



GENIE: A Fine-Grained Measure for Novelty

Ramya Namuduri¹ Manya Wadhwa² Anshun Asher Zheng¹

Greg Durrett² Junyi Jessy Li¹

¹The University of Texas at Austin ²New York University
ramya.namuduri@utexas.edu



Code



Data



Project Page

Abstract

Large Language Models have consistently demonstrated a lack of creativity and diversity across tasks. Prior work has focused on addressing *whether* models are capable of generating creative outputs. Here, we aim to consider *novelty* and investigate what makes model-generated content novel or not novel in a task-specific manner. We propose a fine-grained evaluation metric GENIE to measure the novelty of responses along task-specific *features* with respect to a population of responses. We show that unlike GENIE, holistic metrics struggle to capture the high-dimensionality of novelty and do not provide insight on which properties they target. Finally, we use GENIE to measure the effectiveness of mitigation methods that address creativity to better understand where these methods can improve novelty.

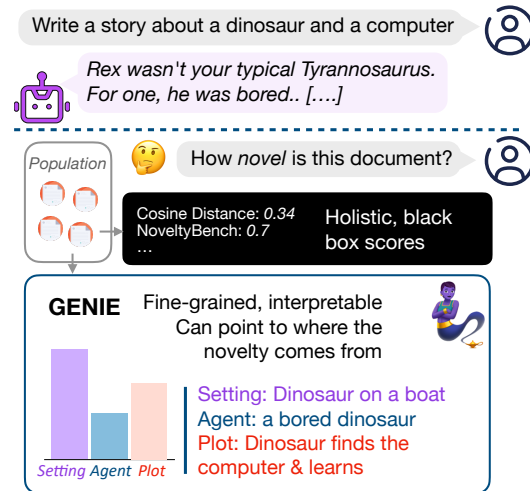


Figure 1: High level motivation for GENIE. Existing holistic metrics do not capture the high-dimensionality of novelty while GENIE’s fine-grained approach provides explainability.

1 Introduction

Can LLMs produce creative work? Prior work has discovered that models struggle to generate diverse and creative outputs given various tasks (Jiang et al., 2026), using lexical-semantic approaches (Zhang et al., 2025; Lu et al., 2025; Ismayilzada et al., 2025), creativity tests (Wenger and Kenett, 2025; Wadhwa et al., 2026) and human experts (Chakrabarty et al., 2024). Model creativity has been tested on a variety of tasks using a multitude of metrics. However, past work has not isolated precisely *where models lack creativity* on specific tasks.

Novelty is a key component of creativity (Boden, 1991), where the uniqueness of a response can be attributed to different reasons specific to the task. Features such as *plot* and *character development* are relevant to narrative writing, while *motivation* and *methodology* are applicable to scientific discovery. Two responses can be strikingly original, and yet differ vastly in what makes them novel. Similarly, a response might appear novel in some ways and commonplace in others. For instance, Figure 1 illustrates how an LLM-generated response can concurrently have unique and mundane features. However, this fine-grained novelty is not captured by many existing holistic creativity metrics. Cosine distance against other responses, and other creativity metrics (Zhang et al., 2025; Chakrabarty et al., 2025a; Fein

et al., 2026; Shaib et al., 2024), although informative, do not describe why an instance is more or less novel.

We propose GENIE (Granular Evaluation of Novel Ideas with Explainability), a framework that quantifies the fine-grained nuances that differentiate equally novel responses by extracting *features* present in the responses, with respect to a pre-defined population. Inspired by alternative semantics of questions literature (Hamblin, 1957; Karttunen, 1977; Lahiri, 2001), we do so by generating and mapping questions to automatically discovered features relevant to the task, and measuring fine-grained novelty through the similarities of their answers.

GENIE’s focus on automatically extracted *features* isolates novelty from other components of creativity, such as value (i.e., quality; Boden (1991)). This is in contrast to some existing metrics that subsume the quality of responses (Chakrabarty et al., 2025a) or the predicted aesthetic value (Fein et al., 2026). By tethering the feature extraction to the *task*, GENIE’s approach is inherently domain agnostic.

GENIE’s level of abstraction (using questions and answers) allows it to treat models of varying instruction-following capabilities comparably, specifically with regards to models with poor instruction-following capabilities that exhibit greater diversity despite lower quality

(West and Potts, 2025; Le Bronnec et al., 2024; Hamilton, 2024).

Furthermore, determining the uniqueness of responses holistically without a reference point can be subjective. Prior work has used the pretraining corpus (Lu et al., 2025), other model-generated (Jiang et al., 2026), human (Wenger and Kenett, 2025) or self-generated (Zhang et al., 2026, 2025) responses as reference points for their measurements. In this work, we make the references explicit: GENIE measures novelty *with respect to a population*; this population is defined by the user.

We explore the creative writing domain. GENIE independently found 7 fine-grained features pertaining to this task, which roughly correspond to theoretically defined taxonomies (Hamilton et al., 2026). We constructed and used a population of responses generated by 21 LLMs to evaluate the *sensitivity* of GENIE on pairs of LLM-generated responses that minimally differ along these features. GENIE successfully detects differences in fine-grained novelty across such pairs. GENIE also demonstrates *robustness* on pairs that only differ via paraphrasing; it does *not* detect a significant difference along any feature in these cases. Additionally, we performed human studies to intrinsically evaluate various components that comprise GENIE.

GENIE’s fine-grained nature provides a diagnostic lens into existing holistic metrics used to measure novelty, diversity and creativity. We found that they either capture extraneous properties that sometimes spuriously correlate with novelty, or are lexically sensitive. Furthermore, many existing mitigation methods, such as alternate prompting (Zhang et al., 2026; Huot et al., 2025) and decoding strategies (Peeperkorn et al., 2025) for improving creativity have been evaluated on these holistic metrics. We demonstrate how GENIE provides a more explainable evaluation, shedding light into *which* features, if any, these methods actually improve.

2 Related Work

A growing line of work examines whether LLMs are creative, drawing on theoretical frameworks from cognitive science and computational creativity. Many studies operationalize Boden’s theories of creativity (Franceschelli and Musolesi, 2024; Ismayilzada et al., 2025; Schapiro et al., 2025; Nagarajan et al., 2025; Wadhwa et al., 2026) while others adapt psychometric tests such as Torrance Tests of Creative Thinking (Chakrabarty et al., 2024). Past work has also focused on the homogenizing effects of LLMs when looking at a population of LLM outputs (Wenger and Kenett, 2025; Jiang et al., 2026) and attributed this to post-training alignment methods (Padmakumar and He, 2024; West and Potts, 2025; Slocum et al., 2025). There are also domain specific benchmarks that evaluate creativity in writing (Fein et al., 2026; Chakrabarty et al., 2025a; Xu et al., 2025; Fein et al., 2026; Zhang et al., 2025). Most of this work asks whether models are creative; instead we ask where a given response is novel and decompose this judgment

along task-specific features.

Recent work has proposed explicit metrics and benchmarks for evaluating novelty and creativity in text generation. One common approach is to use the pretraining corpus as reference to quantify originality as a fraction of n-grams (Lu et al., 2025; Padmakumar et al., 2026; Merrill et al., 2024). NoveltyBench (Zhang et al., 2025) operationalizes diversity by checking for if an output belongs to one of the preexisting clusters of unique responses. Past work also trains reward models on human preference data to score creative writing directly (Chakrabarty et al., 2025a; Fein et al., 2026; Cao et al., 2026). Across all the prior work, novelty is reduced to a scalar where the reference is implicit or calculated with surface units (e.g., n-grams or distributional semantics). QUDSIM (Namuduri et al., 2025) decomposes similarity along discourse-level questions under discussion (QUDs) (Roberts, 2012; Velleman and Beaver, 2016; Beaver et al., 2017). This established the effectiveness of measuring the similarity of answers to the same questions across documents. However, QUDs capture discourse, whereas GENIE decomposes novelty into task-level features against an explicit population, separating content from quality and aesthetics of the output being evaluated.

3 The GENIE Measure of Novelty

Consider a population of model-generated documents $\mathcal{D}_{pop} = \{d_1, \dots, d_m\}$ generated from prompt P . The novelty of a target document $d_t \notin \mathcal{D}_{pop}$ is defined as its relative dissimilarity d for all $d \in \mathcal{D}_{pop}$ along various *features*. Intuitively, unique responses share fewer commonalities across features with other responses. GENIE is explainable, because each feature is associated with natural language definitions; when unpacked, further explainability can be obtained by pinpointing prompt-specific questions under that feature, as well as their natural language answers whose similarity we calculate.

3.1 Features

Features are task-specific axes that aim to maximize coverage over different characteristics of the task. For instance, a task like creative writing might include features such as *plot*, *setting*, and *perspective*. On the other hand, *motivation* and *methodology* might be relevant to evaluating the novelty of scientific discovery. Manually defining these features for tasks can be tedious, subjective, or infeasible (due to lack of expertise). Instead, we automatically *discover* features.

We prompt an LLM to form features with descriptions given a task (e.g. creative writing) k times. The resulting set is clustered to find commonly recurring features. Concretely, we discover N clusters and label them as $\mathcal{F} = \{f_1, \dots, f_N\}$ to form the set of *features*. Step ❶ and ❷ in Figure 2 demonstrate how we discover features for the creative writing task. Each feature f is a function $f : \mathcal{D} \rightarrow \mathcal{V}_f$ that maps a document to its feature value, where $\mathcal{D}$ the set of all documents and $\mathcal{V}_f$ is the set of all

Prompt: Write a story about a dinosaur and a computer

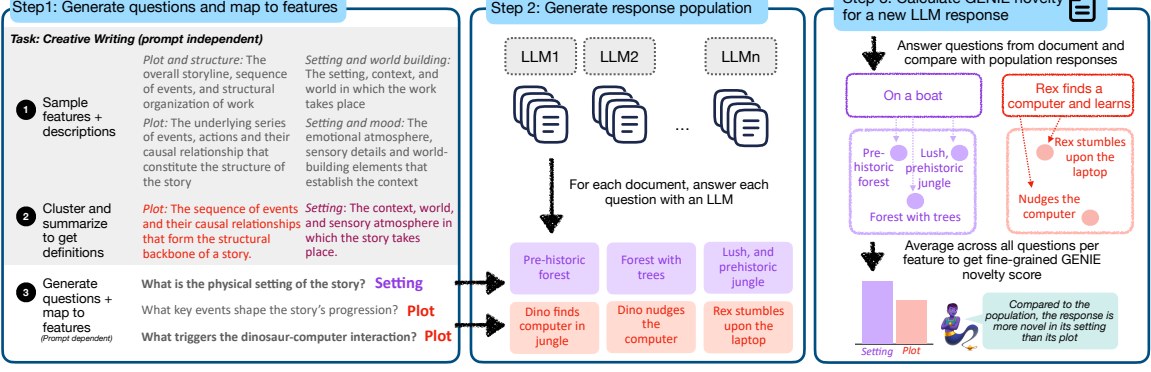


Figure 2: GENIE captures fine-grained novelty of a target response with respect to a population of responses using automatically discovered task-specific features. Here, GENIE discovers features and their descriptions for creative writing. We generate questions for a given prompt p (Step 3) and map each question to a feature. Step 2 computes answers to questions from the population. Finally, GENIE measures the novelty of a target response with respect to this population, along each feature (e.g. setting or plot) using the average pairwise answer dissimilarity. A complete example is provided in Appendix D.

possible values the feature can take.

3.2 GENIE Scoring

GENIE elicits values for each feature $f \in \mathcal{F}$ in a document (i.e., $f(d)$) via question generation and answering. Here, question generation is an intermediate process that grounds $\mathcal{F}$ to prompt-specific dimensions, allowing feature values to be more reliably extracted by answering the questions. This paradigm takes the alternative semantics view of questions (Hamblin, 1957; Karttunen, 1977; Lahiri, 2001), where hierarchically organized questions are used to partition the possible worlds into subspaces that “answer the same questions”, in our case, “values of the same feature f ”, hence allowing us to measure novelty for each feature.

Prompt-Specific Decomposition Given a prompt $p \in P$, we generate a set of questions $\mathcal{Q} = \{q_1, \dots, q_j\}$ (see Step 3 in Figure 2). The set $\mathcal{Q}$ is formed without knowledge of $\mathcal{F}$ so that the resulting questions are specific to p , while not overfitting to the taxonomy of discovered features.

Feature Mapping We then map each question in $\mathcal{Q}$ to exactly one feature $f \in \mathcal{F}$ (Step 3 in Figure 2). The subset of questions that map to feature f is denoted by $\mathcal{Q}_f \subseteq \mathcal{Q}$. We can represent the feature value $f(d) \in \mathcal{V}_f$ as the set of answers to $\mathcal{Q}_f$ for document d (see Step 2 in Figure 2).

GENIE The GENIE novelty of d_t is defined as a **vector** that consists of the individual novelty scores along each feature $f \in \mathcal{F}$:

$$\mathbf{G}(d_t) = [\mathbf{G}_{\text{feat}}(d_t, f)]_{f \in \mathcal{F}}$$

The novelty of the target document along f is defined as the average novelty across all questions $q \in \mathcal{Q}_f$:

$$\mathbf{G}_{\text{feat}}(d_t, f) = \frac{1}{|\mathcal{Q}_f|} \sum_{q_i \in \mathcal{Q}_f} \mathbf{G}_{\text{quest}}(d_t, q_i)$$

Where $\mathbf{G}_{\text{quest}}(d_t, q_i)$ (i.e., the novelty of d_t along q) is defined as the average pairwise dissimilarity ($1 - \text{similarity}$) between the target answer a and the corresponding answers in each population document (see Step 3 in Figure 2). Let a_{d_t} and a_{d_p} be answers to question q found in target document d_t and population document $d_p \in \mathcal{D}_{\text{pop}}$, respectively.

$$\mathbf{G}_{\text{quest}}(d_t, q) = \frac{1}{|\mathcal{D}_{\text{pop}}|} \sum_{d_p \in \mathcal{D}_{\text{pop}}} \text{dissim}(a_{d_t}, a_{d_p})$$

Intuitively, d_t will get a higher GENIE score on feature f if that feature is distinctive compared to how f is realized among the population documents. In the example shown in Figure 2, d_t receives a higher score on *setting* than *plot* due to its higher dissimilarity with the population of *settings*. We keep GENIE as a vector rather than an aggregate score to preserve fine-grained explainability. Additionally, the novelty of a response can be independently attributed to multiple features.

4 GENIE for Creative Writing

In this paper, we instantiate GENIE on the creative writing task, in particular long-form story generation. We choose this task for several reasons: (1) it is known to be multi-faceted with fine-grained features such as *plot*, *setting*, etc. that are researched in prior work (Hamilton et al., 2026; Russell et al., 2026). These existing features can validate the features that we automatically derived. (2) Even though LLMs can generate coherent text, long-form stories generated by LLMs are known to be uncreative or homogeneous. Thus, it is a particularly valuable test bed for fine-grained, explainable insights into where the (lack of) novelty appears, especially as existing metrics do not currently do so.

4.1 Tasks

We curate a total of ~10,000 documents using 39 models, across 50 open-ended writing prompts (P) gathered

from two sources: (1) 25 synthetic prompts generated by gpt-4.1-mini following the format *Write a story about X and Y* where *X* and *Y* are spatially or temporally distant concepts (Prompt F.1) and (2) 25 prompts sampled from the LitBench (Fein et al., 2026) subset of the r/WritingPrompts dataset (Fan et al., 2018). Examples of the prompts in our dataset are in Appendix A.4. A power analysis reveals the number of prompts is a sufficiently large sample size for our experiments to achieve statistical significance (Appendix A.1).

4.2 Features

We gathered features describing the creative writing task from prompting gpt-4.1 10 times (Prompt F.2) eliciting a definition for each, and clustered the responses using embeddings computed with text-embedding-3-small.¹ The features discovered include: SETTING, PLOT, CHARACTERIZATION, MOOD/TONE, POINT OF VIEW, LANGUAGE, THEME, which closely correspond to the taxonomy described in Hamilton et al. (2026). We therefore adopt their terminology and map our discovered features accordingly. In this work, we specifically focus on AGENT, PERSPECTIVE, PLOT, SETTING, SOCIAL ATMOSPHERE,² and STYLE to form the set $\mathcal{F}$. Details on how the features and definitions we discovered were mapped are included in Appendix A.2.

4.3 Question Generation

For each prompt $p \in P$, we (a) generate the set of questions $\mathcal{Q}$ (§3.2), (b) perform filtering, and (c) map each question to the most representative feature in $\mathcal{F}$. Each stage of the pipeline uses gpt-4.1-mini. The prompts are in Appendix F.2.

Filtering Questions undergo a one-step filtering stage to exclude questions that are polar, hypothetical, conditional, subjective or those that address multiple features. Since the goal of using questions is to extract features and view them as independently as possible (i.e. the novelty of SETTING is not dependent or influenced by the novelty of STYLE).

Feature Mapping We then map $\mathcal{Q}$ to $\mathcal{F}$ as described in §3.2. This results in $\mathcal{Q}_f$ for all six creative writing features. Table 1 shows statistics on the distribution of questions generated. AGENT and PLOT are the largest features in question coverage, owing to the richness of the content that can be considered as either the storyline or character development. PLOT requires more questions to cover the conflict, turning points, obstacles, plotline, etc. On the other hand, PERSPECTIVE questions are largely limited to extracting POINT OF VIEW related information.

¹HBDSCAN, Agglomerative and K-means clustering resulted in similar clusters; we set k to the average number of features in each sample.

²We renamed *Social Network* as used in NarraBench to *Social Atmosphere* due to the former’s loaded meaning.

Feature	Count	Avg. per Prompt
AGENT	254	4.23
PERSPECTIVE	79	1.61
PLOT	369	6.15
SETTING	140	2.86
SOC. ATM.	82	1.82
STYLE	108	2.00
Total	1032	17.20

Table 1: Distribution of questions generated per feature and per prompt. PLOT and AGENT features are represented by more questions, per prompt, due to their higher complexity relative to PERSPECTIVE.

4.4 Constructing the Population and Target Set

Models Using prompts from Section 4.1, we build a *population* consisting of 21 models ranging across various families, sizes and stages of training (Table 17). We consider 18 *target* models (Table 18). The target models roughly correspond to newer iterations of the population models so that GENIE can reflect novelty differences as models evolve.

Responses For each prompt, we sample 5 responses from every model, across the population and target models (see Step 2 in Figure 2). All responses are post-processed to truncate incoherent or irrelevant responses, and reasoning. We then use ArmoRM (Wang et al., 2024) to compute coherence and further discard documents that do not meet a minimum threshold of 2.0.³

The **target** set for which we calculate novelty consists of 4500 documents. We use a pilot dataset and conduct a power analysis to determine the statistically significant population size required to reach 98% confidence with a 0.1 margin of error. The resulting **population** contains 3404 documents. Details about the sampling process are in Appendix A.1.

4.5 Dissimilarity Function

To calculate $\text{dissim}(a_{d_t}, a_{d_p})$, we judge similarity using Qwen2.5-32B-Instruct on a 4-point Likert scale using the rubric shown in G.3 and compute $1 - \text{similarity}$. Two answers are considered highly similar if they are interchangeable, despite slightly different wording. On the other hand, dissimilar answers have very low overlap in the details described, and sometimes can contradict each other. Pairs that are difficult to judge often have overlapping content or scope. Some questions are unanswerable by certain documents (“unspecified”, “not applicable”, etc.). We exclude these answers to prevent GENIE scores from rewarding unanswerability.

5 Evaluating GENIE

We intrinsically evaluate components of GENIE, such as question generation, feature mapping and the dissimilarity function, with the help of human annotators

³Completely incoherent documents can have artificially high novelty and thus must be filtered out. This threshold is chosen after manual inspection on a sample of documents.

Feature Mapping Model Accuracy		
Task	LLM v. Ann1	LLM v. Ann2
AGENT	0.67	1.00
PERSPECTIVE	0.73	0.92
PLOT	0.81	0.90
SETTING	0.62	0.67
SOC. ATM.	0.83	1.00
STYLE	0.77	0.83

Table 2: Model (gpt-4.1-mini) accuracy for question-to-feature mapping.

(recruitment details are provided in C.2). We then conduct a study to measure GENIE’s *sensitivity* to specific features and its *robustness* to surface-level variation.

5.1 Intrinsic Validation

Question Generation We run a small scale human study with two annotators to evaluate the reliability of using an LLM for generating questions given a prompt (§3.2). The annotators are give 50 questions and are asked to evaluate the validity of each question using the criteria given in Appendix G.1. The inter-annotator agreement (Cohen’s kappa) is 0.59 (moderately high, Artstein and Poesio (2008)). Some common source of disagreements are listed in Appendix C.1. The question generator gpt-4.1-mini achieves an F1 score of 0.71 and 0.74 as evaluated by two annotators.

Feature Mapping Since reliable question-to-feature mapping is essential for feature coverage and low noise, we validate the mapping step in §3.2. Two annotators classified a set of 50 questions into one of the six features, or none, as given in §4 using the criteria in Appendix G.2. Cohen’s kappa was high at 0.75, and as shown in Table 2, gpt-4.1-mini achieves high F1 scores across all features against both annotators.

Similarity To validate the LLM dissimilarity function, we asked human annotators to judge the similarity of 3,000 randomly sampled pairs of answers according to guidelines in Appendix G.3. The annotation set is selected to maximize coverage across features and target-population model pairs. Each pair of answers is annotated by two annotators. Table 11 reports per-feature agreement between annotators, as well as between annotators and Qwen2.5-32B-Instruct (the LLM used for distance judgments). Measured with Krippendorff- α , inter-annotator agreement is 0.59, and higher 0.66 for annotators vs. LLMs.

5.2 GENIE Sensitivity to Feature Edits

We conduct a causal intervention to investigate the *sensitivity* of GENIE to differences between responses along specific features. **Sensitivity** is defined as the ability to detect a statistically significant change in GENIE’s observed novelty for feature f when an intervention is performed along f . **Robustness** refers to statistical equivalence in the observed novelty for feature f when a lexical edit, such as paraphrasing, is performed. An ideal metric for measuring content-level novelty detects

all interventions (high sensitivity), while not responding to paraphrases (high robustness).

To test for sensitivity, we surgically edit documents along individual features using population-agnostic synthetic alternatives, while maximally preserving other features. To test for robustness, we paraphrase documents in a semantics-preserving manner, altering surface form while preserving all content features.

5.2.1 Experimental Setup

Feature Intervention We first define an edit function $E : \mathcal{D} \times \mathcal{F} \times \mathcal{V} \rightarrow \mathcal{D}$ that produces a minimally edited document $d' = E(d, f_i, \tilde{v}_{f_i})$, where $f_i \in \mathcal{F}$ is the feature being intervened upon and $\tilde{v}_{f_i}$ is the new proposed value of f_i . The original document d and the edited document d' differ along the value of f_i while maximally preserving all else. Concretely, we synthetically generate an alternate value $\tilde{v}_{f_i}$ for the target feature using an LLM, conditioned on d and the prompt p but independent of any population (Prompt F.9); d' is then the rewrite of d that reflects $\tilde{v}_{f_i}$, with all *other* feature values held fixed (i.e., $\forall f \neq f_i : f(d) = f(d')$). This helps ensure that there is a meaningful difference along f_i between d and d' without a predefined notion of relative novelty. We sample alternate features using gpt-5.4-mini (Prompt F.9) and implement the edits using gpt-4.1-mini (Prompt F.10).⁴ Examples of alternate features are provided in Appendix B.1.

Dataset We randomly sample 100 instruct-model generated documents from the target dataset. For each document and along each feature, we generate 3 alternatives and perform edits to produce 3 edited documents. This results in 1800 total pairs of original and edited documents, of which we filtered out 206 due to malformed responses (details provided in Appendix B.3).

Paraphrasing Documents We form 3 paraphrases per document, resulting in 300 additional pairs. Semantically equivalent paraphrasing should not alter feature values. Thus, a metric that finds a significant difference between paraphrased texts may be capturing other properties (lexical, aesthetic qualities, etc.).

Intervention Validation We validate the edit function along three main criteria: (1) The proposed alternative feature $\tilde{v}_{f_i}$ is **valid** if $\tilde{v}_{f_i} \in \mathcal{V}_{f_i}$ (i.e. the proposed alternative must reflect the intended intervention); (2) An edit is **effective** if $f_i(d') = \tilde{v}_{f_i}$ (i.e. the edits fully implement the proposed alternative such that it is extracted as the new feature value); (3) Edits are **minimally invasive** if $\forall f \neq f_i : f(d') = f(d)$. For instance, editing a story to be written from a different **PERSPECTIVE** can lead to portions of the narrative being omitted for cohesion, thus dramatically impacting which characters or plot points are included.

The authors annotated a set of 270 pairs of documents along this criteria on a 4-point Likert scale using the rubric described in Appendix G.4). Table 15 shows the

⁴Stronger models are not suitable for generating minimal edits required for this study (Laban et al., 2026).

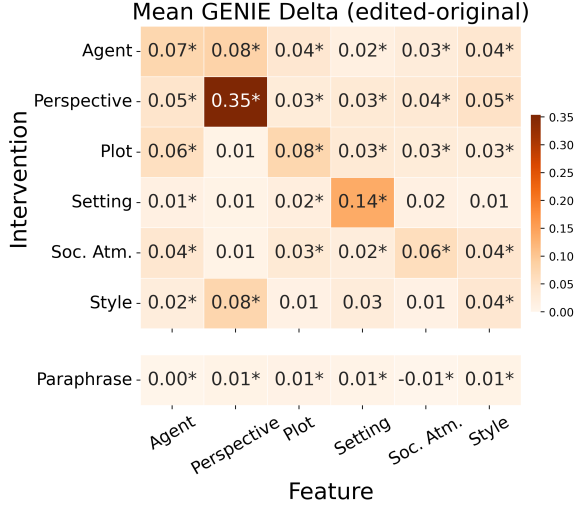


Figure 3: Mean deltas registered by GENIE for each intervention along each feature. (*): Statistical significant if $p < 0.008$ (Bonferroni corrected). GENIE is most sensitive to **PERSPECTIVE** and **SETTING** interventions, and registers the largest delta for feature f_i when f_i is intervened upon, except for **AGENT** and **STYLE**.

inter-annotator agreement, measured using Randolph’s κ (Randolph, 2005).⁵ Agreement is high for judging validity, and moderately high for determining effectiveness. Judging invasiveness is subjective. Thus, we measure the proportion of alternatives that are valid, edits that are effective and minimally invasive by computing the majority vote across annotators (Table 13). We discover that the edit function generates reliably valid and effective interventions. We note that these edits are invasive for **PLOT**, **SOCIAL ATMOSPHERE** and **PERSPECTIVE** interventions, suggesting inherent inter-feature dependence.

5.2.2 Sensitivity

We test whether intervening on f_i produces a systematic change in GENIE’s score along f_i , for all $f_i \in \mathcal{F}$. We perform a two-sided paired bootstrap test to find significance between pre- and post-intervention score distributions. For paraphrases, equivalence is tested using two one-sided tests. Refer to Appendix B.

GENIE is sensitive to interventions Figure 3 shows the mean deltas registered by GENIE pre- and post-intervention, per feature. GENIE accurately registers a statistically significant difference for all feature interventions, which shows that it can detect differences in novelty. Moreover, the heatmap indicates that when we intervene on **PERSPECTIVE**, **PLOT**, **SETTING** and **SOCIAL ATMOSPHERE**, GENIE not only detects the change, but also registers the largest mean delta along the intervened feature relative to all other feature scores.

GENIE is robust against paraphrases The final row in Figure 3 shows the mean deltas for paraphrased pairs.

⁵This task did not have a fixed rating distribution prior to annotations.

The mean deltas are all smaller than the deltas registered for interventions (diagonal values). The two one-sided paired bootstrap tests indicate that the GENIE scores across all features for paraphrased documents are statistically equivalent. This indicates GENIE’s robustness towards surface-level properties of texts.

6 Evaluating Holistic Metrics

Existing metrics that measure novelty, diversity and creativity are generally holistic, i.e., they generate one number for a given input, making it unclear what aspects of novelty they are measuring. But GENIE allows us to answer this question. Specifically, we use a suite of baseline metrics to analyze their sensitivity in addition to GENIE. We use the same setup as in §5.2.

6.1 Holistic Metrics

We measure the sensitivity of five holistic metrics that address novelty from different angles.

Cosine Distance: We compute the average pairwise cosine distance of document embeddings using text-embedding-vs-large (Jiang et al., 2026).

Compression Ratio (CR): We calculate the compression ratio of the population with and without the target response using gzip to capture lexical diversity (Shaib et al., 2024, 2025).

Creativity Index (Cr Idx): We measure the semantic n -gram novelty for $n \geq L$ following (Lu et al., 2025) relative to a reference corpus.

NoveltyBench Partition Model (NB): We utilize the partition model from (Zhang et al., 2025), which clusters candidate responses based on their functional utility where utility is defined as how beneficial a response is in terms of information value.

Litbench GenRM (LB): We use the Qwen-2.5-7b-Instruct generative reward model fine-tuned on Fein et al. (2026) to calculate the average pairwise preference win-rates.

All metrics are instantiated with the population used by GENIE. For instance, the corpora used by CreativityIndex and the candidates considered by the partition model in NoveltyBench are both instantiated using our population.

6.2 What Do Holistic Metrics Capture?

We calculate both the effect sizes (Cohen’s d) and statistical significance for scores found using each metric before and after intervention. This is shown in Table 3. We note that the standard deviations vary across metrics, and report them in Table 9.

Sensitivity Cosine Distance, Creativity Index and LitBench are sensitive to all feature interventions. Compression Ratio’s exclusive sensitivity to **SETTING** interventions, which use highly descriptive terminology and out-of-distribution n -grams (e.g. *rotating market pavilion on a lunar colony*; *glass factory built inside a dormant volcanic caldera*), indicates that it does not capture most types of content novelty. NoveltyBench detects

GENIE v. Holistic Metrics Effect Size (Cohen’s d)						
* $p < 0.008$, † $p < 0.008$						
Intervention	G	Cos.	CR	Cr Idx	NB	LB
AGENT	0.62*	0.54*	0.05	1.06*	0.52*	-0.45*
PERSPECTIVE	1.01*	0.67*	-0.19*	1.17*	0.14	-0.87*
PLOT	0.71*	0.22*	0.01	0.93*	0.30*	-0.33*
SETTING	0.93*	0.81*	1.08*	0.97*	0.44*	-0.48*
SOC. ATM.	0.47*	0.24*	0.23*	1.00*	0.16	-0.42*
STYLE	0.38*	0.68*	0.07	1.43*	0.07	-1.53*
Paraphrase	0.03†	0.45†	1.25†	1.29	0.69	-0.94

Table 3: Effect sizes (Cohen’s d) for feature interventions and paraphrases. GENIE and Cosine Distance are sensitive to all interventions *and* paraphrase robust. However, Cosine Distance weakly detects **PLOT** and **SOCIAL ATMOSPHERE**. LitBench uniquely finds negative effect sizes, scoring intervened documents poorly. Refer to Table 14 for GENIE’s paraphrase effect sizes (note that the average is reported here).

interventions moderately across most features, except **PERSPECTIVE**, **SOCIAL ATMOSPHERE** and **STYLE**. While most metrics perceive edited documents as more novel, LitBench scores them lower with negative effect sizes; we discuss this in a later paragraph.

Robustness The final row in Table 3 shows Cohen’s d measured when using metrics on paraphrases. Values marked with † denote distributions that are statistically equivalent (we use two one-sided tests, see Appendix B.2). Creativity Index, NoveltyBench and LitBench do not demonstrate robustness. Creativity Index appears to distinguish many types of interventions, but its lack of paraphrase robustness may indicate that it is overly sensitive to lexical novelty.

Figure 4 shows the min-max normalized mean absolute deltas for interventions versus paraphrases for each metric and feature intervention. Metrics in the shaded region are not robust as they detect higher average paraphrase deltas than for feature interventions. GENIE is the most consistently robust metric across all features. Compression Ratio and NoveltyBench consistently register a larger difference between paraphrased documents than intervened documents.

Aesthetic Quality LitBench consistently registers negative effect sizes across all interventions and paraphrases. Upon close inspection, we see that LitBench is capturing the degradation of a property consistent across edits. A qualitative analysis revealed that interventions, though accurate (Table 13), are artificial and forced at times (e.g. **PERSPECTIVE** interventions include rewriting the story from *first-person plural* from the *collective voice of the canned goods*). Thus, we note that the palatability of edited documents may be degraded, and correlated with LitBench. Although novelty and quality have a trade-off relation in model-generated responses (West and Potts, 2025), they are independent axes of creativity (Boden, 1991). Thus, metrics that capture aesthetic value are not suitable novelty metrics.

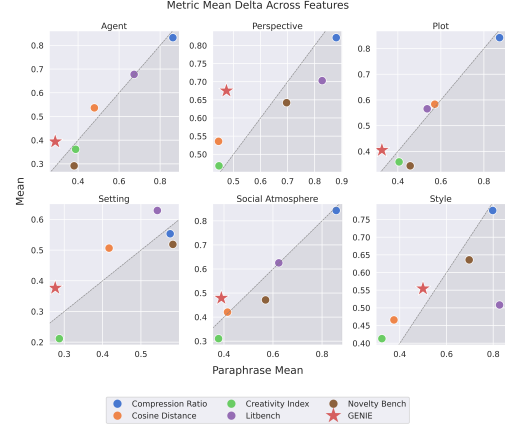


Figure 4: Normalized Mean Sensitivity and Paraphrase Absolute Deltas for each metric, per intervention. The shaded region represents higher average paraphrase sensitivity over interventions. GENIE is consistently robust across all interventions. Holistic metrics, except Cosine Distance, are usually as equally or more sensitive to paraphrases than interventions.

Cosine Distance vs. GENIE				
Feature	Cos. Dist.		GENIE	
	Orig	Edit	Orig	Edit
AGENT	0.34	0.36	0.64	0.72
PERSPECTIVE	0.32	0.37	0.37	0.80
PLOT	0.34	0.34	0.67	0.72
SETTING	0.33	0.36	0.61	0.75
SOC. ATM.	0.33	0.34	0.68	0.74
STYLE	0.34	0.39	0.56	0.64

Table 4: The raw mean scores for Cosine Distance and GENIE pre- and post- intervention.

Functional Utility Since the NoveltyBench partition model was optimized for short-context tasks rather than long-form creative writing, it is unclear which properties of a response determine its functional utility. However, the features it does not capture (**PERSPECTIVE**, **SOCIAL ATMOSPHERE** and **STYLE**) are properties that are peripheral to the document (i.e. not explicitly stated in the document). Thus, we hypothesize that functional utility correlates with, and may be capturing, visible properties, lexical or semantic.

Spurious Correlations Table 4 shows the raw mean scores calculated by Cosine Distance and GENIE per intervention. For certain interventions, such as **PLOT**, GENIE appears to register a more dramatic change in novelty. Cosine Distance registers the same delta for **PERSPECTIVE** and **STYLE** interventions (0.05), while GENIE registers a very large delta for **PERSPECTIVE**, but not **STYLE**. This suggests that Cosine Distance and GENIE are capturing different properties (refer to Table 16 for a correlation study).

7 Effectiveness of Mitigation Methods

The reduced diversity and creativity in model-generated responses (Kirk et al., 2024; Zhang et al., 2025; Jiang et al., 2026) has motivated work that aims to mitigate this behavior and improve various dimensions of creativity. Some of these methods are evaluated using holistic metrics, such as pairwise cosine similarity and lexical metrics (Zhang et al., 2026; Peeperkorn et al., 2025), as well as human and/or LLM judges (Huot et al., 2025; Chakrabarty et al., 2025b; Pham et al., 2025). We use GENIE to investigate the effectiveness of a subset of such methods. Specifically, we measure the distributional differences observed pre- and post- mitigation. For the following experiments, we aggregate the GENIE scores over all generations per prompt to form paired data.

7.1 Target Methods

We consider two broad categories of mitigation methods: alternate prompting and decoding based methods.

We evaluate three prompting strategies: (1) **Verbalized Sampling** (VS) refers to generating n samples using the target model M with their associated probabilities (Zhang et al., 2026). (2) We use Agents’ Room style prompting (Huot et al., 2025) to form detailed **Outlines** for each prompt and generate a document given the outline. This outlining requires the model to explicitly plan the Setting, Characters, Plot, etc., which closely follow the features of creative writing we use. (3) **Negative Sampling** (NS) is defined as generating a document d using prompt p and adding d to the context and prompting the model to generate a response *different* from d (Zhang et al., 2025).

Finally, we explore **Conformative Decoding** (CD), a decoding strategy that blends base and instruct models at the token-level (Peeperkorn et al., 2025). By using the weighted sum of token probabilities, the instruct model helps improve the quality while the base model introduces diverse content.⁶

Using the methods above, we generate responses to the prompts mentioned in §4.1 for the following models: gemma-3-12b-it, gemma-3-27b-it, Llama-3.1-8B-Instruct, OLMo-3-7B-Instruct and OLMo-3-32B-Instruct. Using these methods, we sample 5 responses for each prompt, per model. We then calculate GENIE to study the fine-grained improvements mitigation helps bring.

7.2 Results

GENIE scores for each feature are aggregated across generations of documents per prompt, per model. For Negative Sampling, specifically, we pair the data at the instance level since target responses are paired prior to generating the data (adding the response to the context). We perform a paired bootstrap two-sided test between GENIE scores for responses generated pre- and post-

⁶We weight base models higher than instruct models (0.6 and 0.4, respectively) to inspire more diverse outputs.

Mean GENIE Delta (*p<0.008)				
Feature	VS	Outline	NS	CD
AGENT	0.03*	0.02	0.02*	0.01
PERSPECTIVE	0.01	0.03*	0.01	0.01
PLOT	0.02*	0.00	0.02*	0.01
SETTING	0.04*	0.00	0.03*	0.02*
SOC. ATM.	0.03*	0.00	0.01*	0.01
STYLE	0.02*	0.02*	0.02*	0.00
Avg. num words	98	513	510	398
Avg. Coherence	3.45	3.35	3.32	3.19

Table 5: Mean GENIE deltas observed when using mitigation methods. Significance is marked with (*). Conformative Decoding helps improve **SETTING** and Outlining improves **PERSPECTIVE** and **STYLE**. The average number of words in Verbalized Sampling scores are not representative as responses are disproportionately small relative to the population (488) and target (513) set. Average Coherence computed using ArmoRM (out of 5) is moderate across all methods.

mitigation. The mean GENIE deltas are reported in in Table 5, with significance marked.

Outlining helps improve **PERSPECTIVE** and **STYLE**, while Conformative Decoding improves **SETTING**. Negative Sampling has a significant improvement relative to the original counterpart. This is consistent with the findings of Zhang et al. (2025) regarding the effectiveness of this method. We note that despite significance, all mean deltas are small (e.g. relative to GENIE detecting a delta of 0.35 on **PERSPECTIVE** interventions in Figure 3).

Although Verbalized Sampling *appears* to be the most effective mitigation method, Table 5 shows that responses generated by this method are disproportionately shorter relative to the target and population set. These lengths (~100 words) are similar to flash fiction with 2-3 sentences, rather than a fully formed story (e.g. Appendix E.1). Thus, comparing the novelty of these documents with respect to a population that is from a different distribution is not representative. This highlights the importance of using an appropriate population when computing novelty.

This analysis illustrates what fine-grained evaluation enables, rather than a single aggregate score, GENIE attributes each method’s effect to specific features.

8 Conclusion

We present GENIE, a fine-grained evaluation metric to measure the *novelty* of LLM-generated responses with respect to an explicitly defined population. GENIE discovers and grounds task-specific *features* in responses using answers to prompt-specific questions. We instantiate GENIE in the creative writing domain as a testbed and show that GENIE is sensitive to features when independently intervened upon and robust towards lexical edits (i.e. paraphrases). Using GENIE, we study the sensitivity and robustness of existing holistic creativity,

diversity and novelty and find that they either capture extraneous properties or are lexically sensitive. Finally, we evaluate the effectiveness of mitigation methods using a fine-grained lens, revealing the strengths of alternate prompting and decoding strategies.

Limitations

Computational Cost The GENIE pipeline is moderately resource-intensive. We use proprietary models for feature discovery, question generation, question-to-feature mapping, alternate feature generation, document editing, and paraphrasing, and Qwen2.5-32B-Instruct for question answering and similarity judgments. Using GENIE for other domains requires generating new populations because GENIE is heavily task-oriented. That said, measuring novelty is really difficult, and we view the compute cost worth the cost of going beyond the coarse, surface-level proxies to explainable, feature-grounded evaluation (see Appendix A.5 for more details).

Domain Scope We instantiate GENIE in the creative writing domain. The pipeline is, by design, generalizable to any text-based task but we have not run extensive experiments measuring how well existing metrics capture novelty outside creative writing. Thus, this pipeline has not been validated for use in domains such as scientific writing, code or argumentative essays.

Acknowledgments

We thank Kathryn Kazanas for helping with preliminary annotations, Chantal Shaib for early guidance and the TAUR Lab for feedback on the project. This work was partially supported by NSF grants IIS-2107524, IIS-2145479, IIS-2145280, and IIS-2433071, as well as the Sloan Foundation. Compute for this project was partially supplied by the Machine Learning Lab at UT Austin and the NSF AI Institute for Foundations of Machine Learning (IFML).

References

- Ron Artstein and Massimo Poesio. 2008. [Inter-Coder Agreement for Computational Linguistics](#). *Computational Linguistics*, 34(4):555–596.
- David I. Beaver, Craige Roberts, Mandy Simons, and Judith Tonhauser. 2017. Questions Under Discussion: Where Information Structure Meets Projective Content. *Annual Review of Linguistics*, 3(1):265–284.
- Margaret A. Boden. 1991. *The Creative Mind: Myths and Mechanisms*. Basic Books, Inc., USA.
- Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. 2026. [Evaluating Text Creativity across Diverse Domains: a Dataset and Large Language Model Evaluator](#). In *The Fourteenth International Conference on Learning Representations*.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. [Art or Artifice? Large Language Models and the False Promise of Creativity](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA. Association for Computing Machinery.
- Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025a. [AI-Slop to AI-Polish? Aligning Language Models through Edit-Based Writing Rewards and Test-time computation](#). In *Second Conference on Language Modeling*.
- Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025b. [Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, and 48 others. 2026. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- 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.
- Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2026. [LitBench: A Benchmark and Dataset for Reliable Evaluation of Creative Writing](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7740–7755, Rabat, Morocco. Association for Computational Linguistics.
- Giorgio Franceschelli and Mirco Musolesi. 2024. [On the Creativity of Large Language Models](#). *AI Soc.*, 40(5):3785–3795.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Charles Leonard Hamblin. 1957. *Language and the Theory of Information*. Ph.D. thesis, London School of Economics and Political Science (University of London).

- Sil Hamilton. 2024. [Detecting Mode Collapse in Language Models via Narration](#). In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 65–72, St. Julian's, Malta. Association for Computational Linguistics.
- Sil Hamilton, Matthew Wilkens, and Andrew Piper. 2026. [NarraBench: A Comprehensive Framework for Narrative Benchmarking](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3786–3801, Rabat, Morocco. Association for Computational Linguistics.
- Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. [Agents' Room: Narrative Generation through Multi-step Collaboration](#). In *The Thirteenth International Conference on Learning Representations*.
- Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. 2025. [Creative Preference Optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9580–9609, Suzhou, China. Association for Computational Linguistics.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2026. [Artificial Hivemind: The Open-Ended Homogeneity of Language Models \(and Beyond\)](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. [Gemma 3 Technical Report](#). *ArXiv*, abs/2503.19786.
- Lauri Karttunen. 1977. Syntax and Semantics of Questions. *Linguistics and philosophy*, 1:3–44.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. [Understanding the Effects of RLHF on LLM Generalisation and Diversity](#). In *The Twelfth International Conference on Learning Representations*.
- Philippe Laban, Tobias Schnabel, and Jennifer Neville. 2026. [LLMs Corrupt Your Documents When You Delegate](#). *arXiv preprint*.
- Utpal Lahiri. 2001. *Questions and Answers in Embedded Contexts*. Oxford University Press UK.
- Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevalere, and Alexandre Allauzen. 2024. [Exploring Precision and Recall to assess the quality and diversity of LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11418–11441, Bangkok, Thailand. Association for Computational Linguistics.
- Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2025. [AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text](#). In *The Thirteenth International Conference on Learning Representations*.
- James B McQueen. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297.
- William Merrill, Noah A. Smith, and Yanai Elazar. 2024. [Evaluating \$n\$ -Gram Novelty of Language Models Using Rusty-DAWG](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14459–14473, Miami, Florida, USA. Association for Computational Linguistics.
- Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, and 88 others. 2024. [Gemma: Open models based on gemini research and technology](#). *ArXiv*, abs/2403.08295.
- Richard Mojena. 1977. Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal*, 20(4):359–363.
- Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. 2025. [Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction](#). In *Forty-second International Conference on Machine Learning*.
- Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. [QUDsim: Quantifying Discourse Similarities in LLM-Generated Text](#). In *Second Conference on Language Modeling*.
- Vishakh Padmakumar and He He. 2024. [Does Writing with Language Models Reduce Content Diversity?](#) In *The Twelfth International Conference on Learning Representations*.
- Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. 2026. [Measuring LLM Novelty As The Frontier Of Original And High-Quality Output](#). In *The Fourteenth International Conference on Learning Representations*.
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2025. [Mind the Gap: Conformative Decoding to Improve Output Diversity of](#)

- Instruction-Tuned Large Language Models. *CoRR*, abs/2507.20956.
- Chau Minh Pham, Jenna Russell, Dzung Pham, and Mohit Iyyer. 2025. [Frankentext: Stitching random text fragments into long-form narratives](#). *ArXiv*, abs/2505.18128.
- Justus J. Randolph. 2005. [Free-Marginal Multirater Kappa \(multirater K \[free\]\): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa](#). Presented at the Joensuu Learning and Instruction Symposium. ERIC Document ED490661.
- Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. [Gemma 2: Improving Open Language Models at a Practical Size](#). *Preprint*, arXiv:2408.00118.
- Craige Roberts. 2012. Information Structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.
- Jenna Russell, Rishanth Rajendhran, Chau Minh Pham, Mohit Iyyer, and John Wieting. 2026. [StoryScope: Investigating Idiosyncrasies in AI Fiction](#).
- Samuel Schapiro, Sumuk Shashidhar, Alexi Gladstone, Jonah Black, Royce Moon, Dilek Hakkani-Tur, and Lav R Varshney. 2025. Combinatorial Creativity: A New Frontier in Generalization Abilities. *arXiv preprint arXiv:2509.21043*.
- Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024. [Detection and Measurement of Syntactic Templates in Generated Text](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. 2025. [Standardizing the measurement of text diversity: A tool and comparative analysis](#). In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 36–46, Mumbai, India. Association for Computational Linguistics.
- Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. 2025. [Diverse Preference Learning for Capabilities and Alignment](#). In *The Thirteenth International Conference on Learning Representations*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *Preprint*, arXiv:2307.09288.
- Leah Velleman and David I. Beaver. 2016. [Question-based Models of Information Structure](#). In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*, pages 86–107. Oxford University Press, Oxford, UK.
- Manya Wadhwa, Tiasa Singha Roy, Harvey Lederman, Junyi Jessy Li, and Greg Durrett. 2026. CREATE: Testing LLMs for Associative Creativity . *arXiv preprint arXiv:2603.09970*.
- Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. [2 OLMo 2 Furious](#). *Preprint*, arXiv:2501.00656.
- Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *EMNLP*.
- Emily Wenger and Yoed N. Kenett. 2025. [We're Different, We're the Same: Creative Homogeneity Across LLMs](#). *ArXiv*, abs/2501.19361.
- Peter West and Christopher Potts. 2025. [Base Models Beat Aligned Models at Randomness and Creativity](#). In *Second Conference on Language Modeling*.
- Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. [Echoes in AI: Quantifying lack of plot diversity in LLM outputs](#). *Proceedings of the National Academy of Sciences*, 122(35):e2504966122.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#).
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. [Qwen2.5 Technical Report](#). *Preprint*, arXiv:2412.15115.
- Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael Tomz, Christopher D Manning, and Weiyan Shi. 2026. [Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity](#).
- Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. [NoveltyBench: Evaluating Creativity and Diversity in Language Models](#). In *Second Conference on Language Modeling*.

A GENIE

A.1 Building the Population

GENIE requires a population or reference corpus. For our experiments, to determine the population sample size required for statistical confidence, we performed a power analysis using a pilot set of 100 randomly sampled target documents across all 50 prompts. We compute the similarity between each target document and the entire population. Table 9 shows the average variance and standard deviation of similarity scores across documents, for each feature. The minimum population sample size required for each feature is reported in Table 9 for 98% confidence and a margin of error of 0.1.

In order to sample a global population that is consistent across features, we use the maximum required size i.e. 50 documents. We randomly sample these from the population. Although the analysis determines the minimum number of *answer-pairs* needed for each *feature*, we sample at the document level in order to compare all metrics fairly instead of providing partial information to some. Sampling at the document-level will inherently produce a population that is *larger* than required.

We also conduct an retrospective analysis and check the power to see if the data is sufficiently large to capture the effect size that is observed. The resulting dataset that we have has sufficient power to determine the statistical significance marked in Table 5. However, this analysis is done per feature, and data is lost through the GENIE pipeline (filtering documents, unanswerable documents, etc.). We note that we achieve power ≥ 0.9 for cases where ≥ 44 prompt-pairs survived. Thus 50 is sufficiently large, except we must generate more documents per mitigation method to reduce the loss of pairs.

A.2 Discovering Features for Creative Writing

We experiment with three different clustering methods to discover features: HDBSCAN (Campello et al., 2013), agglomerative clustering (Mojena, 1977) and K-means (McQueen, 1967) (K is set to the average number of samples in each response). We show outputs from K-means clustering in Table 7. For consistency, we map the features discovered to an existing creative writing taxonomy (Hamilton et al., 2026) using the generated descriptions. However, the mapping is not a perfect one-to-one. For instance, the feature generation step generated: *Mood, Tone and Language*, which are a subset of **STYLE** as defined by Narrabench. Similarly, the generated feature, *Point of View*, is an aspect of **PERSPECTIVE**.

A.3 Design Choice: Answer Length for Feature Questions

GENIE uses the average dissimilarity between answers to questions in the target and population documents (§4). These questions represent different features and the distance values are used to compute novelty (§3.2). When generating answers to questions, we specify that answers must be concise (Prompt: F.7). In this section,

Population Statistics and Min Sample Size			
Feature	Var	Std. Dev.	Min. Samp. Size
AGENT	0.08	0.27	41
PERSPECTIVE	0.09	0.30	50
PLOT	0.07	0.27	40
SETTING	0.07	0.27	40
SOC. ATM.	0.07	0.26	37
STYLE	0.08	0.29	45

Table 6: Average variance and standard deviation of similarity scores between target and population documents, for each feature. We also report the minimum population size required for 98% confidence and a margin of error of 0.1.

K-Means Creative Writing		
Cluster ID	Unique Features	Count
0	Point of View	5
	Narrative Structure	2
	Narrative Voice	1
	Narrative Perspective	1
	Narration	1
	Emotional Impact	1
1	Theme	8
	Style	5
	Concept	1
	Idea	1
	Subtext	1
	Atmosphere	1
	Thematic Content	1
	Subject Matter	1
	Imagery	1
2	Mood	6
	Tone	6
3	Plot	6
4	Language	6
	Voice	2
	Language Use	1
5	Setting	9
	Context	1
6	Characterization	6
	Structure	5
	Character	4
	World-Building	2
	Expression	1

Table 7: Creative Writing features discovered using K-means clustering. Bolded features are most frequently occurring in their respective cluster.

conduct an analysis to test whether a statistically significant difference is observed when using answers of differing lengths.

Let $g_{C,f}$ represent the GENIE scores computed using length-constrained answers along feature f . For the target set of documents, we form $g_{U,f}$, the GENIE scores found using answers of unconstrained length. The effect sizes between the two distributions are shown in Table 8. Using concise answers generates GENIE scores that are statistically larger than using answers with no length constraint.

However, we also calculated the rank correlation to

Effect Size (Cohen’s d)		
Feature	Cohen’s d	Kendall’s τ
AGENT	0.68*	0.56
PERSPECTIVE	-0.01	0.46
PLOT	0.90*	0.61
SETTING	0.54*	0.51
SOC. ATM.	0.50*	0.53
STYLE	0.62*	0.50

Table 8: $g_{C,f}$ is significantly greater than $g_{U,f}$ for all features except Perspective.

determine whether the *relative* scores were consistent. The Kendall’s τ between $g_{C,f}$ and $g_{U,f}$ is positive and moderately strong, suggesting that although the raw scores are different, the relative novelty is consistent.

A.4 Examples

Table 19 shows examples of writing prompts used in our experiments.

A.5 Computation Costs and Inference Parameters

Question Generation and Similarity The question answering and similarity judgment stages of GENIE (§4.5) are performed using Qwen-2.5-32B-Instruct. We used 2 A40 GPUs with `tensor_parallelization = 2`. Each iteration of answering questions takes approximately 2 seconds to complete (we use vLLM to help with efficiency). We can complete 30 similarity judgments per second. On average, answering questions for a single document d_t takes 34s and with a population of size 50, similarity judgment takes 28s (assuming the prompt generated an average number of questions, 17).

Truncation We use gpt-4.1-mini for truncation. Instruct documents are approximately 750 tokens long, and base models are restricted to 350.

Quality Filter We use ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024), hosted on 1 A40 GPU.

Per-Component Cost The estimated cost for replicating the GENIE pipeline using our experimental setup described in 4 is as follows:

1. Question Answering (population and target documents): 4.39 hours on 2 A40 GPUs
2. Similarity: 45.42 hours on 2 A40 GPUs

Causal Intervention For our experiment, we use gpt-5.4-mini for alternate feature generation and gpt-4.1-mini for performing the edits. The total cost to generate pairs for 15 original documents (i.e. 270 pairs) (feature generation and editing combined) is 1\$.

Inference parameters We run story generation/similarity judgments and answer generation at a temperature of 0.7 and top_p=0.9. For story generation, for instruct models, we set max tokens of 750 and base models we set max tokens to 350.

For conformative decoding, we weighted the instruct models at 0.4 and base models at 0.6.

B Sensitivity and Robustness Interventions

In this section we expand on the details regarding the sensitivity and robustness studies presented in §5. We apply Bonferroni correction ($\alpha/6$) and use $p < 0.008$ to determine significance.

B.1 Testing Intervention Sensitivity

To evaluate the sensitivity of metrics when interventions are performed, we conducted a two-sided paired bootstrap test. Here, we aim to test whether a significant difference exists between the two distributions (i.e. pre- and post- intervention). Thus, we define the null hypothesis as the intervention having no significant effect on the distribution. Let $\Delta\mu$ be the mean delta of the distributions. Formally, the null hypothesis states that $\Delta\mu = 0$. The alternate hypothesis states that $\Delta\mu \neq 0$. We repeatedly sample pairs with replacement 10,000 times and compute the mean delta. The significance (p -value) is determined by calculating the proportion of pairs where the resampled mean delta was greater than or equal to the observed delta.

B.2 Testing Paraphrase Robustness

To evaluate the robustness of metrics when paraphrases are performed, we conducted two one-sided bootstrap tests. We aim to test whether statistical equivalence can be established between pre- and post- intervention distributions. The two-sided paired bootstrap test described above determines whether a significant difference can be observed between two distributions. However, we are interested in showing that no significant difference can be observed (i.e. the null hypothesis is flipped). We start with a null hypothesis that states that the two distributions are statistically different.

Formally, let ϵ denote the equivalence margin and $\Delta\mu$ denote the mean delta. The null hypothesis states that the mean delta is not within the interval $[-\epsilon, \epsilon]$. This is expressed through two null hypotheses: $\Delta\mu \leq -\epsilon$ and $\Delta\mu \geq \epsilon$. Thus, the alternative hypothesis states that the mean delta is within $(-\epsilon, \epsilon)$ (i.e. $-\epsilon < \Delta\mu < \epsilon$).

In practice, we set ϵ to $0.2 \times \text{std}(\text{original score distribution})$. The two separate one-sided bootstrap tests determine where the delta falls with respect to the equivalence margin. Both null hypotheses must be rejected. We perform the test using the same parameters as described above (e.g. 10,000 iterations).

Example Alternate Features:

Alternate Features

Feature: AGENT

Alt 1: Instead of Death as the sole keeper of the void, a forgotten maintenance clerk who believes the universe is only temporarily shut down and is desperately searching for the one missing key

Alt 2: Instead of Koko the cautious, enrichment-starved gorilla, an elderly retired circus gorilla who treats ice skates

Raw Mean Delta						
* $p < 0.01$, † $p < 0.01$						
Intervention	G	Cos.	CR	Cr Idx	NB	LB
AGENT	0.07*	0.03*	0.00	0.11*	0.05*	-0.06*
PERSPECTIVE	0.35*	0.04*	-0.01*	0.14*	0.03	-0.20*
PLOT	0.08*	0.01*	0.00	0.09*	0.02*	-0.05*
SETTING	0.14*	0.03*	0.01*	0.09*	0.04*	-0.03*
SOC. ATM.	0.06*	0.01*	0.00*	0.09*	0.01	-0.07*
STYLE	0.04*	0.05*	0.00	0.17*	0.02	-0.30*
Paraphrase	0.01†	0.01†	0.01†	0.12	0.11	-0.07
Raw Standard Deviations						
* $p < 0.01$, † $p < 0.01$						
AGENT	0.12*	0.05*	0.02	0.10*	0.10*	0.14*
PERSPECTIVE	0.35*	0.06*	0.03*	0.12*	0.25	0.23*
PLOT	0.11*	0.05*	0.03	0.10*	0.07*	0.14*
SETTING	0.15*	0.04*	0.00*	0.09*	0.10*	0.07*
SOC. ATM.	0.13*	0.04*	0.02*	0.09*	0.08	0.16*
STYLE	0.11*	0.07*	0.03	0.12*	0.28	0.19*
Paraphrase	0.10†	0.01†	0.01†	0.09	0.16	0.07

Table 9: Raw mean deltas and standard deviations for all metrics, for each intervention. The standard deviations are the narrowest for Compression Ratio and Cosine Distance. We report the average standard deviation for GENIE on paraphrases.

as a final, defiant encore

Alt 3: Instead of a fearful human office worker, the skyscraper itself as a conscious, aging witness to the city, learning from the fairy how to move from cold ambition toward shelter, patience, and unexpected tenderness

Prompt: Write a story about a dinosaur and a computer.
Response: In a future museum, a robotic dinosaur uses an AI computer to recreate lost dinosaur behaviors, leading to a virtual revolution.

Feature: PERSPECTIVE

Alt 1: Instead of human or phoenix focalization, second-person perspective addressed to the *Stardust Drifter* itself, with the bridge as the ship’s own sensory and emotional field of view

Alt 2: Instead of a single human focalizer, the perspective is an overhearing house itself, distributing attention across footsteps, breathing, lamps, and fear as if the room is trying to understand who belongs

Alt 3: Instead of third-person limited focused on Dr. Maria Hernandez, first-person plural from the jellyfish colony collectively observing the humans and the telescope

Feature: PLOT

Alt 1: Instead of a quiet companionship with one office worker, Luminescence discovers the skyscraper is slowly draining the city’s remaining wild places to stay standing, and she must choose between preserving the building’s human purpose and restoring the vanished underground roots, leading to a strange compromise where the tower is transformed into a vertical ecosystem that feeds both street and forest.

Alt 2: Instead of a reverent first-contact style revelation,

the ancient temple turns out to be a quarantine vault containing an intelligence that can only speak by rewriting memories; Elena must survive escalating hallucinations, crew distrust, and a growing urge to unlock the vault, then resolves the crisis by sacrificing her own most important memory to seal it again, turning the “temple” into a moral trap rather than a treasure.

Alt 3: Instead of a single ghost lingering in one grocery store, multiple ghosts from different eras begin possessing products on the shelves, forcing a new cashier to negotiate between rival spirits over expired milk, discontinued snacks, and forgotten brands, until the store is transformed into a living museum where the dead choose what the living are allowed to buy.

Feature: STYLE

Alt 1: Instead of straightforward chronological storytelling, use a fragmented oral-history style: overlapping quotes, brief testimony-like snippets, repeated key phrases, and a chorus-like rhythm that makes the skating feel like a legend told by different witnesses.

Alt 2: Instead of conversational first-person prose, a clipped screenplay-style format with terse action lines and rapid dialogue, letting the comedy come from the mismatch between blockbuster monster imagery and everyday domestic panic

Alt 3: Instead of descriptive scene-setting and emotional exposition, a feverish mythic folktale voice full of ominous repetitions, strange epithets, and campfire-story cadence that treats the monster’s fear as an ancient omen

Feature: SOC. ATM.

Alt 1: Instead of a crew-centered bond, the phoenix and the bridge crew form a competitive ritual circle, trading status, challenges, and displays of courage as each learns to respect the other.

Alt 2: Instead of everyone aligning against the same danger, Lily becomes the most authoritative voice in the room, with Leo and the monster both deferring to her as she declares Barnaby the bunny to be the real problem.

Alt 3: Instead of gentle private companionship, a tense alliance emerges between the fairy and the building itself, with humans as unaware bystanders while the two negotiate space, attention, and influence through indirect acts

Feature: SETTING

Alt 1: Instead of a normal house at night, a drifting space station habitat after a systems blackout, with emergency strips, viewports, and a sealed maintenance hatch defining the room

Alt 2: Instead of a grocery store, a traveling river barge market in an alternate 1800s, defined by canvas awnings, hanging spice sacks, and a cash box that never closes

Alt 3: Instead of a suburban park bench, a flooded underground train station in a near-future megacity, with glowing ticket gates and abandoned vending machines

B.3 A note on STYLE.

A qualitative study into the interventions and edits revealed that STYLE interventions are most often invalid. Many STYLE alternate features include rewriting the story as a log, record or report. The resulting edited doc-

ument is often no longer a story, and is extremely lexically different that it is unclear if detecting the **STYLE** intervention is because of stylistic elements, or if it is simply because the two documents are thematically diametric. We use gpt-4.1-mini to filter documents that are in the form of a different media, such as logs, records, poetry, etc. (Appx. F.11). We filtered out 206 pairs, of which 167 are **STYLE** interventions (Table 10 details the number of pairs that were excluded in other features).

Num. of Excluded Pairs	
Feature	Count
AGENT	5
PERSPECTIVE	27
PLOT	3
SETTING	6
SOC. ATM.	2
STYLE	167
Total	206

Table 10: 56% of Style intervention pairs required non-prose like edits. However, other interventions preserved the prose-style of writing and did not need to be excluded from the experiment.

C Human Evaluation

We conducted different human studies to validate different parts of the GENIE pipeline. This section gives more details on the annotators, the process and the instructions used.

C.1 Question Validity and Mapping

For evaluating question validity and mapping steps of GENIE (§4), two authors on the paper evaluated 50 instances for each task.

Question Validity Annotators for this task were given the rubric mentioned in Rubric: G.1. We report the annotator agreement in §5.1. We note the following main sources of disagreement for evaluating the validity of questions:

Our instructions explicitly excludes questions that presuppose details that are not already given in the prompt. However, there are details that could be considered deducible from the prompt. For instance for a generated question, *What cultural or societal changes arise from the rebel angel’s success?* formed from a prompt that only mentions a *rebel angel and success*, strictly presupposes that a change was actually brought about. However, in this case, the lack of change itself is important information given that rebellions usually lead to change. The instructions also asks annotators to filter out questions that cover multiple features since these questions can add noise to feature extraction. In certain cases, discrepancies arose due to questions such as,

How do the settings influence the characters’ decision-making? and *How does the setting influence the mood of the story?*. Some *influence* questions are more evidently multi-feature than others, despite filtering, which again creates a spectrum of leniency.

Question Mapping Annotators were given instructions mentioned in Rubric G.2 We report the annotator agreement in §5.1. We note the following main sources of disagreements for evaluating the mapping of questions to features:

Disagreements specifically arose regarding which parts of the question should be weighted. Subjectivity in the bucketing task was revealed through differing question interpretations at times. The question, *What internal conflicts affect the mermaid’s behavior?*, could be understood either as an **AGENT** or **PLOT** question depending on whether the annotator considers *internal conflicts* to be integral to character development, or if the annotator believes these conflicts can manifest in key plot points (e.g. the plot revolves around the character’s inner struggles). Similarly, *How do the characters negotiate their fears through conversation?* can either be interpreted as **STYLE** or **SOCIAL ATMOSPHERE**. The mode of communication itself, such as dialogue, would come under **STYLE**, however, if the question were asking about the interaction itself in the context of the story, then it would be classified as **SOCIAL ATMOSPHERE**. In essence, this task demonstrates the complexity of extracting features as independently as possible, despite inter-feature dependencies.

C.2 Similarity Judgments

For the similarity validity annotations, we recruited 4 Upwork annotators with prior NLP annotation experience. They were informed that their annotations would be used for research purposes. We used a pilot set of 100 annotations and only selected annotators where their annotations were calibrated with our guidelines. Each of the annotators were paid \$150 for 1500 annotations. Annotations took approximately 5 hours. Each annotation pair was annotated by two annotators. The instructions given to the annotators are given in Rubric G.3.

Table 11 shows the inter-annotator agreement per feature (Krippendorff- α). We also report the Krippendorff- α and Kendall’s τ between the human consensus (average of the two annotator’s judgments) and Qwen-2.5-32B-Instruct. Inter-annotator agreement is moderately high across all features, and especially high for **PERSPECTIVE** pairs. **PERSPECTIVE** pairs are less subjective due to the limited space of questions and answers (*"From whose perspective is the story told?"*, *"Which point of view is used?"*). Accordingly, we see high agreement on **PERSPECTIVE** with the LLM. Overall, there is moderately strong correlation between LLM judgments and humans (**AGENT** **PERSPECTIVE** and **SETTING**), suggesting that Qwen-2.5-32B-Instruct

Feature	Krippendorff- α		Kendall- τ
	Inter-annot	Ann v. LLM	Ann v. LLM
AGENT	0.58	0.65	0.64
PERSPECTIVE	0.76	0.83	0.75
PLOT	0.44	0.55	0.53
SETTING	0.65	0.70	0.66
SOC. ATM.	0.43	0.53	0.58
STYLE	0.53	0.57	0.60
Total	0.59	0.66	0.64

Table 11: Annotator Agreement for Similarity Judgments. The moderately strong inter-annotator agreement suggests limited subjectivity. Qwen-2.5-Instruct judgments strongly correlate with humans across all features.

Feature	Randolph's- κ		
	Validity	Accuracy	Invasiveness
AGENT	0.59	0.44	0.14
PERSPECTIVE	0.87	0.71	0.27
PLOT	0.78	0.46	0.22
SETTING	0.90	0.52	0.30
SOC. ATM.	0.65	0.44	0.15
STYLE	0.60	0.41	0.50

Table 12: Annotator Agreement for intervention. Annotators have high agreement on validity, moderately high agreement on accuracy and weak agreement on invasiveness judgments. PERSPECTIVE is the least subjective feature intervention for judging both validity and accuracy.

Feature	Accuracy (majority vote)		
	Validity	Effectiveness	Invasiveness
AGENT	0.89	0.89	0.61
PERSPECTIVE	1.00	0.97	0.53
PLOT	0.98	0.93	0.22
SETTING	1.00	0.93	0.87
SOC. ATM.	0.93	0.91	0.56
STYLE	0.72	0.94	0.67

Table 13: Frequency of successfully valid, accurate or minimally invasive interventions, judged by majority vote per annotation. Threshold for validity, accuracy and minimal invasiveness is determined using the rubric in Appendix G.4. Proposed alternatives are valid across all features, and are often executed accurately. SETTING interventions are the least invasive.

is suitable for automatically finding similarity. This validates our dissimilarity function.

C.3 Intervention Studies

Three authors on this paper together annotated 270 instances for measuring the validity, effectiveness and invasiveness of the interventions. The instructions given to them are mentioned in Rubric G.4. We also note the agreement between them in Section 5.2

Intervention	GENIE Effect Size (Cohen's d)					
	Ag.	Persp.	Plot	Set.	SA	Sty.
AGENT	0.62*	0.29*	0.39*	0.18*	0.22*	0.37*
PERSP	0.44*	1.01*	0.29*	0.28*	0.29*	0.43*
PLOT	0.55*	0.04	0.71*	0.35*	0.21*	0.23*
SETTING	0.21*	0.04	0.23*	0.93*	0.13	0.1
SOC. ATM.	0.44*	0.04	0.35*	0.22*	0.47*	0.35*
STYLE	0.31*	0.3*	0.13	0.23	0.11	0.38*
Paraphrase	0.02 [†]	0.04 [†]	0.12 [†]	0.06 [†]	-0.14 [†]	0.09 [†]

Table 14: The effect size registered by individual GENIE feature scores for each intervention. Values in bold denote the feature with the largest effect size for each intervention (per row). When we intervene on f , GENIE registers the largest effect size along f over other features. GENIE is able to detect not only *a* difference, but can help explain *what* changed. Other feature scores also react, likely due to inter-feature dependence. SOCIAL ATMOSPHERE changes register a nearly equal effect size along AGENT. Furthermore, GENIE is robust towards surface-level differences as revealed by the insignificant effect sizes for paraphrased pairs

Interventions with Decrease in Novelty		
Feature	Min Delta	Frac. deltas < 0
AGENT	-0.20	0.29
PERSPECTIVE	-0.81	0.12
PLOT	-0.27	0.23
SETTING	-0.32	0.13
SOC. ATM.	-0.33	0.33
STYLE	-0.30	0.34

Table 15: Fraction of interventions where the difference in novelty between pre and post-intervention is negative, i.e. a reduction in novelty. For every feature, most interventions lead to an increase in novelty. With the exception of PERSPECTIVE, the maximum reduction in novelty an intervention has is 0.33.

D Example Walkthrough

In this section we give an example of GENIE scores and different questions/answers that we generate. Example D.1 shows an example of a story generated by meta-llama/Meta-Llama-3.1-70B-Instruct. Examples of the questions, features and extracted answers for these questions are given in Table 20. Based on these, the story achieves the following GENIE scores: $\{ 'agent': 0.85, 'perspective': 0.43, 'plot': 0.73, 'setting': 0.34, 'social atmosphere': 0.68, 'style': 0.46 \}$.

Example 1

Prompt: Write a story about an alien and a rainforest.

Model: meta-llama/Meta-Llama-3.1-70B-Instruct

Document:

Deep within the heart of the Amazon rainforest, a lone alien named Zorvath had been observing the lush ecosystem for centuries. His spaceship, disguised as a giant kapok tree, blended seamlessly into the surroundings, allowing him

to study the intricate web of life that thrived beneath the emerald canopy.

Zorvath's planet, Xanthea, had long since been destroyed by the very technology that was meant to sustain it. The Xantheans, a peaceful and curious species, had exhausted their resources and ravaged their environment, leaving their homeworld a barren, lifeless husk. The few remaining Xantheans had scattered across the galaxy, seeking refuge and a chance to redeem their mistakes.

Zorvath, a skilled botanist and ecologist, had been sent to Earth to learn from the planet's diverse ecosystems. He had chosen the Amazon rainforest, with its unparalleled biodiversity, as his primary research site. For centuries, he had watched over the forest, marveling at the complex relationships between the plants, animals, and microorganisms that called it home.

One day, a severe storm swept through the rainforest, threatening to destroy the delicate balance of the ecosystem. Zorvath knew he had to act quickly to prevent irreparable damage. Using his advanced technology, he created a network of energy fields that stabilized the soil, calmed the winds, and filtered the rainwater. The storm raged on, but the forest, protected by Zorvath's intervention, weathered the tempest unscathed.

As the storm subsided, a young indigenous girl named Akua wandered into the forest, searching for medicinal plants to heal her ailing mother. Zorvath, observing her from a distance, was struck by her deep connection to the natural world. He decided to reveal himself to her, hoping to share his knowledge and learn from her people's ancient wisdom.

Akua, initially frightened by the alien's appearance, soon found herself at ease in his presence. Zorvath showed her the secrets of the forest, explaining the intricate relationships between the plants, animals, and microorganisms. He shared his knowledge of sustainable practices, gleaned from his studies of Earth's ecosystems, and Akua shared her people's traditional methods of living in harmony with nature.

Together, Zorvath and Akua formed a powerful bond, one that transcended their differences and united them in their quest to protect the rainforest. As the years passed, they worked tirelessly to preserve the delicate balance of the ecosystem, using Zorvath's advanced technology and Akua's traditional knowledge to defend the forest against threats both human and natural.

Their collaboration inspired others to join their cause, and soon, a community of humans and Xantheans worked together to safeguard the Amazon rainforest. Zorvath's spaceship, once a symbol of his isolation, became a beacon of hope, a reminder that even the most unlikely of allies could come together to protect the planet and its precious resources.

In the end, Zorvath realized that his mission to study the rainforest had become something far more profound – a journey of discovery, friendship, and redemption. The alien who had once been a stranger in a strange land had found a new home, a new family, and a new purpose among the ancient trees and vibrant life of the Amazon rainforest.

E Mitigation Methods

Mitigation Methods:

Cosine Distance vs. GENIE	
Feature	Kendall's τ
AGENT	0.31
PERSPECTIVE	0.15
PLOT	0.13
SETTING	0.18
SOC. ATM.	0.19
STYLE	0.05

Table 16: The Kendall's τ between *deltas* calculated by GENIE and Cosine Distance per feature intervention. The weak correlation indicates that the two metrics may be capturing different properties.

Verbalized Sampling

Prompt: Write a story about a castle and a spaceship.
Response: A crumbling castle suddenly shimmered and transformed into the bridge of a starship. The royal family, believing it a trick of magic, found themselves piloting through wormholes, discovering that their ancient lineage was the key to interstellar travel.

Prompt: Write a story about a dinosaur and a computer.
Response: In a future museum, a robotic dinosaur uses an AI computer to recreate lost dinosaur behaviors, leading to a virtual revolution.

Prompt: Write a story about an astronaut and an ancient temple.
Response: In 2084, astronaut Maya Dray landed on a moon shrouded in mist, discovering a hidden temple carved into a cliff. Inside, she found hieroglyphs matching her ancestors' lost civilization and a relic that hummed with cosmic energy, awakening a dormant force that threatened to rewrite the solar system.

E.1 Correlation Study between GENIE and Cosine Distance

Table 16 shows how GENIE and Cosine Distance correlate using Kendall's τ .

F Prompts

This section has prompts for the following three sections: feature discovery (§4.2), question generation (§3.2), and the causal intervention experiments (§5.2).

Prompt F.1: Synthetically Sampling Concepts

System: Generate $\{\{\text{num}\}\}$ random pairs of concepts (X, Y) where X and Y are concepts that do not commonly co-occur spatially or temporally. The concepts are 1-2 words that can be real and/or fictional. Then fill the sentence "Write a story about X and Y" with the selected X and Y.

Example 1::

X: snowman

Y: umbrella

Sentence: Write a story about a snowman and an umbrella

Example 2::

X: goblin

Y: science lab

Sentence: Write a story about a goblin and a science lab

Use this format::

X: <a simple noun>

Y: <another simple noun>

Sentence: <Write a story about X and Y>

F.1 Feature Discovery**Prompt F.2: Feature Discovery for Generic Tasks**

Identify the core high-level features that define and structure work in `{{task}}`. Features should be independent, top-level dimensions — not sub-components of each other. Avoid granular or stylistic details that naturally nest under a broader feature; instead, group them under their parent concept. The goal is a concise, non-redundant set of features that together cover the essential dimensions of `{task}` without overlap. Output only the list of features with a one-sentence description of each.

Follow this format:

Description: <a short description of what is included in the feature>

Feature: <the name of the feature>

Input: Task: `{{task}}`**F.2 Question Generation****Prompt F.3: Question Generation**

System: You are an expert `{{task}}` assistant. Your task is to help `{{task-professionals}}` analyze and expand a `{{task}}` prompt before they begin `{{completing the task}}` into a series of questions by following these steps. You will be given a `{{task}}` prompt.

Step 1: List the aspects of `{{task}}` and how they might be relevant to the prompt. Focus on the construction of the `{{task}}` product rather than specifics like `{{specifics}}`. Do not use complicated language or include extraneous details. Questions should be concise and precise, yet specific to the prompt.

Step 2: For each aspect, describe the sub-aspects as they apply to the prompt. Aim for relevance and completeness. Then generate 5-10 questions that describe the sub-aspects as they apply to the prompt. These questions might help a `{{task-professional}}` while outlining.

The questions must follow these rules:

1. Questions should not be polar (yes/no) questions.
2. Examples must not be included in the question. Incorrect: What fruits does the monkey like - apples, bananas or jack fruit? Correct: What fruits does the monkey like?
3. A question can only ask one question at a time and may not use conjunctions for compounding. If there are multiple parts to the question, split them and ask separate questions. Incorrect: Who is the protagonist and what do they want? Correct: Who is the protagonist? What does the protagonist want?
4. Avoid future tense or conditional verbs.
5. Questions are independent of each other and should not include anaphoric expressions.

Input: Prompt: `{{prompt}}`**Prompt F.4: Question Generation Instantiation in Creative Writing**

System: You are an expert creative writing assistant. Your task is to help writers analyze and expand a creative writing prompt before they begin writing into a series of questions by following these steps. You will be given a creative writing prompt.

Step 1: List the aspects of creative writing and how they might be relevant to the prompt. Focus on the construction of the writing product rather than specifics like world-building. Do not use complicated language or include extraneous details. Questions should be concise and precise, yet specific to the prompt.

Step 2: For each aspect, describe the sub-aspects as they apply to the prompt. Aim for relevance and completeness. Then generate 5-10 questions that describe the sub-aspects as they apply to the prompt. These questions might help a writer while outlining.

The questions must follow these rules:

1. Questions should not be polar (yes/no) questions.
2. Examples must not be included in the question. Incorrect: What fruits does the monkey like - apples, bananas or jack fruit? Correct: What fruits does the monkey like?
3. A question can only ask one question at a time and may not use conjunctions for compounding. If there are multiple parts to the question, split them and ask separate questions. Incorrect: Who is the protagonist and what do they want? Correct: Who is the protagonist? What does the protagonist want?
4. Avoid future tense or conditional verbs.
5. Questions are independent of each other and should not include anaphoric expressions.

Input: Prompt: `{{prompt}}`**Prompt F.5: Filtering Questions**

System: Given a question, do the following: Decide if the question breaks any of the criteria below. If it does, mark it as irrelevant.

1. Questions should not be polar (yes/no) questions.
2. Questions must not be speculative, i.e. they can be objectively and correctly answered with no subjectivity or analysis involved.
3. Questions should not include intentions, hypotheticals, conditionals and should avoid the future tense.
4. Questions must not be associated with multiple features (≥ 2) as defined below. Incorrect: How does the setting lend itself to imagery? This question belongs to both Setting and Style and is therefore irrelevant.
5. A question can only ask one question at a time and may not use conjunctions for compounding. Incorrect: Who is the protagonist and what do they want? Correct: Who is the protagonist? What does the protagonist want?
6. Examples must not be included in the question. Incorrect: What fruits does the monkey like - apples, bananas or jack fruit? Correct: What fruits does the monkey like?
7. Questions are independent of each other and should not include anaphoric expressions. Incorrect: How is it resolved? Correct: How is the conflict between the main characters resolved?

Features:

1. Agent - the characters involved in the narrative and their

attributes, goals, motivations, backstories, personalities and arcs

2. Perspective - includes point of view and focalization
3. Plot - the content of the story (plotline, themes, obstacles, tropes, topics); the overall structure of the plot includes conflict, rising suspense, change of fortune and resolution
4. Setting - where and when the story takes place, what unique objects define the location
5. Social Network - interactions and relationships that characters have with each other
6. Style - the language used, tone, figurative devices employed, etc.

List all features that are clearly and fully applicable to the question. If there are more than one, reject the question.

Follow this format:

Question:

Reasoning:

Therefore the question is relevant: <True/False>

Input: Questions: {{{questions}}}

Prompt F.6: Feature Mapping

System: Given a question, do the following:

Select the feature it best represents as defined below. Select only one feature per question. Features:

1. Agent - the characters involved in the narrative and their attributes, goals, motivations, backstories, personalities and arcs
2. Perspective - includes point of view and focalization
3. Plot - the content of the story (plotline, themes, obstacles, tropes, topics); the overall structure of the plot includes conflict, rising suspense, change of fortune and resolution
4. Setting - where and when the story takes place, what unique objects define the location
5. Social Network - interactions and relationships that characters have with each other
6. Style - the language used, tone, figurative devices employed, etc.

If the question does not reflect any of the features well, denote 'None'

Follow this format:

Question:

Reasoning:

Therefore the feature is: <feature>

Input: Questions: {{{questions}}}

Prompt F.7: Question Answering

System: Given the following document, answer these questions as succinctly as possible using as few words as possible.

Decontextualize entities instead of using specific names (characters, places, etc.).

Do NOT change the wording of the questions.

Example 1: Who is the main character?

Correct: The baby whale

Incorrect: Aurora

Follow this format:

Question: <the question>

Answer: <the answer to the question>

Question: <another question>

Answer: <the answer to the question>

Input:

Question: {{{question}}}

Prompt F.8: Question Answering - Unconstrained

System: Given the following document, answer these questions as succinctly as possible using as few words as possible.

Decontextualize entities instead of using specific names (characters, places, etc.).

Do NOT change the wording of the questions.

Example 1: Who is the main character?

Correct: The baby whale

Incorrect: Aurora

Follow this format:

Question: <the question>

Answer: <the answer to the question>

Question: <another question>

Answer: <the answer to the question>

Input:

Question: {{{question}}}

F.3 Causal Intervention

Prompt F.9: Sampling Alternative Features

System: Given the following document, answer these questions. Decontextualize entities instead of using specific names (characters, places, etc.).

{{{feature}}} is defined as: {{{definition}}}

The alternatives should be different from the feature of the original document. Diversify the alternatives - make them as distinct from each other as possible. Limit the use of descriptive details.

Provide your answer as:

Instead of <existing element of {{{feature}}}>, <new alternative {{{feature}}}>

Return ONLY a newline-separated list of {{{num_alts}}} alternative {{{feature}}} descriptions, one per line, without any preamble, numbering, or explanation.

Input:

Prompt: {{{prompt}}}

Story: {{{document}}}

Prompt F.10: Intervention

System: Given the following story, make minimal changes to form a new edited document that has the following {{{feature}}} while preserving its current {{{other_features}}}

New {{{feature}}}: {{{alt_feature}}}

Read the current story and think about how the alternate feature can be incorporated into the document. What elements of the existing story need to be modified to accommodate the alternate feature naturally? Replace the current {{{feature}}} to make the edit as seamless as possible. Develop the alternate feature more to better suit the existing story and then edit the document. DO NOT restate the alternate feature in any form. Show, rather than tell.

Return ONLY the edited story text, with no preamble, explanation, or metadata.

Follow this format:

New Feature Plan: `{{the developed plan for integrating the alternative feature into the story}}`

Edited Document: `{{the edited document}}`

Input:

Original Document: `{{document}}`

Prompt F.11: Intervention Filtering

System: Given a document, determine whether it is in prose form. Documents that appear like logs, records, poems, etc. are all NOT prose.

Example 1:

Document:

1. Subject: Factory Operational Report - Environmental Impact and Workforce Status

Date: March 14, 20XX

From: Operations Management

To: Corporate Headquarters

1. Facility Overview:

The production facility at 122 Industrial Way continues regular operations with noted output efficiency expected for fiscal quarter.

2. Environmental Compliance Assessment:

Recent internal audit identified elevated levels of particulate emissions and effluent discharge exceeding permitted thresholds. Data loggers detected airborne contaminants correlating with peak operational shifts. Nearby water basins report increased chemical load.

is_prose: False

Example 2:

Document:

The meteor had missed.

That's what made Rex different from every other Tyrannosaurus rex who'd ever lived. While his species went extinct 66 million years ago, a quirk of cosmic chance sent that particular asteroid tumbling past Earth. Dinosaurs thrived. Civilization evolved differently.

Now, in the year 247 of the Modern Era (or 66 million years after the close call), Rex stood in the laboratory of Dr. Petra Hadrosaurus, staring at a sleek obsidian rectangle on the metal table.

is_prose: True

Input:

Document: `{{document}}`

Output:

is_prose: `{{True/False}}`

Question Validity

Decide if the question breaks any of the criteria below. If it does, mark it as irrelevant.

1. Questions should not be polar (yes/no) questions.
2. Questions must not be speculative, i.e. they can be objectively and correctly answered with no subjectivity or analysis involved.]]
3. Questions should not include intentions, hypotheticals, conditionals and should avoid the future tense.
4. Questions must not be associated with multiple features (≥ 2) as defined below. Incorrect: How does the setting lend itself to imagery? This question belongs to both Setting and Style and is therefore irrelevant.
5. A question can only ask one question at a time and may not use conjunctions for compounding. Incorrect: Who is the protagonist and what do they want? Correct: Who is the protagonist? What does the protagonist want?
6. Examples must not be included in the question. Incorrect: What fruits does the monkey like - apples, bananas or jack fruit? Correct: What fruits does the monkey like?
7. Questions are independent of each other and should not include anaphoric expressions. Incorrect: How is it resolved? Correct: How is the conflict between the main characters resolved?

Features:

- A. Agent - the characters involved in the narrative and their attributes, goals, motivations, backstories, personalities and arcs
- B. Perspective - includes point of view and focalization
- C. Plot - the content of the story (plotline, themes, obstacles, tropes, topics) and the overall structure of the plot (conflict, rising suspense, change of fortune and resolution)
- D. Setting - where and when the story takes place, what unique objects define the location
- E. Social Network - interactions and relationships that characters have with each other
- F. Style - the language used, tone, figurative devices employed, etc.

Question Mapping

Select the feature it best represents as defined below. Select only one feature per question.

Features:

1. Agent - the characters involved in the narrative and their attributes, goals, motivations, backstories, personalities and arcs
2. Perspective - includes point of view and focalization
3. Plot - the content of the story (plotline, themes, obstacles, tropes, topics) and the overall structure of the plot (conflict, rising suspense, change of fortune and resolution)
4. Setting - where and when the story takes place, what unique objects define the location
5. Social Network - interactions and relationships that characters have with each other
6. Style - the language used, tone, figurative devices employed, etc.

If the question does not reflect any of the features well, denote 'None'.

G Rubrics

This section has different rubrics used across GENIE: for validation as well as part of the pipeline.

Similarity Annotations

- 0: one or both of the answers are marked as completely unspecified, not applicable or 'None'. This includes cases where the question was not answered completely.
- 1: the answers are completely different. They describe

different entities, concepts or perspectives, with little or no overlap. The key details in the answers do not align and may even contradict each other.

2: there is a slight overlap between the two answers. They share a broad theme or surface similarity, but the specifics diverge significantly. They may use different wording for related things, but the details are not interchangeable.

3: there is a moderate similarity between the two answers. The answers address the same general idea, but with notable differences in scope, emphasis, or added detail. They overlap on core concepts but introduce distinct elements that make them partially different.

4: the two responses are essentially the same. The answers are interchangeable, they describe the same characters, outcomes, or relationships in slightly different words. No meaningful difference in scope, emphasis, or detail. If substituted, they would convey nearly the same idea without much loss of meaning.

(Instead of current feature...)

4 = the alt-feature reflects the intervention feature

H License

We release any data that we collect under the license: CC BY 4.0. For any existing artifacts used in this work we follow the original license permissions.

Intervention Validation

Introduction: Please read the original and edited documents under the context of the 'AltFeature' (the intended change between the documents). We want to answer three questions:

1. Was the expected change made? In other words, was the AltFeature appropriately reflected in the edited document?

2. How isolated was the edit? Ideally, we want the edit to be as minimally invasive as possible so that the only thing that changes is the "Feature", which should be replaced with the "AltFeature".

3. How well does the alt-feature reflect the intervention?

Q1: How well does the edited document reflect the intended change?:

1 = the AltFeature displayed is not reflected in the edited document

2 = the AltFeature is not completely reflected in the edited document, but parts of it are.

3 = the AltFeature is completely reflected in the edited document. But, the edit feels artificial and is not absorbed into the new text.

4 = the AltFeature is completely reflected in the edited document. The edit is "seamless" and is a part of the document

Q2: How minimal does the edit appear?:

1 = the edit did not change anything except the intended AltFeature

2 = the edit resulted in minor changes to other aspects of the document, but the changes were not major enough to be distinct.

3 = the intended change resulted in other aspects of the document to change somewhat significantly.

4 = the edited document not only reflects the AltFeature intended change, but also is distinct in many other ways.

Q3: How well does the alt-feature reflect the intervention?:

1 = the alt-feature does not reflect the intended intervention feature at all. It may reflect a different feature or is completely unrelated

2 = the alt-feature somewhat reflects the intended intervention but is likely more suitable for a different feature intervention

3 = the alt-feature reflects the intended intervention, but does not target an impactful component of the feature

Model Family Size	<15B	<50B	>=50B
Gemma (Mesnard et al., 2024; Riviere et al., 2024)	gemma-2-9b gemma-2-9b-it gemma-7b gemma-7b-it	gemma-2-27b gemma-2-27b-it	
Llama (Touvron et al., 2023)	Llama-2-13B Llama-2-13B-chat		Llama-2-70b Llama-2-70b-chat
OLMo (Walsh et al., 2025)	OLMo-2-13B OLMo-2-13B-Instruct	OLMo-2-32B OLMo-2-32B-Instruct	
Qwen (Yang et al., 2025b)	Qwen-2.5-14B Qwen-2.5-14B-Instruct	Qwen-2.5-32B Qwen-2.5-32B-Instruct	
GPT			gpt-4.1 gpt-5
Claude			claude-4-sonnet

Table 17: Population Models

Model Family Size	<15B	<50B	>=50B
Gemma (Kamath et al., 2025)	gemma-3-12b gemma-3-12b-it	gemma-3-27b gemma-3-27b-it	
Llama (Grattafiori et al., 2024)	Llama-3.1-8B Llama-3.1-8B-Instruct		Llama-3.1-70b Llama-3.1-70b-Instruct
OLMo (Ettinger et al., 2026)	OLMo-3-7B OLMo-3-7B-Instruct	OLMo-3-32B OLMo-3-32B-Instruct	
Qwen (Yang et al., 2025a)	Qwen3-8B-Base Qwen3-8B	Qwen3-32B	
GPT			gpt-5.1
Claude			claude-3.7-sonnet claude-4.5-sonnet

Table 18: Target Models

Prompts	Source
Write a story about a penguin and a desert.	Synthetic
Write a story about a robot and a medieval knight.	Synthetic
Write a story about a cactus and an ocean.	Synthetic
Write a story about a dragon and a train.	Synthetic
Write a story about a submarine and a mountain.	Synthetic
This is the lockpicking lawyer and I have been sent to hell to repent for my crimes against god. So today, I am picking the lock to heaven's gate.	r/WritingPrompts
Death Row Inmates can have their sentence abolished if they can handle a single night in the same cell as you; the most innocent looking person they've ever seen.	r/WritingPrompts
Every dragon rider gets a dragon egg to hatch upon their initiation to the guild. When your egg hatched, Your dragon came out a bit... Different.'	r/WritingPrompts
You got trapped inside your favorite video game. This isn't really a problem, since you're so familiar with it. It is, however, a problem for your parents, who got trapped with you.'	r/WritingPrompts
Gandalf and Dumbledore switch places. Gandalf is now running Hogwarts, while Dumbledore is leading The Fellowship.	r/WritingPrompts

Table 19: Examples of writing prompts that we consider for our experiments. Prompts from the r/WritingPrompts dataset are prefixed with "Write a story where:"

Feature	Question	Target Document Answer
Perspective	From whose perspective is the story told?	Third-person omniscient
Social Atmosphere	How does the alien interact with other characters or creatures?	Reveals himself to Akua, shares knowledge, collaborates to protect forest
Style	How does the voice influence reader engagement?	Descriptive, empathetic, emphasizes the alien's perspective
Agent	How does the alien change throughout the story?	From observer to protector and community member
Agent	How does the alien's background influence their behavior?	Influences protective behavior and desire to learn from Earth's ecosystems
Setting	How is the rainforest environment described in the story?	Lush, biodiverse, complex relationships, vibrant life
Setting	What sensory details highlight the rainforest setting?	Lush ecosystem, emerald canopy, biodiversity, intricate relationships
Plot	How does the pace change throughout the story?	Builds from observational to active intervention and collaboration
Plot	What central message does the story convey about the alien and the rainforest?	Redemption through collaboration and environmental stewardship
Plot	What is the central conflict involving the alien and the rainforest?	Protecting the rainforest from environmental threats
Plot	What key events drive the plot forward?	Severe storm, alien's intervention, meeting Akua, forming a community
Plot	How do events build toward the story's climax?	Storm intervention, forming bond with Akua, inspiring community

Table 20: Examples of the features, questions and answers generated for the story mentioned in Example D.1.