CLIN: an Objective Framework for Evaluating Creativity in Short Persian Literary Text

Mohammad Reza Modarres*[1,2], Armin Tourajmehr*[2], Yadollah Yaghoobzadeh†[1,2], Mohammad Taher Pilehvar†[3]
[1]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[2]Tehran Institute for Advanced Studies, Khatam University, Iran
[3]Cardiff University, United Kingdom
Correspondence: mr.modarres@ut.ac.ir

Abstract

Evaluating creativity in large language model (LLM) outputs remains challenging because creativity is multidimensional and human-centered. We examine how reliably LLMs evaluate short literary text in Persian, a low-resource language, across multiple evaluation strategies and prompt formulations. We find that LLM–human agreement varies substantially across dimensions: alignment is stronger for structured TTCT-derived properties such as *Originality*, *Fluency*, and *Elaboration*, but considerably weaker for more subjective dimensions, particularly *Emotion* and *Attractiveness*. Judgments are also sensitive to prompt formulation, while few-shot prompting, ensembling, and multi-agent debate provide no consistent improvement. Motivated by this dimension-dependent behavior, we investigate whether structured creativity dimensions can instead be approximated using simple, interpretable proxy metrics. We introduce CLIN, which evaluates three TTCT-derived dimensions separately using topic-aware novelty for *Originality*, contextual lexical clustering for *Fluency*, and lexical diversity for *Elaboration*. These proxies achieve human alignment comparable to or better than the strongest zero-shot LLM judge in our setting while requiring substantially lower evaluation cost.[1]

1 Introduction

With the rapid advancement of LLMs, creative text generation has received growing attention across applications ranging from story and poetry generation to assistive writing systems (Wen et al., 2023; Zhou et al., 2023; Elzohbi and Zhao, 2023; Yuan et al., 2022). Alongside progress in generation, evaluating creativity has itself become an active research problem, with approaches including psychometric evaluation, LLM-based judging, reference-based comparison, learned evaluators, and automated creativity measures (Chakrabarty et al., 2023; Li et al., 2025; Ismayilzada et al., 2025; Fein et al., 2026; Lu et al., 2026). Despite this progress, substantial disagreement remains over which signals reflect human creativity judgments and when automated evaluations can be trusted.

Creativity is a human-centered and multidimensional construct. Psychometric frameworks such as the Alternative Uses Test (Guilford, 1967, AUT) and Torrance Tests of Creative Thinking (Torrance, 1966, TTCT) address this complexity by separating creativity into complementary dimensions, an approach increasingly adopted in NLP evaluation (Chakrabarty et al., 2023; Tourajmehr et al., 2025). However, evidence on the reliability of LLM creativity evaluators is mixed: model–human agreement varies across criteria (Kim and Oh, 2025; Lyu et al., 2026), reference-based evaluation can improve alignment (Li et al., 2025), specialized evaluators can outperform off-the-shelf LLM judges (Fein et al., 2026), and judgments can be sensitive to relatively minor prompt changes (Lu et al., 2026). We therefore ask how reliably LLM judges recover human assessments across different creativity dimensions and evaluation protocols, and whether structured dimensions can instead be approximated using simple, interpretable measurements.

To address these questions, we conduct a systematic evaluation of LLM-based creativity judgments across multiple experimental settings, including zero-shot evaluation, reference-based comparison, few-shot prompting, ensemble voting, multi-agent debate, and prompt variation. We focus on short literary text written in Persian, a low-resource language that poses additional challenges for both creative generation and evaluation.

Our results show that LLMs achieve moderate alignment with human judgments on relatively structured dimensions such as *Originality*, *Fluency*, and *Elaboration*, but considerably weaker

*,†Equal contribution, equal supervision

[1]Code, data, and additional materials are available at https://smrmodares.github.io/CLiN/.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50879–50893
October 24-29, 2026 ©2026 Association for Computational Linguistics

alignment on more subjective dimensions, particularly *Creativity*, *Attractiveness*, and *Emotion*. More elaborate judging procedures provide no consistent improvement, and LLM-based evaluations remain sensitive to prompt formulation.

These results motivate a different evaluation strategy for the structured dimensions. We introduce CLIN (Creativity as Lexical Ideas, Novelty, and N-grams), an objective and interpretable framework for the TTCT-derived dimensions considered in this work. CLIN evaluates *Originality* through global and topic-relative novelty, *Fluency* through contextual lexical clustering, and *Elaboration* through lexical diversity. Each dimension is evaluated separately rather than collapsed into a single creativity score. Our analysis further shows that these simple, objective, and interpretable proxies can perform comparably to or better than the strongest zero-shot LLM judge in our setting, while requiring substantially lower evaluation cost.

In summary, our contributions are as follows:

- A systematic evaluation of LLMs as creativity judges in short Persian literary text across multiple evaluation strategies and prompt formulations.
- Empirical evidence that LLM-based judgments exhibit limited alignment with human evaluations, particularly for subjective dimensions such as emotion and attractiveness, and remain sensitive to changes in prompt formulation.
- The introduction of CLIN, an objective and interpretable proxy-based framework for evaluating *Originality*, *Fluency*, and *Elaboration*, whose simple lexical and statistical measures achieve alignment with human judgments comparable to or better than the strongest zero-shot LLM judge at substantially lower evaluation cost.

2 Related Work

Creativity as a multidimensional construct. Creativity is commonly treated as a multidimensional construct rather than a single observable property. Psychometric frameworks such as the Alternative Uses Test (Guilford, 1967, AUT), Divergent Association Task (Olson et al., 2021, DAT), and Torrance Tests of Creative Thinking (Torrance, 1966, TTCT) measure complementary aspects of creative behavior, with TTCT distinguishing originality, fluency, flexibility, and elaboration. Recent computational work similarly emphasizes dimensions beyond novelty, including diversity, surprise, quality, and appropriateness (Ismayilzada et al., 2025; Nakajima et al., 2026). Following this view, we evaluate TTCT-derived dimensions separately using dedicated, interpretable proxies.

LLMs as creativity evaluators. LLMs are increasingly used to evaluate creative outputs, but their agreement with humans varies across settings. Prior work reports weak agreement on several creativity dimensions (Chakrabarty et al., 2023), while other studies find stronger consistency but limitations in culturally or contextually dependent judgments (Kim and Oh, 2025). Reference-based evaluation can improve alignment (Li et al., 2025), and LitBench shows that specialized reward models can outperform zero-shot LLM judges (Fein et al., 2026). Recent studies further show substantial disagreement among creativity metrics, sensitivity of LLM judgments to prompt variation (Lu et al., 2026), and criterion-dependent LLM–human agreement (Lyu et al., 2026). Our study complements this work by evaluating the same texts and human-rated dimensions across zero-shot, reference-based, few-shot, ensemble, multi-agent, and prompt-variation settings.

Automated creativity measures. A parallel line of work seeks scalable alternatives to direct human or LLM judgment. CS4 measures story creativity through writing constraints (Atmakuru et al., 2024), PACE evaluates associative creativity through semantic associations (Qiu and Hu, 2025), and recent work has explored inexpensive metrics such as perplexity and corpus overlap (Lu et al., 2026) as well as domain-general evaluation using semantic entropy and multi-agent methods (Sen et al., 2026).

CLIN differs from these approaches in both granularity and objective. Rather than learning a holistic judge or constructing a task-general measure of model creativity, CLIN evaluates individual literary products along predefined human-rated dimensions using separate, transparent proxies: global and topic-relative novelty for originality, contextual lexical clustering for fluency, and lexical diversity for elaboration. Its goal is to determine whether such low-cost, dimension-specific measurements can recover human judgments without requiring a generative evaluator, learned reward model, or reference exemplar.

3 Creativity Evaluation Test

To evaluate creativity in literary text, we build on established psychometric approaches. We adopt the TTCT as a structured foundation, following recent work such as Chakrabarty et al. (2023), who introduced the Torrance Test of Creative Writing (TTCW) for assessing AI-generated stories. Their findings show that LLM-generated stories perform substantially worse than professional human-written stories across TTCW-based creativity assessments, indicating a persistent gap in creative writing performance.

Inspired by this line of work, we assess creativity using three TTCT-derived dimensions: *Originality*, *Fluency*, and *Elaboration*. The *Flexibility* dimension is excluded, as prior studies have shown it to be highly correlated with Fluency and thus redundant in practice (Weiss and Wilhelm, 2022; Käckenmester et al., 2019). Specifically, Originality measures the novelty and non-cliché nature of a text; Fluency captures the diversity of ideas used to convey the topic; and Elaboration evaluates the depth of development and lexical richness. These dimensions provide a structured and interpretable basis for systematic evaluation.

However, TTCT-style metrics alone do not fully capture the holistic and subjective nature of creativity. Prior work by Marco et al. (2024) proposes an alternative rubric grounded in Boden's notion of creativity,[2] incorporating dimensions such as Creativity, Attractiveness, Originality, Anthology Potential, and Own Voice to evaluate short stories. While their framework is well-suited to longer narrative texts, our setting differs substantially: we focus on short literary text, often consisting of a single sentence, where several of these dimensions are less applicable or difficult to assess reliably.

We therefore adopt a complementary approach. Retaining the structured TTCT dimensions as a backbone, we incorporate a subset of subjective dimensions inspired by prior work, specifically *Creativity* and *Attractiveness*, that are compatible with our shorter text format. The Creativity dimension captures overall inventiveness, while Attractiveness reflects aesthetic appeal and reader engagement. In addition, motivated by prior evidence that emotional expression is a core component of creative writing (Kaufman and Sternberg, 2019), as well as findings from Chakrabarty et al. (2023) highlighting LLMs' limitations in modeling emotional depth, we introduce an *Emotion* dimension to assess the intensity and vividness of affective content.

Importantly, we distinguish between structured and holistic evaluation. The TTCT-derived dimensions provide a decomposition of creativity into interpretable components, while the overall *Creativity* score captures a holistic human judgment of creativity as an integrated construct and is not defined as a simple aggregation of the other dimensions. Our goal is twofold: first, to examine how individual TTCT dimensions relate to holistic human perceptions of creativity, and second, to evaluate how well structured, proxy-based metrics align with these subjective judgments. All dimensions are rated following a consistent scoring procedure, in which annotators assign scores based on the rubric definitions provided in Appendix A.1. Both human annotators and LLM evaluators were provided with the same evaluation questionnaire, given in Appendix A.3, to ensure that judgments were based on identical criteria and instructions.

4 Evaluating LLM as Judge

To investigate the reliability of LLMs as evaluators of creativity, we conduct a series of experiments in which LLMs judge literary text under varying evaluation scenarios and prompt formulations. This section first outlines the experimental setup, including the data and scoring procedures, and then presents each experiment along with its results, highlighting the degree of alignment between LLM-based judgments and human evaluations.

4.1 Dataset

Our evaluation dataset consists of 200 short Persian literary texts, including 100 human-authored texts and 100 GPT-3.5-turbo–generated (OpenAI, 2023) texts from the CPers dataset (Tourajmehr et al., 2025). The texts cover five thematic categories: hope, despair, longing, love, and friendship. Including both human-authored and model-generated texts allows us to investigate whether LLM-based evaluation varies for texts originating from different sources.

Five human annotators independently evaluated all texts using a 3-point rating scale (1–3), following the scoring approach of Kim and Oh (2025). Each text was rated along six dimensions as defined in Section 3: *Originality*, *Fluency*, *Elaboration*, *Creativity*, *Attractiveness*, and *Emotion*. Final

[2] Creativity is the ability to come up with ideas that are new, surprising and valuable (Boden, 2004).

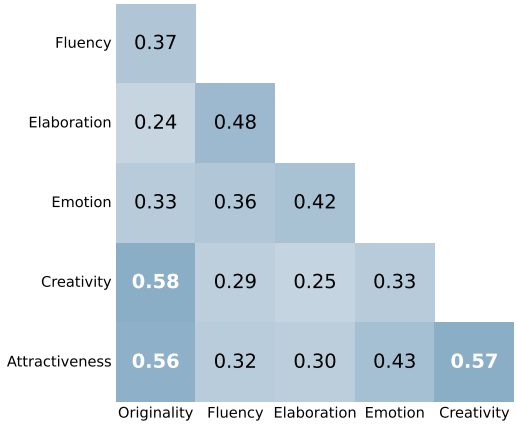


Figure 1: Pairwise Spearman correlations among the six human-rated evaluation dimensions. Values represent Spearman's rank correlation coefficient (ρ) between each pair of dimensions. Only the lower triangle is shown for clarity.

scores for each dimension are obtained via averaging across annotators. The correlations between individual annotators and the aggregated scores are provided in Appendix A.2.

Overall, this process yields a dataset of 200 texts with human average scores, serving as the reference signal for evaluating the alignment of LLM-based judgments across creativity dimensions. Figure 2 showcases some examples of the dataset.

4.2 Relationships Among Human-Rated Dimensions

To further characterize the human annotations, we examine how the six evaluation dimensions relate to one another. We compute pairwise Spearman rank correlations across all 200 texts (Figure 1). This analysis provides a descriptive view of the annotation space, showing which dimensions tend to vary together and which exhibit weaker associations.

The strongest associations involve Originality, Creativity, and Attractiveness. Originality correlates most strongly with Creativity ($\rho = 0.58$) and Attractiveness ($\rho = 0.56$), while Creativity and Attractiveness are similarly associated ($\rho = 0.57$). Thus, within this dataset, holistic creativity judgments tend to co-vary most strongly with perceived originality and aesthetic appeal.

A different pattern emerges for the remaining dimensions. Fluency and Elaboration show a moderate positive association ($\rho = 0.48$), while Emotion is moderately associated with both Elaboration ($\rho = 0.42$) and Attractiveness ($\rho = 0.43$). In contrast, Creativity is more weakly associated with Fluency ($\rho = 0.29$), Elaboration ($\rho = 0.25$), and Emotion ($\rho = 0.33$). These differences indicate that the dimensions are not equally associated with holistic creativity judgments.

Topic	Sentence
عشق Love	دریای قلبم طغیان میکند با امواج چشمات The sea of my heart is overflowing with the waves of your eyes
دلتنگی Longing	یادت پرچم صلحیست، میان ِ شورش ِ این همه فکر Your memory is the flag of peace, amidst the riot of all these thoughts
ناامیدی Despair	پشت این بغض، بیدی نشسته، که خیال میکرد با این بادها نمیلرزه Behind this hatred, sits a willow, which thought it would not tremble with these winds

Figure 2: Sample Persian instances from the dataset.

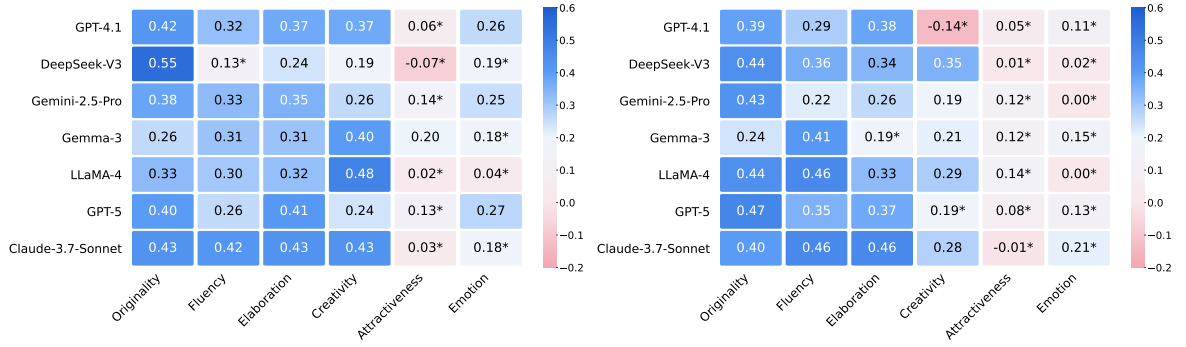
Taken together, the human-rated dimensions are related but not interchangeable. The stronger associations among Originality, Creativity, and Attractiveness reveal one pattern of co-variation, while Fluency, Elaboration, and Emotion exhibit different relationships within the annotation space. These findings are descriptive and should not be interpreted as evidence that the dimensions correspond to distinct latent constructs or causal relationships.

4.3 Experiments

The reliability of LLMs as creativity judges is evaluated across a range of settings designed to test whether increasing evaluation sophistication improves alignment with human judgments. Four experimental configurations are considered, differing in the number of models and their prompting or coordination strategies, ranging from single-model zero-shot evaluation to ensemble methods and multi-agent debate frameworks. This setup enables a systematic examination of whether additional context, interaction, or aggregation improves the reliability of creativity judgments.

Experiments are conducted using seven models: GPT-4.1 (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), Claude 3.7 Sonnet (Anthropic, 2025), LLaMA-4 (Meta AI, 2025), Gemini 2.5 Pro (Google DeepMind, 2025), Gemma-3 (Gemma Team, 2025), and DeepSeek-V3 (DeepSeek-AI, 2025). Each model evaluates the same set of texts across multiple creativity dimensions, and the resulting scores are compared against human annotations.

Agreement between model outputs and human judgments is measured using Spearman's rank correlation for each creativity dimension, providing



(a) Human-authored text

(b) Model-generated text

Figure 3: Spearman correlation between model-assigned scores and human ratings across different dimensions for (a) human-authored text and (b) model-generated text. Values marked with $*$ are *not* statistically significant.

a standard measure of consistency between evaluators. Spearman's correlation is appropriate because our goal is to evaluate whether the relative ordering of model-generated scores aligns with the ordering induced by human annotators, rather than comparing exact score values. The four experimental settings progressively introduce increasing levels of prompting structure and model coordination; each is described in detail below, along with the corresponding results.

4.3.1 Single-judge Evaluation

Each model is first evaluated independently by scoring literary texts across the six proposed dimensions using the same prompts as human annotators (Appendix A.3). Figure 3a shows correlations between model-assigned scores and human ratings for human-authored texts. Among the evaluated models, DeepSeek-V3 achieves the highest agreement with human judgments on *originality*, LLaMA-4 performs best on *creativity*, and Claude 3.7 Sonnet performs best on *fluency* and *elaboration*. Considering performance across the three TTCT-derived dimensions as a whole, Claude 3.7 Sonnet is the strongest overall zero-shot judge.

Despite these relative strengths, no model exhibits satisfactory alignment with human raters on the more subjective dimensions of creativity, namely *attractiveness* and *emotion*. Correlations for these dimensions are consistently low, and for *emotion*, most fail to reach statistical significance.

Figure 3b presents the corresponding analysis for model-generated texts. Claude 3.7 Sonnet achieves the strongest average alignment across the three TTCT-derived dimensions, *originality*, *fluency*, and *elaboration*, although the best-performing model varies by individual dimension.

Dimension	Human	Model-generated	Δ
Originality	0.396	0.401	+0.006
Fluency	0.296	0.364	+0.069
Elaboration	0.347	0.333	−0.014
Emotion	0.196	0.089	−0.107
Creativity	0.339	0.196	−0.143
Attractiveness	0.073	0.073	0.000

Table 1: Average Spearman correlation between LLM-based evaluations and human ratings for human-authored and model-generated texts. Values are averaged across the seven LLM evaluators in Figure 3. Δ denotes the difference between model-generated and human-authored texts.

To further examine whether evaluation behavior differs according to the source of the text, we average the model–human correlations across the seven evaluators for each dimension. As shown in Table 1, alignment is relatively stable across human-authored and model-generated texts for the three TTCT-derived dimensions, with fluency showing the largest increase on model-generated texts. In contrast, agreement with human judgments decreases considerably for overall creativity and emotion when evaluating model-generated texts, while attractiveness remains essentially unchanged. This pattern suggests that LLM-based evaluation is relatively robust to text source for the more structured TTCT-derived dimensions, whereas judgments of holistic creativity and emotion are more sensitive to whether the evaluated text is human- or model-generated.

We further investigate whether providing additional context improves evaluation reliability by introducing a reference-based comparison setting. Instead of assigning absolute scores, models are asked to compare a candidate text against a ref-

erence and select which is better for a given dimension. At a high level, this formulation aims to reduce calibration issues by reframing evaluation as a relative judgment task. However, we find that model outputs are highly sensitive to the choice of reference, and no single reference consistently improves performance across dimensions or instances.

Overall, this setting yields less reliable results than the single-judge setup. Full experimental details and results for the top-performing models (Claude, DeepSeek, and GPT-5) across all six dimensions are provided in Appendix A.7.

4.3.2 Few-shot Prompt Evaluation

The impact of few-shot prompting on the reliability of LLM-based creativity judgments is evaluated by providing models with human-annotated examples that illustrate the expected scoring behavior. Due to cost considerations, we first conducted pilot experiments across multiple models. In these experiments, few-shot prompting generally reduced correlation with human judgments across the evaluated models. Among the few-shot configurations, GPT-4.1 achieved the strongest performance and was therefore selected for the analysis reported here. Consistent with our goal of reporting the strongest achievable performance under practical constraints, subsequent experiments focus on the best-performing configuration in each setting.

Three few-shot configurations are considered: (1) examples with the highest human-assigned scores, (2) examples with the lowest scores, and (3) a mixed set containing high-, medium-, and low-scored examples. Among these, the mixed configuration yields the strongest performance and is therefore reported in Table 2.

Across all configurations, model-assigned scores remain largely unchanged relative to the zero-shot setting, and in several cases show lower alignment with human judgments. These results suggest that few-shot prompting does not improve LLM-based creativity evaluation and may even degrade it.

4.3.3 Ensemble Evaluation

An ensemble-based evaluation is conducted by aggregating model judgments via majority voting. Specifically, an ensemble of three of the strongest models (Claude, Gemini, and GPT-5) is evaluated by comparing aggregated scores with human ratings. The same procedure is applied to a second ensemble of comparatively weaker models: DeepSeek, LLaMA-4, and Gemma-3.

Strategies	Org.	Flu.	Ela.	Emo.	Crea.	Attr.
Few-shot	0.34	0.35	**0.50**	0.29	0.34	0.18*
Majority weak	0.36	0.40	0.28	0.11*	**0.42**	0.03*
Majority strong	0.41	**0.46**	0.41	0.32	0.31	0.06*
Multi-agent	0.42	0.44	**0.45**	0.34	0.37	0.12*

Table 2: Spearman correlation between model-based evaluations and human ratings across evaluation strategies, averaged over human- and model-generated texts. Values marked with $*$ are not statistically significant.

As shown in Table 2, in both cases, correlations are comparable to those of the best individually performing model on each dimension, with neither ensemble improving alignment with human judgments. Performance remains relatively strong for *originality*, *fluency*, and *elaboration*, but consistently weak for *emotion*, *creativity*, and *attractiveness*. These results indicate that simple ensemble strategies are insufficient to overcome the limitations of LLMs in evaluating subjective aspects of creativity.

4.3.4 Multi-agent Debate Evaluation

A multi-agent evaluation framework is implemented using Gemini, GPT-5, and Claude. Models first assign scores independently and then engage in a debate phase, reviewing and commenting on one another's assessments. In cases of disagreement, models may revise their scores; otherwise, initial scores are retained. In addition, models are prompted to provide rationales for their scores; however, incorporating these explanations does not improve alignment with human judgments.

The correlation between the final debate-based scores and the human ratings is measured for both human-authored and model-generated texts. As reported in Table 2, results show no clear improvement over prior evaluation settings. Consistent with previous findings (Choi et al., 2025), multi-agent debate performs similarly to majority voting, with only negligible differences. The debate prompt template and the corresponding evaluation procedure are provided in Appendix A.4.

4.4 Limitations of LLM-based Evaluation

Across the evaluated settings, LLM–human agreement varies substantially across dimensions. Alignment is consistently weaker for emotion and attractiveness than for the structured TTCT-derived dimensions, suggesting that the evaluated LLMs

	Originality	Fluency	Elaboration
Joint	0.67	0.60	0.54
Paraphrased	0.64	0.33	0.36

Table 3: Spearman correlation of model-assigned scores between the separate-question baseline and two alternative prompting settings: joint question presentation and paraphrased input. **Originality** is scored by *DeepSeek-V3*, while **Fluency** and **Elaboration** are scored by *Claude 3.7 Sonnet*, the best-performing model on these dimensions.

are less reliable for these subjective properties in our setting. This pattern is consistent with recent evidence that LLM–human agreement depends on the type of evidence required by an evaluation criterion (Lyu et al., 2026).

As shown in Table 2, majority-vote aggregation achieves broadly comparable alignment to the strongest single-model evaluations, but at increased computational cost. An alternative strategy is to assign each model to the dimension for which it demonstrates the strongest performance. Based on observed relative strengths, DeepSeek is best suited for originality, while Claude performs better on fluency and elaboration.

A more fundamental limitation lies in the sensitivity of LLM-based evaluation to prompt design. In the primary setup, each model is queried separately for its strongest dimension, for example originality for DeepSeek and fluency and elaboration for Claude. When the same questions are instead presented jointly or their phrasing is paraphrased, the resulting evaluations differ substantially despite unchanged underlying criteria.

Table 3 reports agreement under these prompting variations. The effect is particularly pronounced for *fluency* and *elaboration*, demonstrating that even the relative ordering assigned by an evaluator can depend on prompt formulation. This observation is consistent with recent cross-domain evidence of prompt sensitivity in LLM-based creativity evaluation (Lu et al., 2026) and shows that the phenomenon also arises in short Persian literary text.

5 CLIN Framework

Motivated by the dimension-dependent behavior observed in LLM evaluation, we investigate whether the structured TTCT-derived dimensions can be approximated using dedicated, interpretable proxies. CLIN evaluates *Originality*, *Fluency*, and *Elaboration* separately rather than learning a holistic creativity judge or combining the dimensions into a single aggregate score.

Specifically, *originality* is approximated through global and topic-relative novelty, *fluency* through contextual lexical clustering, and *elaboration* through lexical diversity. This formulation differs from approaches that learn holistic evaluators from preferences or construct task-general measures of model creativity: CLIN instead associates each predefined human-rated dimension with a transparent measurement of an individual text. We additionally introduce a quality safeguard for identifying linguistically degraded outputs.

5.1 Originality (Novelty)

Originality reflects the degree to which a text is uncommon or unexpected relative to existing language use. In the creativity literature, it is a core component of creativity and is typically defined in terms of novelty with respect to a reference set (Boden, 2007). Prior work distinguishes between novelty relative to the full body of cultural production and novelty relative to the prior outputs of a specific individual or system (Boden, 2007; Franceschelli and Musolesi, 2023), differing primarily in the scope of comparison.

Following this distinction, originality is decomposed into two complementary components. *Global originality* measures rarity in the language at large, while *local originality* captures novelty relative to a constrained, topic-specific distribution. Dedicated metrics are defined for each component and combined into a unified score, enabling the evaluation of both globally rare expressions and context-dependent novelty. This formulation grounds originality in probabilistic notions of surprise while remaining consistent with established theories of creativity.

Perplexity. Perplexity (PPL) quantifies the statistical unpredictability of a text under a probability model (Jelinek et al., 1977); higher PPL indicates less predictable and potentially more original content. Ismayilzada et al. (2025) previously used PPL as a measure of surprise in preference optimization.

In this work, normalized perplexity is used as a proxy for *global originality*, yielding an objective score in the range $[0, 1]$ for evaluating originality in text. The full formulation and normalization procedure are described in Appendix A.5.

Diversity. Diversity measures how much a sentence deviates from other texts within the same

topic. Sentence embeddings are computed using a multilingual sentence-transformer model[3], and cosine distance from the topic centroid is used as a normalized $[0, 1]$ diversity score.

Perplexity and diversity are combined into a unified novelty score:

$$\textbf{Novelty} = \text{PPL} \times \text{Diversity}$$

where PPL denotes normalized perplexity, capturing global rarity, while Diversity captures topic-specific uniqueness. Their product therefore combines both global and local aspects of originality into a single novelty score.

5.2 Fluency (Lexical Idea)

In the TTCT framework, fluency reflects the ability to generate multiple relevant ideas in response to a prompt. Since explicit identification of "ideas" in free text is not directly observable, we propose a proxy notion of *lexical ideas* to approximate idea production at the semantic level.

Unlike surface-level lexical counting, which treats each word independently, our goal is to capture semantically related words as belonging to the same underlying idea. To this end, each token is mapped into a contextual embedding space using a pretrained language model, where semantically related words are expected to have similar representations.

Formally, given a text T, each token $w_i \in T$ is represented as an embedding $\mathbf{h}_i$. We then apply DBSCAN clustering with cosine distance to these contextual token embeddings:

$$\mathcal{C}(T) = \text{DBSCAN}(\{\mathbf{h}_1, \ldots, \mathbf{h}_m\}; d_{\cos}). \quad (1)$$

DBSCAN groups tokens whose contextual embeddings are sufficiently close in the embedding space, thereby forming clusters of semantically related words. We use DBSCAN because the number of semantic groups may vary across texts and is not known in advance. Unlike clustering methods that require the number of clusters to be specified beforehand, DBSCAN can determine the number of clusters directly from the density structure of the embeddings and can also identify isolated points as noise.

Fluency is then defined as the number of clusters, excluding noise. Each resulting cluster is treated as a proxy for a distinct lexical idea, providing a more meaning-aware approximation of idea production than purely surface-level lexical counting, while allowing the number of inferred ideas to vary naturally across texts.

[3] https://huggingface.co/heydariAI/persian-embeddings

	Org:Novelty	Flu:Lexical Idea	Ela:N-gram
Correlation	0.45	0.46	0.67

Table 4: Spearman correlation between human-rated TTCT dimensions (**Originality, Fluency, Elaboration**) and their corresponding proxy measures (**Novelty, Lexical Idea, N-gram**), averaged over all 200 texts.

5.3 Elaboration (N-gram Diversity)

To approximate the TTCT elaboration component, we use a simplified unigram diversity proxy. Elaboration is measured as the number of unique content-bearing tokens after normalization and stopword removal.

Formally, for a token sequence $X = \{x_1, x_2, \ldots, x_n\}$, we compute the number of distinct filtered tokens as $|\{x_i \in X \mid x_i \notin \text{stopwords}\}|$. This captures lexical diversity and serves as a proxy for elaboration.

5.4 Quality

To ensure linguistic well-formedness and prevent reward hacking, we introduce a quality metric that captures coherence and readability, distinguishing meaningful texts from degenerate sequences that may score highly on creativity-related measures.

Quality is estimated using a contrastive perplexity-based approach that compares a sentence to minimally corrupted variants. Well-formed sentences are expected to be more predictable than their perturbed counterparts, whereas incoherent text exhibits smaller differences.

This formulation is inspired by prior work on contrastive linguistic evaluation, where language models are tested on whether they assign higher probability to well-formed sentences than to minimally different unacceptable variants (Marvin and Linzen, 2018; Warstadt et al., 2020). Full details and validation experiments are provided in Appendix A.6.

The metric is not used in the main analyses, as the dataset contains well-formed texts, but serves as a safeguard evaluated under controlled noise.

5.5 Validating the Individual Proxies

Following the definition of the CLIN proxies, we evaluate whether each proxy captures the relative

Sample 22		
Deep in my heart, melodies of longing rise, revealing my loneliness. In these silent moments, I think of the desire to see you and write our cherished memories in the embrace of longing.		
Org:Novelty	**Flu:Lexical Idea**	**Ela:N-gram**
1.8 / 0.012	2 / 6	2.6 / 17
Sample 23		
When I was with you, the world was filled with the color and scent of love. Now that you're gone, I carry the longing for you in my heart every moment.		
Org:Novelty	**Flu:Lexical Idea**	**Ela:N-gram**
1.2 / 0.008	2 / 6	2.4 / 16

Table 5: Qualitative comparison of average human-assigned TTCT scores and the corresponding CLIN proxy scores for two illustrative samples. Values are reported as human score / proxy score.

differences reflected in human judgments on its corresponding TTCT dimension. Since our goal is to assess rank-order agreement rather than absolute score agreement, we use Spearman's rank correlation.

As shown in Table 4, all three proxies exhibit positive and statistically significant correlations ($p < 0.05$) with their corresponding human-rated dimensions, indicating that they preserve the ranking structure induced by human evaluations. These results provide empirical support for the use of the proposed metrics as proxies for TTCT-based creativity evaluation.

We further compare the CLIN proxies with Claude 3.7 Sonnet, the strongest overall zero-shot LLM judge in Figure 3, on human-authored texts. A paired item-level bootstrap comparison shows that CLIN significantly outperforms Claude on elaboration, while the differences for originality and fluency are not statistically significant. Thus, CLIN achieves comparable performance on originality and fluency and significantly better performance on elaboration. Details are provided in Appendix A.8.

To provide a qualitative illustration, Table 5 presents two generated samples together with their human-assigned scores and corresponding proxy values. Sample 22 receives higher human scores for originality and elaboration than Sample 23, and the corresponding Novelty and N-gram values follow the same ordering. This example illustrates how the proposed proxies capture relative differences that are also reflected in human judgments.

6 Conclusion and Future Work

We investigated creativity evaluation in short Persian literary text from two complementary perspectives. First, we examined LLM-based evaluation across multiple models, prompting strategies, and coordination mechanisms. LLM–human agreement was strongly dimension dependent: the structured TTCT-derived dimensions of *Originality*, *Fluency*, and *Elaboration* showed substantially stronger alignment than *Emotion* and *Attractiveness*, while increasingly elaborate judging procedures did not consistently improve performance and judgments remained sensitive to prompt formulation.

Second, we asked whether these structured dimensions could be evaluated without a generative judge. CLIN associates each dimension with a separate, interpretable proxy based on lexical, semantic, or statistical properties of the text. Despite their simplicity, the proposed proxies achieve human alignment comparable to or better than the strongest zero-shot LLM judge in our setting, while requiring substantially lower evaluation cost.

The results support a dimension-specific approach to creativity evaluation, where some human-rated properties can be approximated with simple, transparent measurements. Future work should test these proxies across languages, longer texts, and other creative tasks, and extend them to more subjective dimensions such as *Emotion* and *Attractiveness*.

Limitations

Despite its contributions, this study has several limitations. First, the analysis is restricted to short literary text, and further experiments are needed to assess how well the findings generalize to longer and more structured creative writing tasks such as stories. Second, while the proposed proxy metrics capture key TTCT-inspired dimensions, they provide only an approximation of creativity and do not fully capture richer subjective aspects such as emotional depth and aesthetic attractiveness. Third, due to resource constraints, certain state-of-the-art large language models were not included. Evaluating the framework with a broader range of evaluator models could provide further insights into its robustness across models. Finally, as this study focuses exclusively on Persian, a low-resource language, extending it to other languages is necessary to assess cross-linguistic robustness and potential

language-specific effects.

References

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code.

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *Preprint*, arXiv:2410.04197.

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

Margaret A. Boden. 2007. Creativity in a nutshell. *Think*, 5:83 – 96.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. arXiv preprint arXiv:2309.14556. To appear in ACM CHI 2024.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *Preprint*, arXiv:2508.17536.

DeepSeek-AI. 2025. DeepSeek-V3-0324 release.

Mohamad Elzohbi and Richard Zhao. 2023. Creative data generation: A review focusing on text and poetry. In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

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. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Gemma Team. 2025. Gemma 3.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent AI model.

J. P. Guilford. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14.

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.

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62.

Wiebke Käckenmester, Antonia Bott, and Jan Wacker. 2019. Openness to experience predicts dopamine effects on divergent thinking. *Personality Neuroscience*, 2:e3.

James C. Kaufman and Robert J. Sternberg, editors. 2019. *The Cambridge Handbook of Creativity*, 2 edition. Cambridge University Press, Cambridge.

Sungeun Kim and Dongsuk Oh. 2025. Evaluating creativity: Can llms be good evaluators in creative writing tasks? *Applied Sciences*, 15(6).

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025. Automated creativity evaluation for large language models: A reference-based approach. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21475–21488, Suzhou, China. Association for Computational Linguistics.

Li-Chun Lu, Miri Liu, Pin Chun Lu, Yufei Tian, Shao-Hua Sun, and Nanyun Peng. 2026. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6329–6352, Rabat, Morocco. Association for Computational Linguistics.

Pengzhao Lyu, Yeun Joon Kim, Hanlin Xiao, and Yingyue Luna Luan. 2026. Why large language models and humans converge and diverge in evaluating creativity. *Preprint*, arXiv:2607.22218.

Guillermo Marco, Julio Gonzalo, M.Teresa Mateo-Girona, and Ramón Del Castillo Santos. 2024. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19654–19670, Miami, Florida, USA. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Kumiko Nakajima, Jan Zuiderveld, and Sandro Pezzelle. 2026. Beyond divergent creativity: A human-based evaluation of creativity in large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2639–2660, Rabat, Morocco. Association for Computational Linguistics.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*.

OpenAI. 2023. Introducing apis for GPT-3.5 Turbo and Whisper.

OpenAI. 2025a. Introducing GPT-4.1 in the API.

OpenAI. 2025b. Introducing GPT-5.

Ziliang Qiu and Renfen Hu. 2025. Deep associations, high creativity: A simple yet effective metric for evaluating large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10872, Suzhou, China. Association for Computational Linguistics.

Tan Min Sen, Zachary Choy Kit Chun, Syed Ali Redha Alsagoff, Nadya Yuki Wangsajaya, Banerjee Mohor, Swaagat Bikash Saikia, and Alvin Chan. 2026. Automated creativity evaluation of language models across open-ended tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23139–23173, San Diego, California, United States. Association for Computational Linguistics.

E. P. Torrance. 1966. *Torrance Tests of Creative Thinking: Directions Manual and Scoring Guide*. Personnel Press, Incorporated.

Armin Tourajmehr, Mohammad Reza Modarres, and Yadollah Yaghoobzadeh. 2025. Evaluating the creativity of LLMs in Persian literary text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14762–14774, Suzhou, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Selina Weiss and Oliver Wilhelm. 2022. Is flexibility more than fluency and originality? *Journal of Intelligence*, 10(4):96.

Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. Grove: A retrieval-augmented complex story generation framework with a forest of evidence. *arXiv preprint*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852. Association for Computing Machinery.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint*.

A Appendix

A.1 Rubric Questions

All creativity dimensions and their descriptions are listed in Table A.1.

A.2 Annotator Agreement

Given the inherently subjective nature of the evaluation task, individual annotators may differ in their judgments and assigned scores. We therefore aggregate the annotators' ratings by averaging their scores to obtain a more stable estimate of the overall human judgment. To assess the reliability of this aggregated evaluation, we compute the two-way random-effects, average-measure intraclass correlation coefficient (ICC(2,k)), where $k = 5$ corresponds to the five annotators. This measure assesses the consistency of the average ratings across annotators and provides an estimate of the reliability of the aggregated human score.

Measure	Originality	Fluency	Emotion	Elaboration	Attractiveness	Creativity
ICC(2,5)	0.395	0.491	0.415	0.557	0.399	0.489

Table 6: Average inter-rater reliability across human-written and model-generated texts, measured using the two-way random-effects, average-measure intraclass correlation coefficient (ICC(2,5)).

A.3 Zero-Shot Prompt Format

The following prompt was used to evaluate Persian sentences in a zero-shot setting, based on the revised rubrics addressing construct clarity and discriminant validity. Evaluators were instructed to assign numeric scores only, according to the criteria specified.

1. **Originality:** The sentence's avoidance of clichés and use of unexpected word combinations.
    - 1: Common phrase/idiom. Contains a recognizable cliché or highly predictable word pair. No surprise.
    - 2: Minor twist on common phrase. Modifies a familiar expression or uses a predictable but not trivial metaphor. Slightly fresh but still familiar.
    - 3: Novel combination. No identifiable cliché. Uses an unexpected predicate or attribute. Genuinely surprising.

Rubric	Description
Originality	To what degree does the text exhibit novel, uncommon, or unexpected ideas and expressions relative to typical responses?
Fluency	To what extent does the text generate a large number of different ideas?
Elaboration	To what extent does the text develop ideas in detail by adding explanations, descriptions, or supporting elements?
Emotion	How strongly and vividly does the text convey emotions and affective states to the reader?
Attractiveness	How appealing, engaging, or aesthetically pleasing is the text when read as a whole?
Creativity (Overall)	To what extent can the text be considered creative?

Table A.1: Creativity-related evaluation rubrics and guiding questions

2. **Fluency (Idea Density):** The number of distinct concepts or clauses integrated smoothly into the sentence.
    - 1: One simple idea. Single clause with one subject-verb-object. No additional modifiers or subordinate clauses that add new information.
    - 2: Two ideas. Two clauses (e.g., one main + one subordinate) or two distinct actions/attributes. Ideas are connected but not crowded.
    - 3: Three or more ideas. Three or more clauses, or a single clause with three distinct informational elements. Still grammatically coherent.

3. **Elaboration (Sensory & Concrete Detail):** Use of specific, sensory, or concrete details that ground abstract ideas.
    - 1: Weak / abstract. Uses only generic adjectives or abstract nouns. No sensory words. Feels vague.
    - 2: Moderate. One or two specific details (e.g., a concrete object name, a smell, a sound). Some sensory input but not sustained.
    - 3: High / vivid. Uses unexpected concrete details across multiple sensory modalities (sight, sound, touch, taste, smell). Descriptions are precise, surprising, or highly evocative.

4. **Emotion:** The sentence's ability to evoke a specific, discrete emotion in the reader.
    - 1: No specific emotion. Neutral, clinical, or purely factual. No emotional vocabulary or evocative imagery.
    - 2: Recognizable but mild. Clearly aims for an emotion (e.g., sadness, joy, fear, nostalgia) but feels muted or generic. Present but not impactful.
    - 3: Highly emotional and impactful. Emotion is unmistakable and strong. Uses visceral imagery, rhythm, or word choice that creates an emotional response.

5. **Attractiveness (Aesthetic Coherence):** The internal harmony between word choice, rhythm, imagery, and tone.
    - 1: Jarring or flat. Words clash in tone (e.g., mixing very formal and very colloquial language without purpose). Rhythm is awkward. No sense of style.
    - 2: Mostly coherent. Tone is consistent. No major clashes. Some rhythm or imagery, but feels ordinary or slightly bland. Readable but not beautiful.
    - 3: Highly harmonious. Word sounds, sentence length, and imagery work together. May use prosody, alliteration, parallelism, or a surprising but fitting metaphor. Feels purposeful and pleasing.

6. **Creativity (Overall):** The sentence as a whole feels novel, surprising, and valuable – an emergent effect beyond the sum of its parts.
    - 1: Not creative at all. Predictable, generic, or assembled from common templates. No surprise. Could have been written by any novice.
    - 2: Somewhat creative. Contains one unexpected element (e.g., a fresh word choice or a small twist) but overall structure is conventional. Mild surprise, limited lasting impression.

- 3: Fully creative. Evokes "I wouldn't have thought of that." Multiple elements interact to produce surprise, and the result feels meaningful, vivid, or insightful. Memorable.

Important: Only provide numeric scores for each criterion (1, 2, or 3). Do not provide explanations.

```
Originality: (number)
Fluency: (number)
Elaboration: (number)
Emotion: (number)
Attractiveness: (number)
Creativity: (number)
```

A.4 Multi-Agent Debate Template



Figure 4: Multi-agent debate process.

For each model participating in the multi-agent debate, the prompt is structured as follows:

> **Evaluation Question:** *{evaluation_question[dimension]}*
>
> Here are the answers from other models in the previous round:
> `{previous_info}`
>
> Your previous answer for this sentence:
> `{last_answers[model]}`
>
> If you consider it necessary, revise your score based on the previous round's feedback and other models' answers.

In this setup, each model first evaluates each sentence independently. In the next round, it observes the previous responses of other agents along with its own prior score and is asked to update or confirm its evaluation, enabling a structured multi-agent discussion that promotes alignment and consistency. If the models disagree in the final round of the multi-agent debate on whether an update should be applied, we determine the final decision using majority voting among the participating models.

A.5 Normalized Perplexity

The per-token negative log-likelihood (NLL) of each sentence is computed under a pretrained language model and converted to perplexity (PPL), which quantifies how well the model predicts a sequence. Formally, perplexity is defined as:

$$\mathrm{PPL}(x) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_{<i})\right), \tag{2}$$

where $P(w_i \mid w_{<i})$ denotes the probability assigned to token w_i given its preceding context.

Since PPL is unbounded and scale-dependent, it is mapped to the fixed range $[0, 1]$ using a clipped min–max normalization. Specifically, given predefined bounds $\mathrm{PPL}_{\min}$ and $\mathrm{PPL}_{\max}$, the raw perplexity is first clipped:

$$\tilde{\mathrm{PPL}}(x) = \mathrm{clip}(\mathrm{PPL}(x), \mathrm{PPL}_{\min}, \mathrm{PPL}_{\max}), \tag{3}$$

and then normalized as:

$$\mathrm{Score}(x) = \frac{\tilde{\mathrm{PPL}}(x) - \mathrm{PPL}_{\min}}{\mathrm{PPL}_{\max} - \mathrm{PPL}_{\min}}. \tag{4}$$

In all experiments, we set $\mathrm{PPL}_{\min} = 1.0$ and $\mathrm{PPL}_{\max} = 1000.0$. This transformation assigns lower scores to highly probable (i.e., conventional) sentences and higher scores to less likely ones, while ensuring numerical stability and robustness to outliers.

A.6 Quality Metric under Controlled Corruption

Quality is computed using a contrastive perplexity-based formulation. Given a sentence s, its perplexity $\mathrm{PPL}(s)$ is calculated using a language model. We generate minimally corrupted variants by applying token deletion and local shuffling, and compute their average perplexity $\overline{\mathrm{PPL}}_{\mathrm{cor}}$. The quality score is then defined as:

$$r = \frac{\overline{\text{PPL}_{\text{cor}}}}{\text{PPL}(s)}, \qquad \text{Quality}(s) = \frac{r}{1+r}. \quad (5)$$

This formulation captures the relative gap between a sentence and its corrupted variants. Well-formed sentences are expected to have substantially lower perplexity than their perturbed counterparts, resulting in higher quality scores. In contrast, incoherent or degenerate sequences tend to exhibit smaller differences, yielding lower scores.

To validate the sensitivity of the metric, we conduct a controlled perturbation experiment in which a fixed proportion of tokens in each sentence is randomly shuffled. The shuffle ratio is varied from 0.0 to 0.5, progressively degrading coherence while preserving the underlying token distribution. For each corruption level, quality scores are computed and averaged across the dataset.

As shown in Table A.2, the metric exhibits a consistent monotonic decrease as the degree of perturbation increases, indicating that it effectively captures disruptions in linguistic well-formedness. This behavior supports its use as a robust proxy for textual quality.

Shuffle Ratio	0.0	0.1	0.2	0.3	0.5
Quality Score	0.848	0.821	0.763	0.699	0.614

Table A.2: Effect of controlled token shuffling on the proposed quality metric. Increasing the proportion of shuffled tokens leads to a consistent degradation in quality scores, demonstrating sensitivity to disruptions in coherence and well-formedness.

A.7 Reference-based Evaluation: Experimental Details and Results

This section provides detailed descriptions and results for the reference-based evaluation setting. We evaluate three of the best-performing models from the main experiments, namely Claude 3.7 Sonnet, DeepSeek-V3, and GPT-5, across all six dimensions.

In this setup, evaluation is formulated as a pairwise comparison task, where the model selects the better text between a candidate and a reference for a given dimension. For each run, reference sentences are randomly sampled from texts within the same topic in the dataset, ensuring topical consistency. No sentence is compared against itself.

To enable comparison with human judgments, we convert absolute human annotations into pairwise preferences, producing comparative human scores. Model decisions are then evaluated by computing correlation with these human-derived comparative judgments.

Results are averaged over three runs with different randomly sampled references and further aggregated across both human-authored and model-generated texts. The final results are summarized in Table A.3.

Model	Org.	Flu.	Ela.	Emo.	Crea.	Attr.
Claude 3.7 Sonnet	0.28	0.21	0.57	0.12	0.18	0.34
DeepSeek-V3	0.18	0.31	0.36	0.16	0.13	0.14
GPT-5	0.31	0.39	0.55	0.21	0.19	0.20

Table A.3: Average correlation between model pairwise judgments and human comparative preferences across six dimensions. Results are averaged over three runs with different randomly sampled references and across both human-authored and model-generated texts.

Dimension	95% CI for $\Delta\rho$	p-value
Originality	$[-0.1097, 0.2976]$	0.3606
Fluency	$[-0.1960, 0.2467]$	0.8313
Elaboration	$[\mathbf{0.0536}, \mathbf{0.4358}]$	**0.0111**

Table A.4: Paired item-level bootstrap comparison between the dimension-specific CLIN components and Claude 3.7 Sonnet on human-authored texts. Differences are defined as $\Delta\rho = \rho_{\text{CLIN}} - \rho_{\text{Claude}}$, such that positive values favor CLIN. Boldface indicates a statistically significant difference at $p < 0.05$.

A.8 Paired Bootstrap Comparison with the Best Zero-shot LLM Judge

We compare the dimension-specific CLIN components with Claude 3.7 Sonnet, the strongest overall zero-shot LLM judge, using a paired item-level bootstrap on the human-authored texts. At each iteration, items are sampled with replacement while preserving the paired human rating, CLIN score, and Claude score for each item. We then compute the Spearman correlation of each evaluator with human judgments and record their difference, defined as $\Delta\rho = \rho_{\text{CLIN}} - \rho_{\text{Claude}}$. The resulting bootstrap distribution is used to estimate the 95% confidence interval and the p-value for the null hypothesis that $\Delta\rho = 0$.

As shown in Table A.4, CLIN significantly outperforms Claude 3.7 Sonnet on elaboration, with a 95% CI of [0.0536, 0.4358] ($p = 0.0111$). For originality and fluency, the differences are not sta-

tistically significant. These results indicate that CLIN performs comparably to the strongest zero-shot LLM judge on originality and fluency, while providing significantly better performance on elaboration.

A.9 Human Annotators

The annotations were performed by graduate students who are native speakers of Persian. After an initial calibration session to ensure a shared understanding of the evaluation criteria, annotators completed the task independently. Annotators were compensated on an hourly basis at a rate aligned with local fair compensation standards, and the annotation process took approximately two hours per annotator. The annotation guidelines are publicly available in the project GitHub repository.

A.10 Intended Use of Data Artifacts

We use publicly available datasets (Tourajmehr et al., 2025) released for research purposes and confirm that our use complies with their intended use and licensing conditions. We release our annotated derivatives under the ACL Community Agreement for the use of materials, maintaining the same research-only usage constraints.