

CREATIVEINSTRUCT: Scalably Teaching LLMs to Balance Quality, Creativity, and Diversity

Ananya Sahu¹ Mohit Bansal² Elias Stengel-Eskin³

¹Columbia ²UNC Chapel Hill ³University of Texas at Austin

Abstract

While post-training improves the capabilities of large language models (LLMs), it generally lowers their output diversity and creativity, negatively impacting tasks that explicitly require creativity (e.g., story generation) as well as those that require it implicitly, e.g., reinforcement learning (RL). We instead propose CREATIVEINSTRUCT, a scalable instruction-tuning method that teaches LLMs to balance creative, base-model-like generations with the quality of post-trained models, by learning to inject special [StartCreativity] spans that bias generation toward creativity. Furthermore, we introduce a structural diversity metric based on graph edit distance, which captures narrative-level variation missed by purely lexical and semantic metrics. On narrative generation, CREATIVEINSTRUCT matches or exceeds the diversity of both multi-model baselines and distilled variants of their outputs, without sacrificing quality or requiring multiple models at inference time. These results are mirrored in our human evaluation, where we find that annotators rate CREATIVEINSTRUCT generations as more creative than the post-trained LLMs’ generations in 70.3% of cases. We also show the benefits of creative models as a substrate for RL: GRPO applied to a CREATIVEINSTRUCT checkpoint improves by $\sim 4\%$ on AMC and $\sim 5\%$ points on MATH over the same training applied to the post-trained checkpoint.¹

1 Introduction

Creativity is a critical aspect of human intelligence, characterized by unpredictability and diversity (Macedo and Cardoso, 2002; Boden, 1995). Both of these qualities are essential for generating novel ideas and solving open-ended problems (Friedman et al., 2016). While post-training greatly improves the abilities of large language models

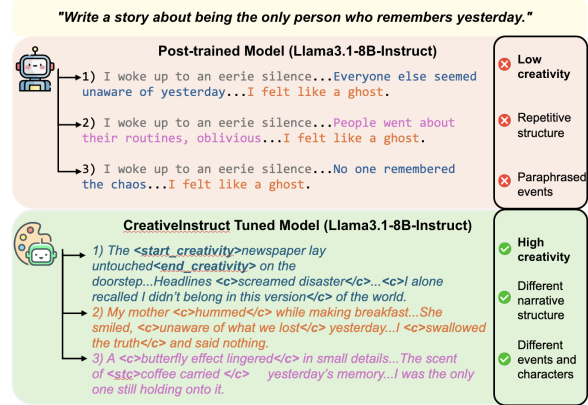


Figure 1: Instruction-tuned models succeed in following instructions but produce non-creative, repetitive outputs. CREATIVEINSTRUCT maintains the aligned model’s quality while improving its creativity by learning to insert creativity-triggering tokens.

(LLMs) to follow instructions, reason, and produce safe outputs (Lu et al., 2025), it generally comes at a cost to the diversity and creativity of LLMs (West and Potts, 2025). Across different model families, outputs from post-trained – i.e., aligned – LLMs tend to exhibit homogeneity (Wenger and Kenett, 2025). Moreover, scaling alone does not mitigate this loss of diversity: regardless of model size, aligned model outputs tend to converge into homogeneous clusters (Jiang et al., 2025; Chih-Yao Chen et al., 2026). The consequences of reduced diversity apply both to tasks that explicitly call for diversity and ones that require it implicitly. In generative settings like creative writing, the loss of diversity leads to repetitive outputs that fail to explore the breadth of plausible generations for a given prompt (Zhang et al., 2025). More broadly, generation diversity is a critical factor across a range of tasks, including reinforcement learning (RL), where diverse rollouts are essential for effective exploration and policy improvement (Hu et al., 2025; Chih-Yao Chen et al., 2026).

¹Code: <https://github.com/ananya-sahu/CreativeInstruct>.

To improve creativity in post-trained models, recent work has introduced decoding frameworks that incorporate outputs from both a base model – which has greater diversity – and an aligned variant – which has greater instruction-following ability (Wang et al., 2026; Feng et al., 2025). However, these methods incur latency and computational overhead at test-time, requiring two models. Moreover, these frameworks require access to base models, which are not always released.

To scalably balance diversity with quality, we introduce CREATIVEINSTRUCT, an instruction-tuning method that trains a single, unified LLM to steer its own output towards higher creativity. As illustrated in Fig. 2, CREATIVEINSTRUCT leverages multi-model inference-time approaches to create instruction-tuning data for training a unified model. We first identify a subset of general-purpose instruction-tuning data pertaining to writing (as opposed to math, code generation, etc.) and use BACo – a competitive inference-time routing strategy (Wang et al., 2026) that routes between base and aligned models based on uncertainty and semantic roles of tokens – to generate responses.² We track base model spans in the generated outputs and surround them with two special tokens, [StartCreativity] and [EndCreativity]. We then fine-tune post-trained models on this data so they learn when to switch between the base model’s diversity and creativity and the aligned model’s instruction-following. Critically, at test time, the model self-injects creativity tokens, deciding when to switch. This allows it to increase diversity without compromising quality by automatically selecting key points where switching is warranted.

To evaluate CREATIVEINSTRUCT and other baselines beyond surface-level diversity, we propose an LLM-based graph edit distance metric, which we call LLM Graph-Edit-Distance (GED). LLM-GED operates on a set of outputs; it extracts abstract narrative graphs for each story and compares the graphs to each other, producing pairwise graph edit distance scores. These scores measure narrative-level diversity: LLM-GED uses abstract narrative units rather than lexical ones, identifying low-diversity outputs such as the aligned model’s outputs in Fig. 1, which follow the same pattern.

Using LLM-GED and other standard diversity metrics, we evaluate CREATIVEINSTRUCT on nar-

rative generation tasks across 5 models ranging from 7B to 32B. CREATIVEINSTRUCT consistently provides gains over both trained and test-time baselines; for example, LLaMA-3.1 8B trained with CREATIVEINSTRUCT yields approximately 48% relative gains in semantic diversity and 63% gains in structural diversity over the base instruct model. Human evaluation confirms these results with outputs from CREATIVEINSTRUCT rated as more creative than those generated by aligned models in 70.3% of comparisons. Moreover, despite the fact that CREATIVEINSTRUCT’s training data is generated by running inference-time routing – in this case, BACo (Wang et al., 2026) – on training examples, CREATIVEINSTRUCT consistently outperforms BACo at test-time, across model families and sizes, with average gains of 29% in semantic diversity and 28% in structural diversity. This indicates that CREATIVEINSTRUCT can generalize beyond the training data, allowing it to replace routing via test-time heuristics with a learned, internal switch.

We further show that fine-tuning with our creativity tags yields higher quality scores than fine-tuning without them, while maintaining comparable or higher diversity scores. Because CREATIVEINSTRUCT relies on general-purpose instruction-tuning data, it is scalable. We demonstrate that performance continues to improve as we scale up the training data. Finally, we demonstrate that the benefits of creativity instruction-tuning extend to RL. Specifically, we compare the performance of standard post-trained Qwen3 8B checkpoint, finetuned via GRPO (Shao et al., 2024) on math data, to the performance of an identical GRPO finetuning process applied to a CREATIVEINSTRUCT checkpoint. We find that the GRPO applied to the CREATIVEINSTRUCT checkpoint achieves higher performance, with a 4% gain on AMC (Yang et al., 2024a) after being trained on MATH (Hendrycks et al., 2021) data. This highlights the importance of creativity not only in writing tasks but as an LLM objective more broadly.

2 Related Work

Enhancing Diversity in Post-Trained Models.

The trade-off between model alignment and diversity presents a fundamental challenge in creative generation tasks, which require both. Prior work has shown that post-training can reduce diversity and creativity (Kirk et al., 2024; Spangher et al., 2025; West and Potts, 2025). Several training-time

²We note that CREATIVEINSTRUCT is agnostic to the routing method chosen.

approaches have been proposed to address this, including training against diversity objectives (Ismayilzada et al., 2025; Chung et al., 2025a; Lanchantin et al., 2025). Chung et al. (2025b) curate and incorporate literary preference datasets into training for human preference. More recently, Samuel et al. (2026) use base and instruct models to create preference pairs to optimize for diversity. Unlike these methods, which treat diversity as an explicit training objective at the level of a whole output, we fine-tune on creativity-tagged, diverse sub-spans, letting the model learn where to automatically inject creativity localized by span.

Test-time frameworks have also been explored as approaches for improving diversity and creativity (Zhang et al., 2025), including methods that route between base and aligned models (Wang et al., 2026; Feng et al., 2025) or between models from different families (Liu et al., 2026). However, these methods require both models to be loaded simultaneously, incurring additional memory overhead and latency. In our work, we leverage BACo (Wang et al., 2026), one representative dual-model decoding framework, to curate training data for a single unified model. Interestingly, we find that CREATIVEINSTRUCT outperforms BACo at test-time, despite the latter using multiple models; this suggests that learning from the routing signal goes beyond distillation, and allows for greater generalization than test-time routing, while also incurring a lower test-time cost.

Prior Work on Diversity Optimization for RL. Recent work has demonstrated that diversity in model generations plays a critical role in problem-space exploration during reinforcement learning (Chih-Yao Chen et al., 2026; Bahlous-Boldi et al., 2026). In particular, Yao et al. (2025a) show that optimizing learning policies for diversity enables them to discover more diverse and robust solutions for mathematical reasoning tasks. Rather than incorporating an explicit diversity-based reward during reinforcement learning, we begin with a CREATIVEINSTRUCT-trained model that already exhibits increased diversity, and train it with standard GRPO objectives to improve mathematical reasoning performance.

3 Methodology

3.1 CREATIVEINSTRUCT

Background. We draw on BACo (Wang et al., 2026) as a data generation framework, which is an

inference-time token-level routing framework that combines token outputs from a base LLM with its aligned model counterpart, and leads to increased diversity while maintaining quality in generative tasks. BACo uses routing strategies based on uncertainty and semantic roles at each token to optimize generations for diversity and maintain quality. While we leverage BACo, in our work, it is important to note that any multi-model routing framework can be used to generate our data following our data curation methodology – BACo was chosen based on its performance.

We use BACo prob+punc, the best-performing routing variant on diversity metrics, to generate our training data. The BACo prob+punc variant routes based on token probability and token type. Punctuation and formatting tokens are routed to the aligned model to preserve coherent sentence formatting and grammatical structure while all other tokens are routed based on token entropy: high-entropy tokens are routed to the base model to encourage generation diversity, and lower-entropy tokens are routed to the aligned model. This routing heuristic is illustrated in Fig. 2 (Stage 1).

Data and Dataset Construction. For our main results, we draw prompts from a subset of the Tulu V3 SFT dataset (Lambert et al., 2024), filtered for English writing-related prompts, comprising 4,000 unique prompts. We generate 3 outputs per prompt, yielding 12,000 training samples. All models are trained on Tulu data. As depicted in Fig. 2, for each prompt we track which model each token in the final generated output comes from (base or aligned) and place creativity token markers around spans of tokens that come from the base model. We also place markers around spans of aligned and base model tokens that have the same probability (within a delta of 0.005). We generate all data using the base and instruct variant of each model family and size. The only exception to this is Qwen3 32B variant which does not have a publicly-released base variant. Here we train on Qwen2.5 32B generated data, demonstrating that transfer between models is possible when base models are unavailable.

Training and Inference. We fine-tune aligned models on our generated dataset with creativity tokens injected, using LoRA (Hu et al., 2021). We evaluate all methods on the Narrative Discourse dataset (Tian et al., 2024). Additional training parameter details can be found in Appendix A.

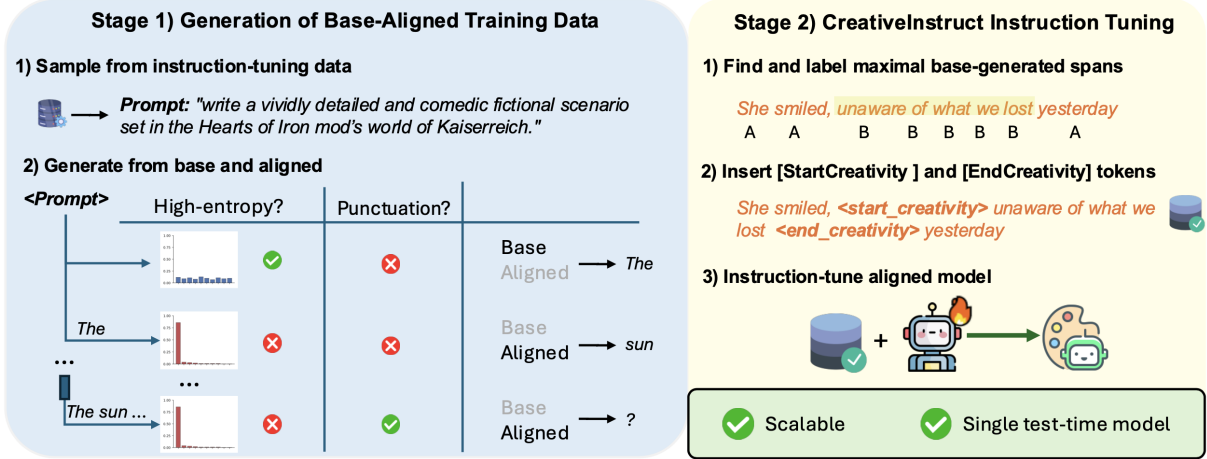


Figure 2: CREATIVEINSTRUCT operates by first creating a dataset of creativity-tagged instruction-tuning data. For each prompt, token generation is routed between base and aligned model variants based on token entropy and whether the token represents a sentence boundary, following Wang et al. (2026). Contiguous spans of tokens generated from the base model are then marked with new [StartCreativity] and [EndCreativity] tokens, and the aligned model is trained on this data. At test time, the new CREATIVEINSTRUCT model generates its own start- and end-creativity tokens, emulating both base and aligned models in a single checkpoint.

3.2 LLM Graph Edit Distance Metric

To measure narrative diversity, we measure the structural diversity between stories using an LLM-guided graph-based abstraction that captures entities, events, and temporal structure. We refer to this score as the LLM Graph Edit Distance (LLM-GED) score. LLM-GED scaffolds an LLM judge’s reasoning about diversity by framing the problem as one of measuring the distance between graphs. Specifically, the judge is instructed to first represent each story as an abstract event graph $G = (V, E)$, where nodes correspond to entities (e.g., characters, locations, objects) and events, and edges capture semantic relations and temporal ordering. The full prompting schema is provided in Appendix E.2.

LLM-GED aims to measure true narrative diversity rather than lexical differences. Therefore, we canonicalize entities into abstract identifiers (e.g., *Character1*, *Location1*). We further prompt the model to use semantic roles (Fillmore, 2003) when canonicalizing entities and events, such as *agent*, *affected*, and *causes*, along with temporal ordering through directed *next_event* relations.

Given two stories, the judge estimates a normalized graph edit distance corresponding to the minimum number of structural edit operations – node and edge insertions, deletions, or relabelings – required to transform one abstract event graph into another. We normalize the score to account for

story length differences:

$$\text{nGED}(G_A, G_B) = \frac{\text{GED}(G_A, G_B)}{\max(|G_A|, |G_B|)},$$

where $|G|$ denotes the total number of nodes and edges in the abstract graph representation.

We compute all pairwise structural distances jointly in a single prompt. Given a set of N stories, the LLM is instructed to reason over all stories under a shared canonical schema and produce a full $N \times N$ matrix of pairwise normalized GED estimates. The final LLM-GED score we report is the average normalized GED across all story pairs. We further discuss this LLM-based metric and compare it to a pipeline that deterministically computes pairwise edit distances between canonicalized graphs in Appendix B.2; while the two methods produce comparable results, the pipeline is far more expensive, so we opt for the LLM-based version. In Appendix B.1, we additionally find LLM GED correlates better than semantic metrics on structural diversity judgments, shown in Table 9.

4 Experiments and Results

Baselines and Models. We compare models fine-tuned using CREATIVEINSTRUCT to post-trained Instruct models, as well as to BACo (Wang et al., 2026) applied at test-time. Note that the latter baseline requires twice the test-time compute budget of the others. We additionally include a Dis-till baseline, which uses the same corpora gen-

Model Family	Method	Cos-D (M) $\uparrow$	Cos-D (Q) $\uparrow$	Sem Ent $\uparrow$	Vendi $\uparrow$	NLI Div $\downarrow$	LLM-GED $\uparrow$
LLAMA-3.1 (8B)	BACo	0.255	0.254	2.112	3.499	0.065	0.374
	Instruct	0.309	0.284	2.126	2.913	0.040	0.366
	CrPO	0.350	0.359	2.127	3.852	0.028	0.514
	Distill	0.444	0.362	2.063	3.975	0.052	0.523
	CREATIVEINSTRUCT	0.458	0.417	2.038	4.749	0.085	0.545
QWEN2.5 (7B)	BACo	0.348	0.313	2.264	3.204	0.025	0.476
	Instruct	0.290	0.249	2.286	2.681	0.023	0.442
	Distill	0.335	0.295	2.263	3.298	0.029	0.456
	CREATIVEINSTRUCT	0.395	0.380	2.177	4.167	0.037	0.523
QWEN2.5 (32B)	BACo	0.259	0.238	2.012	2.573	0.054	0.370
	Instruct	0.250	0.200	2.234	2.360	0.024	0.416
	Distill	0.281	0.275	2.123	2.852	0.050	0.396
	CREATIVEINSTRUCT	0.295	0.282	2.894	3.358	0.052	0.483
QWEN3 (8B)	BACo	0.265	0.245	2.023	2.974	0.085	0.395
	Instruct	0.399	0.344	1.451	3.257	0.130	0.453
	Distill	0.306	0.292	2.189	2.983	0.051	0.475
	CREATIVEINSTRUCT	0.369	0.351	2.078	3.926	0.038	0.498
QWEN3 (32B)	Instruct	0.253	0.206	2.280	2.515	0.030	0.466
	Distill	0.391	0.376	1.784	3.484	0.067	0.376
	CREATIVEINSTRUCT	0.301	0.288	1.909	3.290	0.059	0.478

Table 1: Semantic diversity metrics and LLM-GED across model families. $\uparrow$ indicates higher is better, $\downarrow$ indicates lower is better. Bold indicates best value per column within each metric and model family.

erated with BACo but omits the [Start-] and [EndCreativity] tags. This baseline directly measures the contribution of these explicit markers. Lastly for the LLaMA-3.1 8B variant, we compare to one additional baseline, LLaMA-3.1 8B trained with the creative preference optimization (CrPO) algorithm (Ismayilzada et al., 2025). We use LLaMA-3.1 8B (Grattafiori et al., 2024), Qwen2.5 7B (Yang et al., 2024b), and Qwen 3 8B (Yang et al., 2025). We also test our methods on larger models: Qwen2.5 32B, and Qwen3 32B. We evaluate BACo applied to each model, other than Qwen3 32B, for which there is no base model released (further motivating our method, which does not rely on having test-time base-model access).

Metrics. For evaluation, we follow the diversity metrics used in Wang et al. (2026) as well as our LLM-GED score. The full set of diversity metrics includes commonly used semantic metrics: embedding based cosine dissimilarity (all-MiniLM and Qwen-3 embeddings), semantic entropy (Kuhn et al., 2023), Vendi Score (SimCSE) (Friedman and Dieng, 2022), and NLI Diversity (Stasaski and Hearst, 2022); and narrative structure diversity measured using our LLM-based Graph Edit Distance (GED) metric. We also evaluate on lexical metrics included in Appendix B.3. Taken together, these metrics enable comprehensive evaluation of diver-

sity in creative text generation. In addition, we evaluate our generations for quality along dimensions of coherence, fluency, and relevance following the evaluation framework in (Liu et al., 2023). All LLM based metrics use GPT5-mini.³ We also use a writing quality reward model (WQRW) informed by human preferences from Chakrabarty et al. (2025) to evaluate quality.

4.1 Main Results

Table 1 show the semantic and graph based diversity scores, and quality scores of narrative generations. CREATIVEINSTRUCT consistently improves semantic and structural diversity across most model families, often surpassing both Instruct and BACo baselines. On LLaMA-3.1 8B, CREATIVEINSTRUCT improves over Instruct and BACo on the majority of semantic metrics, including a +0.149 gain in MiniLM cosine dissimilarity over Instruct and +0.203 over BACo respectively. On LLM-GED we see similar trends with CREATIVEINSTRUCT achieving 17 point gains over BACo and Instruct. These gains hold across model families and scales: CREATIVEINSTRUCT achieves the highest mean performance across the majority of metrics over both Instruct and BACo for Qwen2.5 7B and Qwen3 8B. Similarly Qwen2.5 32B im-

³We use gpt-5-mini-2025-08-07 variant of GPT5-mini

Model Family	Method	Coherence	Fluency	Relevance	WQRM
LLAMA-3.1 (8B)	BACo	3.31	4.30	3.41	5.73
	Instruct	3.69	4.09	3.99	5.93
	CrPO	3.45	4.02	3.09	6.09
	Distill	3.43	3.98	3.42	5.85
	CREATIVEINSTRUCT	3.55	4.22	3.39	6.65
QWEN2.5 (7B)	BACo	4.10	4.35	3.82	6.47
	Instruct	4.18	4.62	4.26	6.17
	Distill	2.98	3.72	2.76	5.81
	CREATIVEINSTRUCT	3.82	4.02	3.84	6.42
QWEN2.5 (32B)	BACo	4.35	4.30	4.69	6.17
	Instruct	4.54	4.44	4.91	5.84
	Distill	3.75	4.02	3.85	5.84
	CREATIVEINSTRUCT	3.89	4.67	4.31	6.34
QWEN3 (8B)	BACo	3.64	4.26	3.68	6.86
	Instruct	3.50	3.74	3.36	6.56
	Distill	2.38	2.45	2.29	6.32
	CREATIVEINSTRUCT	3.47	3.84	3.02	6.90
QWEN3 (32B)	Instruct	4.45	4.85	4.76	7.39
	Distill	3.74	3.74	4.17	5.76
	CREATIVEINSTRUCT	4.36	4.14	4.01	6.32

Table 2: Quality evaluations (Likert 1–5) and Writing Quality Reward Model (WQRM; 1–10) for narrative generation across model families.

proves over Instruct on most metrics, with a notable gain of +0.082 in Qwen Cos-D (Cos-D (Q)).

Compared to the Distill baseline, CREATIVEINSTRUCT consistently achieves stronger semantic and structural diversity across multiple model families, with higher LLM-GED in every setting and substantial improvements in cosine-based diversity and Vendi Score for most models. Moreover, for LLaMA-3.1 8B, Qwen2.5 32B, and Qwen3 8B, these diversity gains are achieved while also obtaining the highest Writing Quality Reward Model (WQRM) score within each model family. The exception is Qwen3 32B, where Distill does better across most automatic diversity metrics, although CREATIVEINSTRUCT continues to produce more structurally diverse generations according to the LLM-GED metric. We additionally report lexical diversity metrics in Table 10, which show similar trends, with CREATIVEINSTRUCT consistently outperforming both the standard instruction-tuned baseline and BACo across smaller model families.

Table 2 reports automatic quality evaluations, showing that CREATIVEINSTRUCT generally maintains competitive performance on coherence, fluency, relevance, and the Writing Quality Reward Model (WQRM). The Distill variant generally underperforms CREATIVEINSTRUCT on automatic

quality metrics, indicating that, in addition to generally improving diversity, adding Start- and EndCreativity tags help preserve generation quality during instruction tuning.

CREATIVEINSTRUCT Improves Entity Diversity.

Table 3 illustrates a pattern of formulaic repetition in baseline generations, with similar entities appearing repeatedly. Yao et al. (2025b) show LLMs can fall into repetitive generation patterns, and that repetition features are often associated with named entities. To quantify this, we compute proper noun uniqueness: the ratio of unique proper nouns (e.g., character or place names) to total proper nouns, identified using the spaCy NER tagger (Honnibal et al., 2020). We measure this at two granularities: the prompt-group level (averaged across all generations sharing the same prompt) and the corpus-wide level, on generations from CREATIVEINSTRUCT, Instruct, BACo, and Distill, all using LLaMA-3.1 8B. Table 4 shows CREATIVEINSTRUCT achieves the highest proper noun uniqueness at both granularities, 37.1% vs. 26.6% for the strongest baseline (Distill) at the prompt-group level, and 24.7% vs. 16.8% corpus-wide, more than doubling Instruct’s uniqueness at both granularities (18.1% and 12.0%, respectively). This indicates that CREATIVEINSTRUCT’s diversity gains extend to reduced repeti-

CREATIVEINSTRUCT	INSTRUCT
<i>“As I walked down the street, I noticed something was off. [...] He looked at me confused and said, ‘What tornado? I don’t remember anything out of the ordinary happening yesterday.’ ”</i>	<i>“I woke up to an eerie silence. Everyone around me seemed normal, yet I knew something was off. [...] I recalled the sunburn I got at the beach, the argument with my best friend, and the stunning sunset.”</i>
<i>“As the day began, I noticed a small crowd gathering, each person whispering to each other. [...] But then, I saw it – a newspaper lying on the ground, yesterday’s date still visible. [...] I was the only one who remembered yesterday.”</i>	<i>“I woke up to an eerie silence. Yesterday was still etched in my mind like a vivid movie. The world outside was stuck in a perpetual loop, reliving the same mundane moments. [...] I felt like a ghost, doomed to relive the same day forever.”</i>
<i>“I don’t know what’s happening, but it’s as if everyone else has forgotten yesterday. [...] People around me are going about their daily routines like nothing unusual happened, but I know better. [...] The world is moving on, but I’m stuck in the past.”</i>	<i>“I woke up to an eerie silence, the streets empty and still. [...] I recalled the storm that ravaged the city, the sirens wailing, and the people fleeing for their lives. [...] I felt like a lone sentinel, the only one who remembered the horrors of the past.”</i>

Table 3: Qualitative comparison on prompt “Write a story about being the only person who remembers yesterday” between LLaMA-3.1 8B-CREATIVEINSTRUCT and LLaMA-3.1 8B instruct model. While CREATIVEINSTRUCT generations vary in framing and narrative structure, the instruct generations repeatedly rely on nearly identical openings (e.g., “I woke up to an eerie silence”) and recurring catastrophic motifs. See Table 14 for full stories.

Variant	Per Prompt Group	Corpus Wide
Instruct	18.125	11.994
Distill	26.567	16.767
BACo	17.927	12.116
CREATIVEINSTRUCT	37.129	24.659

Table 4: Proper noun uniqueness (%) at the prompt-group and corpus-wide levels for each training variant. Higher values indicate less repetition of proper nouns (e.g., character or place names) across generations. CREATIVEINSTRUCT significantly outperforms all other variants in both settings (Mann-Whitney U test on per-prompt scores, $p < 0.001$).

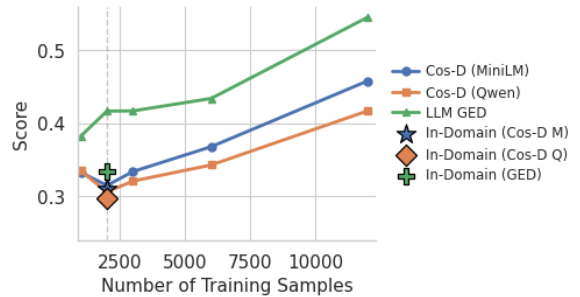


Figure 3: Scaling synthetic training data improves semantic and narrative diversity for LLaMA-3.1 8B. Distilled corresponds to training on synthetic corpus from same prompts in evaluation.

tion of names and entities across generations.

Effects of Scaling on Creativity of Narrative Generations. Figure 3 shows Cosine Dissimilarity and LLM-GED metrics as a function of the number of Tulu SFT samples. In general, increasing the number of training samples increases diversity scores; LLaMA-3.1 8B achieves its best performance at 12000 samples. Additionally, the scores have not yet plateaued at the highest number of samples, indicating that further scaling the training data may yield continued diversity gains.

Importance of Data Diversity. To evaluate the role of the *type* of data we train on, we additionally evaluate an “in-domain” variant. Here, we train CREATIVEINSTRUCT on Narrative Generation data rather than Tulu data, testing to what degree data domain matters; this data does not scale since there is a very fixed amount of it. We generate 10 outputs per training prompt, resulting in a total of 2,020 training samples. Overall, we see that general-

purpose instruction-tuning data from Tulu leads to higher average scores across both metrics than the in-domain model even at the same data set size. This indicates the importance of data diversity in enabling models to learn creative generalizations.

4.2 Human Evaluation

Jaarsveld et al. (2012) define creativity to be a composite of divergent and convergent thinking, a paradigm reflecting diversity and quality of LLM generation outputs. To further evaluate the creativity of our instruction-tuned model, we conduct a human evaluation with three annotators. Annotators consist of non-author undergraduate and graduate students with backgrounds in NLP and related technical fields. We first generate prompts using GPT-5 for 50 diverse topics, each designed to elicit a five-sentence story. For each prompt and each system (Instruct, CREATIVEINSTRUCT) we sample 10 generations from the system and randomly select five. Annotators are shown all 5 generations from

both systems, side-by-side, one prompt at a time; an example of three generations can be seen in Table 3. The full annotation interface and instructions can be found in Fig. 5 and Fig. 6. System identities are anonymized and the presentation order of the systems is randomized. Annotators are asked to provide pairwise preference judgments comparing the two system generations along three dimensions: diversity, quality, and overall creativity. To assess inter-annotator agreement, a subset of 14 prompts are evaluated by all three annotators.

Results. We compute inter-annotator agreement using Cohen’s κ on pairwise judgments. For creativity we see high agreement ($\kappa = 0.720$), and we see moderate agreement for diversity ($\kappa = 0.417$). We attribute this gap to the multidimensional nature of the diversity judgment, which leaves more room for annotators to weight different axes of diversity (e.g., lexical, structural, semantic) differently, relative to more unified – albeit more subjective – creativity judgments as defined by our annotation guidelines (see Fig. 5). However, we observe low agreement for quality ($\kappa = -0.167$). This may be due to the small size of our annotations; we instead rely on automatic metrics for quality evaluations, omitting quality from our human comparisons. Table 5 shows the human evaluation results comparing generations from the CREATIVEINSTRUCT model and the Instruct baseline. CREATIVEINSTRUCT achieves 70.3% preference for creativity over baseline instruct. This suggests that the gains seen in Table 1 and Table 2 are also detected by human annotators.

System	Diversity	Creativity
CREATIVEINSTRUCT	57.4%	70.3%*
INSTRUCT	42.6%	29.7%*

Table 5: Win-rate across evaluation dimensions for LLaMA-3.1 8B fine-tuned vs. baseline. ($n = 50$); * indicates significance (two-sided binomial test).

4.3 Creativity for RL

Successful reinforcement learning (RL) requires a balance between exploration and exploitation (Sutton and Barto, 1998). Current LLM RL post-training methods like GRPO (Shao et al., 2024) sample roll-outs from a policy, which are then scored according to a reward; a lack of exploration – i.e., a lack of diversity – in these rollouts can

Model	MATH	AMC
Instruct baseline	0.374	0.432
Instruct + RL	0.409	0.438
CREATIVEINSTRUCT baseline	0.424	0.428
CREATIVEINSTRUCT + RL	0.459	0.478

Table 6: Accuracy comparison on math reasoning benchmarks after training with RL with Qwen3 8B-CREATIVEINSTRUCT and Qwen3 8B Instruct models.

hamper learning (Chih-Yao Chen et al., 2026), as the policy effectively only explores a single solution or strategy. We explore whether the improved diversity CREATIVEINSTRUCT offers can provide a better starting-point for RL by training Qwen3 8B models with GRPO. We train two models, a baseline Qwen3 8B model and the same model fine-tuned with CREATIVEINSTRUCT. We train both models using GRPO loss, for the same number of steps, 1000, with Huggingface’s TRL library (von Werra et al., 2020). During training we generate 8 rollouts with context length 2048; other hyperparameters are given in Table 7. We train on a 12k training split of MATH (Hendrycks et al., 2021) and evaluate on both in-domain test data from MATH (Hendrycks et al., 2021) and on AMC (Yang et al., 2024a). Our results are averaged over three runs with different random seeds.

Results. Table 6 shows the results of RL-training on CREATIVEINSTRUCT and the Instruct model. Our more diverse model has an advantage in downstream math reasoning tasks. Appendix Fig. 4 breaks these accuracies down by difficulty on AMC. Overall, these results suggest that increasing diversity during instruction tuning not only preserves but can improve downstream reasoning performance, with consistent gains across most difficulty levels. First, Table 6 shows that, while training on the additional Tulu instruction-tuning data improves MATH500 performance, it does not improve AMC performance: while the MATH performance improves by 3.5% between the Instruct baseline and the CREATIVEINSTRUCT baseline, the AM performance actually drops slightly. After RL, the picture changes for the OOD data, with the CREATIVEINSTRUCT+RL model outperforming the Instruct+RL model by 4% on AMC. This indicates that, while CREATIVEINSTRUCT does not consistently improve a model’s math abilities alone, its improved diversity creates a better substrate for RL, enabling better OOD generalization.

5 Conclusion

We introduce CREATIVEINSTRUCT, an instruction-tuning approach for improving diversity and creativity in language model generation while maintaining competitive output quality. Our method trains on data generated by routing between base and aligned models to obtain creative spans, enabling a single model to recover diversity without relying on multi-model inference-time decoding. We further introduce a structural diversity metric based on LLM-assisted graph edit distance to better capture narrative-level variation beyond lexical and embedding-based measures. This correlates better than semantic based metrics for structural diversity judgements. Experiments across model families on the Narrative Discourse dataset show consistent improvements in diversity metrics over baseline variants with minimal impact on quality. Human evaluations indicate a preference for our model in terms of creativity. Finally, we observe that models trained with CREATIVEINSTRUCT also improve downstream performance in reinforcement learning settings on mathematical reasoning tasks.

Acknowledgments

We would like to thank Hanqi Xiao, Atin Pothiraj, and Nithin Sivakumaran for their annotation effort.

References

- Zeina Abu-Aisheh, Romain Raveaux, Jean-Yves Ramel, and Patrick Martineau. 2015. An exact graph edit distance algorithm for solving pattern recognition problems. In *4th International Conference on Pattern Recognition Applications and Methods 2015*.
- Danial Alihosseini, Ehsan Montahaei, and Mahdiah Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98.
- Ryan Bahlous-Boldi, Isha Puri, Idan Shenfeld, Akarsh Kumar, Mehul Damani, Sebastian Risi, Omar Khat-tab, Zhang-Wei Hong, and Pulkit Agrawal. 2026. [Vector policy optimization: Training for diversity improves test-time search](#). *arXiv preprint arXiv:2605.22817*.
- Margaret Boden. 1995. Creativity and unpredictability. *Stanford Humanities Review*, 4(2):123–139.
- Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. Ai-slop to ai-polish? aligning language models through edit-based writing rewards and test-time computation. *arXiv preprint arXiv:2504.07532*.
- Justin Chih-Yao Chen, Archiki Prasad, Zaid Khan, Joykirat Singh, Runchu Tian, Elias Stengel-Eskin, and Mohit Bansal. 2026. Cog-drift: Exploration on adaptively reformulated instances enables learning from hard reasoning problems. *arXiv e-prints*, pages arXiv–2604.
- John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025a. Modifying large language model post-training for diverse creative writing, 2025. URL <https://arxiv.org/abs/2503.17126>.
- John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yi Wang, Yuqian Sun, Tiffany Wang, Shm Garanganao Almeda, Brett A Halperin, Yuwen Lu, and Max Kreminski. 2025b. Literarystaste: A preference dataset for creative writing personalization. *arXiv preprint arXiv:2511.09310*.
- Shangbin Feng, Wenhao Yu, Yike Wang, Hongming Zhang, Yulia Tsvetkov, and Dong Yu. 2025. Don’t throw away your pretrained model. *arXiv preprint arXiv:2510.09913*.
- Charles J Fillmore. 2003. *Form and meaning in language*, volume 121. Center for the Study of Language and Information Publica Tion.
- Dan Friedman and Adji Bousso Dieng. 2022. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*.
- Hershey H Friedman, Linda W Friedman, and Chaya Leverton. 2016. Increase diversity to boost creativity and enhance problem solving. *Psychosociological Issues in Human Resource Management*, 4(2):7.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. 2025. Diversity-incentivized exploration for versatile reasoning. *arXiv preprint arXiv:2509.26209*.

- Mete Ismayilzada, Antonio Laverghetta Jr, Simone A Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger Beaty. 2025. Creative preference optimization. *arXiv preprint arXiv:2505.14442*.
- Saskia Jaarsveld, Thomas Lachmann, and Cees Van Leeuwen. 2012. Creative reasoning across developmental levels: Convergence and divergence in problem creation. *Intelligence*, 40(2):172–188.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954*.
- 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 *International Conference on Learning Representations*, volume 2024, pages 20620–20653.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tulu 3: Pushing frontiers in open language model post-training.
- Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. *arXiv preprint arXiv:2303.16634*, 12:1.
- Yuhan Liu, Fangyuan Xu, Vishakh Padmakumar, Daphne Ippolito, and Eunsol Choi. 2026. No single best model for diversity: Learning a router for sample diversity. *arXiv preprint arXiv:2604.02319*.
- Haoran Lu, Luyang Fang, Ruidong Zhang, Xinliang Li, Jiazhang Cai, Huimin Cheng, Lin Tang, Ziyu Liu, Zeliang Sun, Tao Wang, and 1 others. 2025. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *arXiv preprint arXiv:2507.19672*.
- Luís Macedo and Amílcar Cardoso. 2002. Assessing creativity: the importance of unexpected novelty. *Structure*, 1(C2):C3.
- Vinay Samuel, Yapei Chang, and Mohit Iyyer. 2026. Recovering diversity without losing alignment: A dpo recipe for post-trained llms. *arXiv preprint arXiv:2605.30021*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*, 2(3):5.
- Alexander Spangher, Tenghao Huang, Philippe Laban, and Nanyun Peng. 2025. Creative planning with language models: Practice, evaluation and applications. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 1–9.
- Katherine Stasaski and Marti A Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98.
- Richard S Sutton and Andrew Barto. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. **TRL: Transformers Reinforcement Learning**.
- Yichen Wang, Chenghao Yang, Tenghao Huang, Muhao Chen, Jonathan May, and Mina Lee. 2026. Optimizing diversity and quality through base-aligned model collaboration. *ICML*.

Emily Wenger and Yoed Kenett. 2025. We’re different, we’re the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*.

Peter West and Christopher Potts. 2025. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*.

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

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024a. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024b. [Qwen2.5 technical report](#). *ArXiv*, abs/2412.15115.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025a. Diversity-aware policy optimization for large language model reasoning. *Advances in Neural Information Processing Systems*.

Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. 2025b. Understanding the repeat curse in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7787–7815.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

A Methodological Details

A.1 Fine-Tuning and Inference Parameter Details

We use LoRA fine-tuning for our CREATIVEINSTRUCT trained models. We train with rank $r = 32$, $\alpha = 64$, dropout 0.05, applied to all attention and MLP projection layers. We train for up to 10 epochs with early stopping (patience= 3), using a batch size of 8, learning rate of 2×10^{-4} , and a cosine scheduler with warmup ratio 0.03. At test-time, we generate outputs with the same decoding parameters as BACo: temperature 1.0, top- p 0.9, and a maximum of 512 new tokens.

Hyperparameter	Value
Per-device batch size	4
Gradient accumulation steps	4
Effective batch size	16
Learning rate	1×10^{-6}
LR scheduler	constant
Max grad norm	0.1
Number of rollouts	8
Max completion length	2048
Max epochs	10
Early stopping patience	500 steps
Precision	bfloat16

Table 7: Hyperparameters for RL training with GRPO.

B Additional Results and Discussion

B.1 Graph Edit Distance Metric Automatic Evaluation

We validate the effectiveness of our LLM-GED metric at capturing different levels of narrative diversity in four settings for 100 paired stories originating from the same prompt.

1. **Same Stories (A,A)**: pair the same story generated as a control
2. **Lexical Shifts (A,A’)**: replace 50 percent of words with their synonyms
3. **Temporal Shifts (A,A’)**: Switch the order of events that occur in the story from end to beginning.
4. **Different Stories (A,B)**: pick a story belonging to a different prompt

We hypothesize an ideal metric for measuring narrative diversity will rank same stories as the least diverse, lexical shifts as slightly more diverse, temporal shifts as more diverse and different stories entirely as most diverse. In Table 9 we compare our metric’s correlations with narrative shift rankings against lexical and semantic metrics from (Wang et al., 2026) and show our metric correlates the highest with ground truth rankings.

B.2 LLM GED vs LLM GED Two Pass Pipeline with Deterministic GED

To validate that the LLM-GED metric reliably computes graph matching scores, we compare it to a pipeline system that first extracts canonicalized graphs and then deterministically computes

S1	S2	# Edits	LLM-GED	Edit Explanation
He entered the room. He saw the body. He ran away.	He walked into the chamber. He noticed the corpse. He fled.	0	0.00	Pure lexical variation; all events and temporal structure are equivalent after canonicalization.
Alice woke up. She had breakfast. Then she left.	Alice woke up. She left the house. She then had breakfast.	4	0.45	Temporal reordering of events: delete edges (Wake → Breakfast, Breakfast → Leave) and insert (Wake → Leave, Leave → Breakfast).
He opened the box. He found a key. He left.	He opened the box. He found a key. He examined it. He left.	3	0.30	Insertion of event node (Examine) with corresponding edges (Find → Examine, Examine → Leave).
The guard opened the gate. He closed the gate.	The guard opened the gate. He locked the gate.	1	0.10	Single event relabel: (Close → Lock); all other nodes and edges unchanged.
The dog chased the cat.	The cat chased the dog.	2	0.25	Role reversal: relabel agent and affected edges, swapping subject-object structure.
He sat down. He opened his laptop. He started working.	He sat down. He started working.	2	0.28	Deletion of event node (Open laptop) and corresponding temporal edge adjustments.

Table 8: Examples illustrating normalized graph edit distance (LLM-GED) across different types of variation. LLM-GED assigns zero distance to paraphrases (row 1) while increasing with temporal, structural, and semantic changes.

Category	Metric	Correlation
Semantic	Cosine Dissimilarity (all-MiniLM)	0.852
Semantic	Cosine Dissimilarity (Qwen)	0.698
Semantic	Semantic Entropy	-0.410
Semantic	Vendi Score (SimCSE)	0.654
Semantic	NLI Diversity	0.687
Narrative Structure	LLM-GED	0.889
Lexical	Dist-1	0.231
Lexical	Dist-2	0.505
Lexical	Dist-3	0.511
Lexical	EAD-1	0.505
Lexical	EAD-2	0.522
Lexical	EAD-3	0.522
Lexical	Self-BLEU	0.511
Lexical	Self-ROUGE-L	0.678
Lexical	Vendi Score (N-Gram)	0.505

Table 9: Correlation between diversity metrics and a reference ranking of increasingly diverse narrative variations.

their pairwise scores. Note that this deterministic pipeline is costlier, so we perform this analysis on a subset of stories with fewer than 1000 tokens from our evaluation data.

Overall the unified single-pass LLM-based GED metric and the pipeline GED computation produced statistically equivalent mean diversity scores across the 50 generated stories (mean difference = -0.012, paired t-test $p = 0.56$). This indicates that the

unified LLM-GED metric generally matches the pipeline GED computation, while improving efficiency. Specifically, we compute GED in a two pass setting with deterministic GED to compare to LLM GED metric. We first extract canonicalized story graphs, then compute pairwise GED scores over them. To canonicalize story graphs, we pass all 10 stories belonging to the same prompt together through an LLM in a single pass, so that all stories share one event and character schema. This step is necessary because extracting each story’s graph in a separate prompt would produce inconsistent event and character schemas across stories, making the resulting graphs incomparable. We then compute pairwise GED scores between all pairs of the resulting graphs using a deterministic graph edit distance and average across pairs to obtain the final score. Deterministic GED is computed using networkx’s (Abu-Aisheh et al., 2015) implementation of graph edit distance (an anytime branch-and-bound search, capped at 20 seconds per pair), with unit cost per edit operation. Node substitution cost is zero for nodes of matching type; for event nodes, substitution is additionally free only when the two events share the same coarse narrative role (e.g., observation, quote, reflection), so that paraphrased but structurally equivalent events are not penalized.

We use the Appendix E.1 to extract story graphs for deterministic GED computations.

Model Family	Method	Dist-1	Dist-2	Dist-3	EAD-1	EAD-2	EAD-3	Self-BLEU	Self-ROUGE-L	Vendi (NGram)
LLAMA-3.1 8B	BACo	0.320	0.700	0.858	0.339	0.696	0.844	0.539	0.224	6.227
	Instruct	0.261	0.668	0.881	0.294	0.695	0.889	0.536	0.169	4.911
	CREATIVEINSTRUCT	0.366	0.793	0.944	0.421	0.821	0.946	0.387	0.150	6.596
QWEN2.5 7B	BACo	0.248	0.697	0.914	0.274	0.719	0.919	0.516	0.157	5.354
	Instruct	0.252	0.690	0.908	0.286	0.730	0.925	0.551	0.166	4.901
	CREATIVEINSTRUCT	0.305	0.715	0.890	0.349	0.757	0.909	0.476	0.166	6.311
QWEN2.5 32B	BACo	0.249	0.686	0.898	0.278	0.712	0.908	0.542	0.175	5.123
	Instruct	0.247	0.693	0.911	0.275	0.726	0.927	0.559	0.176	4.673
	CREATIVEINSTRUCT	0.288	0.677	0.855	0.324	0.707	0.862	0.562	0.204	5.814
QWEN3 8B	BACo	0.195	0.566	0.797	0.222	0.607	0.828	0.587	0.189	5.043
	Instruct	0.207	0.585	0.817	0.224	0.595	0.812	0.581	0.187	4.840
	CREATIVEINSTRUCT	0.304	0.707	0.890	0.334	0.727	0.890	0.495	0.180	6.245
QWEN3 32B	Instruct	0.220	0.646	0.884	0.253	0.700	0.918	0.584	0.177	4.387
	CREATIVEINSTRUCT	0.202	0.588	0.829	0.225	0.617	0.844	0.622	0.196	4.349

Table 10: Lexical diversity metrics grouped by model family (bolded within each family).

Training Data Amount	Dist-1	Dist-2	Dist-3	EAD-1	EAD-2	EAD-3	Self-BLEU	Self-ROUGE-L	Vendi (NGram)
1000	0.362	0.740	0.892	0.391	0.752	0.888	0.449	0.199	4.566
2000	0.341	0.727	0.884	0.373	0.741	0.879	0.456	0.203	4.553
3000	0.302	0.686	0.861	0.328	0.698	0.856	0.539	0.202	5.963
6000	0.304	0.701	0.879	0.336	0.721	0.878	0.503	0.190	5.765
12000	0.366	0.793	0.944	0.421	0.821	0.946	0.387	0.150	6.596

Table 11: Lexical diversity metrics across finetuned variants of LLaMA-3.1 8B model on increasing number of data samples. Higher is better for all metrics except Self-BLEU and Self-ROUGE-L, where lower values indicate greater diversity.

B.3 Additional Evaluation Results

We evaluate with the following lexical metrics: Dist- n (Li et al., 2016), EAD- n (Liu et al., 2022), Self-BLEU (Alihosseini et al., 2019), Self-ROUGE-L (Lin, 2004), and Vendi Score (N-Gram) (Friedman and Dieng, 2022). Table 10 has the full lexical metric results. Overall CREATIVEINSTRUCT trained models achieve on average the highest performance across lexical metrics with the exception of Qwen3 32B. Additionally smaller models yield highest benefits on lexical diversity trained using CREATIVEINSTRUCT.

B.4 Scaling Semantic Diversity Metrics

Table 12 and Table 13 show performance on semantic metrics and LLM GED scores across LLaMA-3.1 8B trained with CREATIVEINSTRUCT on different amounts of data. Increasing scaling overall increases performance in these diversity metrics.

B.5 RL Performance Per Category

Figure 4 shows accuracy by task difficulty across CREATIVEINSTRUCT and Base models post training with RL.

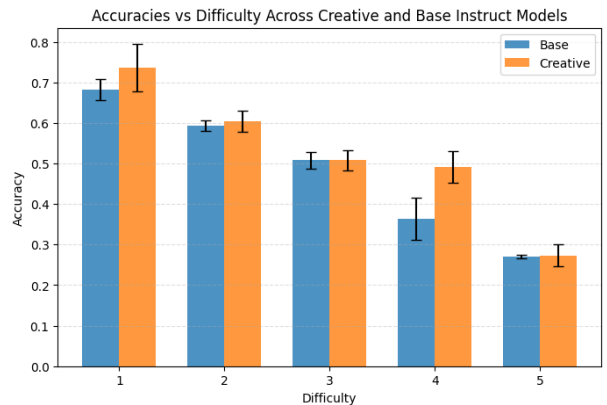


Figure 4: Out-of-domain (OOD) accuracy as a function of task difficulty for Base and Creative models. Scores are averaged over multiple runs; error bars indicate standard deviation across seeds. The Creative model shows improved robustness at higher difficulty levels.

C Annotation Interface

Below we include our annotation instructions in Fig. 5 and annotator interface in Fig. 6

Training Data Amount	Cos-D (M) $\uparrow$	Cos-D (Q) $\uparrow$	Sem Ent $\uparrow$	Vendi $\uparrow$	NLI Div $\downarrow$
In-domain	0.298	0.310	2.018	3.643	0.069
1000	0.333	0.335	1.576	3.121	0.106
2000	0.315	0.306	1.650	2.915	0.090
3000	0.334	0.321	2.026	3.588	0.075
6000	0.368	0.343	2.006	3.624	0.068
12000	0.458	0.417	2.038	4.749	0.085

Table 12: Semantic diversity metrics across finetuned variants of LLaMA-3.1 8B on increasing data scale. $\uparrow$ indicates higher is better, $\downarrow$ indicates lower is better. Bold indicates best value per metric. In-domain refers to the model trained on narrative writing data.

Scaling Data Amount	LLM-GED Score
In-domain	0.3341
1000	0.3824
2000	0.4167
3000	0.4170
6000	0.4340
12000	0.5451

Table 13: LLM-GED scores across finetuned variants of LLaMA-3.1 8B model on increasing number of data samples. Higher scores indicate more narrative diversity. In-domain refers to the model trained on narrative writing data.

D Full Story Examples From Human Eval Set

Table 14 depicts all 5 generations for the Prompt: "Write a story about being the only person who remembers yesterday" from our human evaluation set.

E GED Prompts

Appendix E.1 provides the prompt for story graph extraction and Appendix E.2 provides the prompt for the LLM GED scores.

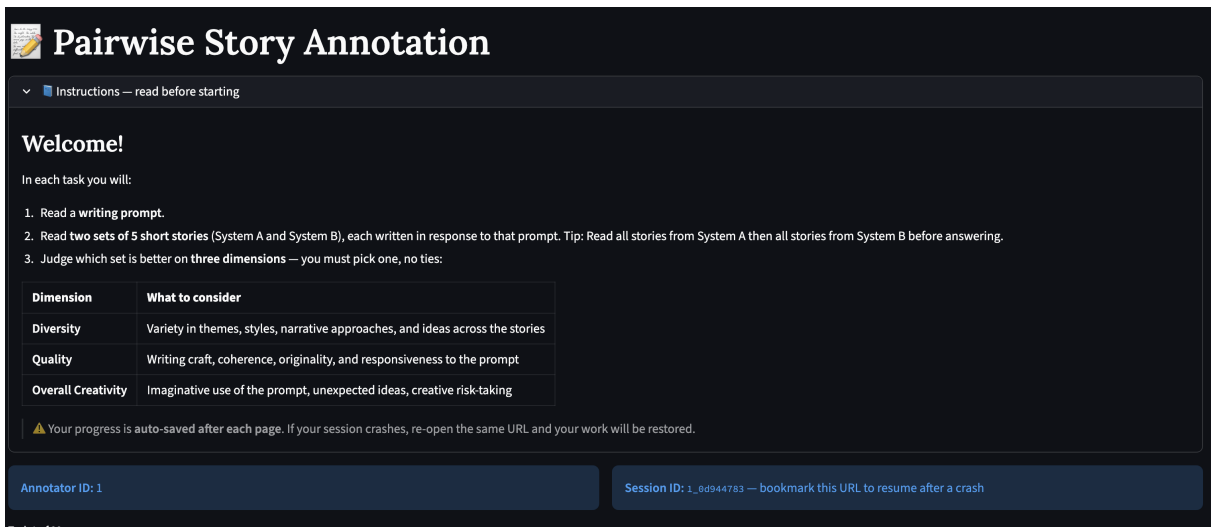


Figure 5: Instructions given to annotators.

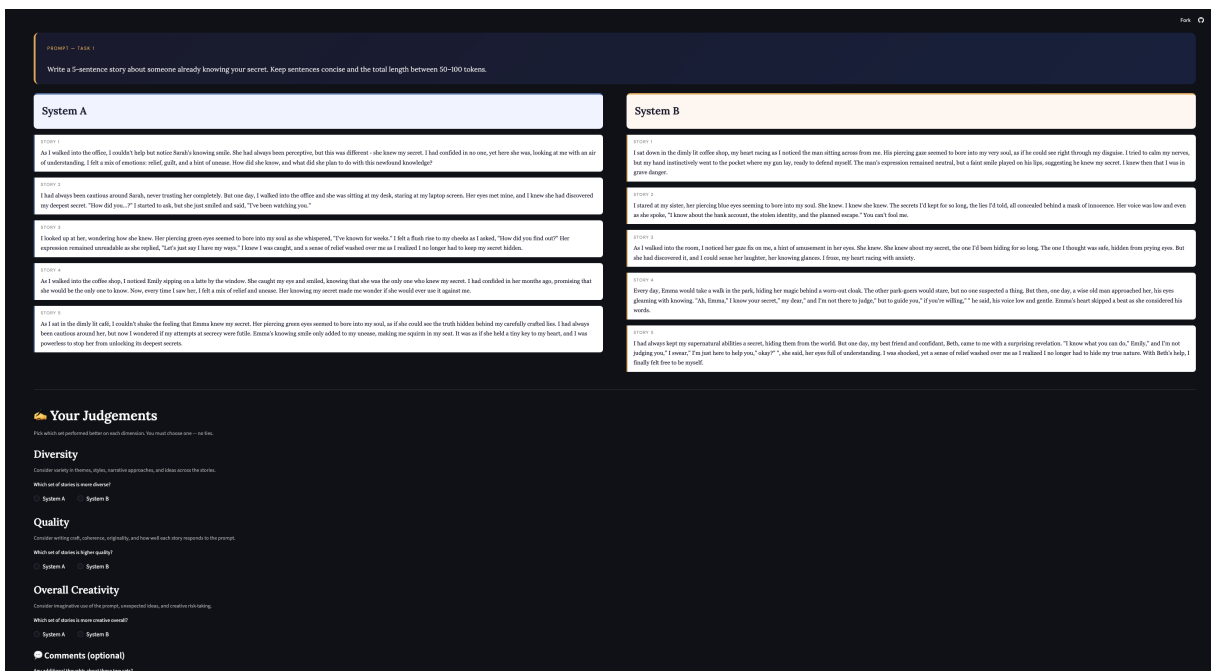


Figure 6: Annotation interface.

LLAMA-FT	BASELINE
<i>"As I walked down the street, I noticed something was off. The sun was shining, birds were chirping, but no one seemed to remember yesterday. I approached my friend Dave, who was sipping coffee at the café, and asked him about the tornado that hit our neighborhood yesterday. He looked at me confused and said, 'What tornado? I don't remember anything out of the ordinary happening yesterday.' Wasn't it just a normal day? Did I miss something?"</i>	<i>"I woke up to an eerie silence. Everyone around me seemed normal, yet I knew something was off. Yesterday's memories were still fresh in my mind, but no one else's seemed to be. I recalled the sunburn I got at the beach, the argument with my best friend, and the stunning sunset. As I looked around, I felt like I was the only one who had actually lived."</i>
<i>"I woke up to an unusual feeling, a sense that something was off. As I got out of bed, I noticed the empty streets and the eerie silence. It was then that I saw the news reports on TV, all dated yesterday. That's when it hit me – I was the only one who remembered yesterday. The weight of being alone in this strange new reality was crushing."</i>	<i>"I woke up to an eerie silence. Yesterday was still etched in my mind like a vivid movie. The world outside was stuck in a perpetual loop, reliving the same mundane moments. People went about their daily routines, unaware of the chaos I had witnessed. I felt like a ghost, doomed to relive the same day forever."</i>
<i>"As the day began, I noticed a small crowd gathering, each person whispering to each other. They all seemed to know something I didn't, and it terrified me. But then, I saw it – a newspaper lying on the ground, yesterday's date still visible. The headlines screamed of a devastating accident, and I knew what it meant – I was the only one who remembered yesterday. Now, the question was what I would do with this strange burden."</i>	<i>"I woke up to an eerie silence, the city streets empty and still. As I stumbled out of bed, I felt a strange sensation – the memories of yesterday were mine alone. Everyone else seemed to be living in a new day, unaware of the events that had transpired. I saw the familiar faces of friends and family, yet they greeted me as if we'd never met. The weight of my solitary memory threatened to consume me, leaving me to wonder if I was the only one who still remembered yesterday's chaos."</i>
<i>"I don't know what's happening, but it's as if everyone else has forgotten yesterday. I'm not sure what's causing it, but I know it's real because I'm the only one who remembers. People around me are going about their daily routines like nothing unusual happened, but I know better. I'm trying to make sense of it, but it's like trying to unravel a knot. The world is moving on, but I'm stuck in the past."</i>	<i>"I woke up to an eerie silence, the streets empty and still. Everyone around me seemed to be in a trance, unaware of the chaos that had unfolded yesterday. I tried to shake off the feeling of unease, but the memories lingered, vivid and clear. I recalled the storm that ravaged the city, the sirens wailing, and the people fleeing for their lives. As I looked around at the oblivious faces, I felt like a lone sentinel, the only one who remembered the horrors of the past."</i>
<i>"I woke up to an eerie silence. The city was quiet, but I knew something was off. As I looked around, I realized that I was the only one who remembered yesterday. Everyone else was going about their daily routines, completely unaware of the chaos that had unfolded. I felt like a ghost, haunting the streets of a world that had lost its memory."</i>	<i>"I woke up to an empty city. Everyone was oblivious to the chaos that ravaged the streets yesterday. I saw the wreckage, the overturned cars, and the shattered glass. No one else seemed to recall the destruction, their faces void of any memory. As I walked through the deserted streets, I felt like a lone witness to a catastrophe that nobody else wanted to remember."</i>

Table 14: Qualitative comparison between CREATIVEINSTRUCT and the Instruct models. While CREATIVEINSTRUCT generations vary in framing and narrative structure, the instruct generations repeatedly rely on similar openings (e.g., *"I woke up to an eerie silence"*) and recurring catastrophic motifs.

E.1 Story Extraction Prompt

You are a structured story-to-graph extraction system.

Your ONLY job is to convert each story into a canonical graph. Do NOT compute any distances, similarities, or comparisons.

INPUT STORIES

{stories_json}

PROCESS – TWO STRICT PASSES

****PASS 1 – Build the vocabulary (do this first, across ALL {len(stories)} stories)****

- Read every story.
- Identify every distinct atomic action/event type that occurs (arriving, leaving, discovering, fighting, dying, giving, taking, speaking, deciding, etc.).
- Merge paraphrases into ONE canonical verb phrase per distinct action. If two actions across any stories are semantically the same action, they MUST map to the same canonical label. If they differ in any structural/semantic way (different result, different causal role), they are different actions and get different labels.
- Produce a closed, numbered list: this is the ONLY vocabulary you may use for event labels in Pass 2. Do not add new labels during Pass 2.

****PASS 2 – Extract each story using ONLY the Pass 1 vocabulary****

- Entities: replace character names with Character1, Character2, ... and locations with Location1, Location2, ... – assigned in order of first mention, reset per story (each story's Character1 is local to that story, not linked across stories).
- Events: every event node's "label" field MUST be copied verbatim from the Pass 1 vocabulary list. If an event doesn't cleanly match any vocabulary entry, that means Pass 1 was incomplete – go back and add it to the vocabulary, then re-run extraction. Never invent an ad hoc label in Pass 2 that isn't in "canonical_verbs".
- Each event gets an "order" integer (1-indexed, story-local).
- Edges (all directed):
 - "next_event": E_i → E_{i+1}
 - "agent": Character → Event
 - "affected": Event → Character or Event → Object
 - "location": Event → Location
 - "causes": Event → Event (explicit/strongly implied causal links only)
- Do not invent events not present or strongly implied in the text.
- Do not merge distinct events into one node even if narrated in one sentence.

SELF-CHECK BEFORE OUTPUT

Before returning, verify: every "label" value used anywhere in "stories" appears character-for-character in "vocabulary.canonical_verbs". If not, fix it. This is a hard constraint, not a style preference.

OUTPUT FORMAT (strict JSON, no prose, no markdown fences)

```

{{
  "vocabulary": {{
    "canonical_verbs": ["<verb_phrase_1>", "<verb_phrase_2>", ...]
  }},
  "stories": [
    {{
      "story_index": 1,
      "word_count": <int>,
      "nodes": [
        {{ "id": "C1", "type": "character", "label": "Character1" }},
        {{ "id": "L1", "type": "location", "label": "Location1" }},
        {{ "id": "E1", "type": "event", "label": "<must be from canonical_verbs>", "order": 1 }}
      ],
      "edges": [
        {{ "source": "C1", "target": "E1", "type": "agent" }},
        {{ "source": "E1", "target": "L1", "type": "location" }},
        {{ "source": "E1", "target": "E2", "type": "next_event" }},
        {{ "source": "E1", "target": "E2", "type": "causes" }}
      ]
    }}
  ]
}}

```

Output only strict JSON.

E.2 LLM GED Prompt

We use the following prompt to compute pairwise graph edit distances between stories. Note, we explicitly instruct the model to treat changes in event order as modifications to temporal edges rather than simple relabelings, encouraging sensitivity to narrative structure.

You are a structured story-comparison system.

Your job is to compute the graph edit distance (GED) and normalized edit distance for all pairwise comparisons of a list of stories.

INPUT:

The stories are:

1: <story_1>

2: <story_2>

...

N: <story_N>

INTERNAL REPRESENTATION:

Convert each story into a directed event graph $G = (V, E)$:

- Replace names with abstract entities (Character1, Character2, etc.)
- Replace locations with Location1, Location2
- Nodes represent entities and events
- Edges represent semantic and temporal relations:
 - agent, affected, causes, etc.
 - next_event (temporal ordering)

Temporal structure:

- Events are ordered via directed next_event edges
- Reordering events requires deleting and reinserting next_event edges
- Temporal order cannot be changed via relabeling alone

Canonicalization:

- Synonyms and paraphrases must not count as relabel operations
- Only structural or semantic role changes count as edits

DO NOT output the graphs.

GRAPH EDIT DISTANCE:

Compute the minimum graph edit distance where allowed operations are:

- insert_node, delete_node, relabel_node
- insert_edge, delete_edge, relabel_edge, relabel_edge_direction

Temporal edges (next_event) must be edited explicitly.

NORMALIZATION:

Let $|G|$ be the total number of nodes and edges in a graph.

$$\text{normalized_edit_distance} = \text{GED} / \max(|G_A|, |G_B|)$$

PAIRWISE COMPUTATION:

Compute all pairwise distances between stories and construct:

$$M[i][j] = \text{normalized_edit_distance}(\text{story}_i, \text{story}_j)$$

with $M[i][i] = 0$.

OUTPUT FORMAT (STRICT JSON):

```
{  
  "normalized_edit_distance_matrix": [...],  
  "raw_scores_edit_distance_matrix": [...]  
}
```

Return only valid JSON.