Creativity Coverage: Human-Grounded Boundaries for Evaluating LLM Creativity

Anonymous ACL submission

Abstract

We introduce creativity coverage, a novel framework for evaluating large language model (LLM) creativity as a boundary rather than a scalar. Unlike existing methods that measure proximity to human creative standards, our approach identifies hard limits: which regions of human creative space can LLMs reach, and which remain beyond their grasp? This formulation aligns with theories of transformational creativity, which emphasize moving beyond known conceptual boundaries rather than performing well within them. We define human creativity boundaries using the distribution of human responses in a shared semantic embedding space, then measure LLM coverage over this space. Across divergent thinking, convergent reasoning, and creative writing tasks, we find that creative boundaries are strongly task-dependent: models achieve high coverage on structured tasks but occupy only a narrow subset of human space in open-ended writing. Our metric correlates with established diversity measures yet provides distinct information. We further identify specific linguistic features—narrative length, lexical specificity, novel entities—that characterize human creativity beyond model reach, offering actionable insights for improving LLM creative capabilities.

1 Introduction

Can large language models (LLMs) be "creative?" A rapidly expanding body of research has attempted to answer this question (Franceschelli and Musolesi; Pandya), but the question itself is complex. First, what do we mean by "creative?" Most researchers pose this more specifically: can LLMs produce textual output that can be regarded as "creative" by human standards? This reformulation clarifies the question but introduces new complications. Creativity, as centuries of cultural studies scholarship have shown (Orwig et al., 2024; Amabile, 1988; Ismayilzada et al., 2025; Hou et al., 2025), is inherently subjective and often in the eye of the beholder. One could simply ask a panel of human evaluators to judge LLM output as "creative" or not, as some have done, but this approach has inevitable limitations of scale and bias.

A more practical approach is to leverage consensus definitions from creativity scholarship and devise quantitative metrics accordingly (Csikszentmihalyi, 1996). The most common measurement approach centers on homogeneity versus diversity. The intuition is straightforward: output with little variance is not "creative" because creativity requires novelty or surprise. Researchers have devised ways to measure homogeneity in LLM output on a continuous scale, where high homogeneity scores indicate low creativity and vice versa (Moon et al., 2025; Murthy et al., 2024). However, scalar measures can only tell us how close an LLM comes to some human standard of creativity.They cannot identify an LLM's creative capacity.

We take a different approach. Rather than measuring how LLM outputs compare to human creative standards on a continuous scale, we identify hard boundaries: creative tasks that humans can accomplish but LLMs cannot. This approach builds on what the influential creativity researcher Margaret Boden identified as the most important yet challenging version of creativity—defining a space completely outside the current known boundary of production (Boden, 1998). Our boundary-based framework tests these limits by establishing what range of creative output humans produce and how much of that range LLMs can cover. Where LLMs fail to reach parts of human creative space marks the boundary of their current creative capacity—not merely a score, but a categorical limit.

Our contributions. This paper makes three contributions to LLM creativity research. First, we introduce a simple but conceptually powerful metric for LLM creativity: *LLM-human coverage*, which

quantifies creativity as boundary space rather than proximity. Drawing on the notion of recall from information retrieval, we measure what proportion of human creative output LLMs can reproduce, identifying the boundary between reachable and unreachable creative space.

Second, we evaluate LLMs' performance on this metric across three distinct categories of creativity tasks—convergent reasoning, divergent reasoning, and aesthetic creativity. We find that different tasks exhibit different creative boundaries, revealing task-specific limitations of LLMs.These boundary measurements suggest targeted pathways for improving LLM creative capacities based on task type.

Third, we contribute to growing evidence that *creativity cannot be treated as a unified capacity* (Corazza, 2016; Runco and Jaeger, 2012). Different creative tasks demand different cognitive processes, and what counts as "creative" varies by domain. Our boundary-based approach reveals these distinctions empirically. Tasks that appear similarly "creative" in scalar terms may have radically different boundaries for what LLMs can accomplish. This has practical implications for both model development and deployment—understanding where hard limits exist helps identify which creative tasks benefit from LLM assistance and which require fundamentally different approaches.

2 Related Work

2.1 Measuring Creativity in LLMs

The most common prior approach to measuring LLM creativity is semantic distance. Researchers prompt an LLM to generate multiple responses to a standard writing or reasoning task, then compute how semantically varied the responses are from each other or from a human benchmark (Anderson et al., 2024; Atmakuru et al., 2024; Bellemare-Pepin et al., 2025; Chen and Ding, 2023; Dinu et al., 2025; Doshi and Hauser, 2024; Zhang et al., 2025). The underlying assumption is that greater semantic diversity in responses signals greater underlying creativity in the LLM , while semantic homogeneity indicates a lack of creativity (Padmakumar and He, 2024; Lee and Chung, 2024). Such metrics typically use embeddings and cosine similarity to quantify semantic distance.

Other researchers measure LLM creativity through concepts like originality, novelty, and surprise, drawing on established creativity theory (Lu et al., 2024; Runco and Jaeger, 2012; Amabile, 1988; Simonton, 2018; Corazza, 2016; Diedrich et al., 2015). These approaches employ either human evaluators or LLM-as-judge methods to assess textual output on Likert scales (He et al., 2025; Hou et al., 2025; Ismayilzada et al., 2025; Zhao et al., 2025). The most rigorous work combines both methods: researchers design codebooks to guide LLM evaluation, then validate these judgments with human raters using the same codebook.

Across all these approaches, creativity is evaluated as a continuous scalar value—whether through similarity scores or numerical ratings of originality or usefulness. While these methods effectively capture relative differences in creative output, *they merely measure proximity to creative standards rather than capacity limits*. A model might score 3.5/5 on originality, but this tells us little about what creative tasks it fundamentally cannot perform. Our work complements these scalar approaches by identifying categorical boundaries: the creative spaces models can and cannot reach.

2.2 Task Specific Creativity Research

Research on LLM creativity typically focuses on three types of tasks: (1) divergent reasoning tasks requiring original or novel thinking (e.g., alternative uses tests (Guilford, 1967)), (2) convergent reasoning tasks with more predictable or fixed solutions (e.g., coding, mathematical problems), and (3) creative writing tasks where originality and surprise are explicitly valued. Much early work optimized LLM performance for single task types, such as fictional story generation (Gómez-Rodríguez and Williams, 2023) or code synthesis (Lu et al., 2025), or problem solving (Boussioux et al., 2024).

More recent research has tested general approaches to measuring creativity across multiple task types, evaluating whether a single metric can capture creative performance across domains (He et al., 2025). However, emerging work challenges the assumption that creativity can be generalized across contexts (Hou et al., 2025; Jain et al., 2025; Lai et al., 2025). Creativity operates differently depending on domain: in short story writing, creativity should be maximized, while in medical diagnosis, excessive creativity may be harmful (Si et al., 2024). This reality suggests that evaluating LLM creativity requires task-specific frameworks rather than universal metrics.

Our work builds on this insight by directly measuring task-specific creativity boundaries. Rather than asking whether a model is "creative" in gen-

eral, we identify which creative tasks models can and cannot accomplish within each domain. By testing convergent reasoning, divergent reasoning, and aesthetic creativity separately, we reveal that creative boundaries vary substantially by task type, suggesting that different cognitive capacities are required for different creative domains.

2.3 Boundary and Coverage Approaches

Leading creativity theorists emphasize that understanding creativity requires identifying boundaries. Boden (1990, 2004) argues that the most significant form of creativity —"transformational creativity"—involves redefining the conceptual space itself, moving beyond existing boundaries rather than merely exploring within them. Similarly, Csikszentmihalyi (1996) conceptualizes creativity as emerging at the intersection of domain boundaries, where individuals push against or transcend established constraints. For both theorists, creativity is fundamentally about boundaries: recognizing them, testing them, and potentially overcoming them.

This boundary-based logic extends beyond creativity theory into cognitive science (Stella et al., 2023). Researchers studying human cognitive capacity consistently measure performance through categorical limits rather than continuous scores. Miller (1956) foundational work on working memory established capacity through what individuals can versus cannot retain—a boundary rather than a scale. Cowan (2001) refined this approach, demonstrating that identifying precise boundaries (the "magical number 4") provides more accurate understanding of cognitive capacity than proximity measures. This principle applies directly to creativity research. To understand creative capacity, we must identify what falls within versus beyond an agent's (whether LLM or human) reach.

Despite these theoretical foundations, boundary-based measurements remain rare in LLM creativity research. One notable exception is Chakrabarty et al. (2024), who evaluate LLM creative capacity using pass/fail tests on standard benchmarks and reveal categorical failures on tasks requiring conceptual blending and constraint satisfaction (Chakrabarty et al., 2024; Tian et al., 2025). However, their approach focuses on specific benchmark tasks rather than providing a generalizable framework for measuring creative boundaries.

Our work builds on this foundation by introducing a systematic coverage-based approach. Rather than testing whether models pass or fail individual creativity tasks, we measure the range of human creative output and quantify how much of that range LLMs can cover. We develop a novel metric called "coverage" based on recall principles from information retrieval (Manning et al., 2008): what proportion of human creative production can a model reach? This approach provides a generalizable method for mapping creative capacity limits across task types, revealing not just whether models fail, but where and how extensively their boundaries differ from human creative range.

3 Methodology

Unlike prior work that measures creativity through explicit scoring functions, we adopt an implicit, human-grounded formulation. We let human responses define the reference baseline, rather than prescribing what counts as creative.

More precisely, for any given prompt, we interpret the distribution of human responses as defining a human reachable semantic region within in an embedding space. This region corresponds to the collection of responses that humans, taken together, regard as valid and meaningful for the task. We refer to this region as the **human creativity boundary.** When a sufficiently large and diverse set of human responses is available, we treat their empirical distribution in embedding space as a stable proxy for the human response region for the task. Using humans as the reference point, we evaluate LLM creativity by analyzing where model-generated responses fall relative to this boundary, which is inferred directly from the data rather than relying on any predefined criteria for quantifying LLM creative capability. Figure 1 illustrates the overall framework, described in detail below.

Data Embedding and Projection. To construct the creativity boundary, we first embed human responses into a semantic space. We use a sentence-level encoder, for example, `all-mpnet-base-v2` (Reimers and Gurevych, 2019; Song et al., 2020), to obtain embeddings for all responses. We have experimented with alternative embedding models, and the results remain consistent (see Appendix).

To improve computational tractability, we apply PCA to the human embeddings and project them to a lower-dimensional subspace. We choose the smallest number of components that explain at least 90% of the variance, yielding a reduced dimensionality d while retaining most of the variability in the embedding space. We additionally

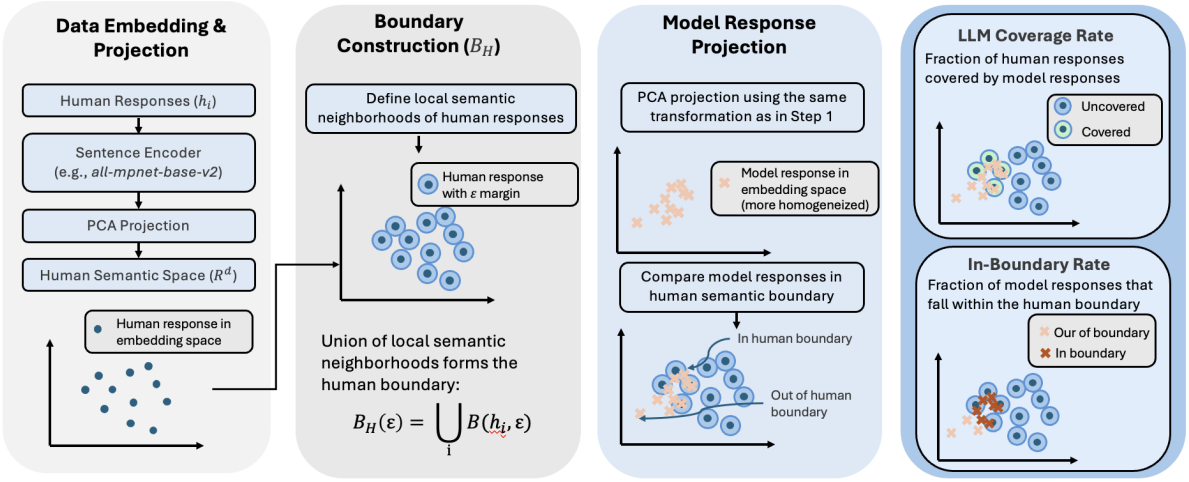


Figure 1: **Overview of the our proposed human-grounded creativity evaluation pipeline.**

verify that our results are robust to reasonable changes of this variance threshold (e.g., 85% and 95%).This defines a shared d-dimensional semantic space $H = \{h_i \in \mathbb{R}^d\}$ in which all subsequent analyses are performed.

We define $M = \{m_j \in \mathbb{R}^d\}$ as the embeddings of model-generated responses, projected into the same PCA space. The PCA transformation is learned exclusively from human responses, and model outputs are projected into this space to ensure a stable semantic representation.

Boundary Construction (B_H)**.** Empirically, human responses exhibit a clear local semantic clustering pattern in embedding space. We therefore approximate the human response region as the union of local neighborhoods around individual human responses. To set the neighborhood size, we estimate a typical within-cluster semantic scale from k-nearest-neighbor (kNN) distances among human embeddings. For each human response h_i, let r_i denote the Euclidean distance to its k-th nearest human neighbor. We set a single global radius ε as a robust quantile of these distances:

$$\varepsilon = Q_q(\{r_i\}),$$

where $Q_q(\cdot)$ denotes the q-th percentile.

Using this radius, we define the human creativity boundary as the union of closed ε-balls centered at each human response:

$$B_H(\varepsilon) = \bigcup_i B(h_i, \varepsilon),$$

where $B(h_i, \varepsilon)$ denotes the closed ball of radius ε around h_i in the PCA embedding space. A model-generated response m_j is considered *in-boundary* if there is at least one human response h_i such that

$$\|m_j - h_i\|_2 \leq \varepsilon.$$

In practice, neighborhood scale is jointly determined by k and q. We fix k and vary q to control scale. We set $k = 15$ to obtain a stable notion of local semantic similarity without being dominated by near-duplicate responses. Across $q \in \{0.50, 0.75, 0.90\}$, downstream results vary only mildly. We report $q = 0.75$ as a representative choice that yields compact but not overly restrictive neighborhoods; Figure 2 qualitatively illustrates the resulting neighborhood granularity.

Semantic Coverage Example

Prompt: *List alternative uses for a rope.*

Human responses:

Reference (anchor):
"A potential use of the rope is to make a hammock."

Neighbors under the 75^{th}-percentile ε:
"A potential use of the rope is to make something."
"A potential use of the rope is a hammock."
"To make a hammock using the rope."

LLM responses:

Covered (in-boundary):
"A potential use of the rope is hammock crafting."

Uncovered (out-of-boundary):
"A potential use of the rope is to make a snare."

Figure 2: **Examples of covered and uncovered LLM responses to the prompt "List alternative uses for a rope."**

Creativity Coverage Metrics. Building on the human creativity boundary defined above, we introduce two complementary metrics that characterize

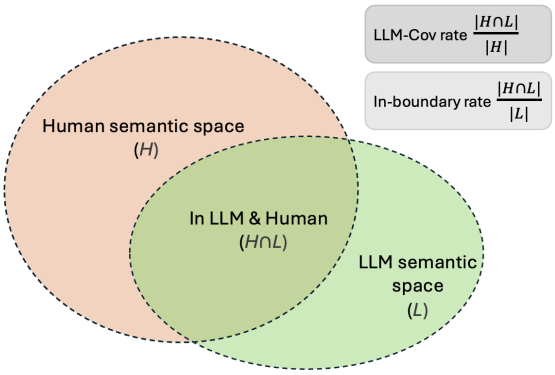


Figure 3: **Geometric interpretation of LLM Coverage Rate and In-Boundary Rate.**

how LLM-generated responses relate to this boundary. Given human response embeddings $H = \{h_i\}$ and LLM response embeddings $M = \{m_j\}$ to the same prompt, and a fixed neighborhood radius ε, we define the following measures.

First, we measure how well model-generated responses collectively cover the human creativity boundary. We call this the **LLM Coverage Rate** (LLM-Cov), defined as

$$\text{LLM-Cov} = \frac{|\{h_i : \exists m_j,\ \|h_i - m_j\|_2 \leq \varepsilon\}|}{|H|}.$$

A human response is *covered* if it lies within the ε-neighborhood of at least one model-generated response. Higher values indicate that model outputs span a larger portion of the human creative space.

Second, we measure how frequently model-generated responses fall within regions observed in human creativity. We call this the **In-Boundary Rate** (IBR), defined as

$$\text{IBR} = \frac{|\{m_j : \exists h_i,\ \|m_j - h_i\|_2 \leq \varepsilon\}|}{|M|}.$$

A model response is *in-boundary* if it lies within the ε-neighborhood of at least one human response. Intuitively, these two metrics provide complementary views of model creativity. LLM Coverage Rate captures how much of the human semantic space is spanned by model-generated responses, while In-Boundary Rate reflects the extent to which model outputs remain within regions observed in human creativity. Together, they characterize whether a model both captures the breadth of human ideas and explores beyond them.

4 Evaluation Procedure

Here, we describe the creativity tasks, evaluated models, and the precise implementation details of our metrics used in the experiments.

Group	Task	Example prompt
Divergent Thinking	AUT	*List alternative uses for a brick.*
	DAT	*Write 10 nouns in English that are as unrelated to each other as possible.*
Convergent Reasoning	CodeNet	*Given an array of integers, find the length of the longest contiguous subarray whose sum is at most K.*
Creative Writing	HP Fanfic Writing	*You are an accomplished author of Harry Potter fan fiction. Write in an immersive narrative voice inspired by Harry Potter that takes place entirely within the existing world of the novels.*

Table 1: **Task families and representative example prompts used in our evaluation.**

4.1 Creativity Tasks Used

We evaluate creativity across task families that capture complementary aspects of creative behavior (Table 1). **Divergent thinking** is assessed using the Alternative Uses Task (AUT), a widely used measure of ideational fluency and originality in creativity research (Góes et al., 2023; Guilford, 1967), and the Divergent Association Task (DAT), which elicits semantically distant concepts grounded in associative theories of creativity (Mednick, 1962; Olson et al., 2021). **Convergent reasoning** is evaluated with CodeNet, where outputs must satisfy strict correctness and structural constraints, yielding a creativity setting that is tightly bounded by program validity (Puri et al., 2021; Ye et al., 2024). **Creative writing** is measured using open-ended story generation prompts, including Harry Potter (HP) fanfiction and general story writing, which have been used in recent work to assess narrative originality, coherence, and long-form generation quality in LLMs (Alfassi et al., 2025; Fan et al., 2018; Mikhaylovskiy, 2023; Tian et al., 2024). Taken together, these tasks align with commonly used creativity benchmarks for LLMs (e.g., Creativity Prism (Hou et al., 2025)), covering complementary dimensions of divergent ideation, constrained problem solving, and open-ended narrative generation.

4.2 Models Evaluated

We evaluate open-source instruction-tuned models with 7B–32B parameters from the Mistral, Llama,

and Qwen families To probe decoding sensitivity, we generate responses at two temperatures, $t \in \{0.3, 1.0\}$, holding all other decoding parameters fixed across models unless otherwise stated.[1]

4.3 Metric Implementations

We embed all human and model responses using `all-mpnet-base-v2` from the Python Sentence-Transformers library, an off-the-shelf sentence embedding model commonly used for semantic similarity (Song et al., 2020). To stabilize neighborhood geometry and to reduce computation cost, we apply PCA and retain the minimum number of components that explain 90% of the variance. We define local neighborhoods using kNN with $k = 15$, which provides a robust notion of local semantic similarity without being dominated by near-duplicate responses. We set the neighborhood radius ε as the 75^{th} percentile of kNN distances, which yields compact, semantically coherent neighborhoods. We also check percentiles $\{50, 75, 90\}$ and find results are stable across this range.

Statistical tests. At various points, we use common statistical tests to measure whether changes in measured metrics are statistically significant. Our null hypothesis is that the metric is equal across both settings, while the alternative is that it is not. We use two-sided paired tests on matched settings, treating each matched configuration (e.g., a model-task pair or a task-temperature pair) as one unit of analysis, then report the resulting p-value.

5 Experimental Results

Here, we present the main results from our evaluation of creativity using our proposed embedding-based human boundary procedure.

5.1 Overall LLM Performance

Overall, we find that the open-source instruction-tuned LLMs we evaluate (Llama, Mistral, and Qwen) have consistently high human coverage (In-Boundary Rate/Human-Cov), close to 1.0 across all models, but much more variable LLM-Cov. These results are summarized in Table 2.

Across tasks, LLM-Cov is highest on the structurally constrained convergent task (CodeNet), consistent with prior work showing more homogeneous model behavior under stronger constraints (Wenger and Kenett, 2025; Keon et al., 2025; Jain et al., 2025), and lowest on creative writing (HP fanfiction). Across model sizes, we do not find evidence that larger models consistently achieve higher LLM-Cov than smaller models when pooling within-family Qwen and Mistral comparisons (p=0.21). In contrast, increasing temperature from t=0.3 to t=1.0 significantly increases LLM-Cov on average (p=1.39×10^{-6}).

Detailed analyses by task family, model family, and decoding temperature, along with statistical tests and effect sizes, are reported in Appendix A.

5.2 Relating LLM-Cov to Common Text Quality Measures: A Case Study on AUT

LLM-Cov measures how much model generations cover the human semantic region in embedding space. To interpret LLM-Cov relative to familiar baselines, we use AUT as a case study and compare it with commonly used statistics from prior creativity benchmarks, including inter-response diversity (Beaty et al., 2014; Olson et al., 2021), word count, average word length, and sentiment polarity.

Table 3 reports LLM-Cov alongside four commonly used response statistics for each model-temperature setting. Inter-response diversity (I-Div) is the average pairwise cosine distance between sentence embeddings within a setting (Beaty et al., 2014; Olson et al., 2021). ALen is mean characters per word, WC is mean words per response, and Sent. is mean VADER polarity (Hutto and Gilbert, 2014). We compute Pearson correlations between LLM-Cov and each statistic across model–temperature settings ($n = 10$, excluding humans). LLM-Cov correlates strongly with inter-response diversity ($r = 0.78$), but shows weak, non-significant associations with average word length ($r = 0.19$), word count ($r = 0.30$), and sentiment polarity ($r = -0.11$). We observe a similar qualitative pattern on DAT task.

Key takeaway: Because inter-response diversity is a widely used proxy for creativity in open-ended generation, this strong alignment validates LLM-Cov as a creativity-related signal. Nevertheless, the boundary-based definition of LLM-Cov provides a more interpretable, human-grounded notion of semantic reach into the human response space.

5.3 Where Do LLMs Diverge From Humans?

Beyond providing an interpretable aggregate metric, LLM-Cov enables a diagnostic question: which human responses are reachable by the model, and

[1] We additionally performed quality checks on a subset of model outputs, applying the same task-specific validity criteria used for human responses.

Model	Divergent Thinking				Convergent Reasoning		Creative Writing	
	AUT (p75)		DAT (p75)		CodeNet (p75)		HP Fanfic (p75)	
	LLM-Cov	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**
$\mathbf{t = 0.3}$								
Qwen2.5-7B	0.289	1.000	0.595	1.000	0.452	1.000	0.251	0.999
Mistral-7B	0.230	0.991	0.353	0.992	0.452	0.603	0.062	0.996
Llama-3.1-8B	0.352	1.000	0.353	1.000	0.724	0.984	0.040	0.771
Qwen2.5-32B	0.286	1.000	0.325	1.000	0.256	1.000	0.090	0.914
Mistral-24B	0.409	1.000	0.153	1.000	0.274	0.978	0.107	1.000
Avg.	0.313	0.998	0.356	0.998	0.432	0.913	0.110	0.936
$\mathbf{t = 1.0}$								
Qwen2.5-7B	0.627	0.997	0.775	1.000	0.716	0.964	0.254	0.966
Mistral-7B	0.664	0.994	0.566	0.676	0.710	0.747	0.094	0.992
Llama-3.1-8B	0.725	0.991	0.755	1.000	0.850	0.872	0.058	0.770
Qwen2.5-32B	0.556	0.990	0.676	0.999	0.374	0.952	0.189	0.926
Mistral-24B	0.812	0.993	0.653	1.000	0.590	0.978	0.144	0.972
Avg.	0.677	0.993	0.685	0.935	0.648	0.903	0.148	0.925
Overall Avg.	0.495	0.996	0.520	0.967	0.540	0.908	0.129	0.931

Table 2: **Models have consistently high Human-COV rates across all model sizes, temperatures, and tasks, but have highest LLM-Cov rates for convergent thinking tasks and lowest for creative writing.** Coverage metrics computed at p75 across divergent thinking, convergent reasoning, and creative writing.

Model	Cov.	I-Div	ALen	WC	Sent.
Human	–	0.32	3.62	5.64	-0.01
Llama-3.1-8B (0.3)	0.35	0.30	3.79	5.52	0.11
Llama-3.1-8B (1.0)	0.73	0.34	3.96	5.37	0.11
Mistral-7B (0.3)	0.23	0.23	4.06	4.02	0.22
Mistral-7B (1.0)	0.66	0.28	3.98	4.23	0.13
Mistral-24B (0.3)	0.41	0.29	3.62	5.06	0.12
Mistral-24B (1.0)	0.81	0.36	3.94	5.04	0.07
Qwen2.5-7B (0.3)	0.29	0.22	3.63	3.01	-0.04
Qwen2.5-7B (1.0)	0.63	0.27	3.70	3.30	0.01
Qwen2.5-32B (0.3)	0.29	0.21	3.96	4.02	0.07
Qwen2.5-32B (1.0)	0.56	0.25	3.97	4.18	0.09

Table 3: **LLM-Cov closely tracks inter-response diversity while remaining largely independent of other response invariants.** AUT task, LLM-Cov (p75). I-Div=inter-response diversity, ALen=avg word length, WC=word count, and Sent=sentiment polarity.

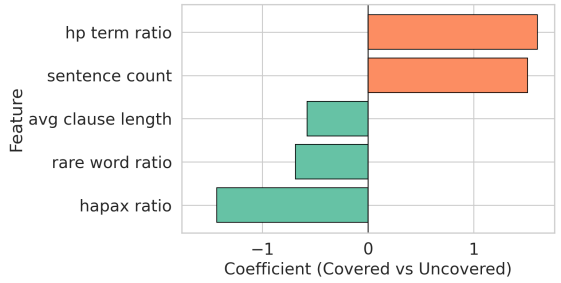
which are systematically missed. We study this on the HP fanfiction task by comparing covered vs. uncovered human responses. We choose this task because its longer narrative responses exhibit richer stylistic and syntactic variation, making it well-suited for linguistic analysis. For each human response, we extract features capturing HP-domain grounding (fraction of HP-specific terms), verbosity (word and character counts), lexical specificity (e.g., rare-word usage), and stylistic markers (e.g., hedging and capitalization). These provide an interpretable basis for contrasting covered and uncovered responses.

We then fit a logistic regression classifier with balanced class weights to predict whether a human response is covered by LLM responses. For simplicity, we focus our analysis on results from Qwen2.5-7B at $t = 0.3$. The dataset contains 251 covered and 749 uncovered responses ($n = 1000$). The classifier achieves 0.69 accuracy, making it a reasonable source of ground truth on which features are most associated with coverage.

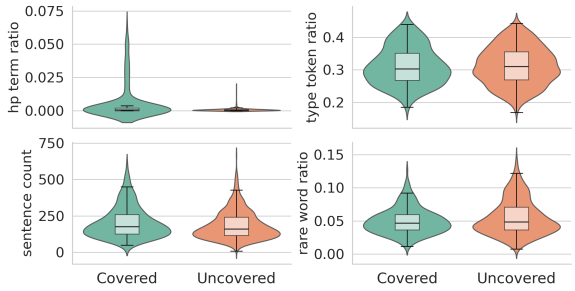
Figure 4 documents where LLM coverage diverges from human responses. Figure 4a shows that coverage is most strongly associated with domain grounding and basic sentence-level structure: responses with HP-term grounding are more likely to be covered, and sentence count is also positively associated with coverage ($p = 0.036$). In contrast, lexical specificity and stylistic complexity (e.g., rare-word ratio) negatively correlate with coverage, indicating that responses with rare words or more specific entities are harder for LLMs to reach.

Figure 4b visualizes these patterns at the distributional level. Covered responses exhibit substantially higher HP-term grounding, whereas uncovered responses tend to be longer and more lexically specific, with higher rare-word usage. By contrast, type-token ratio shows little separation between the two groups in this setting.

Key takeaway: Compared with previous methods, coverage-based metrics distinguish what an



(a) Significant logistic regression coefficients ($p < 0.05$). Hapax ratio is prop. of words appearing only once in a response.



(b) Violin plots comparing the distributions of representative linguistic features.

Figure 4: **Covered and uncovered responses differ in domain grounding and lexical specificity.**

LLM can successfully generate from what it systematically fails to produce. In the HP fanfiction domain, coverage gaps exhibit structure rather than random noise: for instance, covered outputs are more tightly anchored in HP-specific terminology, whereas uncovered outputs more frequently rely on rarer vocabulary. This characterization precisely identifies what the LLM overlooks and indicates where model refinements should be targeted.

6 Discussion

We introduce creativity coverage as a novel method for evaluating the creative capabilities of LLMs. Unlike existing methods that measure creativity on a continuous scale—assigning models a score that reflects proximity to human performance—our framework characterizes creativity as a boundary. This distinction is consequential: scalar metrics tell us how close a model comes to human creativity, but not what lies beyond its reach. Coverage, by contrast, directly identifies the regions of human creative space that models can and cannot access. This boundary-based formulation aligns with foundational theories of creativity. Boden's influential work emphasizes that the most significant form of creativity—transformational creativity—involves moving beyond existing conceptual boundaries rather than merely exploring within them. By operationalizing creativity as coverage over a human-defined semantic region, our method provides a direct empirical test of these limits. Where models fail to cover human responses, we identify not just lower performance, but categorical boundaries that current architectures cannot cross. This framing is particularly valuable for efforts aimed at enhancing LLM creativity: rather than optimizing for higher scalar scores, developers can target specific unreached regions of creative space.

Our findings suggest that creativity coverage complements rather than replaces existing evaluation approaches. In our case study on the Alternative Uses Task, we observe a strong correlation between coverage and inter-response diversity, suggesting that the two metrics capture related aspects of semantic exploration. However, coverage is not reducible to diversity. While diversity measures variation among model outputs, coverage anchors this variation to human-validated regions of semantic space. This distinction matters: a model could generate highly diverse outputs that nonetheless fail to reach the regions humans occupy. Coverage thus provides additional diagnostic information that diversity metrics alone cannot supply.

Beyond validation, our approach enables fine-grained analysis of where models diverge from humans. In the HP fanfiction case study, we identify specific linguistic features—narrative length, lexical specificity, and novel entities—that systematically distinguish covered from uncovered human responses. This level of precision is difficult to achieve with scalar methods, which summarize performance as a single value without indicating which aspects of human creativity remain out of reach. By characterizing the boundary in interpretable terms, our framework offers actionable insights for both understanding model limitations and guiding future improvements.

Finally, our results reinforce the task-dependent nature of creativity. We find that creative boundaries vary substantially across task families: models achieve high coverage on structured tasks like code generation but cover only a narrow subset of human responses in open-ended creative writing. This pattern suggests that different creative domains impose different constraints on model capabilities, and that no single scalar metric can adequately capture this variation. Our coverage-based framework provides a principled method for quantifying these task-specific differences.

7 Limitations

Our study has several limitations that suggest directions for future work.

First, our framework relies on sentence-level embeddings to construct the semantic space in which creativity boundaries are defined. While this encoder captures broad semantic similarity, it may not fully represent all dimensions relevant to creativity, such as narrative structure, stylistic voice, or pragmatic subtlety. Alternative embedding approaches or task-specific encoders could yield different boundary estimates.

Second, the human creativity boundary is only as representative as the human data used to construct it. Our human response sets, while curated for quality, are finite and may not capture the full range of human creative variation. Larger or more diverse human samples could reveal additional regions of creative space missed by our current boundaries.

Third, the epsilon threshold used to define neighborhood boundaries requires a parameter choice (e.g., 50th, 75th, or 90th percentile). While we report results at a fixed threshold for consistency, boundary estimates are sensitive to this choice. Future work could explore adaptive or task-specific threshold selection methods.

Fourth, our detailed analyses—the AUT case study examining coverage-diversity correlations and the HP fanfiction case study identifying linguistic features at the boundary—represent focused investigations rather than comprehensive evaluations across all tasks. While these case studies demonstrate the interpretive value of our framework, we cannot claim that the specific patterns observed (e.g., the role of lexical specificity and narrative length) generalize to all creative domains. Extending these micro-level analyses to additional tasks remains an important direction for future research.

Finally, we evaluate a range of open-source instruction-tuned models spanning 7B to 32B parameters, but do not test larger frontier models or models fine-tuned specifically for creative tasks. The creative boundaries we identify may shift with model scale or specialized training. Evaluating how coverage changes across a broader range of architectures and training regimes would further clarify the nature and stability of these limits.

References

Mistral AI. 2024. mistralai/mistral-7b-instruct-v0.3. Hugging Face model card. Accessed 2026-01-03.

Mistral AI. 2025. mistralai/mistral-small-24b-instruct-2501. Hugging Face model card. Accessed 2026-01-03.

Roi Alfassi, Angelora Cooper, Zoe Mitchell, Mary Calabro, Orit Shaer, and Osnat Mokryn. 2025. Fanfiction in the age of ai: Community perspectives on creativity, authenticity and adoption. *arXiv preprint*. ArXiv:2506.18706 [cs.HC].

Teresa M. Amabile. 1988. A model of creativity and innovation in organizations. *Research in Organizational Behavior*, 10:123–167.

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Creativity and Cognition*, pages 413–425. ArXiv:2402.01536 [cs].

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story-Writing Constraints. *arXiv preprint*. ArXiv:2410.04197 [cs].

Roger E. Beaty, Mathias Benedek, Richard W. Wilkins, Emanuel Jauk, Andreas Fink, Paul J. Silvia, Beate Dunst, and Aljoscha C. Neubauer. 2014. Creativity and the default network: A functional connectivity analysis of the creative brain at rest. *Neuropsychologia*, 64:92–98.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2025. Divergent Creativity in Humans and Large Language Models. *arXiv preprint*. ArXiv:2405.13012 [cs].

Margaret A Boden. 1990. *The creative mind: Myths and mechanisms*. Basic Books.

Margaret A. Boden. 1998. Creativity and artificial intelligence. *Artificial Intelligence*, 103(1):347–356.

Margaret A. Boden. 2004. *The Creative Mind*.

Léonard Boussioux, Jacqueline N. Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R. Lakhani. 2024. The Crowdless Future? Generative AI and Creative Problem-Solving. *Organization Science*, 35(5):1589–1607. Publisher: INFORMS.

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. *arXiv preprint*. ArXiv:2309.14556 [cs].

Honghua Chen and Nai Ding. 2023. Probing the Creativity of Large Language Models: Can models produce divergent semantic association? *arXiv preprint*. ArXiv:2310.11158 [cs].

Giovanni Emanuele Corazza. 2016. Potential originality and effectiveness: The dynamic definition of creativity. *Creativity Research Journal*, 28(3):258–267.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114.

Mihaly Csikszentmihalyi. 1996. *Creativity: Flow and the psychology of discovery and invention*. HarperCollins.

Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C. Neubauer. 2015. The novelty–usefulness tension: Toward a dialectical model of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4):320–337.

Anca Dinu, Andra-Maria Florescu, and Alina Resceanu. 2025. A Comparative Approach to Assessing Linguistic Creativity of Large Language Models and Humans. *arXiv preprint*. ArXiv:2507.12039 [cs].

Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290. Publisher: American Association for the Advancement of Science.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Giorgio Franceschelli and Mirco Musolesi. On the Creativity of Large Language Models.

J. P. Guilford. 1967. *The Nature of Human Intelligence*. McGraw-Hill.

Luis Fabricio Góes, Marco Volpe, Piotr Sawicki, Marek Grses, and Jacob Watson. 2023. Pushing GPT's Creativity to Its Limits: Alternative Uses and Torrance Tests. Publisher: University of Leicester.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Zicong He, Boxuan Zhang, Weihao Liu, Ruixiang Tang, and Lu Cheng. 2025. What Shapes a Creative Machine Mind? Comprehensively Benchmarking Creativity in Foundation Models. *arXiv preprint*. ArXiv:2510.04009 [cs].

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. 2025. CreativityPrism: A Holistic Benchmark for Large Language Model Creativity. *arXiv preprint*. ArXiv:2510.20091 [cs].

C. J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 216–225.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2025. Evaluating Creative Short Story Generation in Humans and Large Language Models. *arXiv preprint*. ArXiv:2411.02316 [cs].

Shomik Jain, Jack Lanchantin, Maximilian Nickel, Karen Ullrich, Ashia Wilson, and Jamelle Watson-Daniels. 2025. LLM Output Homogenization is Task Dependent. *arXiv preprint*. ArXiv:2509.21267 [cs].

Matt Keon, Aabid Karim, Bhoomika Lohana, Abdul Karim, Thai Nguyen, Tara Hamilton, and Ali Abbas. 2025. Galton's Law of Mediocrity: Why Large Language Models Regress to the Mean and Fail at Creativity in Advertising. *arXiv preprint*. ArXiv:2509.25767 [cs].

Clin Lai, Simone Luchini, Nina Lauharatanahirun, and Roger Beaty. 2025. Creative or Uncreative Partner: Comparing Humans and AI in Collaborative Creative Tasks.

Byung Cheol Lee and Jaeyeon (Jae) Chung. 2024. An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, 8(10):1906–1914. Publisher: Nature Publishing Group.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2024. AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. 2025. Benchmarking Language Model Creativity: A Case Study on Code Generation. *arXiv preprint*. ArXiv:2407.09007 [cs].

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

S. A. Mednick. 1962. The associative basis of the creative process. 69:220–232.

Meta. 2024. meta-llama/llama-3.1-8b-instruct. Hugging Face model card. Accessed 2026-01-03.

Nikolay Mikhaylovskiy. 2023. Long text generation challenge. *arXiv preprint*. ArXiv:2306.02334 [cs.CL].

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

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

Sonia K. Murthy, Tomer Ullman, and Jennifer Hu. 2024. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. *arXiv preprint*. ArXiv:2411.04427 [cs].

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. 118(25):e2022340118.

William Orwig, Emma R. Edenbaum, Joshua D. Greene, and Daniel L. Schacter. 2024. The Language of Creativity: Evidence from Humans and Large Language Models. *The Journal of Creative Behavior*, 58(1):128–136. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.636.

Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint*. ArXiv:2309.05196 [cs].

Vivek Pandya. The Age of Generative AI: Over half of Americans have used generative AI and most believe it will help them be more creative | Adobe Blog.

Ruchir Puri, David S. Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir R. Choudhury, Lindsey Decker, Veronika Thost, Luca Buratti, Saurabh Pujar, and Ulrich Finkler. 2021. Project codenet: A large-scale ai for code dataset for learning a diversity of coding tasks. *arXiv preprint*. ArXiv:2105.12655 [cs.SE].

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

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

sentence-transformers. 2020. Sentence-transformers all-minilm-l6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`. Accessed: 2025-12-31.

Sentence-Transformers. 2021. sentence-transformers/all-mpnet-base-v2. Hugging Face model card. Accessed 2026-01-03.

Sentence-Transformers. 2025. Sentence-transformers: Pretrained models. Documentation. Accessed 2026-01-03.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint*. ArXiv:2409.04109 [cs].

Dean Keith Simonton. 2018. Defining creativity: Don't we also need to define the opposite? *Creativity Research Journal*, 30(3):291–294.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*. arXiv. ArXiv:2004.09297 [cs.CL].

Rob Speer. 2016. wordfreq: Zipf frequency scale and `zipf_frequency`. PyPI documentation. Accessed 2026-01-03.

Massimo Stella, Thomas T. Hills, and Yoed N. Kenett. 2023. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, 120(43):e2312911120. Publisher: Proceedings of the National Academy of Sciences.

Qwen Team. 2024a. Qwen/qwen2.5-32b-instruct. Hugging Face model card. Accessed 2026-01-03.

Qwen Team. 2024b. Qwen/qwen2.5-7b-instruct. Hugging Face model card. Accessed 2026-01-03.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are Large Language Models Capable of Generating Human-Level Narratives? *arXiv preprint*. ArXiv:2407.13248 [cs].

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2025. MacGyver: Are Large Language Models Creative Problem Solvers? *arXiv preprint*. ArXiv:2311.09682 [cs].

Emily Wenger and Yoed Kenett. 2025. We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint*. ArXiv:2501.19361 [cs].

Peter West and Christopher Potts. 2025. Base Models Beat Aligned Models at Randomness and Creativity. *arXiv preprint*. ArXiv:2505.00047 [cs].

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. 2024. Assessing the Creativity of LLMs in Proposing Novel Solutions to Mathematical Problems. *arXiv preprint*. ArXiv:2410.18336 [cs].

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

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. 2025. Assessing and Understanding Creativity in Large Language Models. *Machine Intelligence Research*, 22(3):417–436. ArXiv:2401.12491 [cs].

A Extended analysis (task, model family, and temperature).

We further analyze LLM-Coverage metrics patterns along three axes: task family, model family/scale, and decoding temperature. At the task level, coverage varies monotonically with structural constraint: across models and temperatures, LLM-Cov is highest on the convergent reasoning task (CodeNet) and lowest on creative writing (HP fanfiction), while Human-Cov remains high throughout. This implies that generations largely remain within the human semantic region, but explore a substantially smaller fraction of that region for open-ended narrative generation than for constrained tasks, consistent with prior observations that stronger constraints induce more homogeneous model behavior (West and Potts, 2025).

At the model level, we do not observe a uniform scaling benefit. As summarized in Table 2, Qwen2.5-7B and Llama-3.1-8B attain the highest average LLM-Cov across tasks and temperatures in our setting, whereas Qwen2.5-32B is lowest. Within-family paired comparisons underscore heterogeneity: scaling from Qwen 7B to 32B is associated with a decrease in LLM-Cov across matched task–temperature settings (p=0.008), while scaling from Mistral 7B to 24B shows no consistent change (p=0.73).

Temperature exhibits the clearest and most consistent effect. Comparing t=1.0 against t=0.3 within matched model–task settings (n=20 paired comparisons), LLM-Cov increases by 0.237 on average (p=1.39 $\times$ 10^{-6}), with the largest gains on divergent-thinking tasks (AUT: +0.364; DAT: +0.329, averaged over models). Mistral-Small-24B is the most temperature-sensitive overall (average gain +0.314), including a large DAT increase from 0.153 to 0.653. In contrast, Human-Cov remains near 1.0 and does not change significantly under temperature (p=0.238).

B Coverage evaluation settings

This subsection documents the concrete configuration used to compute our two coverage metrics for the paper run: the *in-boundary rate* (the fraction of model outputs that fall inside the human-defined semantic region) and the *LLM coverage rate* (the fraction of the human semantic region reached by model outputs). In brief, we embed responses with an MPNet-based Sentence-Transformers encoder, set the neighborhood radius using the 75th percentile of human kNN radii (p75), and apply PCA fit on human embeddings only, retaining 90% explained variance (var90). We describe this in more detail in the Methods section.

Family	Developer	Checkpoint
Qwen2.5	Qwen team (Alibaba Cloud)	Qwen2.5-7B-Instruct (Team, 2024b)
Qwen2.5	Qwen team (Alibaba Cloud)	Qwen2.5-32B-Instruct (Team, 2024a)
Llama 3.1	Meta	Llama-3.1-8B-Instruct (Meta, 2024)
Mistral	Mistral AI	Mistral-7B-Instruct-v0.3 (AI, 2024)
Mistral	Mistral AI	Mistral-Small-24B-Instruct-2501 (AI, 2025)

Table 4: **Instruction-tuned checkpoints used for coverage evaluation (mpnet, p75, var90).**

We report (A) the evaluated instruction-tuned checkpoints and their provenance, (B) the generation grid together with per-task sample sizes and decoding hyperparameters, and (C) the shared embedding-and-geometry configuration and the prompts used for HP, DAT, AUT, and CodeNet.

A1. Model set. We evaluate five open-source instruction-tuned checkpoints spanning three model families. The Qwen2.5 instruction-tuned checkpoints are released by the Qwen team (Alibaba Cloud) (Team, 2024b,a). Llama 3.1 is released by Meta (Meta, 2024). The Mistral instruction-tuned checkpoints are released by Mistral AI (AI, 2024, 2025). Table 4 lists the checkpoints.

A2. Generation grid, sample sizes, and decoding hyperparameters. For each task subset, we run a fixed temperature grid $t \in \{0.3, 1.0\}$ for every checkpoint in Table 4, yielding a 5-model $\times$ 2-temperature grid per task subset. Table 5 reports the number of human responses and the number of model generations per model$\times$temperature setting used for coverage computation. Minor deviations from nominal generation counts arise from task-specific validity filtering and cleaning. Table 6 records the non-temperature decoding hyperparameters; these are held fixed across checkpoints within each task subset.

A3. Shared embedding and geometric configuration, and prompts. We embed both human and model texts using all-mpnet-base-v2 from Sentence-Transformers (Reimers and Gurevych,

Task subset	Human	Model gen. per config
HP Fanfiction	1000	1000
DAT	1000	1000
AUT	4000	4000
CodeNet	500	500

Table 5: **Sample sizes used for coverage computation (mpnet, p75, var90).**

Task subset	sample	max tokens
HP Fanfiction	True	4096
DAT	True	128
CodeNet	True	512
AUT	True	64

Table 6: **Non-temperature decoding hyperparameters (held fixed across checkpoints within each task subset).**

2019; Song et al., 2020). We fit PCA using only the human embeddings (variance target 0.9, maximum dimension 200) and project model embeddings into the same PCA space. We compute Euclidean distances in this PCA space. We set the neighborhood scale using human kNN radii with k=15 and choose the radius as the 75th percentile (p75). Table 7 summarizes these settings. Table 8 provides the prompts used in this run.

C Appendix A4: Human Data Sources

Human responses for the Alternative Uses Task (AUT) are drawn from a publicly released dataset hosted on the Open Science Framework (OSF): https://osf.io/5cy9n/overview. Human responses for the Divergent Association Task (DAT) are obtained from publicly available sources hosted on OSF and GitHub: https://osf.io/kbeq6/overview and https://github.com/AntoineBellemare/DAT_GPT/tree/main/human_data_dat. Human creative-writing data for the HP fanfiction task are collected from publicly available works hosted on Archive of Our Own (AO3). Human solutions for the convergent reasoning task are taken from the CodeNet dataset (Puri et al., 2021). All human data used in this work are publicly available and no new human subject data are collected.

Component	Setting
Sentence encoder	all-mpnet-base-v2 (Reimers and Gurevych, 2019; Song et al., 2020; Sentence-Transformers, 2021)
PCA fit set	human embeddings only
PCA variance target	0.9
Distance	Euclidean in PCA space
Human k for radii	$k = 15$
Neighborhood radius	$\varepsilon = \text{quantile}(\{r_i\}, 0.75)$

Table 7: **Shared embedding and geometric configuration used in baseline evaluations.**

D HP Fanfiction Diagnostic Analysis (Logistic Regression)

This section documents the HP fanfiction diagnostic classifier used to quantify which linguistic properties are associated with a human response being covered (reachable) versus uncovered under the p75 coverage labeling.

B1. Task, labels, features, and model specification

Label source and cohort. We analyze HP fanfiction human responses under the p75 labeling induced by the coverage evaluation in Section B. The diagnostic results reported correspond to the HP cohort paired with Qwen2.5-7B-Instruct at $t = 0.3$ (Team, 2024b), comprising $N = 1000$ human responses with 251 covered and 749 uncovered instances.

Tokenization and basic counting conventions. We use lightweight regex-based tokenization for feature extraction. Words are sequences matching "[A-Za-z']+" (lowercased). Sentences are segments split on "[.!?]+". Punctuation counts use the set "[.,!?;:]". All features are computed at the response level.

Feature inventory. Table 9 lists the core features used in the logistic regression analysis, grouped by the construct they aim to capture. Missing values are replaced by 0.0 prior to fitting. All features are standardized using z-scoring.

HP lexicon (for hp_term_ratio). We use the following fixed lexicon:

{harry, potter, hermione, ron, weasley, dumbledore, voldemort, snape, draco, malfoy, hogwarts, gryffindor, slytherin, ravenclaw, hufflepuff, quidditch, wand, patronus, horcrux, ministry,

Task	Prompt (verbatim)
HP Fanfiction	**System:** You are an accomplished author of Harry Potter fan fiction. Write in an immersive narrative voice that respects the tone and canon established by J.K. Rowling while exploring new character dynamics and plot developments. **User:** Write a story inspired by Harry Potter that takes place entirely within the existing world of the Harry Potter novels by J.K. Rowling. The story should clearly read like fan fiction: include familiar settings, spells, and characters while inventing a fresh plot. Deliver a complete narrative arc between 1000 and 5000 words, maintaining continuity with canon but introducing new events or perspectives.
DAT	Please enter 10 words that are as different from each other as possible in meaning and usage. Rules: (1) Only single words in English. (2) Only common nouns (things, objects, concepts). (3) No proper nouns (no specific people, brands, or places). (4) No specialised or technical vocabulary; keep words familiar to the general public. (5) Think of the words on your own; do not reference anything you can currently see. Return the words as a numbered list 1–10, one noun per line, with no explanations.
AUT	List #num creative alternative uses for the {object}.
CodeNet	**System:** You are an expert competitive programming tutor who writes clean, efficient Python solutions. When you are given a problem description, you reason through edge cases, design an algorithm, and output fully working Python 3 code. Use standard input/output, avoid external libraries. **User:** You are asked to solve a CodeNet competitive programming task. Problem ID: {problem_id} Problem Statement: {description} Sample Input: {sample_input} Sample Output: {sample_output} Please produce a complete Python 3 program that solves this problem for all valid inputs. Follow these guidelines: (1) Read from standard input and write to standard output. (2) Avoid printing anything other than the final required outputs. (3) Make sure the code handles edge cases implied by the problem statement. (4) Do not include any explanatory comments outside the code. Return only the Python code. Do not wrap it in Markdown fencing.

Table 8: **Prompts used for HP fanfiction, DAT, AUT, and CodeNet in the baseline coverage evaluation.**

auror, muggle, hogsmeade, diagon, grimmauld, weasleys, lumos, expelliarmus, azkaban, phoenix, order, deathly, hallows, death, eater, dark, mark, broom, owl, owlery}.

Classifier configuration and significance reporting. We fit an L2-regularized logistic regression with lbfgs optimization, $C = 1.0$, and max_iter $= 500$. Because the labels are imbalanced, we apply class-balanced weights where the per-class weight is inversely proportional to class frequency:

$$w_c = \frac{N}{K N_c},$$

with N the number of samples, $K = 2$ classes, and N_c the class count. Under $N = 1000$, $N_1 = 251$, and $N_0 = 749$, this yields $w_1 \approx 1.992$ and $w_0 \approx 0.668$. We exclude proper_noun_ratio from the final reported specification. For interpretability, we report standardized coefficients and corresponding Wald-test p-values, highlighting coefficients with $p < 0.05$ (no multiple-comparison correction in the current version).

Fit statistics. For this in-sample diagnostic fit, we obtain accuracy $= 0.690$, F1 $= 0.509$, and ROC-AUC $= 0.753$.

E Robustness Checks

This section checks whether our main coverage patterns are stable to reasonable changes in the representation and boundary construction. We perform two complementary robustness checks. First, we swap the sentence encoder while holding the rest of the pipeline fixed (p75 neighborhood with k=15 and var90 PCA retention), recomputing coverage with a MiniLM encoder (Table 10). Second, we keep the MPNet encoder fixed but adjust the PCA variance-retention target used to define the evaluation space—specifically, we set it to the dimensionality needed to explain 80% or 70% of the variance (referred to as var80 and var70). This alters the effective dimensionality while maintaining the same human-fitted projection procedure. (Tables 11 and 12). Across both checks, the qualitative conclusions remain the same: increasing temperature consistently increases LLM-Cov across tasks, CodeNet remains easiest to cover and HP fanfiction remains hardest, and Human-Cov stays high relative to LLM-Cov, suggesting the differences are driven primarily by breadth of exploration within the human region rather than systematic semantic drift.

Construct	Feature	Definition
HP grounding	hp_term_ratio	Fraction of word tokens that belong to a fixed HP lexicon (listed below).
Verbosity / structure	word_count	Number of word tokens.
	character_count	Number of characters in the raw response string.
	sentence_count	Number of sentence segments after splitting on "[.!?]+".
	average_sentence_length	Mean words-per-sentence (computed over non-empty sentence segments).
	sentence_length_variance	Variance of words-per-sentence over sentence segments.
	average_word_length	Mean characters-per-word over tokens.
Lexical diversity / repetition	type_token_ratio	$\lvert$unique words$\rvert$/word_count.
	trigram_repetition_count	For each trigram, count repeats beyond the first and sum across trigrams.
	self_entropy	Shannon entropy of the empirical word distribution in the response.
Lexical rarity	rare_word_ratio	Fraction of tokens with Zipf frequency < 3 using *wordfreq* (Speer, 2016).
	zipf_log_mean	Mean Zipf frequency of tokens (wordfreq).
	zipf_log_var	Variance of Zipf frequency of tokens (wordfreq).
	hapax_ratio	Fraction of tokens that appear exactly once within the response.
Stylistic / pragmatic markers	uppercase_count	Count of uppercase letters in the raw response string.
	hedge_word_count	Count of hedge terms based on a fixed hedge lexicon.
	punctuation_count	Count of punctuation characters in "[.,!?;:]".
Syntactic proxy	average_clauses_per_sentence	Clauses approximated as (#commas + 1) per sentence, averaged across sentences.

Table 9: **HP diagnostic feature set used for logistic regression. All features are computed at the response level and z-scored prior to fitting.**

Model	Divergent Thinking				Convergent Reasoning		Creative Writing	
	AUT (p75)		DAT (p75)		CodeNet (p75)		HP Fanfic (p75)	
	LLM-Cov	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**
$\mathbf{t = 0.3}$								
Qwen2.5-7B	0.344	0.999	0.456	1.000	0.640	1.000	0.185	0.884
Mistral-7B	0.230	0.944	0.317	1.000	0.444	0.920	0.043	0.732
Llama-3.1-8B	0.327	0.985	0.204	1.000	0.790	0.982	0.099	0.721
Qwen2.5-32B	0.319	1.000	0.225	1.000	0.342	1.000	0.157	0.891
Mistral-24B	0.390	0.971	0.108	1.000	0.290	0.996	0.065	0.898
Avg.	0.322	0.980	0.262	1.000	0.501	0.980	0.110	0.825
$\mathbf{t = 1.0}$								
Qwen2.5-7B	0.608	0.985	0.756	1.000	0.762	0.986	0.196	0.904
Mistral-7B	0.641	0.959	0.706	0.996	0.706	0.884	0.102	0.845
Llama-3.1-8B	0.678	0.971	0.651	0.998	0.884	0.903	0.127	0.778
Qwen2.5-32B	0.492	0.996	0.543	0.995	0.586	0.970	0.234	0.865
Mistral-24B	0.764	0.941	0.487	0.998	0.698	0.982	0.161	0.923
Avg.	0.637	0.970	0.629	0.997	0.727	0.945	0.164	0.863
Overall Avg.	0.479	0.975	0.445	0.999	0.614	0.962	0.137	0.844

Table 10: **Robustness to encoder choice (MiniLM, p75, var90). Coverage metrics recomputed using the Sentence-Transformers all-MiniLM-L6-v2 encoder (sentence-transformers, 2020; Sentence-Transformers, 2025).**

Model	Divergent Thinking				Convergent Reasoning		Creative Writing	
	AUT (p75)		DAT (p75)		CodeNet (p75)		HP Fanfic (p75)	
	LLM-Cov	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**
$\mathbf{t = 0.3}$								
Qwen2.5-7B	0.248	1.000	0.547	1.000	0.504	1.000	0.259	0.999
Mistral-7B	0.204	1.000	0.351	1.000	0.480	0.934	0.086	1.000
Llama-3.1-8B	0.344	1.000	0.349	1.000	0.720	0.992	0.043	0.851
Qwen2.5-32B	0.262	1.000	0.302	1.000	0.278	1.000	0.096	0.966
Mistral-24B	0.377	1.000	0.161	1.000	0.280	0.996	0.117	1.000
Avg.	0.287	1.000	0.342	1.000	0.452	0.984	0.120	0.963
$\mathbf{t = 1.0}$								
Qwen2.5-7B	0.587	0.998	0.774	1.000	0.744	0.994	0.263	0.977
Mistral-7B	0.617	0.993	0.609	0.895	0.744	0.916	0.126	0.997
Llama-3.1-8B	0.710	0.994	0.774	1.000	0.880	0.942	0.068	0.848
Qwen2.5-32B	0.499	0.990	0.678	0.999	0.392	0.962	0.177	0.951
Mistral-24B	0.823	0.996	0.647	1.000	0.626	0.998	0.177	0.988
Avg.	0.647	0.994	0.696	0.979	0.677	0.962	0.162	0.952
Overall Avg.	0.467	0.997	0.519	0.989	0.565	0.973	0.141	0.958

Table 11: **Robustness to PCA variance retention (mpnet, p75, var80). Coverage metrics recomputed with the same sentence encoder (all-mpnet-base-v2).**

Model	Divergent Thinking				Convergent Reasoning		Creative Writing	
	AUT (p75)		DAT (p75)		CodeNet (p75)		HP Fanfic (p75)	
	LLM-Cov	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**	**LLM-Cov**	**Human-Cov**
$\mathbf{t = 0.3}$								
Qwen2.5-7B	0.186	1.000	0.504	1.000	0.514	1.000	0.258	0.999
Mistral-7B	0.183	1.000	0.351	1.000	0.560	0.988	0.086	1.000
Llama-3.1-8B	0.314	1.000	0.303	1.000	0.672	0.998	0.051	0.947
Qwen2.5-32B	0.234	1.000	0.281	1.000	0.286	1.000	0.089	0.980
Mistral-24B	0.344	1.000	0.146	1.000	0.332	1.000	0.115	1.000
Avg.	0.252	1.000	0.317	1.000	0.473	0.997	0.120	0.985
$\mathbf{t = 1.0}$								
Qwen2.5-7B	0.576	0.999	0.751	1.000	0.786	0.998	0.257	0.986
Mistral-7B	0.625	0.994	0.642	0.974	0.776	0.973	0.125	0.999
Llama-3.1-8B	0.699	0.994	0.743	1.000	0.896	0.983	0.074	0.935
Qwen2.5-32B	0.466	0.988	0.629	1.000	0.384	0.968	0.178	0.969
Mistral-24B	0.822	0.994	0.636	1.000	0.666	1.000	0.166	0.993
Avg.	0.638	0.994	0.680	0.995	0.702	0.985	0.160	0.976
Overall Avg.	0.445	0.997	0.499	0.997	0.587	0.991	0.140	0.981

Table 12: **Robustness to PCA variance retention (mpnet, p75, var70). Coverage metrics recomputed with the same sentence encoder (all-mpnet-base-v2).**