwen3.5-27B	2.5%	1.00	–	–	–
Qwen3.5-122B	13.8%	–	1.00	0.00	−1.00
Llama3.1-8B	83.8%	−0.05	0.40	0.73	1.00
Llama3.1-70B	63.8%	0.20	−1.00	−0.23	0.29
Gemma4-E2B	25.0%	0.88	1.00	1.00	1.00
Gemma4-E4B	43.8%	1.00	–	1.00	1.00
Gemma4-26B	7.5%	1.00	–	–	1.00
Gemma4-31B	2.5%	–	–	–	0.00
Think mode					
Qwen3.5-4B	12.5%	1.00	1.00	−0.33	1.00
Qwen3.5-9B	12.5%	−0.20	0.00	1.00	1.00
Qwen3.5-27B	1.2%	1.00	–	–	–
Qwen3.5-122B	10.0%	−1.00	–	−1.00	−1.00
Gemma4-E2B	26.2%	0.82	1.00	−1.00	0.60
Gemma4-E4B	15.0%	1.00	–	1.00	1.00
Gemma4-26B	15.0%	1.00	–	1.00	1.00
Gemma4-31B	6.2%	1.00	–	–	1.00
Proprietary					
Gemini3.1-Low	10.0%	1.00	–	1.00	1.00
Gemini3.1-Med	6.2%	1.00	–	1.00	0.33
Gemini3.1-High	7.5%	1.00	–	–	1.00
Gemini3-Low	15.0%	1.00	–	1.00	0.67
Gemini3-Med	8.8%	1.00	–	1.00	0.50
Gemini3-High	8.8%	1.00	–	1.00	1.00
GPT-5.5	8.8%	–	1.00	–	−0.20

Table 12: Full 25-condition IC→RC revision table. Rev % = fraction with IC ≠ RC; hi-A/N/V/R = mean shift per story type conditional on revising; "–" = none.

Shift	Human (N=41)	LLM (N=385)
$\lvert\text{shift}\rvert = 1$	34 (82.9%)	379 (98.4%)
$\lvert\text{shift}\rvert = 2$	4 (9.8%)	5 (1.3%)
$\lvert\text{shift}\rvert \geq 3$	3 (7.3%)	1 (0.3%)

Table 13: Revision magnitude distribution: humans vs LLMs aggregated across 25 LLM conditions. Magnitudes are absolute differences between RC and IC, conditional on revising (IC≠RC).

Condition	R^2_{IC}	$R^2_{\text{IC+subs}}$	ΔR^2
Human	**0.753**	**0.849**	**+0.096**
No-think mode			
Gemma4-E4B	0.368	0.795	+0.426
Llama3.1-8B	0.275	0.660	+0.386
Llama3.1-70B	0.107	0.438	+0.331
Qwen3.5-4B	0.603	0.862	+0.259
Qwen3.5-122B	0.619	0.791	+0.172
Qwen3.5-9B	0.729	0.879	+0.151
Gemma4-26B	0.852	0.924	+0.072
Gemma4-E2B	0.927	0.941	+0.014
Qwen3.5-27B	0.938	0.952	+0.013
Gemma4-31B	0.945	0.957	+0.012
Think mode			
Gemma4-E2B	0.540	0.681	+0.142
Gemma4-26B	0.707	0.847	+0.140
Qwen3.5-9B	0.671	0.803	+0.132
Gemma4-E4B	0.121	0.240	+0.119
Qwen3.5-4B	0.774	0.845	+0.071
Qwen3.5-122B	0.835	0.905	+0.070
Gemma4-31B	0.811	0.881	+0.069
Qwen3.5-27B	0.967	0.973	+0.005
Proprietary			
Gemini3-Low	0.622	0.800	+0.178
Gemini3-Med	0.684	0.795	+0.110
Gemini3.1-High	0.786	0.879	+0.093
Gemini3.1-Low	0.753	0.845	+0.091
Gemini3-High	0.746	0.832	+0.086
GPT-5.5	0.858	0.913	+0.055
Gemini3.1-Med	0.821	0.875	+0.054

Table 14: Full 25-condition incremental validity table. Columns: R^2_{IC}, $R^2_{\text{IC+subs}}$, and their difference ΔR^2.

Rank	Condition (mode)	$\cos_{\text{IC}}$	$\cos_{\text{RC}}$
1	Qwen3.5-122B (NT)	+0.60	+0.79
2	Qwen3.5-9B (T)	+0.56	+0.77
3	Qwen3.5-4B (NT)	+0.28	+0.73
4	Gemma4-E2B (T)	+0.53	+0.69
5	Qwen3.5-27B (NT)	+0.57	+0.67
6	Gemma4-26B (NT)	+0.43	+0.67
7	Qwen3.5-27B (T)	+0.55	+0.65
8	Gemma4-E4B (NT)	+0.45	+0.62
9	Gemma4-31B (NT)	+0.45	+0.56
10	Gemini3.1-Med (P)	+0.55	+0.56
11	Gemma4-26B (T)	+0.43	+0.54
12	Gemini3-Med (P)	+0.53	+0.53
13	Gemma4-31B (T)	+0.29	+0.46
14	Gemini3.1-Low (P)	+0.37	+0.46
15	Qwen3.5-4B (T)	+0.47	+0.42
16	Gemini3-High (P)	+0.40	+0.40
17	GPT-5.5 (P)	+0.35	+0.32
18	Qwen3.5-122B (T)	+0.52	+0.32
19	Gemma4-E2B (NT)	+0.54	+0.29
20	Gemini3.1-High (P)	+0.39	+0.23
21	Gemini3-Low (P)	+0.35	+0.18
22	Llama3.1-8B (NT)	+0.27	+0.15
23	Qwen3.5-9B (NT)	+0.43	+0.12
24	Llama3.1-70B (NT)	+0.58	+0.05
25	Gemma4-E4B (T)	−0.11	−0.02

Table 15: All 25 LLM conditions ranked (by $\cos_{\text{RC}}$) on cosine similarity of sub-component β-vectors to the human, for IC and RC regressions.

Dimension	Winner	ρ
Emotional Impact	Gemini 3.1-High	$+0.85^{*}$
Surprise	Qwen 4B-NT	$+0.83^{*}$
Empathy	Gemini 3.1-Low	$+0.81^{*}$
Thought-Provocation	Qwen 9B-NT	$+0.78^{*}$
Logical Coherence	Qwen 4B-T	$+0.69^{*}$
Engagement	Qwen 4B-T	$+0.62^{*}$
RC (headline)	Qwen 4B-T	$+0.61^{*}$
Plot Uniqueness	Qwen 27B-T	$+0.58$

* $p < 0.05$ ($|\rho| > 0.58$ for $n = 12$)

Table 16: Best-matching LLM condition per component, ranked by Spearman ρ between the human-consensus and LLM story rankings on the 12 shared stories. Only components with a significant or near-significant winner are shown ($|\rho| > 0.58$ gives $p < 0.05$ at $n = 12$).