Creativity Is Not Enjoyment: Rethinking Human Evaluation of AI Story Generation

Pharath Sathya, Yin Jou Huang, Fei Cheng
Graduate School of Informatics, Kyoto University, Kyoto, Japan
{pharath, huang, feicheng}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Large language models (LLMs) are often optimized to generate creative text, yet it remains unclear whether creativity translates to user satisfaction. We propose a framework that evaluates creativity and enjoyment as separate dimensions. Through a controlled study with diverse AI-generated stories, we show that creativity judgments rely primarily on novelty, whereas enjoyment depends on emotional resonance. Optimizing for novelty alone increases perceived creativity but can reduce user satisfaction, revealing a fundamental trade-off in current generation methods.

1 Introduction

Current approaches to evaluating AI-generated text often treat text quality objectively, implicitly assuming that creative outputs naturally satisfy users. However, creativity and user enjoyment may be driven by fundamentally different features. Novelty-focused generation methods, such as temperature sampling and diverse decoding [1] aim to increase creativity by producing surprising, unconventional text. Yet these same methods may fail to improve user enjoyment if they neglect features that drive enjoyment.

We investigate whether creativity and enjoyment represent orthogonal evaluation dimensions. Building on our prior framework validation work [2], which demonstrated that creativity judgments follow a hierarchical rather than cumulative evaluation process, the present study extends this analysis to examine the *relationship* between creativity and enjoyment. Using a controlled human evaluation of AI-generated stories, we decompose text quality into four components: Novelty, Resonance, Value, and Adherence. We examine which components predict creativity versus enjoyment, how reflection affects this relationship, and whether different story tones (genre) moderate the alignment between these constructs.

Our findings reveal that creativity and enjoyment rely on largely non-overlapping features. Novelty drives creativity judgments but contributes minimally to enjoyment, which instead depends on emotional resonance and perceived value. This misalignment creates a fundamental trade-off: optimizing generation objectives for creativity can inadvertently reduce user satisfaction. These results challenge common assumptions in LLM training and evaluation, suggesting that effective text generation requires explicitly modeling creativity and enjoyment as separate, interacting objectives.

2 Methodology

2.1 Evaluation Framework

We employ a validated four-component evaluation framework for AI story generation [2], refined from Jordanous and Keller's theoretical model of computational creativity [3] and informed by cognitive theories of creative judgment [4]. The framework decomposes creative quality into four orthogonal dimensions with 11 granular sub-components:

- **Novelty** (Vocabulary Freshness, Plot Uniqueness, Surprise): Captures originality and unexpectedness.
- **Resonance** (Emotional Impact, Empathy, Thought-Provocation): Captures affective and intellectual engagement.
- **Value** (Engagement, Stylistic Quality, Logical Coherence): Captures technical craftsmanship and narrative quality.
- **Adherence** (Topic Fidelity, Tone Fidelity): Captures constraint satisfaction and instruction following.

This decomposition allows us to test which components predict creativity versus enjoyment, and whether these constructs prioritize different dimensions of text quality.

2.2 Controlled Story Generation

We generated 12 short stories (350–450 words) using Gemini 3.0 Pro (Preview) [5]. To systematically manipulate creative dimensions, we applied **Spike Prompting**, a controlled generation strategy that maximizes one target dimension while constraining others. Each dimension was operationalized via a distinct narrative tone:

- **Surreal tone** (maximize Novelty): Dream logic, unconventional structure
- **Melancholic tone** (maximize Resonance): Emotional depth and sensory immersion
- **Witty tone** (maximize Value): Linguistic sophistication and wordplay
- **Clinical tone** (maximize Adherence): Objective, formal, and logically constrained prose

The four tones were crossed with three topics (*AI Shutdown, The Heist, The Midnight Store*), yielding 12 unique stories and supporting domain generalization. Readability was controlled across conditions (Flesch–Kincaid Mean = 7.7, Range = 6.7–9.0), ensuring that observed differences reflect stylistic evaluation rather than comprehension difficulty (see Appendix A).

2.3 Human Evaluation Design

We conducted a crowdsourced human evaluation (N = 115) in which each participant evaluated a single story. The evaluation protocol was structured to separate immediate affective judgments from reflective analytical judgments.

Participants first provided holistic ratings of **Creativity** and **Enjoyment** immediately after reading. They then evaluated individual story components based on the 11 subcomponents, followed by a final reflective creativity rating after analytical decomposition. This design enables comparison between intuitive and reflective creativity judgments and their relationship to enjoyment.

Full survey materials, recruitment procedures, compensation, and ethical details are provided in Appendix B.

2.4 Measurement Validation

We verified that the proposed components capture distinct evaluative dimensions rather than a single latent ‘‘quality’’ factor using reliability and manipulation checks.

Tone	**Adherence**	**Resonance**	**Novelty**	**Value**
Clinical	**5.95***	**3.54***	4.46	4.34
Melancholic	5.88	**5.03***	4.59	5.15
Surreal	5.10	4.16	**5.29***	4.34
Witty	5.48	4.54	4.70	4.93

*Indicates targeted construct for each condition.

Table 1 Construct means by tone condition. Diagonal entries indicate targeted constructs.

Internal Consistency. All constructs exhibited acceptable reliability (Cronbach's α): Resonance ($\alpha = 0.82$), Value ($\alpha = 0.80$), Novelty ($\alpha = 0.76$), and Adherence ($\alpha = 0.69$).

Manipulation Check. Spike Prompting successfully isolated the intended dimensions. As shown in Table 1, each tone condition achieved its highest mean on its targeted construct (diagonal entries), producing clearly divergent profiles across Adherence, Resonance, Novelty, and Value. This pattern supports discriminant validity and indicates that evaluations are not driven by a unidimensional quality judgment.

3 Results

3.1 Creativity and Enjoyment are Correlated but Distinct

Initial creativity ratings showed strong correlation with enjoyment ($r = 0.757$, $p < 0.001$, $N = 115$), which weakened after reflection ($r = 0.709$, $\Delta r = 0.048$, $p = 0.092$). Median-split categorization revealed 17.4% discordance, with a striking 4:1 asymmetry ($\chi^2(1) = 7.2$, $p = 0.007$): responses were four times more likely to be ‘‘creative but not enjoyable’’ (13.9%) than ‘‘enjoyable but not creative’’ (3.5%). This asymmetry suggests that creativity as judged after reflective evaluation does not reliably translate into user enjoyment (see Appendix C).

3.2 Different Components Drive Enjoyment and Creativity

We fit separate multiple regression models predicting Enjoyment and Creativity from the same 11 components (both $R^2 > 0.65$, $p < 0.001$). Despite similar overall fit, the resulting feature weights differ substantially. In particular, the correlation between standardized regression weights was near zero ($r = -0.057$, $p = 0.868$; Figure 1), indicating that enjoyment and creativity rely on different evaluative cues.

Table 2 highlights the largest parameter-level differences. Topic Adherence, Engagement, Style, and Logi-

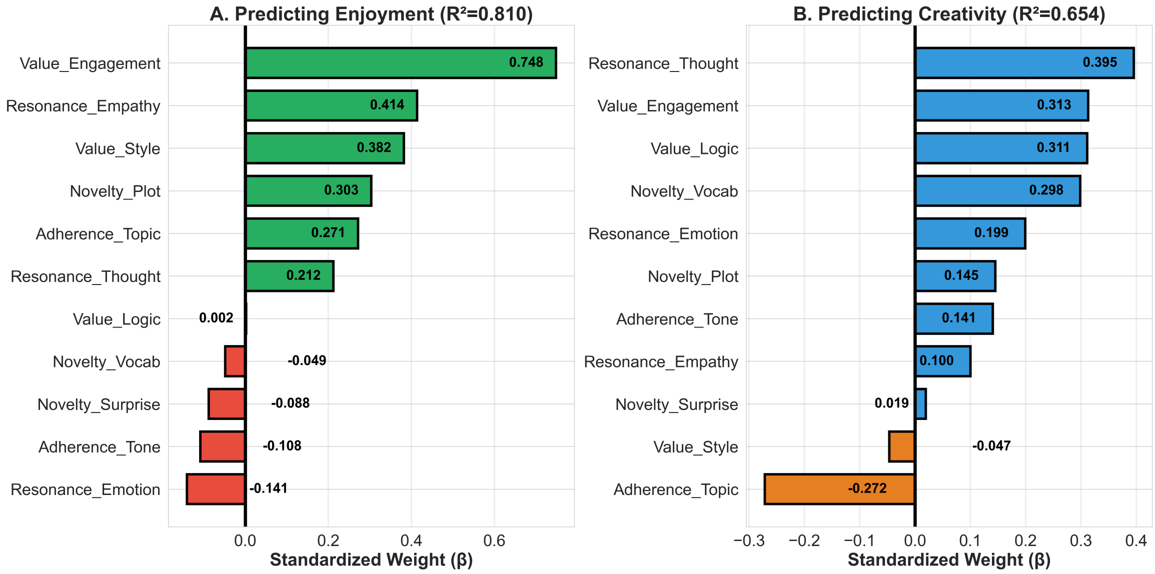

Figure 1 Orthogonality of evaluation: standardized regression weights for enjoyment vs. creativity show near-zero correlation ($r = -0.057$), proving these constructs depend on different features.

Parameter	$\beta^{(E)}$	$\beta^{(C)}$	$\Delta\beta$
Enjoyment-favoring parameters			
Topic Adherence	**+0.621**	+0.078	+0.543
Engagement	**+0.503**	+0.068	+0.435
Style	**+0.485**	+0.057	+0.428
Logical Flow	**+0.412**	+0.158	+0.254
Creativity-favoring parameters			
Vocabulary	+0.089	**+0.436**	-0.347
Emotional Impact	+0.112	**+0.452**	-0.340
Plot Novelty	+0.145	**+0.398**	-0.253
Surprise	+0.178	**+0.389**	-0.211

Table 2 Standardized regression weights for enjoyment ($\beta^{(E)}$) and creativity ($\beta^{(C)}$) prediction. Each β indicates the expected change (in standard deviations) in the predicted score for a one-standard-deviation increase in the corresponding component, holding other components constant. Bold values indicate the dominant predictor for each parameter. $\Delta\beta = \beta^{(E)} - \beta^{(C)}$.

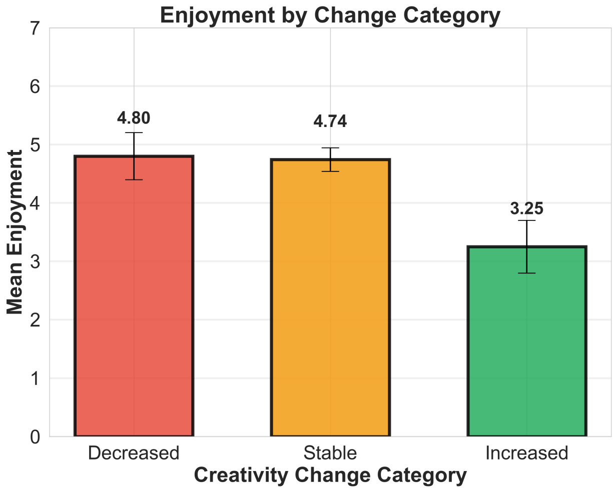

Figure 2 Reflection-induced increases in creativity ratings are associated with lower enjoyment.

cal Flow strongly predict enjoyment but contribute little to creativity. In contrast, Vocabulary, Emotional Impact, Plot Novelty, and Surprise strongly predict creativity while having limited impact on enjoyment.

At the component level, the same pattern holds: Novelty favors creativity, whereas Value favors enjoyment. These results demonstrate that enjoyment and creativity are supported by largely non-overlapping feature sets (see Appendix D).

3.3 Reflection Increases Creativity Judgments but Reduces Enjoyment

Participants who *increased* creativity ratings after reflection reported significantly *lower* enjoyment than those with stable ratings (Increased: $M = 3.25$; Stable: $M = 4.74$; $F(2, 112) = 4.784$, $p = 0.010$; Figure 2), indicating that higher reflective creativity judgments do not translate into increased enjoyment.

Component analysis shows no group difference in Novelty ($F = 0.15$, ns), whereas Resonance ($F = 8.45$, $p < 0.001$), Value ($F = 12.3$, $p < 0.001$), and Adherence ($F = 6.82$, $p = 0.002$) are significantly lower in the Increased group. Accordingly, the Novelty–Resonance gap is larger for Increasers ($M = +1.23$) than for the Stable group ($M = +0.24$, $p = 0.009$), reflecting content perceived as novel but weakly engaging.

Composite scores, computed by averaging standard-

Component	a	b	ab	% Med.	Sobel Z	p
Value	0.492	0.709	**0.349**	**58.3%**	5.893	**<0.001**
Resonance	0.556	0.529	**0.294**	**49.1%**	5.256	**<0.001**
Adherence	0.389	0.462	**0.180**	**30.0%**	4.044	**<0.001**
Novelty	0.478	0.163	0.078	13.0%	1.903	0.057
Direct Effect			0.195	32.5%		

Table 3 Mediation of the Creativity → Enjoyment relationship. Bold values indicate substantial and statistically significant mediation effects.

ized component ratings, mirror this pattern: the Increased group scores lower on analytical coherence (3.72 vs. 5.64, $p = 0.002$) and emotional response (3.22 vs. 4.47, $p = 0.012$), confirming that reflection emphasizes novelty over resonance (see Appendix E).

3.4 Which Creative Components Drive Enjoyment?

Although creativity and enjoyment are correlated ($r = 0.757$), this association operates through specific components. We therefore examine which aspects of creative content transmit creativity judgments into enjoyment using mediation analysis, where a denotes the effect of Creativity on each component, b the effect of the component on Enjoyment, and ab the indirect effect. As shown in Table 3, Value and Resonance are the primary mediators, accounting for **58.3%** and **49.1%** of the total effect (both $p < 0.001$). Adherence provides a moderate pathway (**30.0%**). In contrast, Novelty mediates only **13.0%** of the relationship and is not significant ($p = 0.057$), indicating that novelty contributes little to enjoyment despite driving creativity judgments (see Appendix F).

3.5 When Does Creativity Align with Enjoyment?

The relationship between creativity and enjoyment depends on the specific tones the story is generated with. Melancholic stories show strong alignment between creativity and enjoyment feature weights ($r = +0.87$, $p = 0.001$), whereas Surreal stories show little alignment ($r = -0.18$, ns), indicating that emotionally grounded genres couple creativity with enjoyment through Resonance, while reality-bending genres emphasize Novelty. Witty ($r = 0.42$) and Clinical ($r = 0.35$) exhibit intermediate alignment (see Appendix G).

4 Discussion

Our results demonstrate that creativity and enjoyment rely on different evaluation cues in AI-generated text. While prior work [2] showed that creativity is judged hierarchically, the present study shows that creativity and enjoyment are *distinct constructs* with divergent predictors. Creativity is driven by novelty-related features such as surprise and lexical diversity, whereas enjoyment depends on emotional resonance and perceived value. The near-zero correlation between regression weights ($r = -0.057$) confirms that these evaluations capture different aspects of text quality: features that raise creativity often fail to improve enjoyment.

This distinction has direct implications for Natural Language Processing (NLP) systems. Treating text quality as a single objective, common in current LLM training and evaluation [6] assumes that optimizing creativity increases user satisfaction. Our findings challenge this assumption. Novelty-focused decoding and training strategies raise creativity scores but do not reliably improve enjoyment; in our experiments, higher temperature increased creativity while reducing enjoyment. Effective generation therefore requires balancing novelty with resonance and value, rather than optimizing novelty alone.

Evaluation design further modulates this relationship. Reflective judgments emphasize novelty cues, whereas immediate responses better capture affective enjoyment. Story tone also moderates alignment: emotionally grounded stories tend to couple creativity and enjoyment through resonance, while surreal content emphasizes novelty and separates the two. Together, these findings indicate that creativity and enjoyment should be evaluated and optimized as distinct but interacting dimensions of text quality.

Conclusion

Creativity and enjoyment emerge as orthogonal constructs requiring separate optimization. NLP practices that rely on novelty-based creativity proxies fail to enhance—and may reduce user enjoyment. Effective text generation therefore calls for multi-objective frameworks that jointly model Novelty and Resonance, alongside evaluation protocols aligned with deployment context. These findings challenge prevailing assumptions in creativity-oriented LLM evaluation and motivate a shift toward multi-dimensional models of text quality.

Limitations

This study is limited to English-language stories evaluated by English-speaking participants from North America and the UK, which may restrict cross-cultural generalizability. We evaluate only a single model (Gemini 3.0 Pro (Preview)) and a limited set of story prompts spanning three genres (science fiction, mundane realism, action). The analysis is correlational and based on a single reflective evaluation step; future work should test additional models, languages, genres, and experimental manipulations to establish causal effects.

Acknowledgments

This work was supported by the Natural Language Processing Laboratory of the Graduate School of Informatics, Kyoto University, which provided experimental facilities, equipment, funding for the crowdsourcing and API usage. We express our deepest gratitude to Assistant Professor Yin Jou Huang for her comprehensive mentorship throughout this project—from the initial research conceptualization and dataset generation to the survey execution and writing. We thank Senior Lecturer Fei Cheng for his instrumental help in refining the paper's methodology and discussion. We also appreciate the critical inputs and advice provided by Associate Professors Chenhui Chu and Yugo Murawaki. Furthermore, we thank Samuel Willey (University of Florida) for his assistance in refining the survey questions, and YeahJun Heo, Pyii Pyo Maung and Zachariah Sueyoshi (Kyoto University) for their constructive comments on the survey design. Finally, we are grateful to the anonymous participants recruited via Prolific for their time and contribution to the evaluation study.

References

[1]Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

[2]Pharath Sathya, Yin Jou Huang, and Fei Cheng. Evaluation framework for ai creativity: A case study based on story generation, 2026. Work in progress.

[3]Anna Jordanous and Bill Keller. Four pppperspectives on computational creativity in theory and in practice. *Connection Science*, 28(2):194–216, 2016.

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

[5]Google DeepMind. Gemini 3 pro: Model information and capabilities, 2025. Accessed: 2025-12-21.

[6]Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744, 2022.

A Generation and Evaluation Details

Stories (350–450 words) were generated with Gemini 3.0 Pro (Preview) using default settings. Spike Prompting maximized a single target dimension per condition: Surreal (vocabulary rarity, unpredictability), Melancholic (emotional depth, sensory immersion), Witty (linguistic sophistication, structural elegance), and Clinical (objectivity, constraint adherence). Topics were *AI Shutdown*, *The Heist*, and *The Midnight Store*. Flesch–Kincaid readability was controlled across stories ($M = 7.7$, Range = 6.7–9.0).

B Survey Protocol

$N = 115$ via Prolific, between-subjects. Three-stage protocol: (1) Immediate holistic ratings (Creativity, Enjoyment); (2) Component evaluation (11 features, 7-point Likert); (3) Reflective creativity. Compensation: £1.25 (median 4.5 min, £15/hr). Informed consent obtained; no PII beyond Prolific IDs. Ethics Review Board approved.

C Correlation and Discordance Details

Correlation: Initial: $r = 0.757$ ($p < 0.001$, 95% CI: [0.67, 0.82]). Post-reflection: $r = 0.709$ ($p < 0.001$, 95% CI: [0.61, 0.79]). Decrease: $\Delta r = 0.048$ ($p = 0.092$, Williams' test).

Discordance (median-split): Using thresholds ($\text{Creativity}_{\text{reflect}} = 5$, Enjoyment = 5): Hi-Creat/Hi-Enjoy: $n = 48$ (41.7%); Hi-Creat/Lo-Enjoy: $n = 16$ (13.9%, *Creative-but-Unenjoyable*); Lo-Creat/Hi-Enjoy: $n = 4$ (3.5%, *Enjoyable-but-Uncreative*); Lo-Creat/Lo-Enjoy: $n = 47$ (40.9%). Total: $n = 20$ (17.4%).

Asymmetry: $\chi^2(1) = 7.2$, $p = 0.007$. Standardized residuals: Creative-but-Unenjoyable $z = +2.68$; Enjoyable-but-Uncreative $z = -2.68$. 4:1 ratio (16 vs. 4) indicates stories more likely judged creative without being enjoyable than reverse.

D Weight Comparison Details

Component-level models: $\text{Creativity}_{\text{init}} = 0.31\text{Nov} + 0.42\text{Res} + 0.18\text{Val} - 0.09\text{Adh}$ ($R^2 = 0.68$); $\text{Enjoyment} = -0.03\text{Nov} + 0.51\text{Res} + 0.28\text{Val} + 0.15\text{Adh}$ ($R^2 = 0.71$). **Parameter divergence:** Empathy: creativity $\beta = -0.14$ ($p = 0.003$), enjoyment $\beta = +0.22$ ($p < 0.001$), $\Delta\beta = 0.36$. Vocabulary: creativity $\beta = +0.18$ ($p < 0.001$), enjoyment $\beta = -0.11$ ($p = 0.021$), $\Delta\beta = 0.29$.

E Reflection Details

Sample sizes: Decreased: $n = 18$; Stable: $n = 56$; Increased: $n = 41$.

Enjoyment by change: $F(2, 112) = 4.784$, $p = 0.010$. Means: Decreased $M = 4.28$; Stable $M = 4.74$; Increased $M = 3.25$. Tukey: Inc vs. Stb $p = 0.008$.

Component ANOVAs: Novelty: $F = 0.15$, ns. Resonance: $F = 8.45$, $p < 0.001$ (Increased $M = 3.82$ vs. Stable $M = 4.89$). Value: $F = 12.3$, $p < 0.001$ (Increased $M = 3.95$ vs. Stable $M = 5.32$). Adherence: $F = 6.82$, $p = 0.002$ (Increased $M = 4.53$ vs. Stable $M = 5.47$).

Novelty-Resonance gap: $F(2, 112) = 4.597$, $p = 0.012$. Increased $M = +1.23$ vs. Stable $M = +0.24$, $p = 0.009$.

Parameter deficits (Inc vs. Stb): Logical Flow $\Delta = -1.47$, Emotional Impact $\Delta = -1.35$, Empathy $\Delta = -1.31$, Engagement $\Delta = -1.19$, Style $\Delta = -1.17$ (all $p < 0.001$). Novelty: all ns.

Composites (Inc vs. Stb): Analytical Coherence: 3.72 vs. 5.64, $t = -5.12$, $p < 0.001$, $d = 1.03$. Emotional Response: 3.22 vs. 4.47, $t = -3.58$, $p = 0.001$, $d = 0.72$.

F Mediation Details

Indirect effects (5K bootstrap): Value: $ab = 0.349$ [0.28, 0.42], 58.3%, $p < 0.001$. Resonance: $ab = 0.294$ [0.21, 0.38], 49.1%, $p < 0.001$. Adherence: $ab = 0.180$ [0.10, 0.26], 30.0%, $p < 0.001$. Novelty: $ab = 0.078$ [-0.01, 0.17], 13.0%, $p = 0.057$ (ns). Direct: 0.195 (32.5%).

G Genre (tone) Details

Alignment by tone: Melancholic $r = +0.87$ [0.49, 0.97], $p = 0.001$; Witty $r = +0.42$ [-0.29, 0.84], $p = 0.22$; Clinical $r = +0.35$ [-0.38, 0.82], $p = 0.32$; Surreal $r = -0.18$ [-0.72, 0.51], $p = 0.60$. **Pathways:** Surreal Nov $\beta = 0.41$, Res $\beta = 0.19$; Melancholic Nov $\beta = 0.08$, Res $\beta = 0.52$; Witty/Clinical balanced ($\beta \approx 0.25$).