Where Models Concentrate and Humans Spread: A Coverage Framework for Distributional Pluralism in Open-Ended Generation

Zini Yang [1] **Richard Jean So** [2] **Emily Wenger** [3 1]

Abstract

We introduce a coverage-based framework for evaluating distributional pluralism in open-ended generation, asking whether outputs from large language models (LLMs) cover the empirical distribution of human responses across diverse contributors and communities. Building on Sorensen et al.'s distributional pluralism, we instantiate it as a geometry-based coverage problem: given a sample of legitimate human responses, we estimate a human response space without relying on pre-specified groups, opinions, or value dimensions. This framework allows us to evaluate not only whether LLMs generate plausible responses, but also where their outputs concentrate and which regions of human variation remain uncovered. We construct an empirical human response boundary in a shared sentence-embedding space and evaluate model outputs along two complementary metrics: how often they remain inside the boundary (IBR) and how much of the human response distribution they cover (LLM-Cov). Across tasks, LLMs show substantially lower coverage than a human-to-human reference, with the gap concentrated in peripheral regions of the human distribution. We further show that model outputs concentrate in central, high-density regions of the human distribution, while under-covered regions in the narrative task are structured. In HP fanfiction, models more easily reach canon-visible and stylistically regular writing while missing more implicit, irregular, and community-specific forms of expression, illustrating how open-ended generation can produce pluralistic under-representation even when outputs remain plausible.

[1]Department of Computer Science, Duke University, Durham, NC, USA [2]Department of English, Duke University, Durham, NC, USA [3]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. Correspondence to: Zini Yang <zini.yang@duke.edu>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Can LLMs represent heterogeneous human views, values, and forms of expression? Pluralistic alignment frames this as a central question for AI evaluation, and a growing body of work takes it up from multiple angles (Sorensen et al., 2024; Xie et al., 2025; Kirk et al., 2024). In opinion surveys and value-laden settings, pluralism most often appears as disagreement across stances: different groups give different answers, and a pluralistic system should reflect that distribution rather than collapse it to one side or to an average. Much empirical work on pluralistic alignment has focused on this regime, operationalizing plurality through stance distributions, group-conditioned preferences, or annotator disagreement (Santurkar et al., 2023; Meister et al., 2025; Novis-Deutsch et al., 2025).

These approaches operationalize plurality top-down, through pre-specified groups, opinions, or value dimensions, and require enumerating the relevant axes of variation in advance. In creative and open-ended settings, pluralism takes a different form. Two writers asked to extend the same story do not disagree on facts; they extend it in different directions. Their cultural backgrounds, reading histories, and the literary or fan communities they write within shape which characters they care about, which gaps in the source they want to fill, and which aesthetic conventions feel natural. Harry Potter fanfiction is a familiar instance: a shared fictional world supports thousands of mutually legitimate continuations, organized around different pairings, subgenres, emotional registers, and relationships to canon. Plurality here is not primarily a plurality of opinions. It is a plurality of legitimate *realizations*: many distinct outputs can be valid for the same prompt, and what counts as a meaningful difference is shaped by genre, community, and stylistic norms.

We argue that this calls for an evaluation metric for pluralistic alignment that takes the empirical human response distribution as its reference, rather than a pre-specified set of options or values. We ask: how much of this empirical distribution can a model reach, and does it remain inside that distribution while doing so? This is distributional pluralism for open-ended tasks, instantiated bottom-up from observed human responses. We approach this question with

a geometry-based framework: we embed human responses in a shared semantic space, treat the union of their local neighborhoods as an empirical human response boundary, and evaluate model outputs along two complementary axes. *In-Boundary Rate (IBR)* measures how often model outputs fall inside the boundary, and *LLM Coverage (LLM-Cov)* measures how much of the human responses the model reaches. We apply this framework to AUT, DAT, and Harry Potter fanfiction, spanning constrained short-form ideation and open-ended narrative writing.

Our contributions This paper makes three contributions to pluralistic alignment and open-ended generation. First, we extend distributional pluralism to open-ended generation. Where existing pluralistic-alignment work often measures plurality over predefined opinions, values, demographic groups, or preference labels, we study settings where plurality appears as different *realizations* of the same prompt, shaped by many underlying factors. This reframes pluralistic alignment as a question of whether models preserve the breadth of human expression, and motivates our human-grounded coverage framework.

Second, we introduce a coverage framework with two complementary metrics, IBR and LLM-Cov, that distinguish human-plausibility from pluralistic reach, together with a Human-to-Human reference that calibrates what coverage looks like when a population samples from itself. Across two ideation tasks (AUT, DAT) and Harry Potter fanfiction, current LLMs achieve high IBR while recovering only a fraction of the Human-to-Human reference, and the gap is concentrated in peripheral regions of the human distribution.

Third, the framework lets us ask not just how much of the human distribution models reach, but which part. Through a covered-versus-uncovered analysis on Harry Potter fanfiction, we show that the human responses models reach are systematically more canon-anchored, more dialogue-marked, and more aligned with canonical mixed-gender pairings than the unreached ones, indicating that models reach the more canon-visible and stylistically regular part of the human distribution. Coverage-based evaluation thus diagnoses not only whether models collapse, but which human possibilities they leave out.

2. Related Work

2.1. Pluralistic Alignment and Distributional Pluralism

Our framing builds on recent work on pluralistic alignment, which argues that language models should not collapse heterogeneous human values, viewpoints, preferences, and forms of expression into a single averaged response (Sorensen et al., 2024; Xie et al., 2025; Kirk et al., 2024). Sorensen et al. (2024) distinguish several operationalizations of pluralism, including Overton pluralism, steerable pluralism, and distributional pluralism. In this taxonomy, distributional pluralism asks whether the distribution of model outputs, across samples or users, matches the distribution of a target population. A growing literature studies whose opinions current models reflect (Santurkar et al., 2023; Novis-Deutsch et al., 2025), how to benchmark distributional alignment and heterogeneous preferences (Kirk et al., 2024; Meister et al., 2025; Poole-Dayan et al., 2026; Nie et al., 2026; Zhang et al., 2026), and how to incorporate public input (Anthropic, 2023), cultural variation (AlKhamissi et al., 2024; Masoud et al., 2025), or modular pluralist architectures (Feng et al., 2024; Fu et al., 2026; Xu et al., 2025) into alignment pipelines.

Most empirical work in this area operationalizes pluralism through pre-specified categories: opinions, values, demographic groups, preference labels, or annotator disagreement. This framing is well suited to opinion- and value-laden settings, where plurality is largely *propositional*: different people hold different positions on a question, and a pluralistic system should reflect that distribution. Related work on perspectivism and annotator disagreement similarly treats disagreement not as noise to be collapsed, but as a meaningful signal of human variation (Plank, 2022; Mostafazadeh Davani et al., 2022; Uma et al., 2021; Basile et al., 2021). These perspectives have strongly shaped pluralistic alignment by showing that aggregation can erase the very population-level variation models are meant to represent.

In open-ended generation, however, plurality often takes a different form: it is not primarily a distribution over stances, labels, or values. Legitimate responses may instead differ in style, genre, interpretation, trope, community convention, or narrative direction. We call this *realizational pluralism*: plurality over the many valid ways a prompt can be realized. In the taxonomy of Sorensen et al. (2024), the closest category is distributional pluralism, but our setting changes the reference object. Rather than assuming a target distribution over pre-specified groups, opinions, or value dimensions, we instantiate distributional pluralism bottom-up from observed human responses. This makes pluralistic evaluation applicable to domains such as ideation, creative writing, and fanfiction, where the relevant axes of variation are often implicit, stylistic, and community-specific rather than predefined.

2.2. Scalar Evaluation of Creativity and Open-Ended Generation

A large literature evaluates LLMs on open-ended tasks using scalar or model-relative criteria such as semantic diversity, novelty, originality, surprise, usefulness, and human judgment (Anderson et al., 2024; Atmakuru et al., 2024; Bellemare-Pepin et al., 2026; Chakrabarty et al., 2024; Chen

& Ding, 2023; Dinu et al., 2025; Doshi & Hauser, 2024; Hou et al., 2025; Ismayilzada et al., 2025; Zhang et al., 2025; Zhao et al., 2025). A common approach is to prompt a model to generate multiple responses to a writing, reasoning, or ideation task, then compute how semantically varied those responses are from each other or from a human benchmark. Such methods often use embeddings and cosine similarity to quantify semantic distance, treating greater dispersion as evidence of greater diversity or creativity (Padmakumar & He, 2024; Lee & Chung, 2024).

Other work draws on established creativity theory, measuring model outputs through concepts such as originality, novelty, surprise, usefulness, elaboration, and quality (Runco & Jaeger, 2012; Amabile, 1988; Simonton, 2018; Corazza, 2016; Diedrich et al., 2015). These approaches often rely on human evaluators or LLM-as-judge methods to assign numerical ratings to generated text (He et al., 2025; Hou et al., 2025; Ismayilzada et al., 2025; Zhao et al., 2025). The most careful studies use codebooks to guide evaluation and then validate automated judgments against human ratings.

From a pluralistic standpoint, these methods share a limitation: they typically measure the quality or dispersion of model outputs without directly asking how much of the empirical human response distribution those outputs reach. A model can appear diverse relative to its own samples, or score highly on average originality, while still concentrating near the center of what humans actually produce.

2.3. Plural Human Response Spaces in Open-Ended Tasks

Open-ended generation includes divergent thinking tasks such as the Alternative Uses Task and the Divergent Association Task, as well as short story writing, narrative generation, research ideation, and fanfiction (Guilford, 1967; Mednick, 1962; Olson et al., 2021; Si et al., 2024; Alfassi et al., 2025). Prior work increasingly emphasizes that creativity is not a single, domain-invariant capacity, but is shaped by task structure, genre, and evaluative context (Baer, 1998; Baer & Kaufman, 2005; Hou et al., 2025; Jain et al., 2025; Lai et al., 2025). This task-dependence matters for pluralistic evaluation: the form plurality takes depends on the task.

In constrained ideation tasks, plurality may appear as different uses, associations, or conceptual links. In narrative writing, plurality is more deeply structured by genre conventions, voice, pacing, character focus, relationship dynamics, and community memory. Fanfiction provides a particularly clear example. A shared fictional world can support many mutually legitimate continuations, organized around different pairings, subgenres, emotional registers, and relationships to canon (Alfassi et al., 2025). These differences are not merely variations in quality. They define different regions of a human response space.

Recent work on LLM-assisted creativity suggests that models can improve individual-level fluency, elaboration, or perceived creativity while narrowing the collective diversity of responses (Anderson et al., 2024; Doshi & Hauser, 2024; Lee & Chung, 2024; Kumar et al., 2025; Wenger & Kenett, 2025). From this perspective, creative homogenization is not only a stylistic concern but also a distributional pluralism failure: models may produce plausible responses while failing to cover peripheral or community-specific regions of human expression.

2.4. Boundary-Based and Coverage-Oriented Perspectives

Boundary-based thinking has a long history in creativity theory. Boden (1990; 2004; 1998) characterizes creativity in terms of movement within, exploration of, or transformation of conceptual spaces, rather than improvement along a single scalar dimension. Csikszentmihalyi (1996) similarly emphasizes that creative production is shaped by domains, constraints, and communities of evaluation. This boundary-based view suggests that understanding open-ended generation requires more than assigning a quality score; it requires asking what regions of a possibility space an agent can reach.

In the LLM setting, boundary-like evaluations appear in constrained story-writing benchmarks, pass/fail creativity tests, and task-specific stress tests that reveal where models fail to satisfy conceptual, stylistic, or constraint-based requirements (Atmakuru et al., 2024; Chakrabarty et al., 2024; Tian et al., 2025). These approaches are valuable because they expose categorical failures that scalar ratings may obscure. However, they usually focus on whether a model passes a particular benchmark or satisfies a particular constraint, rather than estimating the shape of the broader human response distribution.

Coverage-oriented work on creative homogenization comes closer to our goal. Embedding-dispersion analyses, diversity-growth metrics, and pairwise similarity comparisons show that LLM outputs can be more homogeneous than human outputs at the population level (Anderson et al., 2024; Doshi & Hauser, 2024; Wenger & Kenett, 2025). Other recent work cautions that homogenization is task-dependent and that generic diversity penalties may not improve all tasks (Jain et al., 2025). These findings point toward the need for task-conditioned coverage frameworks rather than a single universal diversity score.

Our work builds on boundary and coverage perspectives but changes the reference object: instead of measuring diversity only among model outputs, we use observed human responses to define an empirical human response boundary. We then evaluate model outputs along two complementary axes: whether they remain inside this boundary, and how

much of the human response distribution they cover. A model can be fluent, plausible, and even diverse relative to itself while still failing to cover peripheral regions of human expression. Our framework is designed to identify precisely this failure mode: plausibility without pluralistic breadth.

3. Methodology

Unlike prior work that evaluates open-ended generation through explicit scoring functions, we adopt an implicit, human-grounded formulation. We let human responses define the reference space, instead of prescribing the relevant dimensions of pluralism in advance.

More precisely, for any given prompt, we interpret the distribution of human responses as defining a human reachable semantic region within 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 **empirical human response 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 model outputs by analyzing where they fall relative to this boundary, which is inferred directly from the data rather than relying on predefined groups, opinions, or value dimensions. Figure 1 illustrates the overall framework, described in detail below.

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

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 verify that our results are robust to reasonable changes of this variance threshold. 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 empirical human response 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. We use $q = 0.50$ as the main setting and examine alternative percentile choices as robustness checks.

Distributional Pluralism Coverage Metrics. Building on the empirical human response boundary defined above, we introduce two complementary metrics that characterize how LLM-generated responses relate to the human response distribution. Given human response embeddings $H = \{h_i\}$ and LLM response embeddings $M = \{m_j\}$ for the same prompt, and a fixed neighborhood radius ε, we define the following measures.

First, we measure how much of the human response distribution is reached by model-generated responses. 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 reach a larger portion of the empirical human response space.

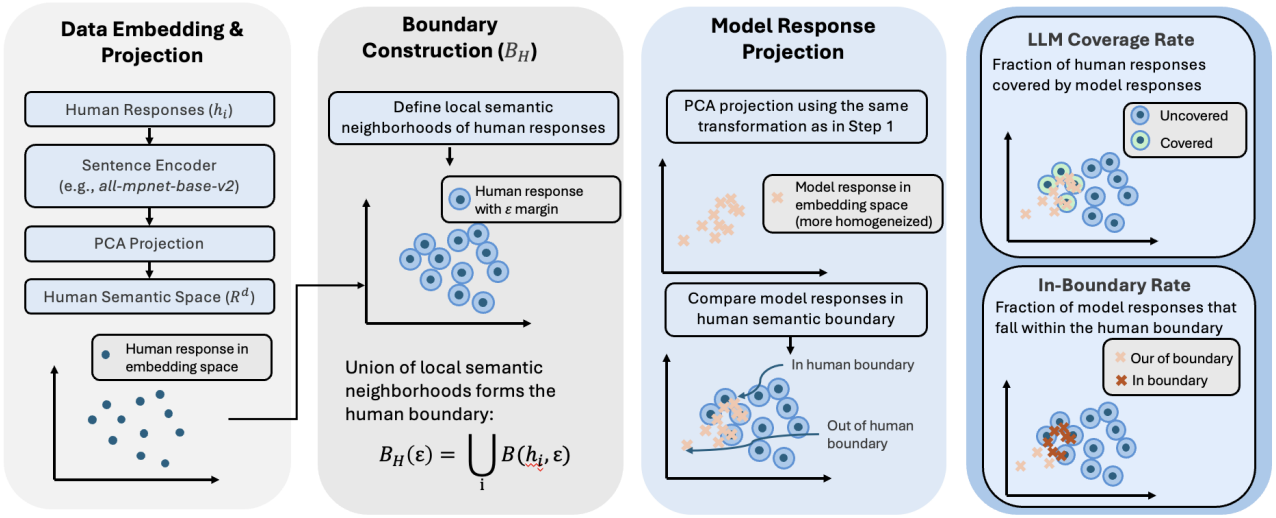

Figure 1. **Overview of our human-grounded coverage framework.**

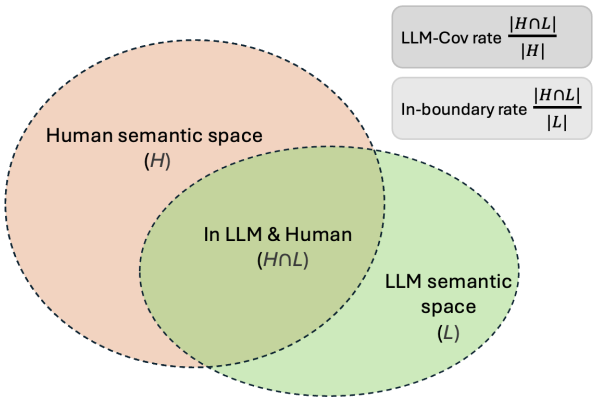

Figure 2. **Geometric interpretation of LLM-Cov and IBR.** LLM-Cov measures how much of the human response distribution is reached by model outputs, while IBR measures how often model outputs remain inside the empirical human boundary.

Second, we measure how often model-generated responses remain inside the empirical human response boundary. 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. Higher values indicate that model outputs remain in regions occupied by human responses.

Together, these two metrics separate pluralistic breadth from human-legible plausibility. LLM-Cov asks whether model outputs cover the range of observed human responses, while IBR asks whether those outputs remain within the empirical human boundary. This distinction is central to our framework: a model can produce plausible responses with high IBR while still being pluralistically narrow with low LLM-Cov.

Task	Abbreviated prompt
AUT	*List alternative uses for a brick.*
DAT	*Write 10 nouns in English that are as unrelated to each other as possible.*
HP Fanfic	*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. **Open-ended ideation and narrative tasks used in our evaluation.**

4. Evaluation Procedure

Here, we describe the open-ended generation tasks, evaluated models, and metric implementation details used in our experiments.

4.1. Tasks

We evaluate distributional pluralism across three open-ended generation settings in (Table 1). The Alternative Uses Task (AUT) and Divergent Association Task (DAT) are constrained short-form ideation tasks, while Harry Potter fanfiction is a long-form narrative task.

AUT is a widely used divergent-thinking task that asks participants to generate alternative uses for common objects, capturing ideational flexibility and originality (Guilford, 1967). **DAT** elicits semantically distant concepts and is grounded in associative theories of creativity (Mednick, 1962; Olson et al., 2021). **HP Fanfic** provides a community-structured narrative setting in which legitimate responses can vary by pairing, trope, register, relationship to canon, and fandom convention; recent work on fanfiction and AI

highlights the relevance of fan communities, authenticity, and creative norms in this domain (Alfassi et al., 2025). Together, these tasks allow us to evaluate whether model coverage gaps are uniform across open-ended 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.

4.3. Metric Implementations

We embed all human and model responses using `all-mpnet-base-v2` from the SentenceTransformers, 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 50^{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 coverage-based evaluation of distributional pluralism on the three open-ended generation tasks.

5.1. Overall LLM Performance

Overall, the open-source instruction-tuned LLMs we evaluate (Qwen, Llama, and Mistral families) usually fall inside the empirical human boundary but reach only a narrow region of it. They achieve consistently high IBR across tasks and temperatures, but substantially lower and more variable LLM-Cov, well below our Human-to-Human reference, where a held-out human sample is evaluated against the remaining human responses using the same procedure.

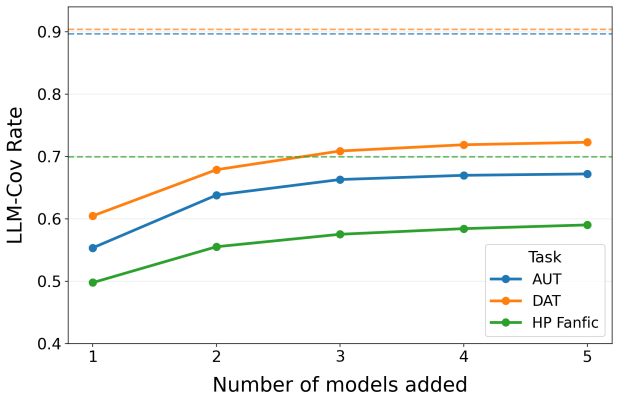

Figure 3. **Sequential model addition yields rapidly diminishing gains in LLM-Cov and remains below the Human-to-Human reference on every task.**

Models are human-plausible, but not broadly pluralistic. These results are summarized in Table 2.

Across model sizes, we do not find evidence that larger models consistently achieve higher LLM-Cov than smaller ones within the same family. Decoding temperature has a clearer effect: increasing temperature from t=0.3 to t=1.0 substantially raises LLM-Cov on every task while leaving IBR largely unchanged, indicating that higher-temperature sampling broadens model reach within the human distribution rather than pushing outputs outside it.

Population-level homogenization. A natural follow-up question is whether the coverage gap is shared across models, or whether different models cover complementary regions of the human distribution that together approach Human-to-Human coverage. Figure 3 shows that ensembling does not close the gap. Across all three tasks, union model covered area (LLM-Cov) rises sharply with the first model and then plateaus well below the Human-to-Human reference, with additional models contributing little new human response space. This pattern is consistent with prior evidence that LLMs tend to be homogeneous with one another rather than covering independent regions of the human response space (Wenger & Kenett, 2025).

5.2. Coverage by Human Centrality

We next ask how LLM coverage is distributed across the human response space. To do this, we divide human responses into ten centrality deciles using human-human distances: central responses sit close to many other human responses, while peripheral responses are farther from the rest of the human distribution.

Figure 4 shows that LLM-Cov is concentrated near the center of the human space and falls steeply toward the periphery. The Human-to-Human reference shows the same downward shape, confirming that peripheral responses are intrinsically harder to cover from any starting point. The model drop-off,

Model	AUT		DAT		HP Fanfic	
	IBR	LLM-Cov	IBR	LLM-Cov	IBR	LLM-Cov
$t = 0.3$						
Qwen2.5-7B	1.000	0.074	1.000	0.368	0.699	0.255
Ministral-8B	0.960	0.184	0.740	0.052	0.484	0.038
Llama-3.1-8B	0.928	0.172	1.000	0.116	0.621	0.100
Qwen2.5-32B	0.997	0.149	0.998	0.104	0.433	0.346
Mistral-24B	0.936	0.256	0.992	0.033	0.857	0.333
Avg.	0.964	0.167	0.946	0.135	0.619	0.214
$t = 1.0$						
Qwen2.5-7B	0.897	0.323	1.000	0.604	0.664	0.377
Ministral-8B	0.840	0.553	0.714	0.318	0.575	0.290
Llama-3.1-8B	0.847	0.480	0.963	0.497	0.559	0.264
Qwen2.5-32B	0.943	0.306	0.996	0.455	0.650	0.477
Mistral-24B	0.828	0.550	0.967	0.314	0.783	0.421
Avg.	0.871	0.443	0.928	0.438	0.646	0.366
Overall Avg.	0.918	0.305	0.937	0.286	0.632	0.290
Human-to-Human	0.908	0.896	0.924	0.904	0.746	0.700

Table 2. **Models achieve consistently high IBR across tasks and temperatures, but substantially lower LLM-Cov than the Human-to-Human reference.** Coverage metrics are computed at $p50$ across all the tasks.

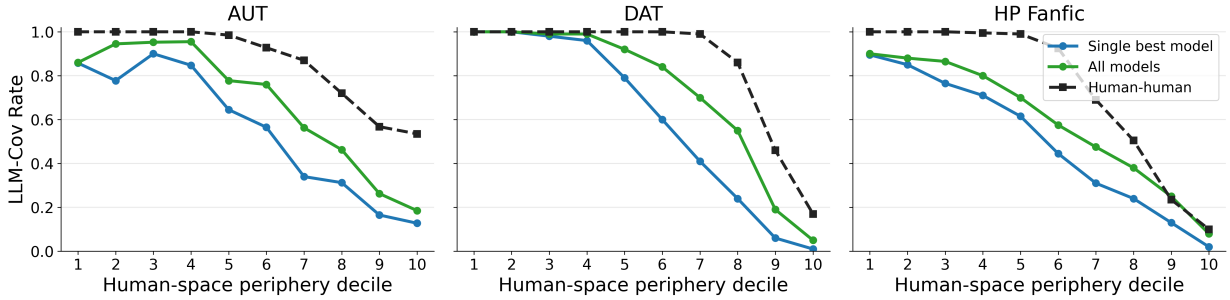

Figure 4. **Models under-cover peripheral regions of the human response distribution.** LLM-Cov drops sharply toward less central human responses, while the Human-to-Human reference declines more gradually and remains higher across the distribution.

however, is much sharper than the human one, and the gap to the Human-to-Human reference is largest in the periphery.

The uncovered region is therefore not randomly distributed; it is systematically concentrated in less central parts of the human distribution. We use HP fanfiction in the next section as a case study to examine what kinds of human responses these uncovered regions actually contain.

5.3. What Models Reach and What They Miss

Setup. We use Mistral-Small-24B-Instruct-2501 at t=1.0 and apply the HP fanfiction $p50$ boundary in MPNet+PCA space. Each AO3 excerpt is labeled COVERED if it lies within ε of at least one model output, and UNCOVERED otherwise. We compare the two groups along three axes: canonical anchoring, style and rhythm, and relationship orientation.

The covered and uncovered excerpts differ in clear ways. Covered excerpts look more obviously like Harry Potter fanfiction: they mention more familiar settings, use more canon-specific words, and are more likely to name HP characters. They also contain slightly more dialogue and show less variation in sentence length. Uncovered excerpts are

Axis	Variable	Diff.	p
Canonical anchoring	HP setting present	+0.169	2.1×10^{-13}
	Canon vocab present	+0.153	3.3×10^{-11}
	HP character present	+0.075	2.5×10^{-6}
	Non-HP characters present	−0.076	2.1×10^{-5}
Style and rhythm	Mean sentence length	−1.61	3.6×10^{-6}
	Sentence-length variance	−224.9	8.5×10^{-6}
	Dialogue ratio	+0.029	0.003
Relationship	Romance present	−0.056	0.007
	M/M relationship	−0.149	3.3×10^{-12}
	M/F relationship	+0.110	6.0×10^{-7}

Table 3. **Covered HP fanfiction is more canon-visible, while uncovered writing shows more fandom-specific relationship patterns.** Diff. is computed as Covered minus Uncovered.

less obviously tied to the original books on the surface. They are more likely to introduce non-HP characters and show more variation in sentence rhythm. This suggests that the model more easily reaches writing that clearly signals "Harry Potter" through familiar names, places, and style, while missing writing that moves farther away from those surface cues.

The relationship patterns add another layer. Uncovered excerpts contain more romance overall and many more M/M relationships, while covered excerpts contain more M/F relationships. This suggests that the model is better at reaching relationship patterns closer to the original canon, and worse at reaching relationship patterns that are more specific to fan communities.

6. Discussion

Our results show a consistent separation between plausibility and pluralistic breadth. Across tasks, LLM outputs often remain inside the empirical human response boundary, but cover far less of the human response distribution than a matched Human-to-Human reference. This means that current models are not primarily failing by producing invalid responses; they are failing by representing only a narrow subset of valid human possibilities, especially in open-ended generation test.

This gap is not resolved by simple scaling or aggregation. Larger models do not consistently improve coverage, higher temperature broadens reach without closing the gap, and sequentially combining models yields diminishing returns. These results suggest that current instruction-tuned LLMs tend to occupy overlapping regions of the human response space rather than complementary ones. For pluralistic alignment, this matters because pluralism cannot be assumed to emerge simply from sampling more outputs or pooling multiple similar models.

The missed regions are also structured. Model coverage is concentrated in central, high-density areas of the human response distribution and drops sharply toward the periphery. In HP fanfiction, the covered subset is more canon-visible, dialogue-marked, and stylistically regular, while the uncovered subset is less surface-canon marked, more rhythmically irregular, and more shaped by fandom-specific relationship patterns. Thus, low coverage is not just a geometric artifact; it corresponds to meaningful forms of human variation that models under-represent.

These findings extend pluralistic alignment beyond settings where the relevant alternatives are known in advance. Existing work often studies distributions over predefined opinions, values, demographic groups, or preference labels. Open-ended generation requires a complementary view: pluralism can also appear as different legitimate realizations of the same prompt. A bottom-up coverage framework makes this form of pluralism measurable without specifying the relevant axes of variation beforehand.

Future work. This framework gives us a natural way to study open-ended pluralism more broadly. Rather than only looking at distributions over opinions or preferences, we can ask about distributions over expression: style, genre, voice, interpretation, and community convention. To do this well, we will need richer representations than the sentence embeddings we use here, which may miss dimensions like discourse structure, narrative voice, and community-specific meaning. We will also need broader human reference sets, so we can ask whose variation models cover and whose they leave out. Coverage can also be used to test interventions, like alternative decoding strategies, post-training methods, or community-specific adaptation, by asking whether they actually expand model reach into under-covered regions, not just increase model-internal diversity.

7. Limitations

Our framework has several limitations. First, it relies on sentence-level embeddings to construct the empirical human response boundary. These embeddings capture broad semantic similarity, but may miss dimensions important for open-ended cultural generation, such as narrative structure, voice, genre convention, pragmatic nuance, and community-specific meaning. Different encoders or hybrid representations could yield different boundary estimates.

Second, the boundary is only as representative as the human data used to construct it. Our datasets provide concrete reference distributions, but they are finite and cannot capture the full range of human variation. Larger and more diverse human samples may reveal additional regions of human response space.

Third, the boundary depends on a neighborhood-radius choice. We use $p50$ for consistency and report robustness checks, but absolute IBR and LLM-Cov values can vary with this threshold. Future work could explore adaptive or task-specific boundary selection.

Fourth, our HP fanfiction analysis is diagnostic rather than exhaustive. The covered-versus-uncovered features are proxies for richer forms of fandom-specific meaning, and we do not claim that the same patterns generalize to all open-ended domains. We also evaluate open-source instruction-tuned models from 7B to 32B parameters, but not larger frontier models or models specifically trained for creative or community-grounded generation.

References

Alfassi, R., Cooper, A., Mitchell, Z., Calabro, M., Shaer, O., and Mokryn, O. Fanfiction in the age of ai: Community perspectives on creativity, authenticity and adoption, 2025. URL http://arxiv.org/abs/2506.18706. arXiv:2506.18706 [cs.HC].

AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., and Diab, M. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12404–12422, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.671. URL https://aclanthology.org/2024.acl-long.671/.

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

Anderson, B. R., Shah, J. H., and Kreminski, M. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, pp. 413–425, 2024. doi: 10.1145/3635636.3656204. URL https://doi.org/10.1145/3635636.3656204.

Anthropic. Collective constitutional ai: Aligning a language model with public input. https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input, 2023. Anthropic research post.

Atmakuru, A., Nainani, J., Bheemreddy, R. S. R., Lakkaraju, A., Yao, Z., Zamani, H., and Chang, H.-S. CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story-Writing Constraints, 2024. URL http://arxiv.org/abs/2410.04197. arXiv:2410.04197 [cs].

Baer, J. The case for domain specificity of creativity. *Creativity Research Journal*, 11(2):173–177, 1998.

Baer, J. and Kaufman, J. C. Bridging generality and specificity: The amusement park theoretical (apt) model of creativity. *Roeper Review*, 27(3):158–163, 2005.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 15–21, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.bppf-1.3/.

Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., Bengio, Y., and Jerbi, K. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279, 2026. doi: 10.1038/s41598-025-25157-3. URL https://doi.org/10.1038/s41598-025-25157-3.

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

Boden, M. A. Creativity and artificial intelligence. *Artificial Intelligence*, 103(1):347–356, 1998. doi: 10.1016/S0004-3702(98)00055-1. URL https://www.sciencedirect.com/science/article/pii/S0004370298000551.

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

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., and Wu, C.-S. Art or Artifice? Large Language Models and the False Promise of Creativity, 2024. URL http://arxiv.org/abs/2309.14556. arXiv:2309.14556 [cs].

Chen, H. and Ding, N. Probing the Creativity of Large Language Models: Can models produce divergent semantic association?, 2023. URL http://arxiv.org/abs/2310.11158. arXiv:2310.11158 [cs].

Corazza, G. E. Potential originality and effectiveness: The dynamic definition of creativity. *Creativity Research Journal*, 28(3):258–267, 2016. doi: 10.1080/10400419.2016.1195627.

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

Diedrich, J., Benedek, M., Jauk, E., and Neubauer, A. C. The novelty–usefulness tension: Toward a dialectical model of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4):320–337, 2015. doi: 10.1037/aca0000021.

Dinu, A., Florescu, A.-M., and Resceanu, A. A Comparative Approach to Assessing Linguistic Creativity of Large Language Models and Humans, 2025. URL http://arxiv.org/abs/2507.12039. arXiv:2507.12039 [cs].

Doshi, A. R. and Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290. URL https://www.science.org/doi/10.1126/sciadv.adn5290.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pp. 4151–4171, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.240. URL https://aclanthology.org/2024.emnlp-main.240/.

Fu, Y., Son, S., and Bogunovic, I. Overton pluralistic reinforcement learning for large language models, 2026. URL https://arxiv.org/abs/2602.20759. arXiv:2602.20759 [cs.LG].

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

He, Z., Zhang, B., Liu, W., Tang, R., and Cheng, L. What Shapes a Creative Machine Mind? Comprehensively Benchmarking Creativity in Foundation Models, 2025. URL http://arxiv.org/abs/2510.04009. arXiv:2510.04009 [cs].

Hou, Z. J., Zhang, B. A., Lu, Y., Baghel, B. K., Brei, A., Lu, X., Jiang, M., Brahman, F., Chaturvedi, S., Chang, H.-S., Khashabi, D., and Li, X. L. CreativityPrism: A Holistic Benchmark for Large Language Model Creativity, 2025. URL http://arxiv.org/abs/2510.20091. arXiv:2510.20091 [cs].

Ismayilzada, M., Stevenson, C., and Plas, L. v. d. Evaluating Creative Short Story Generation in Humans and Large Language Models, 2025. URL http://arxiv.org/abs/2411.02316. arXiv:2411.02316 [cs].

Jain, S., Lanchantin, J., Nickel, M., Ullrich, K., Wilson, A., and Watson-Daniels, J. LLM Output Homogenization is Task Dependent, 2025. URL http://arxiv.org/abs/2509.21267. arXiv:2509.21267 [cs].

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera-Gomez, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html.

Kumar, H., Vincentius, J., Jordan, E., and Anderson, A. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2025. doi: 10.1145/3706598.3714198. URL http://arxiv.org/abs/2410.03703. arXiv:2410.03703 [cs].

Lai, C., Luchini, S., Lauharatanahirun, N., and Beaty, R. Creative or Uncreative Partner: Comparing Humans and AI in Collaborative Creative Tasks, 2025. URL https://osf.io/ey7u4_v1.

Lee, B. C. and Chung, J. J. An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, 8(10):1906–1914, 2024. doi: 10.1038/s41562-024-01953-1. URL https://www.nature.com/articles/s41562-024-01953-1. Publisher: Nature Publishing Group.

Masoud, R. I., Liu, Z., Ferianc, M., Treleaven, P., and Rodrigues, M. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025. URL https://aclanthology.org/2025.coling-main.567/.

Mednick, S. A. The associative basis of the creative process. *Psychological Review*, 69(3):220–232, 1962. doi: 10.1037/h0048850.

Meister, N., Guestrin, C., and Hashimoto, T. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 24–49. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.2. URL https://aclanthology.org/2025.naacl-long.2/.

Mostafazadeh Davani, A., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449. URL https://aclanthology.org/2022.tacl-1.6/.

Nie, S., Omoomi, K., Flek, L., Zhao, Z., and Welch, C. PerSpectra: A scalable and configurable pluralist benchmark of perspectives from arguments. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dyooGJcKJg. ICLR 2026 poster.

Novis-Deutsch, N., Elyoseph, T., and Elyoseph, Z. How much of a pluralist is chatgpt? a comparative study of value pluralism in generative ai chatbots. *AI & Society*, 2025. URL https://link.springer.com/article/10.1007/s00146-025-02450-3.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., and Webb, M. E. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118 (25):e2022340118, 2021. doi: 10.1073/pnas.2022340118.

URL https://www.pnas.org/doi/10.1073/pnas.2022340118.

Padmakumar, V. and He, H. Does Writing with Language Models Reduce Content Diversity?, 2024. URL http://arxiv.org/abs/2309.05196. arXiv:2309.05196 [cs].

Plank, B. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.731/.

Poole-Dayan, E., Wu, J., Sorensen, T., Pei, J., and Bakker, M. A. Benchmarking overton pluralism in llms. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=f2VxF4QIx1. ICLR 2026 poster.

Reimers, N. and Gurevych, I. 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)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Runco, M. A. and Jaeger, G. J. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Si, C., Yang, D., and Hashimoto, T. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, 2024. URL http://arxiv.org/abs/2409.04109. arXiv:2409.04109 [cs].

Simonton, D. K. Defining creativity: Don't we also need to define the opposite? *Creativity Research Journal*, 30(3): 291–294, 2018. doi: 10.1080/10400419.2018.1488195.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*. arXiv, 2020. doi: 10.48550/arXiv.2004.09297. URL http://arxiv.org/abs/2004.09297. arXiv:2004.09297 [cs.CL].

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Tian, Y., Ravichander, A., Qin, L., Bras, R. L., Marjieh, R., Peng, N., Choi, Y., Griffiths, T. L., and Brahman, F. MacGyver: Are Large Language Models Creative Problem Solvers?, 2025. URL http://arxiv.org/abs/2311.09682. arXiv:2311.09682 [cs].

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. doi: 10.1613/jair.1.12752. URL https://www.jair.org/index.php/jair/article/view/12752.

Wenger, E. and Kenett, Y. We're Different, We're the Same: Creative Homogeneity Across LLMs, 2025. URL http://arxiv.org/abs/2501.19361. arXiv:2501.19361 [cs].

Xie, Z., Wu, J., Shen, Y., Jain, R., Xia, Y., Li, X., Chang, A., Rossi, R. A., Yu, T., Kumar, S., Majumder, B. P., Shang, J., Ammanabrolu, P., and McAuley, J. J. A survey on personalized and pluralistic preference alignment in large language models. In *Proceedings of the Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lSWOMjonL7. COLM 2025.

Xu, S., Leng, Y., Yu, L., and Xiong, D. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6859–6877, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.naacl-long.350/.

Zhang, L. H., Milli, S., Jusko, K. L., Smith, J., Amos, B., Bouaziz, W., Revel, M., Kussman, J., Sheynin, Y., Titus, L., Radharapu, B., Yu, J., Sarma, V., Rose, K., and Nickel, M. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4NtoAVqfhA. ICLR 2026 poster.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., and Ippolito, D. Noveltybench: Evaluating language models for humanlike diversity, 2025. URL https://arxiv.org/abs/2504.05228.

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., Du, Z., Guo, Q., Li, L., and Chen, Y. Assessing and Understanding Creativity in Large Language Models. *Machine Intelligence Research*, 22(3):417–436, 2025. doi: 10.1007/s11633-025-1546-4. URL http://arxiv.org/abs/2401.12491. arXiv:2401.12491 [cs].

A. Reproducibility Details

This section documents the prompts, models, decoding settings, and sampling conventions used throughout the paper. Our goal is to make every coverage measurement reported in the main text reproducible from publicly available models without ambiguity about generation parameters, prompt formatting, or human reference construction. Methodology part summarizes the end-to-end pipeline, from raw human and model responses to the in-boundary rate (IBR) and LLM-coverage (LLM-Cov) metrics.

A.1. Task Prompts

For each task we use a fixed prompt template applied identically across all models and temperatures, so that any differences in coverage reflect the model rather than the prompt. Prompts were designed to elicit the same kind of response from LLMs that humans give in the corresponding psychometric task, while keeping the output format easy to parse.

Alternative Uses Task (AUT). AUT is a divergent-thinking task in which participants list creative uses for an everyday object. We instantiate it with the object `rope` and ask the model to produce a numbered list of short verb phrases.

> *You are helping brainstorm creative uses. Generate {per_call} distinct, feasible uses for the object '{object_name}'. Respond with 1–5 word verb phrases only (no full sentences, no extra nouns), number the list 1 through N, and avoid repeats or explanations.*

Each item is then normalized into the canonical frame "A potential use of the rope is . . . " before sentence embedding, which prevents trivial surface differences in how the model formats list items from affecting embedding geometry.

Divergent Association Task (DAT). DAT asks participants to produce a set of mutually unrelated common nouns; semantic distance among the chosen words is taken as an index of divergent thinking. We use the standard 10-word version:

> *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. (3) No proper nouns. (4) No specialized or technical vocabulary. (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.*

HP fanfiction. For Harry Potter fanfiction we use a minimal prompt that does not prescribe genre conventions, tropes, or pairings, so that any narrowing in model output reflects the model rather than the prompt.

> ***System.*** *You are a creative fiction writer.*
> ***User.*** *Write one Harry Potter fanfiction story. Keep it around 500 to 1000 tokens. Output only the story.*

Outputs that fell outside the 500–1000 token window were discarded. This filtering removes both truncated generations and very long responses that would otherwise dominate the embedding geometry.

A.2. Models, Decoding, and Generation Counts

We evaluate five open-source instruction-tuned LLMs spanning two model families and a 7B–32B parameter range (Table 4). All models are publicly available through HuggingFace and were run using the default chat template provided with each tokenizer.

Table 4. **Model identifiers used for generation.**

Short name	HuggingFace identifier
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`
Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct`
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`
Ministral-8B	`mistralai/Ministral-8B-Instruct-2410`
Mistral-24B	`mistralai/Mistral-Small-24B-Instruct-2501`

All main results use two decoding temperatures, $t = 0.3$ (lower entropy, more concentrated) and $t = 1.0$ (higher entropy, more exploratory), while holding all other decoding parameters fixed across models. For AUT and DAT we use nucleus sampling with `top_p` $= 0.95$ and rely on the model's default `max_tokens`; for HP fanfiction we use the local generation helper with the same nucleus sampling and filter outputs to the 500–1000 token window.

For each (model, task, temperature) configuration, we collect $2{,}000$ generations after de-duplication. Initial generation pools were larger ($2{,}100$–$3{,}000$ depending on the model) and pruned to a uniform $2{,}000$ to keep per-model sample sizes comparable in all coverage calculations. The two temperatures and five models give 10 model–temperature configurations per task, used throughout the main paper's averages.

A.3. Human Reference Construction

The empirical human response boundary is constructed from a human reference pool whose size and provenance differ by task. For AUT we use $4{,}000$ human responses to the `rope` prompt sampled from publicly available divergent-thinking datasets. For DAT we use $1{,}000$ human responses. For HP fanfiction we use $2{,}000$ AO3 excerpts after token-length filtering (500–1000 tokens) and duplicate removal, drawn so that no single author or single fic dominates the pool.

For each task, the Human-to-Human reference is computed by a $50/50$ random split of the human pool with a fixed seed: one half defines the boundary, and the other half is treated as the held-out query set, evaluated against the boundary using the same procedure as for model outputs. This keeps the boundary and the query set at the same sample size.

B. Robustness of the Coverage Framework

The main paper relies on a small number of methodological choices: which sentence encoder is used, how dimensionality reduction is performed, what neighborhood radius defines the empirical human boundary, how the Human-to-Human reference is calibrated, and how many human and model samples are needed to estimate coverage reliably. This section reports robustness checks for each of these choices. To avoid confounding hyperparameter sensitivity with task-specific data artifacts, all robustness experiments are conducted on AUT, which has the largest human reference pool ($4{,}000$) and the most clearly defined response space. Across every setting we examine, the qualitative pattern reported in the main paper persists: model outputs are reliably in-boundary, but cover substantially less of the human distribution than the Human-to-Human reference.

B.1. Encoder, PCA, and Boundary Sensitivity

We test whether the main pattern is robust to three methodological choices: the sentence encoder used to embed responses, the PCA variance threshold used to project the embedding space, and the boundary percentile that controls neighborhood tightness. Each is a plausible target for reviewer concern, since alternative settings might in principle change the framework's verdict.

The main paper uses `sentence-transformers/all-mpnet-base-v2`, a 768-dimensional encoder that performs well on standard semantic similarity benchmarks and is widely used as an off-the-shelf representation in the sentence-transformers ecosystem. As a robustness check we additionally run `all-MiniLM-L6-v2`, a lighter-weight encoder from the same family with different dimensionality and training data. For PCA we sweep three variance thresholds covering the range typically reported in coverage work: 85%, 90% (the main-paper setting), and 95%. For the boundary percentile we sweep four values—$p25$, $p50$ (the main-paper setting), $p75$, and $p90$—to observe how coverage scales as the boundary moves from tight to permissive.

Table 5 reports task-averaged IBR and LLM-Cov across all ten model–temperature configurations under each setting, together with the per-configuration minimum and maximum LLM-Cov to indicate spread across models.

Encoder choice. Switching from MPNet to the smaller MiniLM model shifts both metrics modestly. IBR drops from 0.918 to 0.870 and LLM-Cov from 0.305 to 0.296. The relative gap between IBR and LLM-Cov remains large under both encoders, and the spread across model–temperature configurations is comparable. This suggests that the coverage gap we report is a property of model outputs rather than of the particular encoder used to compare them.

PCA variance threshold. Varying the retained-variance threshold from 85% to 95% changes the embedding subspace dimensionality from roughly 75 to 158 components but leaves the coverage results largely unchanged. The spread between

Table 5. **AUT robustness across encoder, PCA threshold, and boundary percentile.** Each row reports mean IBR and LLM-Cov across the ten model–temperature configurations, with the minimum and maximum LLM-Cov across configurations. The qualitative pattern (high IBR, substantially lower LLM-Cov) holds across every setting examined.

Configuration	IBR	LLM-Cov	Min Cov.	Max Cov.
Encoder (PCA 90%, $p50$)				
MiniLM	0.870	0.296	0.090	0.526
MPNet	0.918	0.305	0.074	0.553
PCA variance threshold (MPNet, $p50$)				
85%	0.924	0.307	0.076	0.569
90%	0.918	0.305	0.074	0.553
95%	0.891	0.297	0.058	0.557
Boundary percentile (MPNet, PCA 90%)				
$p25$	0.272	0.049	0.000	0.130
$p50$	0.918	0.305	0.074	0.553
$p75$	0.994	0.519	0.194	0.803
$p90$	0.999	0.680	0.336	0.901

configurations remains within ± 0.02 on both metrics. This shows that low-variance embedding directions, which a higher PCA threshold preserves, do not change the conclusion that LLM-Cov is much lower than the Human-to-Human reference.

Boundary percentile. The boundary percentile has the largest expected effect, since it directly controls how permissive the boundary is. As the percentile increases from $p25$ to $p90$, the radius ε grows and both IBR and LLM-Cov increase monotonically, as expected. The important observation is that across this entire range, the qualitative pattern in the main paper is preserved: model LLM-Cov remains substantially below the Human-to-Human reference at any fixed percentile (the H2H comparisons are in Appendix B.3). The framework's verdict is therefore not an artifact of the specific percentile chosen.

B.2. Human-to-Human Saturation

A natural concern about the Human-to-Human reference is that it might be inflated by a small reference pool, with every held-out human response trivially having a near-neighbor in a tightly clustered reference set. If true, this would weaken the main paper's central comparison.

We test this directly. For each pool size $n \in \{100, 250, 500, 1000, 2000\}$ on AUT, we sample n human responses as the reference pool and an equal-size held-out set as queries, then compute coverage using the same procedure as in the main paper. We repeat this five times per n with different random sub-samples and report mean and standard deviation in Table 6.

Table 6. **AUT Human-to-Human coverage saturation** under MPNet+PCA90 and the $p50$ boundary. As the reference pool grows, H2H coverage decreases monotonically and approaches the value reported in the main paper.

Reference n	H2H Cov.	SD	Replicates
100	1.000	0.000	5
250	0.994	0.002	5
500	0.981	0.004	5
1000	0.944	0.006	5
2000	0.908	0.000	5

Several things are visible in this table. First, at very small pool sizes ($n = 100$) the H2H reference is essentially saturated at 1.000; this is the regime where a small held-out human set is trivially covered by a tightly clustered reference set, and it is exactly the artifact a reviewer might worry about. Second, as n grows, H2H coverage decreases monotonically and the per-replicate variance shrinks. By $n = 2000$ the reference has stabilized at ≈ 0.908, the value reported in the main paper. This tells us the H2H number we compare against is not an artifact of under-sampling; it is the value the reference pool converges to as human variation is more fully represented. Model LLM-Cov, in contrast, is well below 0.908 at every model

and temperature, so the gap is not eliminated by a larger reference set.

B.3. Boundary Quantile Calibration

To make the relationship between boundary percentile and Human-to-Human coverage explicit, Table 7 reports H2H LLM-Cov on AUT across a fine grid of quantile choices. This calibration table serves two purposes. First, it lets readers mentally re-baseline our coverage numbers at any quantile of interest without re-running the full pipeline. Second, it confirms that the H2H curve is smooth and monotonic, so the boundary radius is not a discontinuity-inducing hyperparameter near the operating point.

Table 7. **Boundary quantile calibration on AUT** (MPNet, PCA 90%). The Human-to-Human reference increases smoothly with the boundary quantile, providing a continuous baseline against which model LLM-Cov can be compared.

Boundary quantile q	**AUT H2H Cov.**
0.10	0.377
0.20	0.559
0.25	0.622
0.30	0.672
0.40	0.814
0.50	0.908
0.60	0.959
0.70	0.979
0.75	0.987
0.80	0.992
0.90	0.997

At $q = 0.10$ the boundary is very tight and even matched human samples only cover 0.377 of each other; by $q = 0.90$ the boundary is so permissive that almost any held-out human response has a near-neighbor in the reference pool. The main paper's $p50$ choice sits in the informative middle of this curve, where the boundary is tight enough to differentiate models but permissive enough that human variation itself is well represented.

B.4. KNN k Sensitivity

The boundary radius ε is set as a quantile of the distance to each human response's k-th nearest human neighbor. The main paper uses $k = 15$, chosen to provide a stable notion of local semantic similarity without being dominated by near-duplicate responses. A natural question is whether the coverage pattern is sensitive to this choice—in particular, whether very small k (making the radius noisy and possibly too small) or very large k (making the radius reflect inter-cluster rather than within-cluster geometry) would change the main result.

To check this, we sweep $k \in \{5, 10, 15, 20, 30\}$ while holding the encoder, PCA variance threshold, and boundary quantile fixed at their main-paper values. Table 8 reports task-averaged IBR and LLM-Cov across the ten AUT model–temperature configurations at each k.

Table 8. **AUT robustness to the KNN parameter** k under MPNet+PCA90 and the $p50$ boundary. Both IBR and LLM-Cov increase monotonically with k, as expected, but the qualitative pattern is preserved throughout: at every k, mean LLM-Cov remains far below the AUT Human-to-Human reference of 0.908.

k	**Mean IBR**	**Mean LLM-Cov**
5	0.587	0.171
10	0.754	0.248
15	0.918	0.305
20	0.938	0.340
30	0.962	0.388

Two observations from this sweep are worth emphasizing. First, the expected directionality holds: small k produces tighter radii, making both IBR and LLM-Cov lower, while larger k relaxes both. Second, even at $k = 30$—substantially looser than

the main-paper setting—the mean LLM-Cov of 0.388 remains well below the H2H reference of 0.908. In other words, no choice of k in this range brings model coverage close to the human-to-human ceiling. The coverage gap is therefore not a function of the KNN neighborhood size, and the main-paper choice of $k = 15$ sits in a stable region of the curve where IBR has already approached its plateau but LLM-Cov is still far from saturating.

B.5. Model Output Size Sensitivity

A complementary concern is whether the LLM-Cov values we report reflect a true narrowness of the model distribution or simply insufficient sampling of model outputs. If we sample only a small number of model generations, even a broad-distribution model would appear to cover little of the human space. We test this directly by subsampling each model–temperature output set on AUT.

For each subsample size $n_M \in \{500, 1000, 1500, 2000\}$ we draw 10 random subsamples per model–temperature configuration without replacement and compute IBR and LLM-Cov using the same procedure as in the main paper. Table 9 reports the mean across all subsamples and all ten model–temperature configurations.

Table 9. **AUT robustness to model output sample size** under MPNet+PCA90 and the $p50$ boundary. LLM-Cov increases sublinearly with the number of model outputs and is saturating well below the human-to-human reference; IBR is essentially flat, as expected for a property of individual outputs rather than of the output set as a whole.

Model outputs n_M	**Mean IBR**	**Mean LLM-Cov**
500	0.918	0.205
1000	0.916	0.239
1500	0.916	0.258
2000	0.917	0.269

Two patterns are visible. First, IBR is essentially flat across all subsample sizes (0.916–0.918), which is expected: IBR is the average per-output property of remaining in-boundary, so subsampling estimates it directly and converges quickly. Second, LLM-Cov rises from 0.205 at $n_M = 500$ to 0.269 at $n_M = 2000$, but the marginal gain shrinks rapidly: doubling from 1000 to 2000 outputs only increases LLM-Cov from 0.239 to 0.269. Extrapolating from this trajectory, the increase from 2000 outputs to a much larger sample would not bring LLM-Cov anywhere close to the H2H reference of 0.908. The LLM-Cov values reported in the main paper therefore do not understate model breadth in any meaningful way; the coverage gap reflects a genuine narrowness of the model output distribution rather than under-sampling of model generations.

C. HP Fanfiction Covered-vs-Uncovered Diagnostics

The main paper presents a focused covered-vs-uncovered comparison on HP fanfiction. This appendix expands that analysis along three dimensions: it reports the full set of features we measured (including non-significant ones omitted from the main paper), documents how each feature is operationalized, and shows that the covered-vs-uncovered distinction is reliably learnable from these features—confirming that the coverage gap corresponds to real, structured differences in human cultural production rather than to embedding noise.

C.1. Setup

We define COVERED and UNCOVERED groups relative to `Mistral-Small-24B-Instruct-2501` at t=1.0 on the HP fanfiction $p50$ boundary in the MPNet+PCA space. A human excerpt is COVERED if at least one model output from this configuration lies within ε of it, and UNCOVERED otherwise. Under this definition, of the $2{,}000$ human excerpts in the held-out reference pool, $1{,}211$ (60.6%) are COVERED and 789 (39.4%) are UNCOVERED. The size of the UNCOVERED group means the analysis is reasonably well powered, with hundreds of observations per group for both binary and continuous features.

We focus on Mistral-24B at $t = 1.0$ because it is the highest-coverage single configuration on HP fanfiction, which makes the covered-vs-uncovered split a stringent test: features that distinguish the two groups under this model are unlikely to be artifacts of an exceptionally narrow generator.

C.2. Full Variable Comparison

Table 10 reports all 19 features we measure, organized into three axes: canonical anchoring, style and rhythm, and relationship orientation. Continuous variables report group means with Mann–Whitney U p-values; binary indicators report group rates with chi-square p-values. The Δ column gives the covered-minus-uncovered difference, so positive values mean a feature is more common in covered excerpts.

Table 10. **Full covered-vs-uncovered comparison** for HP fanfiction. Δ is Covered minus Uncovered: positive values indicate features more common in covered excerpts, and negative values indicate features more common in uncovered excerpts. Significant features (at $p < 0.001$) appear in every group, indicating that the coverage gap is structured rather than random.

Variable	Cov.	Uncov.	Δ	p
Canonical anchoring				
HP setting present (rate)	0.538	0.369	+0.169	2.1×10^{-13}
HP setting / 1k words (mean)	2.36	1.38	+0.98	1.4×10^{-15}
Canon vocab present (rate)	0.553	0.401	+0.153	3.3×10^{-11}
Canon vocab / 1k words (mean)	2.48	1.72	+0.76	7.4×10^{-12}
HP character present (rate)	0.893	0.817	+0.075	2.5×10^{-6}
Non-HP characters (rate)	0.152	0.228	−0.076	2.1×10^{-5}
Implicit HP-world (rate)	0.990	0.939	+0.051	1.6×10^{-10}
Style and rhythm				
Mean sentence length (mean)	15.47	17.08	−1.61	3.6×10^{-6}
Sentence-length variance (mean)	125.1	350.0	−224.9	8.5×10^{-6}
Dialogue ratio (mean)	0.388	0.359	+0.029	0.003
Quoted spans (mean)	18.91	17.04	+1.87	2.3×10^{-4}
Punctuation density (mean)	0.174	0.168	+0.006	2.5×10^{-5}
Relationship orientation				
Romance present (rate)	0.703	0.759	−0.056	0.007
Family present (rate)	0.382	0.360	+0.022	0.354
Friendship present (rate)	0.339	0.327	+0.012	0.626
Ship tension present (rate)	0.499	0.535	−0.036	0.125
M/M relationship (rate)	0.258	0.407	−0.149	3.3×10^{-12}
M/F relationship (rate)	0.394	0.284	+0.110	6.0×10^{-7}
F/F relationship (rate)	0.047	0.066	−0.019	0.087

The full table makes two points beyond what the main paper highlights. First, the canonical-anchoring axis is reinforced by features the main paper does not show, including "implicit HP-world knowledge" (the rate at which an excerpt presupposes the HP setting without exposition), which is 0.990 in covered versus 0.939 in uncovered. Models are slightly less likely to reach excerpts that gesture at the HP world implicitly rather than naming it. Second, several relationship features are statistically null, including *family*, *friendship*, and *ship tension*. This null pattern is informative: it tells us that the covered-vs-uncovered difference along the relationship axis is not about whether relationships appear in general, but specifically about gender configuration of romantic pairings. M/M is over-represented in uncovered excerpts; M/F is over-represented in covered ones. F/F is too rare in the data to show a reliable difference.

C.3. Variable Operationalization

Table 11 documents how each variable group is operationalized. The canonical-anchoring and style features are deterministic, computed by rule-based matchers and text statistics applied uniformly across both groups. The relationship and fandom-internal features are LLM-judge binary annotations obtained by prompting an instruction-tuned model with each excerpt and a fixed classification prompt; we treat these annotations as approximate labels rather than ground truth, and the classifier diagnostics in Appendix C.5 report results both with and without LLM-judge features to isolate their contribution.

The deterministic features are independent of any model in the loop and can be reproduced exactly given the keyword lists, which we include in the released code. The LLM-judge features are stochastic in principle, but in practice produce stable group-level rates across re-runs because the binary classification questions (e.g., "is this romance?") are coarse enough to be reliable even when individual labels are noisy.

Table 11. **Variable operationalization** for the HP covered-vs-uncovered diagnostics.

Group	Variables	Operationalization
Canonical anchoring	`setting_present`, `setting_count`, `setting_per_1k`	Rule-based match against canonical HP setting terms, such as Hogwarts, Gryffindor, Slytherin, Diagon Alley, Hogsmeade, Ministry, Grimmauld Place, Privet Drive, Burrow, and Azkaban.
Canon vocabulary	`canon_vocab_present`, `canon_vocab_count`, `canon_vocab_per_1k`	Rule-based match against HP-world terms, such as wand, spell, potion, Quidditch, Auror, Death Eater, Horcrux, Patronus, Animagus, Muggle, pureblood, and Order.
Characters	`character_present`	Rule-based match against an enumerated list of HP character names and surnames.
Style and rhythm	`word_count`, `sentence_count`, `mean_sentence_len`, `sentence_len_var`, `fragment_ratio`, `dialogue_ratio`, `quote_span_count`, `punctuation_density`, `ellipsis_density`	Deterministic text statistics from sentence segmentation, word tokenization, quoted spans, punctuation counts, ellipses, and sentence-length distributions.
Relationship orientation	`relationship_type_romance`, `relationship_type_family`, `relationship_type_friendship`, `relationship_gender_mm`, `relationship_gender_mf`, `relationship_gender_ff`, `ship_tension_present`	LLM-judge binary annotations on the 2,000 human excerpts.
Fandom-internal signals	`characters_outside_hp`, `hp_world_implicit`	LLM-judge binary annotations for non-HP characters and for whether the text presupposes the HP world without explicit exposition.
Affect and narrative function	`comfort_*`, `angst_*`, `subversion_*`	Rule-based keyword proxies for affective and narrative-function cues.

C.4. Univariate Logistic Regression

To quantify the per-feature contribution to the covered-vs-uncovered distinction, Table 12 reports univariate logistic regression coefficients predicting UNCOVERED status from each standardized feature. Positive β indicates that higher feature values increase the probability of being uncovered; negative indicates the opposite. The odds ratio per one standard deviation gives the effect size in a more interpretable scale, and in-sample AUC indicates how predictive each feature is on its own. We report the 12 most significant variables.

Two observations from the univariate table are worth noting. First, the strongest single predictors are roughly evenly split between canonical-anchoring features (HP setting, canon vocab, type-token ratio, implicit HP-world) and relationship-orientation features (M/M, M/F). This balance suggests that neither axis is dominating the covered-vs-uncovered split: both surface canonicity and community-internal relationship convention contribute independently. Second, no single feature is highly predictive on its own—in-sample AUCs cluster between 0.52 and 0.60. This is expected: the covered-vs-uncovered distinction is multidimensional, and only combinations of features (Appendix C.5) recover stronger predictive performance.

C.5. Classifier Diagnostics

A potential objection to the covered-vs-uncovered analysis is that the two groups might differ only in irreducible embedding noise rather than in interpretable cultural features. We address this by asking whether the binary covered-vs-uncovered label is recoverable from human-interpretable features alone. If interpretable features can predict coverage status well above chance, the coverage gap reflects structured cultural differences rather than random embedding artifacts.

We fit logistic regression classifiers under six feature sets (Table 13) and report 5-fold cross-validated AUC in Table 14.

Several patterns are visible. First, every feature set predicts UNCOVERED status well above chance, with cross-validated AUC ranging from 0.616 (LLM-judge features alone) to 0.740 (combined expanded). This is direct evidence that the coverage gap corresponds to structured differences in interpretable features, not random embedding noise. Second, sentence-

Table 12. **Univariate logistic regression coefficients predicting UNCOVERED status.** Features are standardized to unit variance. OR/SD is the odds ratio per one standard deviation, and AUC is in-sample. The strongest single features are HP setting presence, type-token ratio, and M/M relationship orientation, all consistent with the canonical-anchoring and relationship axes highlighted in the main paper.

Variable	β	OR/SD	AUC	p
HP setting present	-0.343	0.71	0.584	2.0×10^{-13}
Type-token ratio	-0.346	0.71	0.597	4.4×10^{-13}
M/M relationship	$+0.317$	1.37	0.575	3.3×10^{-12}
Canon vocab present	-0.309	0.73	0.576	3.0×10^{-11}
Implicit HP-world	-0.319	0.73	0.525	1.0×10^{-8}
M/F relationship	-0.236	0.79	0.555	5.2×10^{-7}
HP character present	-0.213	0.81	0.538	2.3×10^{-6}
Non-HP characters	$+0.193$	1.21	0.538	1.8×10^{-5}
Dialogue ratio	-0.153	0.86	0.539	8.6×10^{-4}
Quoted-span count	-0.150	0.86	0.549	1.4×10^{-3}
Punctuation density	-0.136	0.87	0.556	4.4×10^{-3}
Romance present	$+0.129$	1.14	0.528	5.8×10^{-3}

Table 13. **Feature sets used in classifier diagnostics.**

Feature set	Description
LLM-JUDGE ONLY	The 9 LLM-judge binary annotations: relationship types, gender configurations, ship tension, non-HP characters, and implicit HP-world.
DETERMINISTIC ONLY	Rule-based features only: settings, canon vocabulary, character mentions, and deterministic text statistics.
SENTENCE/LEXICAL ONLY	Sentence- and word-level statistics, including sentence lengths, dialogue ratio, quote spans, punctuation, and type-token ratio.
COMBINED (PAPER-FRIENDLY)	LLM-judge annotations plus the rule-based features used in the main paper.
COMBINED (FULL)	All deterministic and all LLM-judge features.
COMBINED EXPANDED	All features above plus auxiliary affect and narrative-function proxies.

and lexical-level features alone reach $\mathrm{CV\ AUC} = 0.716$, demonstrating that purely surface stylistic differences are already substantially predictive of coverage status, independent of LLM-judge annotations. Third, combining LLM-judge features with rule-based ones gives only a modest improvement over either alone, suggesting that relationship-orientation features carry some signal not captured by canonical-anchoring rules but that the two feature families are partly redundant.

The headline number for the paper's conceptual claim is the $\mathrm{CV\ AUC} = 0.740$ achieved by the full feature set: covered-vs-uncovered status is approximately three-quarters recoverable from interpretable cultural features, confirming that the coverage gap reflects substantive cultural differences in human production rather than artifacts of the embedding geometry.

Table 14. **Classifier diagnostics for predicting UNCOVERED status.** $n = 2{,}000$. CV AUC is mean $\pm$ standard deviation across 5-fold cross-validation. The chance AUC baseline is 0.500. Every feature set exceeds chance, and the combined feature set reaches CV AUC $= 0.740$, indicating that the coverage gap is structured along interpretable cultural features rather than random.

Feature set	**In-sample AUC**	**CV AUC**
LLM-JUDGE ONLY	0.631	0.616 ± 0.018
DETERMINISTIC ONLY	0.678	0.667 ± 0.009
SENTENCE/LEXICAL ONLY	0.726	0.716 ± 0.020
COMBINED (PAPER-FRIENDLY)	0.708	0.693 ± 0.015
COMBINED (FULL)	0.710	0.693 ± 0.011
COMBINED EXPANDED	0.766	0.740 ± 0.011