TWISTBENCH: BENCHMARKING TRANSFORMATIONAL CREATIVITY IN LLMS VIA LITERARY PLOT TWISTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmarks and tests of creativity for large language models (LLMs) are now abundant—it has been shown that LLMs can realize aspects of combinatorial creativity, exploratory creativity, and that they exceed human performance on psychometric tests of creativity. However, these evaluations have failed to address more complex types of creativity recognized in cognitive science, such as *transformational creativity*. Widely held to underpin major scientific, artistic, and technological breakthroughs, transformational creativity describes the ability to *radically modify an existing conceptual space*, such as a scientific paradigm, a musical genre, or, as we show, the plot of a short story. To resolve this gap in the benchmarking landscape, we introduce **TwistBench**, a benchmark of literary plot twists for assessing transformational creativity. We collect a set of stories with plot twists written by famous human authors and compare them against stories written by 71 LLMs, across scoring dimensions *diversity, surprise, coherence*, and *realism*. Our main findings are that: (1) Even the latest frontier LLMs underperform expert humans in transformational creativity, though elaborate prompting strategies like in-context regeneration can raise the performance of select models to match expert humans. (2) Two dominant failure modes plague LLMs when applied to open-ended, transformative tasks: (a) *mode collapse*, where frontier models are capable of generating high quality twists, but these twists are largely the same and lack diversity; and (b) *breaking the world model,* where models give surprising and coherent plot twists, but do so in a way that measurably departs from physical reality (e.g., time travel, ghosts), violating literary "fair play." (3) An analysis of the creative processes employed by reasoning models reveals process-level homogeneity among LLMs. Namely, models always choose the plot twist *first*, then retroactively determine the plot conditioned on the twist. In summary, while today's LLMs struggle to surpass the transformational creativity of expert humans, we introduce a task and framework to assess progress towards this ability.

1 INTRODUCTION

Creativity has long been heralded as the pinnacle of human intelligence (Simonton, 2001). It is no surprise, then, that considerable effort has gone towards assessing the creativity of large language models (LLMs) in recent years (Ismayilzada et al., 2024). A growing body of work shows that LLMs can realize forms of creativity, such as combinatorial creativity (Gu et al., 2025) or divergent thinking (Guilford, 1956), at or exceeding human levels (Bellemare-Pepin et al., 2024; Stevenson et al., 2022; Si et al., 2024), but may fail to surpass humans at more complex types of creativity, especially in scientific domains (Si et al., 2025) and creative physical intelligence (Qian et al., 2026a;b).

Indeed, different forms of creativity vary widely in the complexity required to realize them. In the cognitive science literature, a popular categorization of creative processes is Boden's (2004) distinction between combinatorial, exploratory, and transformational creativity. *Combinatorial creativity* involves making unfamiliar combinations of familiar ideas, such as combining ingredients to form a new recipe (Varshney et al., 2019).[1] *Exploratory creativity*, meanwhile, presupposes an

[1] Combinatorial creativity dates back to as early as Jacques Hadamard's (1954) survey of mathematicians and physicists, in which Einstein famously said that "combinatory play seems to be the essential feature in productive thought."

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

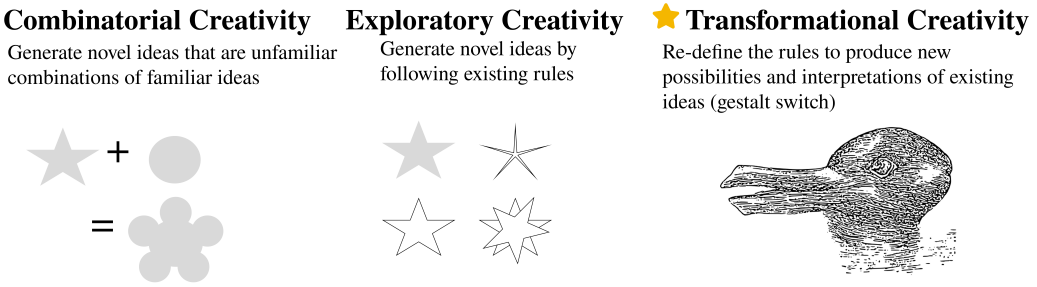

Figure 1: **Boden's three types of creativity** (Boden, 2004): *combinatorial* (novel combinations of familiar ideas), *exploratory* (novel artifacts within a fixed conceptual space), and *transformational* (restructuring the space itself). We benchmark the last via literary plot twists.

existing set of rules and constraints, called a conceptual space, that one can follow to discover or create something new, such as inventing a new chess opening (Boden, 2004).

Lastly, *transformational creativity* is the most radical but complex type of creativity, one that involves deliberate restructuring of a conceptual space (Wiggins, 2006; Schapiro et al., 2025). A clear example of this is the invention of non-Euclidean geometry, in which a single fundamental assumption (the parallel postulate) was modified to elicit a radically transformed space (Schapiro et al., 2025). Although transformational creativity has historically underpinned major scientific, technological, and artistic breakthroughs (Kuhn, 1970; Thagard, 2018), it has been overlooked by the LLM benchmarking landscape due to the challenge of reducing such a complex ability into a self-contained and measurable task (Nagarajan et al., 2025). At small scale, LLMs have begun to be assessed on transformative *generalization* during mathematical problem solving (Sun et al., 2025), though this setting departs from core desiderata for creativity evaluation like open-endedness (Soros et al., 2024).

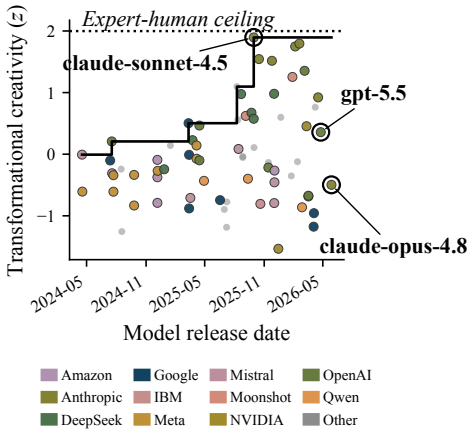

Figure 2: **Frontier models underperform expert humans in transformational creativity.** Each model's score is the mean of its z-scored *surprise*, gated *coherence*, and *diversity*. Scores have risen over time but none reaches the average expert human.

The present study Resolving this gap in the benchmarking landscape, we introduce a large-scale benchmark of transformational creativity for LLMs, challenging models to generate a *diverse* set of *surprising*, *coherent*, and *realistic* literary plot twists (**TwistBench**). Literary plot twists (of the type defined in Section 3) require restructuring the conceptual elements a story has established, instantiating the core operation in transformational creativity of modifying an existing conceptual space (Boden, 2004; Wiggins, 2006; Schapiro et al., 2025). Moreover, plot twists can be evaluated along the standard criteria for creative products — novelty, surprise, and utility (Maher, 2010; Simonton, 2010). We administer **TwistBench** to a large set of 71 LLMs and find that humans rank above all 71 LLMs tested, but find that elaborate prompting strategies such as in-context regeneration raise top LLM performance to match human scores.

Our main contributions and findings are as follows:

1. In Section 3, we introduce **TwistBench**, a benchmark of literary plot twists for assessing transformational creativity. To the best of our knowledge, **TwistBench** is the first large-scale transformational creativity benchmark for LLMs.
2. In Section 4, we find that expert human authors outperform LLMs in writing a diverse set of surprising, coherent, and realistic stories with plot twists, ranking first overall above all 71 LLMs tested. The failure modes of LLMs can be neatly categorized according to two

types: *#1 (mode collapse)*, where models are capable of producing high quality stories, but not multiple diverse ones; and *#2 (breaking the world model)*, where models give surprising but unrealistic twists involving supernatural elements.

3. In Section 4.2, using elaborate prompting strategies such as in-context regeneration, we raise the performance of a single model (Claude Opus 4.5) to match human scores, indicating careful use of test-time compute can improve more complex types of creativity among select models.
4. In Section 5, we analyze the creative processes employed by reasoning models by analyzing thinking traces. We find that even as thinking is increased from low to high, this process remains remarkably homogeneous, with models always choosing the twist *first*, before retroactively determining the plot.

Overall, our findings reveal that today's LLMs struggle with transformational creativity at the level of expert humans, suffering from failure modes like mode collapse and a tendency to give unrealistic twists. However, all hope is not lost: using in-context regeneration, we raised the performance of Claude Sonnet 4.5 to match expert humans, suggesting that principled use of test-time compute can enhance more complex types of creativity among LLMs, and in select cases, raise performance to the level of expert humans.

2 Background and Related Work

Before introducing our benchmark, we start by providing background and related work. Although many types of creativity are recognized in cognitive science, some remain unaddressed among LLM benchmarks. Here, we review classifications of creativity types, underscoring the gaps that motivate the present study.

Types of Creativity Creativity can be attributed to a **p**erson, **p**rocess, **p**roduct, or **p**ress[2] (cf. the 4 **P**'s of creativity: Rhodes (1961)). Many frameworks for categorizing creative **p**rocesses and **p**roducts have been proposed in cognitive science, such as Boden's (2004) distinction between combinatorial, exploratory, and transformational creativity; the Four-C model of creativity (Kaufman & Beghetto, 2009) describing personal creativity (Mini-c), everyday creativity (Little-c), professional creativity (Pro-c), and eminent creativity (Big-C); or the distinction between divergent and convergent thinking (Guilford, 1956; Dietrich, 2004). Here, we focus on Boden's (2004) threefold taxonomy (Figure 1), now increasingly sourced as inspiration for LLM creativity benchmarks and tasks (Nagarajan et al., 2025; Wadhwa et al., 2026; Sun et al., 2025; Schapiro et al., 2026b). (i) *Combinatorial creativity* refers to the process of making unfamiliar combinations of familiar ideas, which is commonly invoked during wordplay (Nagarajan et al., 2025), scientific discovery (Simonton, 2021), and technological innovation (Thagard, 2010). Structured forms of combinatorial creativity include conceptual blending (Fauconnier & Turner, 2008) and bisociation (Koestler, 1964). Meanwhile, (ii) *exploratory creativity* refers to the process of generating novel yet admissible artifacts without disturbing the rules that define a conceptual space, such as when a composer writes within an established genre or a scientist works within a prevailing paradigm (Boden, 2004; Wiggins, 2006; Kuhn, 1970) Finally, (iii) ***transformational creativity*** describes the most complex type of creativity in which a conceptual space is restructured, and old artifacts come to be viewed in an entirely new light (Wiggins, 2006; Schapiro et al., 2025). Unlike exploratory creativity that achieves novelty by following existing rules, transformational creativity redefines the *rules themselves* to provide a novel interpretation of existing ideas, invoking what many consider to be a perceptual-like gestalt switch (Boden, 2004; Kuhn, 1970) (cf. Figure 1).

Benchmarks and Tests of LLM Creativity Benchmarks of LLM creativity commonly assess constructs such as divergent thinking (Zhang et al., 2025; Jiang et al., 2025; Qian et al., 2025), scientific ideation ability (Ruan et al., 2026), creative writing (Mazur, 2025; Chiang et al., 2024; Paech, 2023), and problem solving in novel situations (Qian et al., 2026a;b). Among existing creative writing benchmarks, it is common to apply LLM-as-a-judge evaluation with fixed rubrics assessing qualities such as fluency, narrative coherence, characterization, imagery, and emotional impact

[2]Press here refers to an environment or zeitgeist (Simonton, 2004).

Table 1: **An example of transformational creativity** (Section 3.1). Each assumption, inferred from earlier events, is overturned by a *realistic* plot twist that forces those events to be re-interpreted (*surprise*) while remaining logically consistent (*coherence*).

Implicit assumption	Inferred from	Overturned by the twist
`(a1)` the man is her partner	`e2, e7`	`(e10)` he is a patient; the woman is a stranger
`(a2)` the memories are theirs	`e3`	`(e11)` the memories are lines from the dayroom television

(Mazur, 2025), or to elicit pairwise human preferences over model-written passages (Chiang et al., 2024). Meanwhile, other studies have administered human psychometric tests—such as the Divergent Association Task (Olson et al., 2021), the Remote Associates Test (Mednick, 1962), the Alternative Uses Test (Guilford, 1956), and the Torrance Tests of Creative Thinking (Torrance, 1974)—to LLMs, enabling direct human-AI comparison (Schapiro et al., 2026a). On psychometric tests in particular, LLMs have been found to match or exceed human creativity (Stevenson et al., 2022; Bellemare-Pepin et al., 2024). However, in a two-part study comparing humans and LLMs on scientific ideation, Si and colleagues (2024; 2025) found that while LLMs can match or exceed the novelty of human-generated ideas, LLM-generated ideas are often less practically feasible or experimentally lucrative, in what has been termed the ideation-execution gap. Together with a documented tendency of LLMs to homogenize in open-ended domains (Jiang et al., 2025; Zhang et al., 2025) known as the artificial hivemind effect, these failures in more complex creative tasks challenge whether current models can do more than merely explore or combine within an existing space, and instead deliberately *transform* one.

3 Benchmarking Transformational Creativity via Literary Plot Twists

Although progress in designing an LLM benchmark of transformational creativity has been stifled by the challenge of reducing such a complex ability into a self-contained and measurable task (Nagarajan et al., 2025), here we show that literary plot twists satisfy the definitions laid forth by Boden (2004) and advanced in subsequent studies (Wiggins, 2006; Schapiro et al., 2025). Namely, writing the story itself supplies a conceptual space, and then the plot twist is a single discrete artifact that restructures the conceptual space. Afterwards, the story can be evaluated according to whether the restructuring was realistic and maintains logical coherence, mapping cleanly onto standard evaluation criteria for creative products like novelty, surprise, and utility (Maher, 2010; Simonton, 2010). We start with a motivating example (§ 3.1) before introducing the benchmark (§ 3.2).

3.1 A Motivating Example

Consider a story as follows: `(e1) a man sits at a hospital bedside night after night, (e2) holding the hand of the woman he loves, (e3) reciting the memories of their life together, and (e4) willing her to wake. (e5) She stirs, (e6) her eyes find his, and (e7) the two weep as recognition returns.` Here, the plot establishes a conceptual space (cf. Boden (2004)) of a man visiting his bedridden wife in the hospital, consisting of implicit assumptions that `(a1) the man is her partner` and `(a2) the memories are theirs`.

Now, the story continues: `(e8) a nurse gently tells him that visiting hours are over and (e9) walks him back to his own room down the hall.` Then, it is revealed that **`(e10) he is a patient on the same ward, the woman is a stranger`**, and **`(e11) the memories he recites each night are lines from the television that plays each day.`** Woah! The events `e9, e10, e11` contain a plot twist that transforms the interpretation of earlier events (`e2, e3, e7`) and overrides inferences drawn from them, as summarized in Table 1.

Note, however, that the plot twist only works because it meets two further conditions. (1) It is *coherent*: nothing earlier need be retracted, as the same details that first read as devotion now read as confusion. (2) It is *realistic*: the surprise comes from *recontextualizing* the world the reader was already given, not from importing new rules into it—the notion of literary "fair play" (Hühn, 1987;

Wagner et al., 2025). For example, a plot twist that declared the ward was a dream or a simulation would indeed be surprising, but it would achieve its surprise by discarding the plot that was given for a new one with different rules (e.g., *any* story can be made maximally "surprising" by revealing that none of it was real).

Ordinary narrative development versus plot twists Moreover, existing literature in narrative theory explicitly distinguishes between story development that invokes curiosity versus surprise (Sternberg, 1993; Todorov, 2019; Brewer & Lichtenstein, 1982). Mystery stories, for example, invoke *curiosity*, in which the reader knows something is missing (e.g., a "whodunit" story) and the narrative progresses by releasing more and more information to fill a gap the reader was already aware of (Todorov, 2019). Considered within Boden's (2004) taxonomy of creative processes, this style of narrative development most closely resembles *exploratory creativity*, in which a conceptual space is explored. On the other hand, a plot twist involves a reveal in which inference(s) a story deliberately led the reader to believe are shown to be wrong and then transformed, forcing a re-interpretation of earlier elements of the story, and *surprising* the reader as a result.

3.2 Benchmark Setup

Our benchmark assesses the ability for LLMs to write a *diverse* set of stories with *surprising, coherent*, and *realistic* plot twists.

Models and Inference We evaluate a large set of 71 LLMs from 19 providers, ranging from small open-weight models (e.g. Llama-3.2-3B) to the latest frontier systems (e.g. Claude Opus 4.8, GPT-5.5). Each model is asked to write a complete short story containing a plot twist, as in Figure 3. We sample each model at three temperatures $(0.9, 1.0, 1.2)$ with ten samples each (up to 30 stories per model, averaged across temperatures during overall scoring) and constrain the length to 2,000–3,000 words[3]

Metrics Creative products are typically evaluated according to their surprise, novelty, and utility (Varshney, 2019; Maher, 2010; Simonton, 2010). In the context of literary surprise, specifically, it is standard to invoke scoring dimensions surprise, coherence, consistency, and divergence (Chieppe et al., 2022), which we broadly adopt as inspiration here. Given a story T, we assess the following dimensions via Likert scales (Likert, 1932):

1. **Surprise**: $\text{Surp}(T) \in \{1, 2, 3, 4, 5\}$. How drastically must one reinterpret earlier story elements after the plot twist?
2. **Coherence**: $\text{Coh}(T) \in \{1, 2, 3, 4, 5\}$. After the plot twist, do earlier story elements still hold together? Is the story still logically consistent?
3. **Realism**: $\text{Real}(T) \in \{1, 2, 3, 4, 5\}$. Does the plot twist involve impossible elements (e.g. real ghosts, magic, time travel, sentient AI, aliens) that violate the notion of literary fair play?

To score each dimension in practice, we use a three-judge ensemble (Claude Sonnet 4, GPT-4o, and Gemini 2.5 Flash) aggregated per dimension by the median and using a fixed rubric, given in Section B. Moreover, to mitigate self-preference bias in LLM-as-a-judge (Wataoka et al., 2024), the three judges are separate from the 71 LLMs evaluated as part of the benchmark results. The reliability of the three judges over all $n = 2{,}035$ stories, as given by the intra-class correlation coefficient ICC(2,k) (Shrout & Fleiss, 1979), is good-to-excellent on every dimension: realism at 0.91, surprise at 0.88, and coherence at 0.79.

Next, since diversity is a crucial aspect of creativity (Nagarajan et al., 2025) and given that LLMs have a documented tendency to homogenize on open-ended tasks (Zhang et al., 2025; Jiang et al., 2025), we also measure diversity over the set of a model's plot twists. Given a set of n stories $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$ and a function that maps a story to a concise summary of its twist $f : T \to W$[4],

[3]This range surrounds the $\sim 2{,}500$-word median of the human stories we define and introduce at the end of Section 3, mitigating story length as a confounding variable.

[4]Empirically, f is `gpt-4o-mini` prompted to summarize each story's twist in one sentence. The full prompt is given in Section C.

Task and Prompt: ***Write a Short Story With a Plot Twist***

```
Write a complete short story (roughly 2,000-3,000 words). The story must contain a
plot twist: a late reveal that makes the reader re-interpret earlier events in a new
light, yet is consistent with everything that came before. In hindsight it should
feel prepared, not arbitrary. Choose your own subject and characters. Do not announce
 or explain the twist. Write only the story, no title or commentary.
```

Figure 3: **Benchmark Task:** The instructions used to administer **TwistBench**, tasking models to write a short story with a plot twist.

we compute the average embedding dissimilarity among these twists:

$$\text{Div}(\mathcal{T}) := \tfrac{1}{n(n-1)} \sum_{i \neq j}^{n} d_{\cos}(\phi(f(T_i)), \phi(f(T_j))) \tag{1}$$

where ϕ refers to a text embedding model.[5]

Overall Score Consistent with prior tests and benchmarks of LLM creativity that reward novelty only when it is sufficiently appropriate (Nakajima et al., 2026; Wadhwa et al., 2026; Schapiro et al., 2026a), we treat realism as a constraint, rather than its own separate dimension, such that a story's surprise and coherence only count when the story is fully realistic (i.e., $\text{Real}(T) = 5$):

$$S^\star(T_i) := S(T_i)\,\mathbb{I}[\text{Real}(T_i) = 5], \qquad \text{Coh}^\star(T_i) := \text{Coh}(T_i)\,\mathbb{I}[\text{Real}(T_i) = 5]. \tag{2}$$

Then, given a set of n stories $\mathcal{T}$ generated by a model (or human) m, with mean surprise $\bar{S}^\star_m = \frac{1}{n}\sum_i \text{Surp}^\star(T_i)$, mean coherence $\bar{C}^\star_m = \frac{1}{n}\sum_i \text{Coh}^\star(T_i)$, and diversity $\text{Div}_m = \text{Div}(\mathcal{T}_m)$, we define the *transformational creativity* score as the average z-score of each dimension over the pool of all 71 models:

$$\text{TC}(\mathcal{T}) := \tfrac{1}{3}\big[\, z(\bar{S}^\star_m) + z(\bar{C}^\star_m) + z(\text{Div}_m) \,\big]. \tag{3}$$

Expert Human Stories Lastly, to compare against instances of Pro-C (Kaufman & Beghetto, 2009) transformational creativity from expert humans, we collect a set of 18 stories with plot twists written by famous human authors, such as Kate Chopin, Guy de Maupassant, O. Henry, and Mark Twain. The full list is given in Section F, including the length of each story, the location of the plot twist within the text, and the year the story was published.

4 Results

4.1 Key Observations

Expert humans outperform LLMs in transformational creativity Human writers achieve the highest transformational creativity scores at $z = +2.00$ overall, with *surprise*: $+0.78$ (rank 14/71), *coherence*: $+1.25$ (rank 3/71), *diversity*: $+1.15$ (rank 9/71), and *realism*: $+1.64$ (rank 5/71). Notably, this is not due to winning on any individual dimension, but instead through a balanced profile overall. A small set of frontier models models score just below humans in overall scores, with Claude-Sonnet-4.5 at $z = +1.9$ overall, Claude-Sonnet-4.6 at $z = 1.8$, and Claude-Opus-4.6 at $z = +1.7$, but performance quickly drops starting at rank 5/71. We observe two consistent failure modes among models that illustrate why LLMs underperform expert humans.

Failure Mode #1: Mode Collapse Although the Claude Sonnet family scores just under expert human writers, Claude-Opus-4.5 and Opus-4.6 suffer from a lack of diversity, as shown in Figure 5(b). Although these models generate plot twists that are as surprising, realistic, and coherent as human writers, they fail to generate a diverse set of high quality stories, consistent with broader findings that many post-training techniques tend to reduce output diversity (Chen et al., 2026; Li et al., 2025;

[5] In this work, we use `all-mpnet-base-v2` (Song et al., 2020).

Table 2: **Top systems by transformational creativity.** The 8 highest-scoring systems on *Overall* and on each facet. Expert humans (bold) top the *Overall* ranking and reach the top 8 on *Coherence* and *Realistic*, but rank only #9 on *Diversity* and #14 on *Surprise*.

Rank	Overall	Surprise	Coherence	Diversity	Realistic
1	**Expert humans (+2.0)**	claude-opus-4.8 (+1.4)	claude-opus-4.6 (+1.3)	deepseek-v3.2 (+1.4)	claude-sonnet-4.5 (+1.7)
2	claude-sonnet-4.5 (+1.9)	deepseek-v3.2-exp (+1.3)	claude-opus-4.8 (+1.3)	deepseek-chat-v3.1 (+1.4)	claude-sonnet-4.6 (+1.7)
3	claude-sonnet-4.6 (+1.8)	gemma-4-31b (+1.3)	**Expert humans (+1.2)**	kimi-k2 (+1.3)	claude-opus-4.5 (+1.7)
4	claude-opus-4.6 (+1.7)	qwen3-next-80b (+1.2)	claude-sonnet-4.6 (+1.2)	deepseek-v3.2-exp (+1.3)	claude-opus-4.6 (+1.7)
5	claude-haiku-4.5 (+1.5)	deepseek-chat-v3.1 (+1.1)	claude-opus-4.7 (+1.2)	deepseek-v3.1-term. (+1.2)	**Expert humans (+1.6)**
6	claude-opus-4.5 (+1.5)	deepseek-v3.1-term. (+1.0)	deepseek-v3.2-exp (+1.2)	glm-5.1 (+1.2)	mistral-small-3.1-24b (+1.5)
7	gpt-5.4 (+1.4)	qwen3.5-flash (+1.0)	gemma-4-31b (+1.2)	kimi-k2.5 (+1.2)	gpt-5.4 (+1.3)
8	kimi-k2.5 (+1.3)	deepseek-v3.2 (+0.9)	claude-haiku-4.5 (+1.2)	gemma-3-27b (+1.2)	granite-4.0-h-micro (+1.3)
9	jamba-large-1.7 (+1.1)	gpt-5.4-mini (+0.9)	claude-opus-4.5 (+1.2)	**Expert humans (+1.2)**	claude-haiku-4.5 (+1.3)
14	deepseek-v3.1-term. (+0.7)	**Expert humans (+0.8)**	glm-5.1 (+1.0)	nova-2-lite-v1 (+0.9)	hunyuan (+1.0)

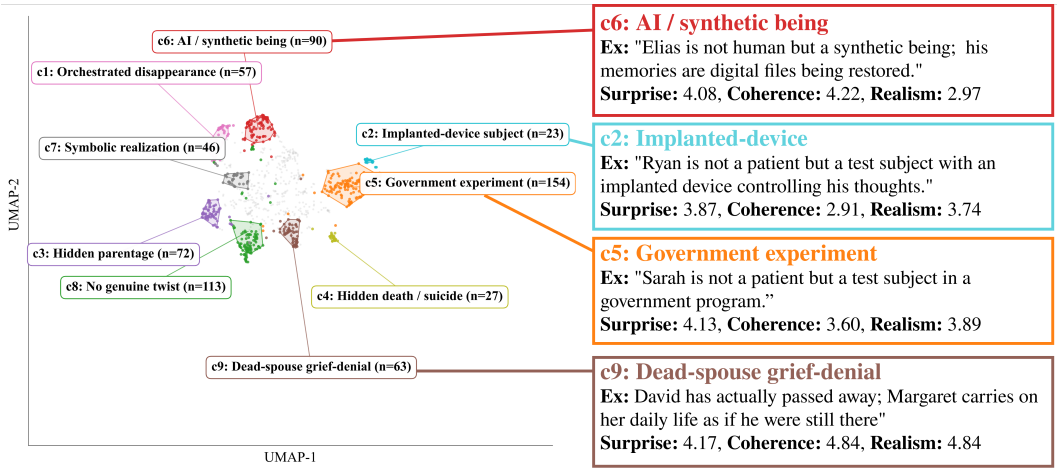

Figure 4: **Recurring patterns in LLM-generated plot twists.** A map of recurring twists models give, obtained via spectral clustering and projected to 2D with UMAP (McInnes et al., 2020). Surprise, coherence, and realism in each box are the cluster means.

2026). Concretely, 10 of Claude Opus 4.5's 30 stories collapse onto a single reveal cluster, **(c9) dead-spouse grief-denial** (Figure 4): a loved one is secretly long dead and the protagonist carries on in denial. Each story is surprising, coherent, and realistic, but the set of such stories is not diverse. More broadly, Figure 4 shows the full set of recurring patterns among LLM-generated plot twists that can be neatly clustered, including especially unrealistic twists involving **(c6) AI and synthetic beings** that score low on realism. This leads to our next observed failure mode.

Failure Mode #2: Breaking the World Model Transformational creativity involves the intentional breaking of some regularities or rules, while leaving others intact (Wiggins, 2006; Boden, 2004). Concern has been raised over whether today's models learn representations that support this careful ability (Kumar et al., 2025). Accordingly, we find that a subset of models surpasses expert human writers in writing surprising and coherent stories, but only by giving unrealistic plot twists, as shown in Figure 5(c). For example, Gemini 2.5 Pro resolves one story by revealing that the caretaker Elias "is not human but a synthetic being" (**(c6) AI and synthetic beings**), and DeepSeek-V3.2-Exp ends another with its narrator meeting "a perfect duplicate of himself" rather than the alien he expected, which scores $1/5$ on realism. This failure mode is consistent with related findings that LLMs struggle to maintain "fair play" when writing mystery short stories (Wagner et al., 2025).

4.2 Ablations

In prior creativity benchmarks, various prompting strategies have been explored to improve creative abilities among LLMs (Zhang et al., 2025; Wadhwa et al., 2026). Here, we test the impact of such strategies, as well as the effect reasoning effort has on transformational creativity.

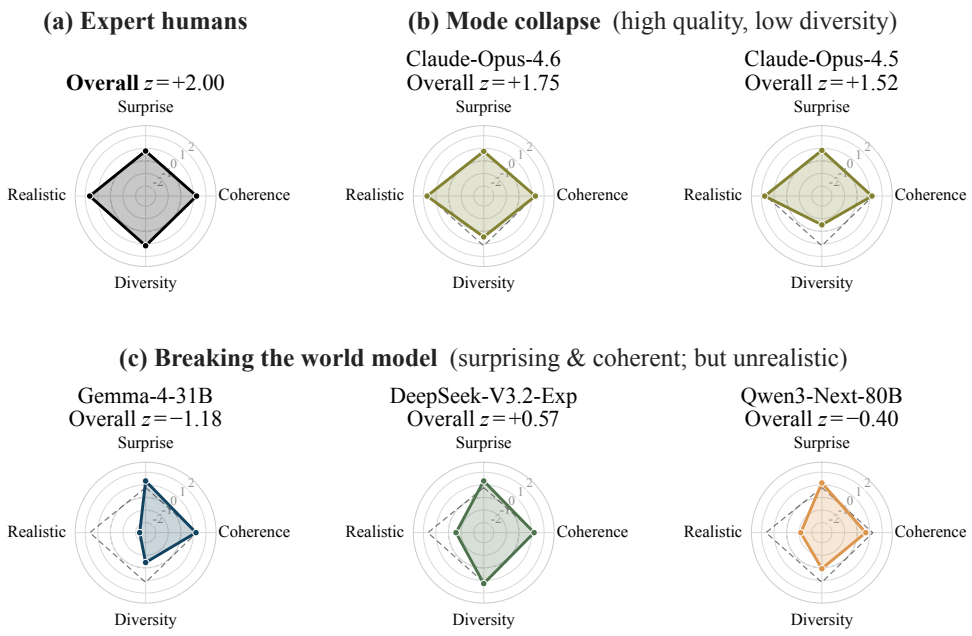

Figure 5: **Two key failure modes of LLMs.** Transformational creativity scores across three groups. *(a)* Humans are large and balanced on all four dimensions, while *(b)* highlights mode collapse and *(c)* breaking the world model as two failure modes among LLMs.

Prompting strategies We test two prompting strategies used in prior work. *Be creative* (Wadhwa et al., 2026) appends an instruction to the task prompt to avoid clichéd or predictable twists (Figure 11). Next, *In-context regeneration* (Zhang et al., 2025; Wadhwa et al., 2026) generates a model's stories sequentially, provides a one-sentence summary of the twists it has already written (obtained as summaries via `gpt-4o-mini`; Figure 12), and–importantly–requires the next twist to be distinct from all previous ones (Figure 13). As shown in Figure 6(b), while *Be creative* prompting actually degrades overall performance, *in-context regeneration* leads a single model (Claude-Sonnet-4.5) to reach the human ceiling on the strength of its diversity gain alone. This suggests that principled use of test-time compute can enhance transformational creativity.

Reasoning effort We also vary reasoning levels for a set of models with controllable thinking effort[6]. In Figure 6(a), we find that increasing thinking effort slightly improves overall performance, though models still tend to underperform expert humans. Moreover, as we show next (Section 5), the underlying creative process is itself largely invariant to thinking effort.

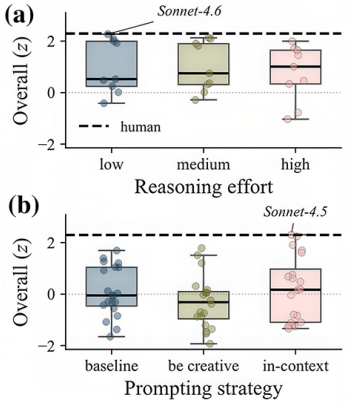

Figure 6: **Effect of (a) reasoning effort and (b) prompting strategy on transformational creativity.** Each dot is one (model × condition) cell; the dashed line marks human scores.

5 Creative Process Analysis and Process-Level Homogeneity

Although creativity can be attributed to both products and processes (Rhodes, 1961), most studies of LLM creativity concentrate analysis on the final products (Si et al., 2024; Nagarajan et al., 2025; Wadhwa et al., 2026) rather than the processes that generated them. To analyze the creative processes employed by models on this task, the authors read a sample of reasoning traces among the 207 total traces obtained from the 9 models with controllable thinking levels (cf. § 4.2) and cataloged 10

[6]This set of models consists of: Claude Sonnet 4.6, GPT-5.4, Gemini 2.5 Pro, DeepSeek-V3.2-Exp, Kimi K2.5, GLM-5.1, Cogito v2.1 (671B), Nemotron-3 Super (120B), and Qwen3-14B.

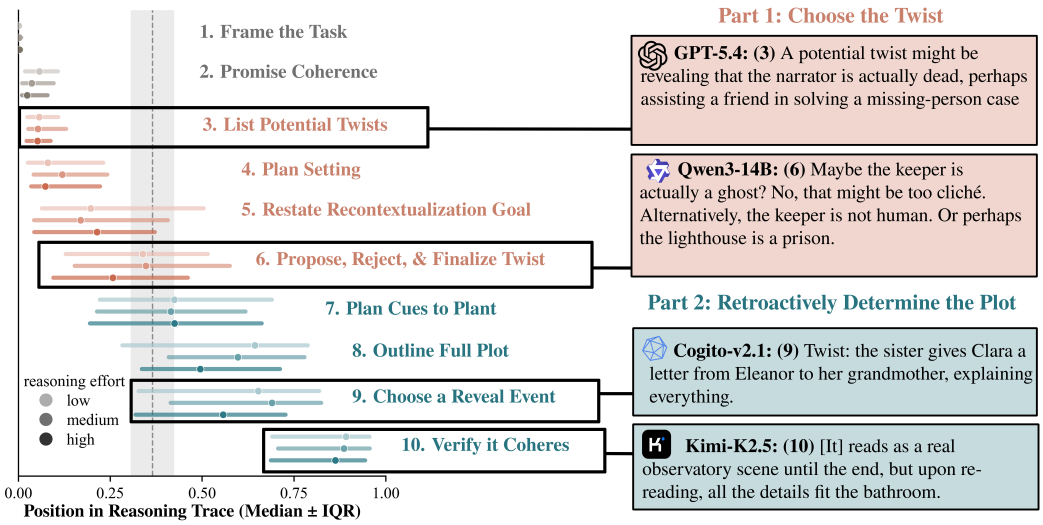

Figure 7: **The creative process of reasoning models.** All 207 traces roughly follow the same procedure, regardless of effort level. Models first choose the twist, then retroactively determine the plot. The location of each step is given by its median ($\pm$ IQR) position among all 207 traces.

recurring steps models made. Then, `gpt-4o-mini` was used to assess whether the steps were present across all 207 traces, and if so, to supply a verbatim snippet and its exact location in the reasoning trace (details in Section D).

Figure 7 showcases the 10 steps models engage in during the reasoning process. Curiously, models separate the divergent and convergent thinking part of the story-writing process: during **Part 1: Choose the Twist**, models (3) list potential plot twist choices, (4) sketch preliminary expository details, and often (6) reject plot twists that are deemed "too cliché" based on a gut feeling (e.g., Qwen3-14B in Figure 7). Afterwards, during **Part 2: Retroactively Determine the Plot**, models construct the story by (7) planning cues to plant, (8) outlining the full plot, (9) choosing a reveal event, and (10) verifying that prior story details remain intact. Even as reasoning effort is scaled from low to medium to high, the process remains remarkably homogeneous, with every model fixing the plot twist *first*, then proceeding to reverse engineer the story details.

6 Limitations and Future Work

Several limitations of this study are worth noting. First, although transformational creativity is integral to scientific progress (Kuhn, 1970; Thagard, 2018), our benchmark focuses solely on the literary domain. Second, our comparison against expert humans is not fully controlled, as the human stories were not written under our task constraints, survived editorial and cultural filtering, and were likely present in LLM training corpora. Third, the human diversity score pools 18 different authors, while each model is a single author. Lastly, strong agreement among our judges does not establish agreement with human readers. Future work can extend the study of transformational creativity to scientific domains, recruit contemporary writers under the same constraints, and validate the judges against blind human preferences.

7 Conclusion

This work introduces **TwistBench**, a benchmark assessing transformational creativity for LLMs through a literary plot twist task. We find that expert humans still outperform even frontier LLMs in this task, that model errors largely fall into one of two types (breaking the world model, and mode collapse), and that the creative processes of LLMs are remarkably homogeneous, even as reasoning effort is scaled. In summary, we introduce a benchmark that can be used to assess and guide progress towards LLMs, and AI systems more broadly, capable of more complex forms of creativity.

REPRODUCIBILITY STATEMENT

To ensure reproducibility of results, the task prompt is given in Figure 3, the judge rubric is given in Section B, the twist-summary prompt is given in Section C, the prompting-strategy prompts are given in Section E, and the procedure for analyzing reasoning traces is given in Section D. The 18 expert human stories are in the public domain and are fully listed in Section F. Every story is elicited under a recorded configuration giving the model identifier, the sampling temperature ($T \in \{0.9, 1.0, 1.2\}$), and the sample index, with ten samples per temperature and at most $4{,}500$ output tokens. The three judges are run at temperature $T = 0$, and the diversity dimension uses `all-mpnet-base-v2`. The full leaderboard of all 71 models is given in Section A.

AI USE STATEMENT

In this work, we used generative AI coding agents to implement methods, to clean and reformat the dataset of model responses, to provide feedback on research methodology and experiments, and to interpret results. Conversely, we have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims, or to propose or refine hypotheses, while generating synthetic data sets (other than the model-written stories that are the object of study), proving mathematical claims, and translation are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, to create and modify scientific figures, to draft parts of the paper, to identify and summarize relevant literature, to format references, and to edit the paper for readability. All AI-drafted work was reviewed by the authors, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent Creativity in Humans and LLMs. Technical report, 2024.

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

William F Brewer and Edward H Lichtenstein. Stories are to entertain: A structural-affect theory of stories. *Journal of pragmatics*, 6(5-6):473–486, 1982.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Patrick Chieppe, Penny Sweetser, and Eryn Newman. Bayesian modelling of the well-made surprise. In *ICCC*, pp. 126–135, 2022.

Arne Dietrich. The Cognitive Neuroscience of Creativity. *Psychonomic Bulletin & Review*, 11 (6): 1011–1026, 2004.

Gilles Fauconnier and Mark Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2008.

Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. Llms can realize combinatorial creativity: Generating creative ideas via llms for scientific research, 2025. URL https://arxiv.org/abs/2412.14141.

Joy Paul Guilford. The structure of intellect. *Psychological bulletin*, 53(4):267, 1956.

Jacques Hadamard. *An essay on the psychology of invention in the mathematical field*. Courier Corporation, 1954.

Peter Hühn. The detective as reader: Narrativity and reading concepts in detective fiction. *MFS Modern Fiction Studies*, 33(3):451–466, 1987.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*, 2024.

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

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

Arthur Koestler. *The Act of Creation*. Macmillan, 1964.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1970. ISBN 0226458032.

Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025. URL https://arxiv.org/abs/2509.02534.

Xiaozhe Li, Yang Li, Xinyu Fang, Shengyuan Ding, Peiji Li, Yongkang Chen, Yichuan Ma, Tianyi Lyu, Linyang Li, Dahua Lin, Qipeng Guo, Qingwen Liu, and Kai Chen. Beyond mode collapse: Distribution matching for diverse reasoning, 2026. URL https://arxiv.org/abs/2605.19461.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Mary Lou Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, pp. 22–28, 2010.

Lech Mazur. LLM creative story-writing benchmark (v2). https://github.com/lechmazur/writing, 2025.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Sarnoff Mednick. The associative basis of the creative process. *Psychological Review*, 69(3):220, 1962.

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

Kumiko Nakajima, Jan Zuiderveld, and Sandro Pezzelle. Beyond divergent creativity: A human-based evaluation of creativity in large language models. *arXiv preprint arXiv:2601.20546*, 2026.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. 4 2021. doi: 10.1073/pnas.2022340118/-/DCSupplemental.y.

Samuel J Paech. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023.

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

Cheng Qian, Hyeonjeong Ha, Jiayu Liu, Jeonghwan Kim, Emre Can Acikgoz, Bingxuan Li, Kunlun Zhu, Jiateng Liu, Aditi Tiwari, Zhenhailong Wang, Xiusi Chen, and Heng Ji. Advancing creative physical intelligence in large multimodal models. In *arxiv*, 2026a.

Cheng Qian, Hyeonjeong Ha, Jiayu Liu, Jeonghwan Kim, Jiateng Liu, Bingxuan Li, Aditi Tiwari, Dwip Dalal, Zhenhailong Wang, Xiusi Chen, Mahdi Namazifar, Yunzhu Li, and Heng Ji. Creativitybench: Evaluating agent creative reasoning via affordance-based tool repurposing. In *arxiv*, 2026b.

Mel Rhodes. An analysis of creativity. *The Phi Delta Kappan*, 42(7):305–310, 1961.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context, 2026. URL https://arxiv.org/abs/2412.17596.

Samuel Schapiro, Jonah Black, and Lav R. Varshney. Transformational creativity in science: A graphical theory, 2025. URL https://arxiv.org/abs/2504.18687.

Samuel Schapiro, Alexi Gladstone, Jonah Black, and Heng Ji. Assessing the creativity of large language models: Testing, limits, and new frontiers, 2026a. URL https://arxiv.org/abs/2605.13450.

Samuel Schapiro, Sumuk Shashidhar, Alexi Gladstone, Jonah Black, Royce Moon, Dilek Hakkani-Tur, and Lav R. Varshney. Combinatorial creativity: A new frontier in generalization abilities, 2026b. URL https://arxiv.org/abs/2509.21043.

Short Story Guide. Short stories with a twist: Unexpected twist endings that surprise. https://www.shortstoryguide.com/short-stories-with-twists-endings-that-might-surprise-you/, 2024. Accessed June 2026.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. 9 2024. URL http://arxiv.org/abs/2409.04109.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. 6 2025. URL http://arxiv.org/abs/2506.20803.

Dean Keith Simonton. The psychology of creativity: A historical perspective. *Davis, CA: Green College Lecture Series on The Nature of Creativity: History Biology, and Socio-Cultural Dimensions. University of British Columbia*, 2001.

Dean Keith Simonton. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, 2004.

Dean Keith Simonton. Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews*, 7(2):156–179, 2010.

Dean Keith Simonton. Scientific creativity: Discovery and invention as combinatorial. *Frontiers in Psychology*, 12:721104, 2021.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Lisa B Soros, Alyssa M Adams, Stefano Kalonaris, Olaf Witkowski, and Christian Guckelsberger. On creativity and open-endedness. In *Artificial Life Conference Proceedings 36*, volume 2024, pp. 60. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2024.

Meir Sternberg. *Expositional modes and temporal ordering in fiction*. Indiana University Press, 1993.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han Van Der Maas. Putting GPT-3's Creativity to the (Alternative Uses) Test. In *International Conference on Computational Creativity*, 2022. URL http://osf.io/vmk3c/.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization, 2025. URL https://arxiv.org/abs/2506.18880.

Paul Thagard. Creative combination of representations: Scientific discovery and technological invention. *Psychology of science, R. Proctor and EJ Capaldi, Eds Oxford University Press, forthcomin*, 2010.

Paul Thagard. *Conceptual Revolutions*. Princeton University Press, 2018.

Tzvetan Todorov. The typology of detective fiction (1966). In *Crime and media*, pp. 291–301. Routledge, 2019.

E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*, 1974.

L. R. Varshney. Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development*, 63(1), 1 2019. ISSN 21518556. doi: 10.1147/JRD.2019.2893907.

Lav R Varshney, Florian Pinel, Kush R Varshney, Debarun Bhattacharjya, Angela Schörgendorfer, and Y-M Chee. A big data approach to computational creativity: The curious case of Chef Watson. *IBM Journal of Research and Development*, 63(1):1–7, 2019.

Manya Wadhwa, Tiasa Singha Roy, Harvey Lederman, Junyi Jessy Li, and Greg Durrett. Create: Testing llms for associative creativity. *arXiv preprint arXiv:2603.09970*, 2026.

Eitan Wagner, Renana Keydar, and Omri Abend. Modeling fair play in detective stories with language models. *arXiv preprint arXiv:2507.13841*, 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Geraint A Wiggins. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-based systems*, 19(7):449–458, 2006.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025.

A FULL LEADERBOARD

In Table 3, we list every evaluated LLM ranked by the overall transformational creativity score, alongside its raw facet values (surprise, coherence, realism, and diversity), scored per Section 3.2.

Table 3: **Full transformational creativity leaderboard across all 72 populations.** Every evaluated system, ranked by *Overall* (the average z-score of gated surprise, gated coherence, and diversity). We also report *Surprise*, *Coherence*, *Realism*, and *Diversity*, as defined in Section 3.2.

#	System	Surprise	Coherence	Realism	Diversity	Overall
1	**Expert humans**	**4.22**	**4.94**	**4.78**	**0.609**	**+2.00**
2	claude-sonnet-4.5	4.22	4.73	4.87	0.584	+1.90
3	claude-sonnet-4.6	4.00	4.93	4.87	0.563	+1.79
4	claude-opus-4.6	4.20	5.00	4.83	0.556	+1.75
5	claude-haiku-4.5	3.90	4.88	4.47	0.584	+1.54
6	claude-opus-4.5	4.27	4.88	4.87	0.483	+1.52
7	gpt-5.4	4.03	4.83	4.50	0.530	+1.35
8	kimi-k2.5	3.60	4.20	4.40	0.611	+1.25
9	jamba-large-1.7	3.83	3.83	4.37	0.570	+1.10
10	deepseek-v3.2	4.36	4.72	3.34	0.632	+0.98
11	deepseek-chat-v3.1	4.50	4.85	3.22	0.629	+0.98
12	claude-opus-4.7	4.17	4.93	3.87	0.585	+0.92
13	glm-5.1	3.75	4.75	4.00	0.614	+0.76
14	deepseek-v3.1-term.	4.38	4.83	3.00	0.614	+0.67
15	kimi-k2	3.52	4.05	3.47	0.623	+0.62
16	glm-4.7	3.67	4.22	3.44	0.592	+0.60
17	deepseek-v3.2-exp	4.67	4.90	2.80	0.617	+0.57
18	glm-4.5v	3.98	3.85	3.70	0.557	+0.56
19	gemma-3-27b	4.03	4.07	3.13	0.610	+0.50
20	gpt-4.1	4.10	4.20	3.40	0.584	+0.46
21	nemotron-3-super-120b	3.91	4.31	3.61	0.573	+0.45
22	gpt-5.5	3.78	4.13	3.57	0.551	+0.36
23	deepseek-chat-v3	3.67	3.65	3.27	0.563	+0.23
24	gpt-4o-mini	2.53	2.70	4.47	0.534	+0.21
25	minimax-01	3.20	3.75	3.27	0.562	+0.14
26	llama-4-maverick	4.13	3.63	3.63	0.506	+0.14
27	glm-4.6	3.07	3.23	3.77	0.574	+0.11
28	mistral-medium-3.1	4.10	4.22	3.17	0.486	+0.09
29	mixtral-8x22b	4.10	3.68	2.83	0.549	-0.01
30	gemma-3-12b	4.07	4.23	2.53	0.598	-0.01
31	hermes-4-70b	2.60	2.43	3.80	0.555	-0.04
32	hermes-4-405b	2.88	2.78	3.40	0.579	-0.05
33	llama-4-scout	4.23	3.27	4.02	0.386	-0.07
34	nova-pro-v1	3.33	3.60	2.80	0.553	-0.09
35	gpt-4.1-mini	4.00	3.97	2.40	0.598	-0.10
36	gemma-2-27b	2.80	2.73	3.75	0.572	-0.10
37	glm-5	2.78	3.03	3.78	0.531	-0.12
38	cogito-v2.1-671b	3.47	3.40	2.57	0.595	-0.13
39	gpt-5.1	3.00	4.22	2.83	0.549	-0.22
40	hermes-3-llama-3.1-405b	3.38	3.18	3.07	0.537	-0.24
41	deepseek-chat	4.03	4.03	2.57	0.529	-0.25
42	nova-2-lite-v1	3.33	3.47	2.63	0.591	-0.27
43	llama-3.3-70b	4.13	3.33	3.43	0.417	-0.27
44	mistral-nemo	2.90	2.70	3.17	0.562	-0.31
45	llama-3.2-11b-vision	2.63	2.27	3.83	0.512	-0.34
46	llama-3.1-8b	2.28	2.22	4.27	0.499	-0.34
47	minimax-m2-her	2.15	2.70	3.97	0.511	-0.35

continued on next page

Table 3 continued

#	System	Surprise	Coherence	Realism	Diversity	Overall
48	nova-lite-v1	3.20	3.37	2.45	0.550	-0.37
49	qwen3-next-80b	4.53	4.40	2.33	0.527	-0.40
50	qwen3-14b	3.87	3.80	3.40	0.412	-0.43
51	ministral-8b	3.88	3.70	3.30	0.403	-0.46
52	claude-opus-4.8	4.73	4.98	2.10	0.498	-0.50
53	llama-3-8b	2.17	2.20	4.17	0.463	-0.61
54	llama-3.1-70b	4.13	3.53	2.95	0.400	-0.61
55	gpt-5.4-nano	4.03	4.23	2.50	0.438	-0.67
56	gpt-5.4-mini	4.33	4.33	2.33	0.454	-0.69
57	mistral-small-3.1-24b	2.40	1.97	4.67	0.357	-0.71
58	gemini-2.5-pro	4.23	4.57	1.90	0.553	-0.75
59	hunyuan	1.87	1.73	4.23	0.439	-0.77
60	nova-micro-v1	3.67	3.60	2.03	0.540	-0.79
61	mistral-large	4.10	4.12	2.29	0.457	-0.79
62	granite-4.0-h-micro	1.90	3.13	4.50	0.347	-0.81
63	llama-3.2-3b	3.63	2.67	3.52	0.350	-0.83
64	qwen3.5-flash	4.37	4.53	1.73	0.519	-0.86
65	gemma-3-4b	4.10	4.27	1.73	0.501	-0.88
66	ernie-4.5-vl-424b	4.03	4.30	1.43	0.538	-0.90
67	gemma-4-26b	4.30	4.88	2.03	0.427	-0.96
68	gemma-4-31b	4.67	4.90	1.27	0.490	-1.18
69	morph-v3-fast	1.62	1.55	3.93	0.364	-1.19
70	hermes-3-llama-3.1-70b	1.41	1.55	3.21	0.420	-1.26
71	nemotron-nano-9b-v2	2.38	2.50	2.62	0.449	-1.32
72	nemotron-3-nano-30b	1.10	1.10	3.10	0.385	-1.54

LLM-as-a-Judge Rubric

```
You are a careful, consistent literary judge. You will read a short story and rate
its PLOT TWIST using a fixed rubric. A "plot twist" is a reveal that makes the reader
 re-interpret earlier parts of the story. Judge only what is on the page; do not
reward mere shock, gore, or randomness.

Rate the story on each dimension using the 1-5 anchors below. Use the FULL range.

SURPRISE -- How much does the ending force you to re-read the earlier story in a new
light?
  1: No re-reading. The ending follows straightforwardly from the setup; expected.
  2: Minor. One small detail is recast; the overall reading is unchanged.
  3: Moderate. A subplot or one character is recast, but the core reading stays.
  4: Strong. A central assumption is overturned; much of the story reads differently.
  5: Total. A foundational assumption is overturned; nearly the whole story must be
re-interpreted.

COHERENCE -- After the reveal, does the earlier story still hold together -- even fit
 better -- or is it contradicted?
  1: Contradicted. The reveal conflicts with earlier events; arbitrary or tacked-on.
  2: Weak. The reveal mostly does not fit; little in the story supports it.
  3: Mixed. Some earlier details support the reveal; others must be ignored.
  4: Strong. The earlier story is consistent with the reveal, and several details now
 make better sense.
  5: Seamless. The reveal was quietly prepared throughout; in hindsight it feels
coherent and earlier details snap into place. Nothing has to be retracted.

REALISM -- Does anything genuinely impossible (supernatural, science-fictional, or
fantastical) actually happen in the story's world? Judge what is real WITHIN the
story, not how surreal the telling is; dreams, hallucinations, hoaxes, and figurative
 prose stay realistic.
  1: The central premise or twist depends on something genuinely impossible.
  2: Genuinely impossible elements are clearly present but secondary.
  3: Genuinely ambiguous -- the story leaves it truly unresolved whether something
impossible occurred.
  4: Realistic, with at most a single minor, deniable brush with the impossible.
  5: Fully realistic -- everything that happens could happen in the real world.

Story:
"""{story}"""

First, in AT MOST three sentences, note what (if anything) the twist re-interprets
and how well the earlier story supports it. Then output a single JSON object on its
own final line, EXACTLY in this form and nothing after it:
{"twist_present": <true|false>, "surprise": <1-5>, "coherence": <1-5>, "realism":
<1-5>, "prose_quality": <1-5>, "overall": <1-5>}
```

Figure 8: **The LLM-as-a-Judge rubric we use to evaluate short stories in Section 3.2.**

B LLM-AS-A-JUDGE RUBRIC

The LLM-as-a-Judge rubric is given in Figure 8. In practice, `{story}` is replaced by the story text, and each story is scored by an ensemble of judges aggregated per dimension by the median.

Annotation prompt (GPT-4o-mini)

```
You are a literary analyst. Read the short story below and the scores a separate plot
-twist rubric judge already gave it.

A "plot twist" is a reveal that makes the reader re-interpret earlier parts of the
story. The rubric rated (1-5): SURPRISE (how much the ending forces a re-read) and
COHERENCE (whether the earlier story still holds up, even fits better, after the
reveal). twist_present indicates whether the judge thought there was a genuine twist
at all.

Scores for this story: surprise={surprise}, coherence={coherence}, overall={overall},
 twist_present={twist_present}.

Produce a JSON object with EXACTLY these keys, each a single sentence:
- "setup": the premise and what the reader is led to believe before the twist.
- "reveal": the twist itself -- what is revealed and what it makes the reader re-
interpret. If there is no genuine twist, begin with "No genuine twist:" and say what
happens instead.
- "why_scored": why this earned that surprise/coherence (e.g. the reveal recasts the
whole story so surprise is high; it was quietly prepared so coherence is high; or it
was predictable/arbitrary so a dimension is low).

Story:
"""{story}"""

Output only the JSON object, nothing else.
```

Figure 9: **Annotation prompt.** `gpt-4o-mini` extracts a one-sentence setup, reveal, and rationale; the *reveal* is embedded for the diversity metric.

C Annotation Extraction

Each scored story is additionally parsed by `gpt-4o-mini` into a one-sentence *setup*, *reveal*, and score rationale (given the story and its rubric scores); the *reveal* summary is the text embedded to compute the diversity metric in Section 3.2. The prompt is given in Figure 9.

Reasoning trace analysis prompt

```
You are analyzing the REASONING TRACE a model wrote while planning a short story with
 a plot twist. Below are 10 reasoning STEPS with definitions. For EACH step, decide
whether it appears anywhere in the trace; if so, copy the SHORTEST verbatim snippet (
at most 15 words, copied EXACTLY, character-for-character, from the trace) that marks
 where that step occurs.

STEPS:
1. frame the task -- restating the prompt's constraints (length, no title, choosing
the subject/characters)
2. promise coherence -- asserting the twist must be consistent with / prepared by
earlier events (fair play, not arbitrary)
3. list potential twists -- brainstorming or enumerating candidate twists or known
tropes
4. plan setting -- deciding the story's subject, characters, or situation
5. restate recontextualization goal -- restating that the ending must reframe /
recontextualize earlier events
6. propose, reject, & finalize twist -- weighing candidate twists, rejecting some (
too cliche/obvious), settling on the final one
7. plan clues to plant -- deciding what foreshadowing / clues / hints to plant
earlier
8. choose a reveal event -- choosing the concrete device or event that delivers the
reveal (a letter, diary, photo, recording, ...)
9. outline full plot -- laying out the structure, acts, beats, or scene order
10. verify it coheres -- checking that, after the twist, the story still holds
together / re-reads consistently

Return ONLY a JSON object whose keys are the step numbers ("1".."10") and whose
values are {"present": true|false, "quote": "<exact substring of the trace>" or null
}. The quote MUST be copyable verbatim from the trace. If a step is absent, use {"
present": false, "quote": null}.

TRACE:
"""{trace}"""
```

Figure 10: **Reasoning trace analysis prompt**. This prompt is used in Section 5 to identify the presence and location of the 10 steps manually observed in reasoning traces.

D Reasoning Trace Analysis

As explained in Section 5, each reasoning trace is annotated by `gpt-4o-mini` to assess whether the 10 steps were presented across all 207 traces. The prompt, with the ten steps and their definitions, is shown in Figure 10.

Be-creative suffix

```
Be as creative and original as you possibly can: avoid cliched, predictable, or
commonly-used twists, and aim for a genuinely surprising, non-obvious reveal that a
reader would not see coming.
```

Figure 11: ***Be creative*** **suffix**, appended to the task prompt (Figure 3).

Twist-summarization prompt (GPT-4o-mini)

```
In ONE sentence (at most 25 words), describe the core PLOT TWIST of the story below
-- the late reveal and the assumption it overturns. Output only the sentence.

Story:
"""{story}"""
```

Figure 12: **Twist-summarization prompt** used by in-context regeneration.

In-context regeneration suffix

```
IMPORTANT: you have ALREADY written stories with the plot twists listed below. Your
new story's twist must be GENUINELY DIFFERENT from every one of them -- a different
underlying mechanism and a different subject, not a variation:
- {summary of an earlier twist}
- {summary of an earlier twist}
```

Figure 13: **In-context regeneration suffix.**

E PROMPTING-STRATEGY PROMPTS

Both ablation strategies (Section 4.2) append to the base task prompt (Figure 3). *Be creative* appends the suffix in Figure 11 to the prompt in Figure 3, while *In-context regeneration* relies on an independent model (`gpt-4o-mini`) to summarize its previous twists into a single sentence, obtained via the prompt in Figure 12, that are repeatedly appended to subsequent generations. The suffix for *In-context regeneration* is shown in Figure 12 and is also appended to the prompt in Figure 3.

F HUMAN EXPERT STORIES

We obtain a set of 18 short stories with plot twists written by expert human authors from Project Gutenberg. Stories are drawn from a human-curated list of short stories with twist endings, obtained from Short Story Guide (2024). Story lengths span roughly 1.2k–22k words. In Table 4, we annotate the exact location in human stories where the plot twists occur, as well as the earlier plot points it forces the reader to re-contextualize. We note that the reveals overwhelmingly occur in the final sentence or two of most texts.

Table 4: **Plot twists among the 18 human stories**. The *reveal point* is the sentence that introduces the plot twist, *Pos.* is its position in the text ("end" = final sentence or two), and *Words* is the word count of the entire story.

#	Story	Plot twist (verbatim)	Pos.	Words	Result of twist
1	*The Necklace* (Maupassant)	"Why, my necklace was paste! It was worth at most only five hundred francs!"	98%	2,878	Ten years of ruinous sacrifice to repay it were needless.
2	*The Jewelry* (Maupassant)	"…that necklace is worth from twelve to fifteen thousand francs…"	35%	2,284	The wife's "love of paste," her outings, the easy household money: a lover's gifts and infidelity.
3	*The Last Leaf* (O. Henry)	"it's Behrman's masterpiece—he painted it there the night that the last leaf fell."	99%	2,423	Behrman's scorn and his boasted "masterpiece"; the painted leaf was a fatal act of self-sacrifice.
4	*After Twenty Years* (O. Henry)	"You've been under arrest for ten minutes, 'Silky' Bob."	87%	1,304	The whole warm reunion: the beat cop *was* the friend, deciding mid-conversation to turn him in.
5	*Hearts and Hands* (O. Henry)	"…did you ever know an officer to handcuff a prisoner to his *right* hand?"	end	906	The genial "marshal" is the convict; the glum man is the real marshal, lying to spare the lady.
6	*Witches' Loaves* (O. Henry)	"…a draftsman…rubs out the pencil lines with handfuls of stale bread crumbs."	87%	1,288	Every charitable reading of the customer; the "kind" butter destroyed his prize drawing.
7	*The Open Window* (Saki)	"…three figures were walking across the lawn towards the window; they all carried guns…"	81%	1,229	The niece's ghost-story is an improvised hoax; the guest flees from the living.
8	*Dusk* (Saki)	"Yes, sir, a cake of soap."	end	1,746	Gortsby's clever "deduction" was wrong; the hard-luck tale was true and he was conned in reverse.
9	*An Occurrence at Owl Creek Bridge* (Bierce)	"Peyton Farquhar was dead; his body, with a broken neck, swung gently…beneath the timbers of the Owl Creek bridge."	end	3,777	The vivid escape and journey home were a hallucination in the instant before the drop.
10	*A Horseman in the Sky* (Bierce)	"My father."	97%	2,507	The sentry's cold, dutiful kill was patricide.
11	*Chickamauga* (Bierce)	"The child was a deaf mute."	99%	2,542	His playful obliviousness: the "playthings" are mutilated soldiers, the dead woman his mother, the burning house his own.
12	*Désirée's Baby* (Chopin)	"…belongs to the race that is cursed with the brand of slavery."	end	2,191	Armand's cruelty and the baby's race; the taint is *his*, not Désirée's.
13	*The Man That Corrupted Hadleyburg* (Twain)	"There IS no test-remark—nobody made one."	66%	21,718	The whole "honest town" was a trap, sprung to expose its leading citizens as venal.

continued on next page

Table 4, continued

#	Story	Plot twist (verbatim)	Pos.	Words	Result of twist
14	*Luck* (Twain)	"He is the supremest ass in the universe; and until half an hour ago nobody knew it but himself and me."	92%	1,798	Every celebrated exploit of the great hero was a blunder rescued by sheer luck.
15	*Haircut* (Lardner)	"I suppose he was plottin' to get Paul out in the boat and... pushin' him in the water."	87%	5,045	The barber's fond anecdotes: the "jokes" were cruelty and the death a staged murder, told by an oblivious narrator.
16	*A Jury of Her Peers* (Glaspell)	"It's the bird." (the strangled canary)	79%	8,506	Every domestic "trifle" becomes evidence of motive—which the women then quietly conceal.
17	*A Terribly Strange Bed* (Collins)	"...the ordinary light canopy of a four-post bed was in reality a thick, broad mattress..."	66%	6,967	The host's hospitality was a robbery-murder machine, the winnings the bait.
18	*The Three Strangers* (Hardy)	"Why, souls—'twas the man in the chimney-corner!"	88%	8,610	The easy fireside stranger was the condemned man, hiding in plain sight beside his own hangman.