LLM Review: Enhancing Creative Writing via Blind Peer Review Feedback

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often struggle with creative generation, and multi-agent frameworks that improve reasoning through interaction can paradoxically hinder creativity by inducing content homogenization. We introduce LLM Review, a peer-review-inspired framework implementing Blind Peer Review: agents exchange targeted feedback while revising independently, preserving divergent creative trajectories. To enable rigorous evaluation, we propose SciFi-100, a science fiction writing dataset with a unified framework combining LLM-as-a-judge scoring, human annotation, and rule-based novelty metrics. Experiments demonstrate that LLM Review consistently outperforms multi-agent baselines, and smaller models with our framework can surpass larger single-agent models, suggesting interaction structure may substitute for model scale.

1 Introduction

Large Language Models (LLMs) have achieved strong performance across natural language processing tasks (Lappin, 2024; Zhang et al., 2024b; Yi et al., 2024), and are increasingly deployed in multi-agent systems for reasoning and coordination (Tran et al., 2025). However, these systems are optimized for correctness rather than creativity. Prior work finds that LLMs tend to reproduce familiar patterns rather than generate genuinely novel ideas (Mohammadi, 2024; Chakrabarty et al., 2024; Li et al., 2025), and existing single-agent approaches such as decoding strategies, prompt engineering, and post-training optimization (Yang et al., 2023; Potraghloo et al., 2025; Matan and Velvizhy, 2025) yield surface-level diversity rather than substantive conceptual novelty.

Human creativity is fundamentally social, emerging through discussion, critique, and iterative refinement (Nijstad and Paulus, 2003). Recent multi-agent frameworks such as debate and discussion show improvements in reasoning and output diversity (Du et al., 2023; Lu et al., 2024; Summers-Stay et al., 2023), sharing an implicit assumption: more interaction yields better outcomes. We argue this assumption breaks down for creativity. Research on group brainstorming shows that interactive groups often produce fewer and less original ideas than individuals working independently, due to production blocking and convergent tendencies (Diehl and Stroebe, 1987; Larey and Paulus, 1999). Recent work further demonstrates homogenization effects when humans collaborate with LLMs (Anderson et al., 2024). We propose a different view: creativity is not improved by more interaction, but by the right information flow constraints. Creative novelty requires divergence, the ability to explore different trajectories without converging on share patterns (Gillebaart et al., 2013). Existing frameworks repeatedly expose agents to each other's evolving outputs, inadvertently encouraging alignment and limiting semantic exploration.

We introduce **LLM Review**, a framework that enhances creativity by constraining rather than maximizing information flow through a mechanism we call Blind Peer Review. Inspired by double-blind academic reviewing, agents provide targeted feedback on peers' initial drafts but revise independently, without seeing how peers respond to the same feedback. This information asymmetry lets agents benefit from external critique while preserving independent creative trajectories. To evaluate our approach, we introduce **SciFi-100**, a science fiction writing dataset, together with a unified evaluation framework combining LLM-as-a-judge scoring, human annotation, and rule-based metrics capturing lexical and semantic novelty against a corpus of canonical science fiction (Colton, 2008).

Our contributions: (1) **LLM Review**, a framework that enhances creativity by constraining information flow through Blind Peer Review; (2) **SciFi-100**, the first science fiction writing dataset with a unified evaluation framework combining LLM-as-a-judge, human annotation, and rule-based novelty metrics; (3) LLM Review outperforms baselines, with smaller models exceeding larger single-agent models, showing that interaction structure can offset model scale.

2 Related Work

Multi-Agent LLMs Recent work has explored multi-agent frameworks built on Large Language Models (LLMs) to improve factuality, reasoning, and task performance through structured interaction, including role-based workflows, debate and critique protocols, persona-driven interaction, and large-scale orchestration (Yao et al., 2022; Chen et al., 2024; Hong et al., 2023; Qian et al., 2024a; Du et al., 2023; Chan et al., 2023; Tseng et al., 2024; Wang et al., 2025). These systems are primarily evaluated on goal-directed benchmarks and focus on task success, autonomy, or efficiency (Li et al., 2024; Ye et al., 2025; Qian et al., 2024b; Dang et al., 2025; Zhang et al., 2024a). In contrast, our work studies creative writing and shows that interaction *structure*, rather than interaction frequency or scale, plays a critical role: strategically restricting information flow helps preserve divergent creative trajectories.

LLM Creativity Creativity in language models has been studied across tasks such as literary composition, metaphor generation, and alternative use generation, primarily in single-LLM settings (Gómez-Rodríguez and Williams, 2023; Chakrabarty et al., 2024; DiStefano et al., 2025; Stevenson et al., 2022). Prior approaches improve creativity through decoding strategies, post-training optimization, or inference-time prompting (Ghazvininejad et al., 2017; Keskar et al., 2019; Wei et al., 2025; Chung et al., 2025; Lagzian et al., 2025). While effective, these methods largely operate at model or decoding level; our work instead frames creativity as a socially grounded, multi-agent process driven by structured discussion.

Creativity Evaluation Evaluating creativity is inherently subjective, and prior work relies on human judgments or LLM-based evaluators for scalability (Gómez-Rodríguez and Williams, 2023; Chakrabarty et al., 2024; Feng et al., 2025; Zheng et al., 2023). Automatic proxies are often used to capture complementary aspects such as diversity and semantic novelty relative to a reference corpus (Zhang et al., 2021; Peeperkorn et al., 2024). Following established views of creativity as balancing novelty and value (Colton, 2008; D'Souza, 2021), we adopt a combined evaluation framework that integrates rule-based novelty metrics with LLM-as-a-judge assessment.

3 Methodology

3.1 Task Definition

Given a science-fiction writing prompt $x \in \mathcal{X}$, the goal is to generate a short story y (approximately 300 words) that is coherent, creatively strong, and of good quality. We study both single-agent generation and multi-agent discussion frameworks that iteratively improve drafts through structured interaction. Unless otherwise specified, all frameworks use the same base writer model (the model under evaluation) and the same role-played writer personas to control for style and diversity effects.

3.2 SciFi-100 Data Curation

To curate high-quality science fiction prompts that can be used to evaluate model performance, we structure a systematic process to create a dataset that matches creative writing attributes. We first identify ten central aspects (see Appendix A) of creative writing based on our experiences as human writers as well as foundational insights from narratology, literary theory, and creative writing pedagogy (Genette, 1980; Pound, 2013; Freytag, 1895; Forster, 1927; Rosenblatt, 1994; Burroway et al., 2022; Csikszentmihalyi, 1990; Chatman and Chatman, 1978; Ricoeur, 2004; Tannen, 2005). For each aspect, we query LLMs (Hurst et al., 2024) to generate twenty unique prompts by specifically instructing the model to create scenarios where a science fiction narrative can unfold. After the model generates 200 prompts (20 prompts per writing aspect), we manually select and revise 10 prompts per aspect (100 in total) to ensure diversity and thematic relevance. The dataset's balanced distribution across aspects ensures a comprehensive evaluation of creative dimensions. To our best knowledge, SciFi-100 is the first dataset designed to assess scientific fiction writings of LLMs.

3.3 Multi-Agent Role-Play Setup

For multi-agent frameworks, we instantiate $N{=}3$ writer agents. Following prior work on role-play and diversity of thought (Camacho, 2016; Lu et al., 2024), each agent is assigned a persistent persona (e.g., Humanistic Writer, Futuristic Writer, Ecological Writer) that remains fixed across all rounds. We repeatedly restate these roles in prompts to encourage consistent viewpoints and reduce homogenization. All frameworks share the same formatting constraints (story-only outputs, no commentary) to minimize evaluation noise.

3.4 Compared Frameworks

We compare our proposed framework, LLM Review, against the following baselines.

Single Agent A single LLM is prompted once to write a ∼300-word science-fiction story for the given prompt. This baseline captures the base model's inherent creative writing capability under zero-shot prompting.

LLM Teacher LLM Teacher, inspired by classroom-style role-play prompting (Camacho, 2016), models a teacher-student loop in which a teacher agent provides guidance and critique to student writers. The framework proceeds in three phases: the teacher offers high-level advice, students draft stories and receive aggregated feedback, and students revise to produce final outputs. This baseline represents a simple extension of role-play prompting with critique, but its teacher-centered, one-to-many feedback can encourage convergence toward similar revisions.

LLM Debate LLM Debate (Du et al., 2023) structures interaction as proposal and critique, where agents present candidate drafts and challenge each other's content (logic gaps or weak originality), followed by refinement. Unlike other role-play-based frameworks, LLM Debate doesn't assign explicit personas to agents. We adapt the debate protocol for creative writing by focusing critiques on plausibility, novelty, and narrative quality.

LLM Discussion LLM Discussion (Lu et al., 2024) is a three-phase multi-agent framework (Initiation, Discussion, Convergence) built on top of Du et al. (2023) with role-play. Agents iteratively read others' drafts and update their responses accordingly. This baseline represents structured multi-agent collaboration without explicit critique roles.

3.5 LLM Review

Inspired by the iterative peer-review process in academic writing, we propose *LLM Review*, a structured feedback loop designed to improve creativity through *distributed critique* and *private revision*. The key idea is to let agents both *create* and *act as reviewers* for each other, then revise using feedback without seeing peers' revised drafts.

Phase 1: Compose Each agent independently writes an initial story draft conditioned only on the prompt and its persona.

Phase 2: Review Each agent reviews peers' drafts and provides targeted feedback (e.g., originality, world-building opportunities, speculative consistency, stronger imagery, character depth). Agents then revise their own draft using (a) their initial draft and (b) the received feedback.

Originality constraint During revision, agents do **not** see peers' revised drafts (only the initial drafts and feedback). This design reduces homogenization and preserves independent creative trajectories across rounds.

Figure 1 illustrates the interaction flow of LLM Review and the corresponding baselines.

3.6 LLM-as-a-Judge Evaluation

LLMs have demonstrated the ability to approximate human judgment and effectively evaluate content, achieving results comparable to human evaluators, even on creativity tasks (Lu et al., 2024). Therefore, we adopt an LLM-as-a-judge technique to assess the creativity of the generated stories, utilizing `gpt-4o` (Hurst et al., 2024) as the judging model.

In our evaluation pipeline, each story is individually assessed on five key aspects derived from the literature on science fiction writing: **Scientific Concept Integration**, **Speculative Logic**, **Character Depth**, **Immersive World-Building**, and **Ethical and Philosophical Themes**. These evaluation aspects reflect the consensus of prior work in narratology, speculative fiction studies, creativity research, and narrative ethics (Canavan and Suvin, 2016; Chatman and Chatman, 1978; Forster, 1927; Le Guin, 2015; Nussbaum, 1988) For each aspect, the LLM assigns a score between 0 and 5, where 0 indicates potential plagiarism with poor quality, and 5 indicates highly creative writings of good quality. To ensure consistency and minimize variability, we evaluate each story three times and report the average score.

3.7 Human evaluation

To validate that our LLM-as-a-Judge scores reflect human preferences, we recruit nine student annota-

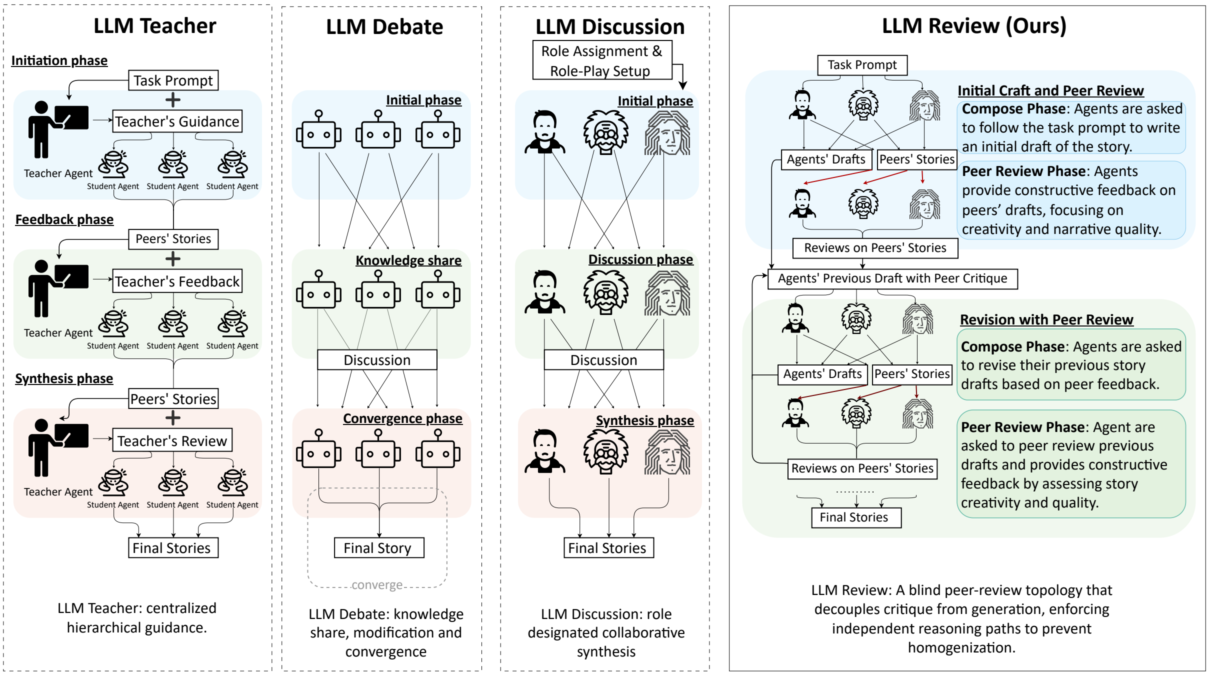

Figure 1: Comparison of multi-agent framework, from single-agent zero-shot writing to multi-agent frameworks. A single LLM generates a story in one pass without feedback, while LLM Teacher, LLM Debate, and LLM Discussion introduce hierarchical guidance, discussion, or role-based collaboration. LLM Review (ours) adopts a blind peer-review topology that decouples critique from generation, enabling independent revisions and reducing homogenization.

tors to evaluate the stories generated on SciFi-100 by the LLM-Review framwork with Llama-3.2-3B model. Annotators rate each story using the same five criteria as in Section 3.6 and the same 0-5 Likert rubric. For each story and criterion, we average the nine ratings to obtain a single human consensus score. We report the human score distribution (mean $\pm$ std over stories) and in Table 4.

3.8 Rule-based Evaluation

To evaluate the creativity of generated content, we design a rule-based evaluation framework with two components. First, we measure an intrinsic dimension: the absolute token-level diversity of the generated text. Second, we measure novelty relative to traditional science fiction, and decompose it into semantic novelty and lexical novelty. We use the SFGram dataset (Schaetti, 2018), which contains 1003 classic science fiction novels, as the reference corpus representing traditional science-fiction writing. Our framework provides a systematic and quantifiable multi-dimensional evaluation of creativity, defined as follows:

3.8.1 Absolute Diversity

We quantify the intrinsic diversity of generated text using token-level surprisal, i.e., the negative log-probability of each generated token under the model distribution (Hale, 2001; Demberg and Keller, 2008). For a generated sequence of length L, the average surprisal is computed as:

$$S_{\text{avg}} = -\frac{1}{L}\sum_{j=1}^{L} \log p(x_j \mid x_{<j}), \tag{1}$$

where x_j denotes the generated token at position j, and $p(x_j \mid x_{<j})$ is the model-assigned probability of that token given the preceding context. Concretely, we obtain next-token probabilities from the model outputs at each position and compute surprisal for the realized token x_j. We normalize by L to mitigate the influence of sequence length.

Compared to entropy, which summarizes uncertainty over next-token candidates, surprisal focuses on the information content of the tokens the model actually generates, and thus provides a practical intrinsic signal for our evaluation.

3.8.2 Lexical Divergence

We measure lexical divergence at the unigram word level using Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951):

$$D_{\text{KL}}(p\|q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{2}$$

Here, $q(x)$ denotes the unigram word distribution estimated from the SFGram corpus, and $p(x)$ denotes the unigram word distribution of the generated text. We use $D_{\mathrm{KL}}(p\|q)$ to quantify how the generated lexical distribution departs from the reference; larger values indicate greater lexical deviation and serve as a proxy for lexical novelty. We apply additive smoothing to avoid zero probabilities.

3.8.3 Semantic Divergence

Nearest-neighbor semantic similarity We measure semantic novelty by embedding generated sentences and the SFGram reference corpus, and computing cosine similarity in the embedding space. We embed SFGram using the `all-mpnet-base-v2` Sentence Transformers model (Reimers and Gurevych, 2019). Since the encoder has a 512-token input limit, we split SFGram into chunks of approximately 250 words and embed each chunk. For a generated sentence embedding $\mathbf{u}$, we compute cosine similarity (Salton and McGill, 1983) to every reference chunk embedding $\mathbf{v}$:

$$\text{Cosine Similarity} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}. \tag{3}$$

We take the maximum similarity over reference chunks as nearest-neighbor overlap and report semantic novelty as $1 - \max$ Cosine Similarity, so larger values indicate higher novelty.

Embedding Volume Gain In addition to nearest-neighbor semantic similarity, we measure how broadly a generated story spreads in the embedding space relative to the reference corpus. While nearest-neighbor similarity captures local overlap with the corpus, the volume gain provides a distribution-level view of semantic spread. Let $\Sigma_{\text{ref}} = \Sigma(E_{\text{ref}})$ denote the covariance of SFGram chunk embeddings, and let $\Sigma_{\text{ref}\cup\text{story}} = \Sigma(E_{\text{ref}} \cup E_{\text{story}})$ denote the covariance after adding the story chunk embeddings. We summarize multivariate scatter using the log-determinant of the covariance, $\log\det(\Sigma)$, i.e., the log of the generalized variance (Wilks, 1932; Rencher, 1998); geometrically, $\det(\Sigma)$ is proportional to the squared volume of the covariance ellipsoid in embedding space. We define the embedding volume gain as:

$$\Delta_{\text{vol}} = \log\det(\Sigma_{\text{ref}\cup\text{story}}) - \log\det(\Sigma_{\text{ref}}). \tag{4}$$

Larger Δ_{vol} indicates that adding the story expands the embedding-space coverage beyond the reference corpus. For numerical stability, we compute $\log\det(\Sigma + \epsilon I)$ with a small ϵ.

Our rule-based metrics primarily capture diversity and novelty relative to the reference corpus, rather than quality dimensions such as coherence or readability. Hence, larger deviation or dispersion may sometimes reflect increased randomness rather than better creative writing, so we interpret these metrics alongside LLM-as-a-judge scores for complementary quality-aware assessment.

4 Experiments

4.1 Experimental Setup

We evaluate all frameworks on SciFi-100. For each prompt, the target output is a single story of approximately 300 words. Unless otherwise stated, multi-agent frameworks use $N{=}3$ writer agents with fixed personas (Section 3.3). We use consistent output-format constraints across all methods (story text only) to reduce off-format generations.

Decoding For the main comparisons, we use top_p$= 0.9$ and temperature$= 0.9$ (default settings in our implementation). We additionally study the effect of decoding hyperparameters on the agent's performance.

Rounds and compute All multi-agent frameworks run for three iterative rounds. Each experiment is conducted on 4 x NVIDIA A100 Tensor Core GPUs (80GB).

4.2 Compared Methods

We report results for: (1) Single Agent, (2) LLM Discussion, (3) LLM Debate, (4) LLM Teacher, and (5) LLM Review (ours).

4.3 Models

We evaluate our framework across a diverse set of state-of-the-art instruction-tuned models, covering both open-weights and proprietary frontier families. For the Llama family, we utilize **Llama 3.2** (Grattafiori et al., 2024), specifically the `Llama-3.2-1B-Instruct` (llama 1b) and `Llama-3.2-3B-Instruct` (llama 3b) variants, to assess performance on lightweight, edge-class models. We also include the **Qwen 2.5** series (Yang et al., 2025), employing `Qwen2.5-1.5B-Instruct` (qwen 1.5b) and `Qwen2.5-3B-Instruct` (qwen 3b), to verify generalization across different model architectures. For the closed-source frontier baseline, we use **gpt-4o** (Hurst et al., 2024).

Framework	LLM-as-a-Judge Evaluation					Rule-Based Evaluation			
	Concepts	Logic	Characters	World-Building	Ethics	Surprisal	KL Div.	1-Cos Sim.	Volume Gain
Single Agent	3.62$\pm$1.14	3.62$\pm$1.13	3.41$\pm$1.11	3.63$\pm$1.11	3.40$\pm$1.14	0.476$\pm$0.271	2.233$\pm$0.012	0.345$\pm$0.002	0.0138$\pm$0.0002
LLM Teacher	3.71$\pm$0.87	3.73$\pm$0.87	3.61$\pm$0.88	3.73$\pm$0.87	3.43$\pm$0.92	0.509$\pm$0.260	2.356$\pm$0.012	0.393$\pm$0.002	0.0160$\pm$0.0002
LLM Debate	3.76$\pm$0.22	3.76$\pm$0.21	3.74$\pm$0.24	3.65$\pm$0.29	3.80$\pm$0.37	0.540$\pm$0.456	2.501$\pm$0.032	0.424$\pm$0.006	0.0175$\pm$0.0003
LLM Discussion	3.94$\pm$0.26	3.89$\pm$0.33	3.90$\pm$0.25	3.74$\pm$0.44	3.73$\pm$0.55	**0.584$\pm$0.303**	2.597$\pm$0.037	0.435$\pm$0.009	0.0208$\pm$0.0004
LLM Review (Ours)	**3.98$\pm$0.24**	**4.00$\pm$0.22**	**3.96$\pm$0.28**	**4.04$\pm$0.26**	**4.00$\pm$0.34**	0.573$\pm$0.316	**2.638$\pm$0.034**	**0.441$\pm$0.008**	**0.0211$\pm$0.0004**

Table 1: Comparison of LLM-as-a-judge and rule-based creativity evaluations on the LLaMA-3.2-3B model.

5 Results

5.1 Creativity Evaluation Results

We evaluate creativity from two complementary perspectives: an LLM-as-a-judge rubric assessing creativity-aware writing quality across five science-fiction dimensions, and a rule-based suite measuring intrinsic diversity (token-level surprisal) and novelty relative to a reference corpus via lexical (KL divergence) and semantic divergence (nearest-neighbor overlap and embedding-space volume gain against SFGram). Our main comparison uses LLaMA-3.2-3B as the writer model (Table 1).

Main comparison Across both evaluations, multi-agent frameworks outperform the single-agent baseline, with LLM Review consistently ranking highest. LLM-as-a-judge results show improvements across all five dimensions with lower score variability, indicating more robust creativity-aware writing quality. Rule-based metrics exhibit a consistent ordering, with LLM Review achieving the strongest novelty signals, followed by Discussion, while Teacher and Debate show more conservative lexical and semantic deviation from SFGram. The agreement between rubric-based judgments and automatic novelty metrics supports the interpretability of the rule-based evaluation for comparing frameworks.

Mechanism analysis The observed performance differences reflect how each framework structures interaction, feedback, and information flow. The consistent gains of LLM Review in LLM-as-a-judge scores highlight the role of explicit, targeted critique in improving creativity-aware writing quality: unlike Discussion and Debate, which expose agents to peers' drafts without structured revision guidance, or Teacher, which provides centralized feedback that can steer agents toward similar revision targets, LLM Review decentralizes critique by requiring agents to deliver concrete peer-level feedback. At the same time, improvements in rule-based novelty metrics stem from controlled information exchange: while Discussion and Debate repeatedly condition agents on peers' evolving outputs, encouraging alignment in phrasing and themes, LLM Review shares only independent critiques while preserving independent creative trajectories. Together, this design balances guidance and independence, yielding stronger and more stable lexical and semantic novelty signals than other multi-agent baselines.

Generalization across model families and scales Table 2 summarizes LLM Review's performance across writer model families and scales. While using GPT-4o as the writer achieves the highest LLM-as-a-judge scores, its weaker rule-based novelty signals suggest potential self-preference when the writer and judge come from the same model family (Panickssery et al., 2024), motivating reliance on rule-based metrics as a complementary reference. Across model scales within the same framework, rule-based novelty remains relatively stable, whereas LLM-as-a-judge scores increase with model size, indicating that interaction structure primarily shapes novelty while scaling mainly improves writing quality. Consistent with this, Table 3 shows a favorable structure-scale trade-off: LLM Review with a smaller writer can outperform a larger single-agent baseline, supporting distributed peer feedback as an effective and relatively compute-efficient alternative to model scaling.

5.2 Human alignment

To assess whether our LLM-as-a-judge evaluation reflects human preferences, we conduct a human rating study on outputs produced by the LLM Review framework with LlaMA-3.2-3B model. We recruited nine annotators to score each story using the same five evaluation dimensions and the same Likert rubric as in our LLM-as-a-judge setup as in Section 3.6. We average the nine ratings to obtain a single human consensus score for each story and dimension, and compare it against the corresponding LLM-as-a-judge score. We quantify

	LLM-as-a-Judge Evaluation					Rule-Based Evaluation		
Model	**Concepts**	**Logic**	**Characters**	**World-Building**	**Ethics**	**KL Div.**	**1-Cos Sim.**	**Volume Gain**
gpt-4o	4.03±0.09	4.07±0.18	4.01±0.08	4.22±0.29	4.3±0.38	2.570±0.004	0.388±0.001	0.0068±0.0001
llama 3b	3.98±0.24	4.00±0.22	3.96±0.28	4.04±0.26	4.00±0.34	2.638±0.034	0.441±0.008	0.0211±0.0004
llama 1b	3.85±0.47	3.81±0.50	3.60±0.54	3.75±0.50	3.65±0.64	2.627±0.025	0.449±0.007	0.0213±0.0004
qwen 3b	3.52±0.49	3.90±0.29	3.89±0.38	3.79±0.41	3.79±0.69	2.254±0.008	0.386±0.001	0.0206±0.0002
qwen 1.5b	3.27±0.73	3.41±0.73	3.30±0.69	3.39±0.70	3.47±0.87	2.412±0.012	0.329±0.002	0.0196±0.0001

Table 2: LLM-as-a-judge and rule-based creativity evaluations across different model families for the LLM Review framework. We exclude surprisal when comparing different base models since it is model-dependent (defined under each model's distribution $p(\cdot)$) and thus not directly comparable across models.

	LLM-as-a-Judge Evaluation					Rule-Based Evaluation		
Framework	**Concepts**	**Logic**	**Characters**	**World-Building**	**Ethics**	**KL Div.**	**1-Cos Sim.**	**Volume Gain**
Single Agent (qwen 3b)	3.09±0.68	3.25±0.67	3.24±0.46	3.13±0.61	3.18±0.82	2.205±0.013	**0.346±0.002**	0.0148±0.0006
LLM Review (qwen 1.5b)	**3.27±0.73**	**3.41±0.73**	**3.30±0.69**	**3.39±0.70**	**3.47±0.87**	**2.412±0.012**	0.329±0.002	**0.0196±0.0001**
Single Agent (llama 3b)	3.62±1.14	3.62±1.13	3.41±1.11	3.63±1.11	3.40±1.14	2.233±0.012	0.345±0.002	0.0138±0.0002
LLM Review (llama 1b)	**3.85±0.47**	**3.81±0.50**	**3.60±0.54**	**3.75±0.50**	**3.65±0.64**	**2.627±0.025**	**0.449±0.007**	**0.0213±0.0004**

Table 3: LLM-as-a-judge and rule-based creativity evaluations comparing smaller LLM Review models to larger single-agent baselines within the same model family.

alignment using (i) ICC(A,1) to measure absolute agreement, (ii) Bland-Altman bias and 95% limits of agreement (LoA) to evaluate calibration and typical per-story discrepancies, and (iii) Pearson's r to capture linear association and ranking consistency between human and judge scores and the result table is shown in Table 4. The definitions of the above metrics are shown in the Section C.

5.3 Ablation study

We conduct ablation studies to understand the sensitivity of *LLM Review* to key design choices: decoding hyperparameters, number of iterative rounds, and number of participating agents. These experiments use LLaMA-3.2-3B as the writer model unless otherwise noted.

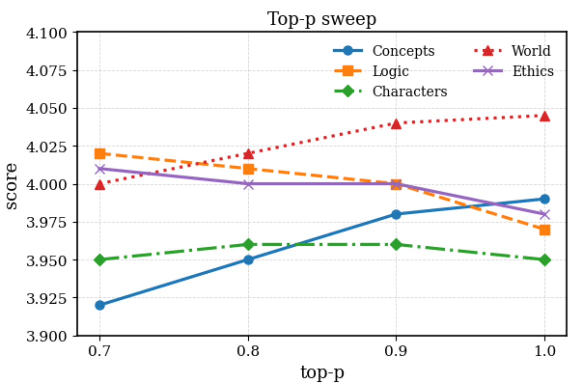

Figure 2: The average score of 5 LLM-as-a-judge evaluation aspects with different Top-p decoding methods.

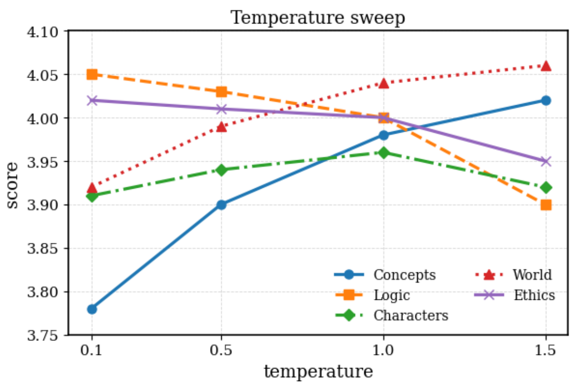

Figure 3: The average score of 5 LLM-as-a-judge evaluation aspects with different temperatures.

Decoding experiments We study the effect of stochastic decoding hyperparameters on creative writing quality by sweeping top-p and temperature. As shown in Figure 2, increasing top-p from 0.7 to 1.0 consistently improves scores for Scientific Concept Integration and Immersive World-Building, suggesting that allowing a broader candidate pool encourages richer idea exploration and setting construction. In contrast, Speculative Logic and Ethical/Philosophical Themes exhibit a mild downward trend, indicating a trade-off between creativity and structural coherence at higher sampling entropy. Character Depth remains relatively stable across different top-p values. Figure 3 shows that increasing temperature leads to a stronger trade-off: higher temperatures substantially boost Concepts and World-Building but noticeably degrade

Framework	Concepts	Logic	Characters	World-Building	Ethics
LLM Review (Human Scored)	3.98±0.23	3.99±0.20	3.94±0.21	4.02±0.23	3.96±0.27
ICC(A,1)	0.62	0.58	0.59	0.65	0.61
Bias	0.016	0.012	-0.007	0.009	-0.018
LoA low	-0.321	-0.303	-0.345	-0.318	-0.429
LoA high	0.352	0.328	0.332	0.336	0.394
Pearson r	0.679	0.610	0.607	0.689	0.647

Table 4: Human evaluation of LLM Review (llama 3B) using the same 0-5 Likert rubric and the same five dimensions as LLM-as-a-judge. We report mean±std of the averaged human scores and alignment between human consensus and LLM-as-a-judge via ICC(A,1), Bland-Altman bias and 95% limits of agreement (LoA), and Pearson's r. Human ratings are consistently high across dimensions (mean 3.94-4.02). The judge shows moderate absolute agreement with humans (ICC(A,1)=0.58-0.65) and consistent linear association (Pearson's r=0.607-0.689). Bland-Altman analysis indicates negligible systematic bias (all |bias| ≤ 0.018), suggesting that LLM-as-a-judge scores are well-calibrated to human ratings and track human judgments in both level and ranking.

Logic and, to a lesser extent, Ethics, while Character Depth peaks around temperature $= 1.0$ before declining. Overall, these results highlight the tension between creativity and coherence in stochastic decoding and motivate our choice of a moderate top-p (≈ 0.9) and mid-range temperature as default settings.

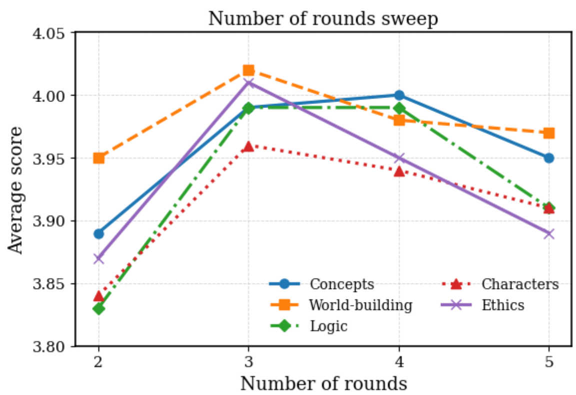

Figure 4: Number of execution rounds vs the average score of 5 LLM-as-a-judge evaluation aspects.

Number of rounds We examine the effect of iterative rounds in Figure 4. Most dimensions achieve peak performance with three rounds of discussion. Although Concepts and Logic continue to improve slightly until the fourth round, they decline thereafter. We adopt $R = 3$ rounds as our default setting to balance performance and efficiency.

Number of agents Figure 5 shows that with three-round discussions, using three agents yields the best overall performance, which declines with further additions. Unlike increasing rounds, which selectively benefits certain dimensions, adding more agents uniformly degrades all metrics, likely due to feedback dilution. We use $N = 3$ agents for further experiments.

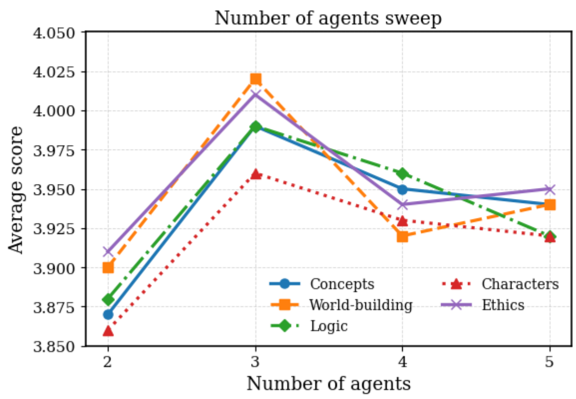

Figure 5: Number of agents vs the average score of 5 LLM-as-a-judge evaluation aspects.

6 Conclusion

Our results challenge the assumption that more interaction yields better outcomes in multi-agent LLM systems. While convergence benefits tasks with verifiable ground truth, creativity requires divergence. LLM Review succeeds by disentangling feedback (targeted critique) from exposure (observing peers' outputs), where agents receive peer critique but never see how others revise. This asymmetry preserves independent creative trajectories while still benefiting from external feedback. A key finding is that smaller models using our framework outperform larger single-agent models, suggesting interaction structure may be a more compute-efficient lever than model scaling for creative tasks.

Limitations

Our evaluation focuses on short-form science fiction writing; generalization to other creative domains (poetry, long-form fiction, music) may require domain-specific metrics and reference corpora. Our rule-based novelty metrics measure divergence from a fixed reference corpus (SFGram) and do not by themselves guarantee meaningful creativity; we therefore interpret them jointly with quality-oriented LLM-as-a-judge scores. Our human study uses nine student annotators on a single configuration; professional writers might assess differently. Finally, LLM Review requires approximately 9× the inference cost of single-agent generation, though this can be offset by using smaller models.

Ethical consideration

This work includes human evaluation of machine-generated text; annotators provided informed consent, no personal identifying information was collected, and results are reported in aggregate. Generated content may reflect biases present in underlying language models, particularly in speculative narratives, and automated critique could reinforce shared assumptions. The proposed framework is intended for research use and should be deployed with human oversight in practical applications.

References

Brenda R Anderson, Jasneet H Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 413–425.

Janet Burroway, Elizabeth Stuckey-French, and Ned Stuckey-French. 2022. *Writing fiction: A guide to narrative craft*. University of Chicago Press.

Maria Camacho. 2016. David kelley: From design to design thinking at stanford and ideo. *She Ji: The Journal of Design, Economics, and Innovation*, 2(1):88–101.

Gerry Canavan and Darko Suvin. 2016. Metamorphoses of science fiction.

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*, pages 1–34.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Seymour Benjamin Chatman and Seymour Chatman. 1978. *Story and discourse: Narrative structure in fiction and film*. Cornell university press.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*.

Simon Colton. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, volume 8, page 7. Palo Alto, CA.

Mihaly Csikszentmihalyi. 1990. The domain of creativity.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, and 1 others. 2025. Multi-agent collaboration via evolving orchestration. *arXiv preprint arXiv:2505.19591*.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3):497–509.

Paul V DiStefano, John D Patterson, and Roger E Beaty. 2025. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, 37(4):555–569.

Richard D'Souza. 2021. What characterises creativity in narrative writing, and how do we assess it? research findings from a systematic literature search. *Thinking skills and creativity*, 42:100949.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Kehua Feng, Keyan Ding, Tan Hongzhi, Kede Ma, Zhihua Wang, Shuangquan Guo, Cheng Yuzhou, Ge Sun, Guozhou Zheng, Qiang Zhang, and 1 others. 2025. Sample-efficient human evaluation of large language models via maximum discrepancy competition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10913–10947.

Edward Morgan Forster. 1927. *Aspects of the Novel*. Harcourt, Brace.

Gustav Freytag. 1895. *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs.

Gérard Genette. 1980. *Narrative discourse: An essay in method*, volume 3. Cornell University Press.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48.

Marleen Gillebaart, Jens Förster, Mark Rotteveel, and Astrid CM Jehle. 2013. Unraveling effects of novelty on creativity. *Creativity Research Journal*, 25(3):280–285.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Arash Lagzian, Srinivas Anumasa, and Dianbo Liu. 2025. Multi-novelty: Improve the diversity and novelty of contents generated by large language models via inference-time multi-views brainstorming. *arXiv preprint arXiv:2502.12700*.

Shalom Lappin. 2024. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*, 33(1):9–20.

Timothy S Larey and Paul B Paulus. 1999. Group preference and convergent tendencies in small groups: A content analysis of group brainstorming performance. *Creativity Research Journal*, 12(3):175–184.

Ursula K Le Guin. 2015. *Steering the craft: A twenty-first century guide to sailing the sea of story*. Houghton Mifflin Harcourt.

Ao Li, Yuexiang Xie, Songze Li, Fugee Tsung, Bolin Ding, and Yaliang Li. 2024. Agent-oriented planning in multi-agent systems. *arXiv preprint arXiv:2410.02189*.

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025. Automated creativity evaluation for large language models: A reference-based approach. *arXiv preprint arXiv:2504.15784*.

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

P Matan and P Velvizhy. 2025. A comprehensive review of supervised fine-tuning for large language models in creative applications and content moderation. In *2025 International Conference on Inventive Computation Technologies (ICICT)*, pages 1294–1299. IEEE.

Behnam Mohammadi. 2024. Creativity has left the chat: The price of debiasing language models. *arXiv preprint arXiv:2406.05587*.

Bernard A. Nijstad and Paul B. Paulus. 2003. Group creativity: Common themes and future directions. In *Group Creativity: Innovation through Collaboration*. Oxford University Press.

Martha Nussbaum. 1988. Love's knowledge. *Perspectives on Self-Deception*, pages 488–514.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. 2025. Top-h decoding: Adapting the creativity and coherence with bounded entropy in text generation. *arXiv preprint arXiv:2509.02510*.

Ezra Pound. 2013. *ABC of Reading*. New Directions Publishing.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng

Su, Xin Cong, and 1 others. 2024a. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and 1 others. 2024b. Scaling large language model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alvin C Rencher. 1998. *Multivariate statistical inference and applications*, volume 635. Wiley New York.

Paul Ricoeur. 2004. *The rule of metaphor: The creation of meaning in language*. Routledge.

Louise M Rosenblatt. 1994. *The reader, the text, the poem: The transactional theory of the literary work*. SIU Press.

Gerard Salton and Michael McGill. 1983. Introduction to modern information retrieval.

Nils Schaetti. 2018. Sfgram: a dataset containing thousands of scienc-fiction books and novels. https://github.com/nschaetti/EchoTorch.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*.

Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. 2025. Megaagent: A large-scale autonomous llm-based multi-agent system without predefined sops. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4998–5036.

Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025. Igniting creative writing in small language models: Llm-as-a-judge versus multi-agent refined rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17171–17197.

Samuel S Wilks. 1932. Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Siheng Chen, and Jing Shao. 2025. Mas-gpt: Training llms to build llm-based multi-agent systems. *arXiv preprint arXiv:2503.03686*.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2024a. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. *arXiv preprint arXiv:2410.02506*.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2024b. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

LLM-as-a-Judge evaluation prompt for each aspect

System: You are an expert evaluator for creative science fiction writing.

User:
You are an expert evaluator for creative science fiction writing. Evaluate the STORY independently on each of the following aspects grounded in science fiction writing literature:

Scientific Concept Integration
- How well futuristic/scientific ideas are introduced, explained as needed, and woven into plot, setting, and character actions (not just name-dropped).
- Look for clarity, relevance, and meaningful impact on the story's events.

Speculative Logic
- The internal consistency and plausibility of the speculative elements given the story's established rules.
- Cause-and-effect coherence, logical constraints, and avoidance of convenient contradictions.

Character Depth
- Distinct, believable characters with motivations, agency, and emotional/psychological complexity.
- Growth, conflict, or meaningful choices that feel earned.

Immersive World-Building
- A vivid, coherent story world with concrete details (culture, environment, technology, institutions) that support immersion.
- World details should serve the narrative rather than overwhelm it.

Ethical and Philosophical Themes
- Presence and depth of ethical, societal, or philosophical questions typical of strong science fiction.
- Nuance, originality, and integration into narrative stakes (not purely didactic).

Scoring scale (0-5) for each of the above aspects:
- 0: Strong indicators of potential plagiarism OR extremely low-quality/derivative writing with minimal original content.
- 1: Very weak; major flaws, thin originality, little coherence or craft.
- 2: Weak-to-fair; some promising ideas but underdeveloped or inconsistent execution.
- 3: Competent; clear strengths with noticeable but non-fatal weaknesses.
- 4: Strong; well-executed, creative, and cohesive with only minor issues.
- 5: Exceptional; highly creative, polished, and deeply integrated execution across the aspect.

STORY:
{story}

SCORE:

Table 5: Prompt template for different LLM-as-a-Judge evaluation aspects (Scientific Concept Integration , Speculative Logic , Character Depth , Immersive World-Building , Ethical and Philosophical Themes) and each aspect is evaluated independently. Human annotators use the same prompt except for the system prompt part.

A SciFi-100 Overview

Figure 6 shows the data distribution of our SciFi-100 and Table 6 shows example prompts from each aspect of creative writing.

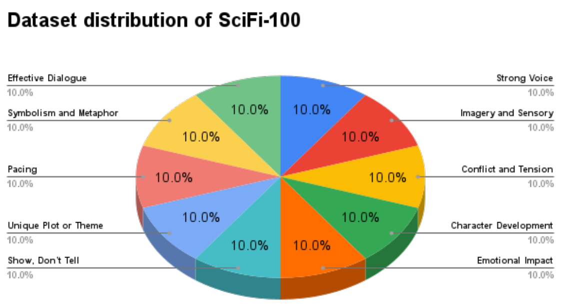

Figure 6: Dataset distribution of SciFi-100.

B LLM-as-a-Judge Evaluation Prompts

Table 5 shows our prompts for LLM-as-a-Judge evaluation on five different criterias for scientific fiction creativity writing.

C Alignment Metrics

For each story $i \in \{1, \ldots, N\}$ and evaluation dimension d, let $h_{i,d} \in [0, 5]$ denote the human consensus score (averaged over annotators) and $s_{i,d} \in [0, 5]$ denote the LLM-as-a-judge score. We compute alignment *separately for each dimension* d. For clarity, we fix a dimension d and omit the subscript d below, writing h_i and s_i.

(1) ICC(A,1): absolute agreement (two-way random effects, single measurement) We treat the human consensus and the LLM judge as two "raters" ($k = 2$) that score the same N targets (stories). Define the score matrix $X \in \mathbb{R}^{N\times k}$ by

$$x_{i1} = h_i, \quad x_{i2} = s_i.$$

Let the row mean, column mean be:

$$\bar{x}_{i\cdot} = \frac{1}{k}\sum_{j=1}^{k} x_{ij}, \quad \bar{x}_{\cdot j} = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \tag{5}$$

And the grand mean be:

$$\bar{x}_{\cdot\cdot} = \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k} x_{ij} \tag{6}$$

The ANOVA mean squares are

$$MS_R = \frac{k}{N-1}\sum_{i=1}^{N}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 \tag{7}$$

$$MS_C = \frac{N}{k-1}\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 \tag{8}$$

$$MS_E = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2}{(N-1)(k-1)} \tag{9}$$

The intraclass correlation coefficient for absolute agreement, single measurement is

$$\mathrm{ICC(A,1)} = \frac{MS_R - MS_E}{MS_R + \frac{(N-1)k-N}{N}MS_E + \frac{k}{N}MS_C} \tag{10}$$

Higher ICC indicates stronger absolute agreement (1 is perfect agreement); values can be close to 0 (weak agreement) or even negative when disagreement dominates.

(2) Pearson correlation: linear association and ranking consistency Pearson's r between $\{h_i\}_{i=1}^{N}$ and $\{s_i\}_{i=1}^{N}$ is

$$r = \frac{\sum_{i=1}^{N}(h_i - \bar{h})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{N}(h_i - \bar{h})^2}\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2}} \tag{11}$$

where $\bar{h} = \frac{1}{N}\sum_{i=1}^{N} h_i$ and $\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$. This metric captures whether stories that humans score higher also tend to receive higher judge scores.

(3) Bland-Altman: calibration via bias and 95% limits of agreement (LoA). Define the per-story difference and (optionally) the per-story mean as

$$\Delta_i = s_i - h_i \tag{12}$$

The **bias** (mean signed difference) is

$$\mathrm{bias} = \bar{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\Delta_i \tag{13}$$

and the sample standard deviation of differences is

$$SD_\Delta = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\Delta_i - \bar{\Delta})^2} \tag{14}$$

Assuming the differences are approximately normally distributed, the **95% limits of agreement** are

$$\mathrm{LoA}_{\mathrm{low,high}} = \bar{\Delta} \pm 1.96\, SD_\Delta \tag{15}$$

which estimate the interval in which the judge-human discrepancy Δ_i is expected to fall for about 95% of stories.

Aspect of Creative Writing	Example Prompt
Strong Voice	Write from the perspective of a starship mechanic on a long-haul journey who discovers a hidden compartment. Let their tone be equal parts suspicion, curiosity, and a touch of rebellious glee.
Imagery and Sensory Details	Capture the awe and danger as a spacecraft crew observes a neutron star up close, each sensation amplifying the thrill and peril of the sight.
Conflict and Tension	Show us a researcher trying to persuade a skeptical alien committee to support a high-risk, high-reward interspecies research project.
Character Development	Tell the tale of a researcher overcoming a deep-seated fear of spacewalks to collect crucial samples from an asteroid.
Emotional Impact	Convey the exhilaration of a scientist who finally finds evidence of life on another planet after decades of searching, a discovery beyond their wildest dreams.
Show, Don't Tell	Capture a character's unquenchable curiosity about alien species through their late nights, stacks of research, and endless notebook scribbles.
Unique Plot or Theme	Depict a futuristic lab with live-streamed alien dissections, complete with real-time public commentary and reactions.
Pacing	Imagine the team dynamic as anticipation builds for a groundbreaking discovery presentation, followed by a subdued, introspective moment afterward.
Symbolism and Metaphor	Use the image of a sealed, heavily guarded lab door to symbolize the secrecy and exclusivity surrounding a groundbreaking experiment.
Effective Dialogue	Imagine a heated conversation between two scientists debating the ethical implications of weaponizing a newly discovered alien technology.

Table 6: Example prompts from SciFi-100 by aspects of creative writing.

D Potential Risks

The proposed framework may amplify harmful or biased narratives present in underlying language models, as multi-agent critique and revision can reinforce shared assumptions rather than surface alternative viewpoints. In addition, LLM Review could be misused to automate large-scale creative content generation, contributing to content flooding and reducing the visibility of human authorship. Finally, its human-inspired design may encourage over-attribution of agency or originality to machine-generated outputs, highlighting the need for careful deployment and human oversight.

E The Use of Large Language Models (LLMs)

LLM was used only to aid writing quality (proofreading and polishing grammar) and generate the SciFi-100 dataset. No ideas, claims, methods, results, or references are generated by LLMs. All content decisions and revisions are made by the authors.

F Human Evaluation Protocol: Participant Instructions, Recruitment,

F.1 Instructions Given to Participants

Study overview You are invited to take part in a research study about evaluating short science-fiction stories. In this task, you will read a set of short stories (approximately 300 words each) and rate them on several quality/creativity-related dimensions. The stories you will read are machine-generated.

What you will do For each story, you will provide **five separate ratings** (integers from 0 to 5) according to the criteria below: (This part is same as Table 5 so we skipped here.)

Important guidelines

- Provide your **independent judgment**. There are no right or wrong answers.
- Use the **full 0-5 scale** when appropriate.
- Rate the story **as written**; do not assume missing details unless implied by the text.
- Do **not** spend time proofreading grammar; focus on the five criteria above.

- If you are unsure between two scores, choose the one that best matches the rubric definitions.

Risks and sensitive content notice This is a minimal-risk study. However, because the content is science fiction, some stories may include fictional depictions of conflict, danger, or other potentially sensitive themes. If you feel uncomfortable at any time, you may stop immediately or skip a story without penalty.

Privacy and data handling We record only your story ratings for analysis. We do not ask you to provide personal identifying information as part of the ratings task, except what may be needed to administer compensation (if applicable). We report results only in aggregate.

F.2 Recruitment

We recruited **nine student annotators** to rate machine-generated stories produced for prompts from SciFi-100. Participants were recruited via university mailing. Inclusion criteria were: (i) age 18 or older, (ii) proficient in English reading comprehension, and (iii) willingness to read and rate short science-fiction stories.