Constrained Semantic Decompression for Persian Proverb Story Generation via PAND

Anonymous ACL submission

Abstract

Proverbs encode moral and cultural knowledge in highly compressed forms, making their expansion into narratives a challenging test of abstract semantic grounding. We frame proverb-conditioned short story generation for children as a constrained semantic decompression task, where LLMs must transform implicit moral principles into coherent narrative realizations. Focusing on Persian as a mid-resource language, we introduce the Proverb Aligned Narrative Dataset (PAND), pairing proverbs with human-written stories and explicit meanings. Through a systematic evaluation of proprietary and open-weight models under multiple prompting regimes, we uncover a consistent decompression gap: models often produce fluent stories that fail to preserve the underlying moral abstraction. While explicit reasoning improves alignment, we show that feedback-guided generation strengthens semantic grounding and enables smaller open-weight models to rival or exceed single-pass state-of-the-art baselines in creative narrative quality. Our findings highlight fundamental limitations in current LLMs ability to ground abstract meaning and point to interaction as a key mechanism for improving faithful narrative generation.

1 Introduction

A proverb is a brief, widely recognized saying that encapsulates cultural wisdom, morals, and traditional views in a fixed and memorable form (Mieder, 2004). In many cultures, particularly within the Persian tradition, such abstract maxims are closely intertwined with storytelling; behind each proverb lies a narrative that explicates its moral intent. While recent advances in natural language processing (NLP) have demonstrated strong capabilities of large language models (LLMs) in creative writing (Gómez-Rodríguez and Williams, 2023; Teleki et al., 2025), transforming a compact proverb into a coherent narrative poses a challenge that goes beyond standard text generation.

We formalize this transformation as a *semantic decompression task*, requiring the expansion of implicit moral abstractions into concrete characters, settings, and plot events. Unlike open-ended story generation benchmarks that primarily emphasize fluency (Chiang and Lee, 2023) or creativity (Atmakuru et al., 2024), this task probes a models ability to ground abstract semantics by synthesizing novel narrative realizations that preserve the proverbs moral core. The challenge becomes more pronounced under additional constraints such as short story length and children's audience suitability, framing proverb-to-story generation as a *constrained semantic decompression* problem that requires both semantic faithfulness and controlled narrative realization (Valentini et al., 2023). These conditions remain largely underexplored in prior work, which has focused predominantly on English, adult-oriented, or unconstrained narratives (Zhu et al., 2023; Yang and Jin, 2024).

In mid-resource languages such as Persian, the scarcity of high-quality aligned data limits the effectiveness of post-training and fine-tuning, making inference-time strategies and in-context learning particularly important. To study semantic decompression under these conditions, we introduce the Proverb Aligned Narrative Dataset (PAND), a curated corpus pairing Persian proverbs with human-written short stories and explicit proverb meanings. Using this resource, we investigate the semantic decompression capabilities of both state-of-the-art (SOTA) proprietary models and smaller open-weight models across three regimes illustrated in Figure 1: (i) *Pure*, which tests raw abstraction capacity and the utility of explicit reasoning; (ii) *Surface-Assisted*, which introduces lexical cues to assess reliance on surface signals; and (iii) *Feedback-Guided*, an iterative framework that refines narratives through critique to test whether interaction can deepen semantic grounding.

To evaluate semantic decompression, we adopt a

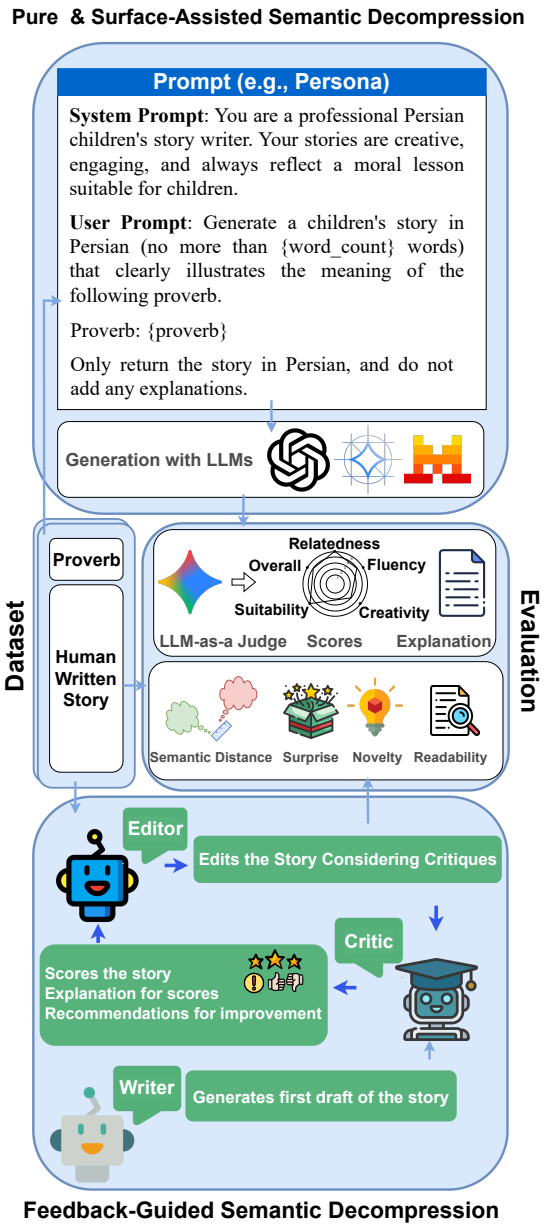

Figure 1: The Semantic Decompression Framework. The top section illustrates pure and surface-assisted generation. The middle section depicts the evaluation pipeline comparing generated stories to human baselines via a calibrated LLM-as-a-judge and automatic structural metrics. The bottom section details the feedback-guided regime where Writer, Critic, and Editor personas iteratively refine the narrative.

multifaceted evaluation strategy that combines subjective assessment via a calibrated LLM-as-a-judge with automatic structural metrics. Rather than treating story generation as a monolithic quality judgment, we explicitly disentangle semantic faithfulness from surface realization. This allows us to identify cases where models produce linguistically fluent narratives that nonetheless fail to preserve the proverbs moral abstraction. We further analyze narrative expansion and pragmatic grounding to ensure that generated stories are not only coherent but also meaningfully instantiated and appropriate for the target constraints.

Our results reveal a consistent decoupling between surface fluency and semantic depth. While explicit prompting improves grounding, open-weight models struggle to achieve faithful semantic decompression compared to state-of-the-art baselines. We further find that feedback-guided refinement can substantially mitigate this gap, enabling smaller models to correct abstraction errors and, in some settings, exceed single-pass baselines in creative narrative quality. Automatic metrics corroborate these findings, showing structural shifts toward higher semantic distance and surprise, consistent with deeper narrative instantiation.

While our experiments focus on Persian proverbs, we view semantic decompression as a general abstraction-to-realization challenge that applies to other compressed knowledge forms, including fables, legal maxims, and moral heuristics.

Our contributions are as follows:

- We formalize proverb-conditioned short story generation for children as a **constrained semantic decompression** task, offering a principled perspective on abstract-to-concrete generation in LLMs.
- We introduce **PAND**, a curated dataset of Persian proverbs paired with human-written stories and explicit meanings, designed to evaluate moral grounding in a culturally grounded, mid-resource setting.
- We provide an empirical analysis of semantic decompression in Persian and identify a **decompression gap**, where smaller open-weight models prioritize surface fluency over faithful narrative grounding.
- We show that **feedback-guided semantic decompression** improves abstraction grounding, enabling open-weight models to compensate for limited abstraction capacity and surpass single-pass state-of-the-art baselines in creativity.

2 Related Work

Proverbs and Moral Reasoning Prior NLP work on proverbs and related figurative expressions,

such as idioms and similes, has largely focused on interpretation tasks, including detection, paraphrasing, and meaning prediction (Zhou et al., 2021; Chakrabarty et al., 2022; Tayyar Madabushi et al., 2022; Cecilia Liu et al., 2024; Magdy et al., 2025; Khoshtab et al., 2025; Banou et al., 2025). While LLMs often recover figurative meanings, recent studies show that they struggle to apply these abstractions appropriately in context, revealing a gap between abstract understanding and grounded use (Attia et al., 2025).

Related work on moral reasoning has explored both moral extraction and narrative generation from abstract or visual inputs (Guan et al., 2022; Marcuzzo et al., 2025; Rezapour et al., 2025). However, these approaches typically rely on explicit moral statements, whereas proverbs encode moral content implicitly. The closest prior work, *ePiC* (Ghosh and Srivastava, 2022), studies proverb-conditioned narrative generation in English using fine-tuned models. We extend this line by framing proverb-to-story generation as a *constrained semantic decompression* task in Persian, requiring models to expand implicit, culturally grounded morals into coherent narratives for children via prompting.

Automatic Story Generation Recent advances in LLMs have also improved automatic story generation, with extensive work on hierarchical and planning-based approaches that separate high-level content planning from surface realization (Fan et al., 2018, 2019; Yao et al., 2019; Goldfarb-Tarrant et al., 2020). More recent frameworks such as *Re3* (Yang et al., 2022), *Agents' Room* (Huot et al., 2025), and *DOME* (Wang et al., 2025) further decompose story generation into multi-stage or multi-agent processes.

Another line of work emphasizes iterative or feedback-based refinement. *Self-Refine* (Madaan et al., 2023) demonstrates that a single LLM can generate feedback to improve its own outputs, while methods such as *CritiCS* (Bae and Kim, 2024), *SWAG* (Pei et al., 2024), and *Dramaturge* (Xie et al., 2025) incorporate external critics or structured guidance. These approaches, however, primarily target long-form English narratives and do not address the challenge of expanding abstract semantic inputs, such as proverbs, into short, semantically faithful narratives in mid-resource languages like Persian.

Story Generation Evaluation Early work on story generation evaluation focused on surface-level narrative properties such as coherence, redundancy, and linguistic fluency (Chhun et al., 2022). More recent studies have shifted toward creativity-oriented evaluation, comparing LLM-generated narratives to human writing using both human judgments and automatic metrics (Chakrabarty et al., 2023; Johnson et al., 2023; Orwig et al., 2024; Tian et al., 2024; Marco et al., 2024, 2025; Lu et al., 2025). Despite high fluency, these works often reveal limitations in stylistic diversity and originality, and largely evaluate generation in open-ended settings that prioritize unconstrained imagination over semantic grounding.

In contrast, we evaluate story generation as a constrained grounding problem rather than an open-ended creative task. Our framework explicitly disentangles semantic faithfulness from surface narrative quality. We align our subjective evaluation with Ghosh and Srivastava (2022) and adapt automatic structural creativity metrics from Ismayilzada et al. (2024) to Persian, enabling assessment of moral grounding and audience constraints alongside creative narrative expansion.

3 Dataset

In this section, we describe the steps of constructing our Persian gold dataset called PAND, which contains 150 instances[1]. Each instance consists of a proverb, its meaning, and a human-written short childrens story corresponding to that proverb. To the best of our knowledge, this is the first Persian resource that combines these three elements in a single, structured corpus. Table 4 in Appendix A presents an example of our dataset.

3.1 Data Collection

We collected the dataset in two stages:

1. **Proverb-Story Pairs:** We collect pairs of Persian proverbs and short stories for children by crawling several publicly available Persian literary and educational websites, including *LELB Society*[2] and *Kanoon*[3]. The full list of links to these websites is provided in Table 3 of Appendix A.

2. **Proverb Meanings:** For each proverb, we separately collect semantic meanings from

[1]The dataset and code will be released on publicly accessible platforms.
[2]https://lelb.net/
[3]https://www.kanoon.ir/

reputable online Persian resources, including *Abadis*[4], an online dictionary, and *Daneshchi*[5], an educational platform. These meanings provide explicit explanations of the proverbs moral or metaphorical content. They are used by our LLM-as-a-judge framework, described in Section 5.1, to assess the relatedness between generated stories and the intended moral meanings of the proverbs.

3.2 Data Preprocessing and Cleaning

To improve data quality, we preprocess and clean the raw data by removing irrelevant proverb-story pairs, discarding stories shorter than 50 words or longer than 1500 words, normalizing punctuation, and correcting half-space and paragraph structure issues common in Persian text. The text normalization prompt is provided in Figure 25 of Appendix I. The cleaned dataset contains 150 proverb-story pairs covering 116 unique proverbs. For proverbs with multiple stories, we retain a single reference human-written story per proverb for our experiments, selected by the highest Overall LLM-as-a-judge score, with ties broken by the sum of remaining metrics (see Section 5.1). Detailed statistics of the final PAND dataset appear in Table 2 of Appendix A.

4 Proverb Decompression Strategies

We formalize proverb-conditioned story generation as mapping a compressed proverb P to an expanded short narrative S. We treat this mapping as a **constrained semantic decompression** problem, where the generated story must preserve the proverbs moral abstraction while satisfying length and child-appropriateness constraints. To probe this capability, we employ prompting strategies organized into three regimes that differ in the degree of external guidance, ranging from unassisted generation to iterative refinement. Our goal is to identify strategies that enable models to produce narratives that are both creative and morally faithful, approaching the quality of human-written stories.

We evaluate these regimes across a set of proprietary and open-weight LLMs with strong Persian support, including GPT-4.1 (OpenAI et al., 2024b), Gemma 3 (12B and 27B) (Team et al., 2025), and Mistral Small 3.2 24B (Team, 2025a). Additional details regarding model selection, prompting settings, and computational cost are provided in Appendix B. The following subsections describe each regime in detail.

4.1 Pure

In this regime, prompts include only the proverb and a word limit, with no access to the corresponding human-written stories in PAND. This setting isolates a models intrinsic capacity to interpret an abstract moral and realize it as a narrative, without relying on surface cues from reference texts. We employ four story-agnostic prompting strategies: (i) **Zero-Shot**, which directly instructs the model to generate a creative short story from the proverb; (ii) **Moral CoT**, a chain-of-thought approach (Wei et al., 2022) that first elicits the proverbs moral before story generation; (iii) **Outline CoT**, a plan-and-write strategy in which the model first produces a three-act narrative outline (Aristotle, 1902) and then realizes it as a story; and (iv) **Persona**, which prompts the model to adopt the role of a professional Persian childrens story writer.

4.2 Surface-Assisted

In this regime, we augment the proverb and word limit with hints derived from human-authored stories in PAND. These hints provide limited surface-level cues that can steer generation toward the reference narrative, but also introduce a risk of memorization. We consider two story-hinted prompting strategies: (i) **Prefix-Conditioned**, which supplies the first two sentences of the reference story as a continuation cue; and (ii) **Cue-Words**, which supplies a small set of keywords from the reference story that the model must include, extracted using the prompt shown in Figure 23 of Appendix I.

4.3 Feedback-Guided

To examine whether semantic decompression can be improved through interaction, we introduce an iterative refinement framework. As shown in Figure 1, this regime decomposes generation into three personas: *Writer*, *Critic*, and *Editor*. This multi-turn process mirrors a human editorial workflow, in which an initial draft is iteratively critiqued and revised. By formalizing this workflow, the framework enables models to correct semantic decompression errors through structured feedback rather than relying on single-pass generation.

The process begins with the *Writer* producing an initial story conditioned only on the input proverb,

[4] https://abadis.ir/
[5] https://www.daneshchi.ir/

serving as a baseline attempt (Figure 13 in Appendix I). The *Critic* then evaluates this draft using the subjective criteria described in Section 5.1, providing structured scores and natural-language feedback on narrative coherence and alignment with the proverbs moral abstraction, along with targeted revision suggestions (Figure 21 in Appendix I). Finally, the *Editor* rewrites the story by integrating the original draft and critique, explicitly regrounding the narrative where abstraction failures occur (Figure 19 in Appendix I). This interaction forms a feedback loop that can be repeated for multiple refinement iterations. Implementation details and algorithmic pseudocode are provided in Appendix C. Full prompt templates for all regimes are provided in Appendix I.

5 Evaluation Methods

We evaluate proverb-conditioned story generation using a hybrid framework that combines subjective judgments, LLM-based evaluation, and automatic structural metrics. This combination is designed to capture complementary aspects of semantic grounding, narrative quality, and creative expansion.

5.1 Subjective Evaluation of Story Quality

Evaluating proverb-conditioned story generation requires assessing not only surface narrative quality but also the faithful grounding of abstract moral content. Following the evaluation framework of *ePiC* (Ghosh and Srivastava, 2022), we define five subjective metrics that capture different interpretive dimensions of semantic decompression. All criteria are rated on a 5-point Likert scale:

Relatedness Measures how faithfully the generated story conveys the meaning or moral of the given proverb (1: unrelated; 5: fully faithful).

Creativity Assesses the originality and narrative interest of the story beyond literal restatement (1: repetitive or uninteresting; 5: highly creative and engaging).

Fluency Evaluates grammatical correctness and readability (1: frequent errors and poor readability; 5: grammatically correct and easy to read).

Suitability for Children Rates the appropriateness of content, language, and conveyed message (1: inappropriate; 5: highly suitable).

Overall Provides a holistic assessment of narrative quality (1: very low; 5: excellent).

While human evaluation offers the highest fidelity, it does not scale efficiently. To enable large-scale comparison across models and prompting regimes, we adopt an **LLM-as-a-Judge** framework (Li et al., 2025). Because proverb interpretation requires cultural grounding, we first validate the judge against a human gold standard. Three graduate students annotated a controlled set of 40 paired human-written and GPT-4.1-generated stories (80 total). Annotation reliability was assessed using the Intraclass Correlation Coefficient (ICC(3,k)) (Shrout and Fleiss, 1979). As shown in the Human–Human column of Table 1, annotators exhibit strong agreement on key metrics, including Relatedness (0.88) and Creativity (0.83), confirming the reliability of our evaluation protocol (see Appendix F).

To scale evaluation while preserving cultural sensitivity, we employ a calibrated **LLM-as-a-Judge**.

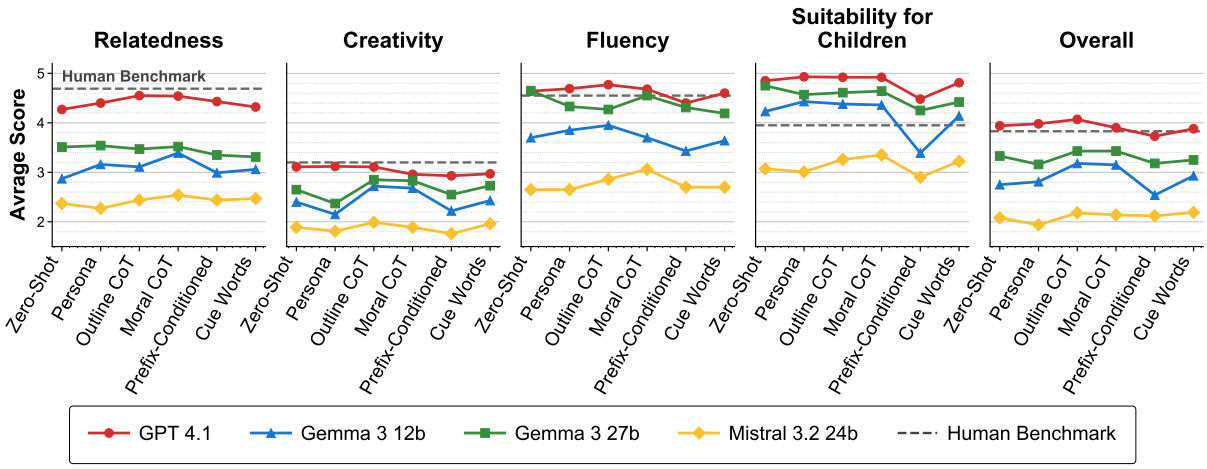

Figure 2: Average score of LLMs given by our Judge (Gemini 2.5 Pro) on different metrics for different prompts.

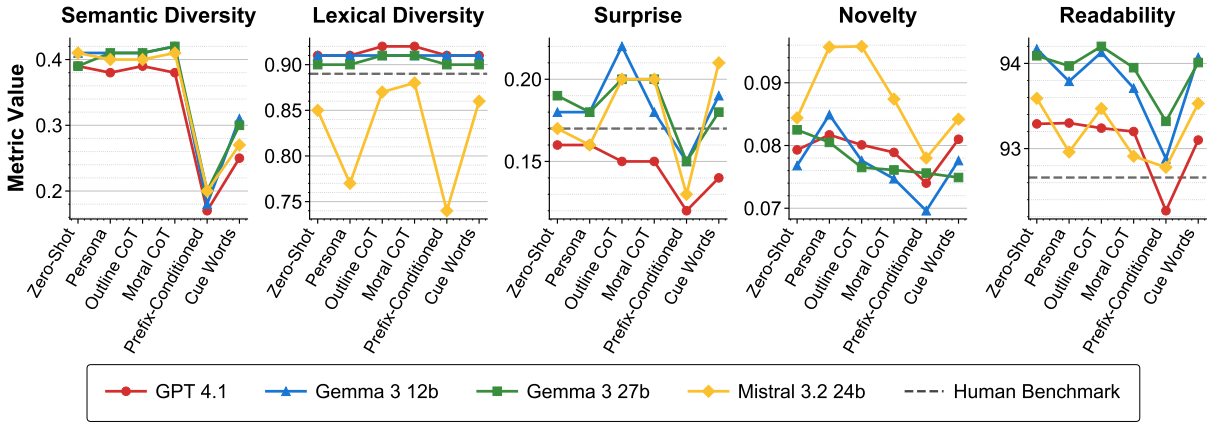

Figure 3: Performance of different models with different prompts from the lens of automatic metrics.

Metric	Human–Human Agreement	Gemini–Human Agreement		
		Mean	Median	All
Relatedness	0.88	0.74	0.72	0.87
Creativity	0.83	0.55	0.51	0.79
Fluency	0.48	0.43	0.38	0.53
Suitability for children	0.76	0.68	0.69	0.78
Overall	0.82	0.72	0.66	0.83

Table 1: Inter-annotator agreement (ICC3k) between human annotators and between Gemini 2.5 Pro and human annotators.

To mitigate the risk that generic models misinterpret the cultural nuances of Persian proverbs, we conduct a calibration study comparing multiple state-of-the-art and open-weight models against a human gold standard. Appendix D details this analysis, including candidate judges and aggregation strategies. Table 1 reports agreement metrics for our selected evaluator, Gemini 2.5 Pro (Comanici et al., 2025). Under the "All" aggregation setting, where the judge acts as a fourth independent annotator, Gemini 2.5 Pro achieves strong agreement with humans, with correlations of 0.87 for Relatedness and 0.83 for Overall quality, supporting its use as a reliable and scalable proxy for human judgment in subsequent experiments.

5.2 Automatic Structural Metrics

To complement subjective and judge-based evaluation, we employ fully automatic structural metrics adapted from Ismayilzada et al. (2024) to quantify semantic aspects of creativity. These metrics capture properties that are difficult to assess reliably through Likert-scale judgments alone by leveraging embedding-based representations to measure semantic variation. Formal definitions, mathematical formulations, and implementation details are provided in Appendix G.

Diversity We quantify diversity along two complementary dimensions. *Lexical diversity* measures surface-level variation using the ratio of unique n-grams to total n-grams ($n \in \{1, \ldots, 5\}$). *Semantic diversity* captures embedding-based distance between a generated story and its corresponding human-written reference for the same proverb, reflecting the extent to which a models semantic decompression diverges from human narrative realizations.

Novelty Novelty measures the semantic uniqueness of a generated story relative to a broader narrative context (Maher, 2010; Runco and Jaeger, 2012). We compute novelty by comparing a storys semantic representation against a background corpus consisting of all other model-generated stories for the same prompt, as well as human-written stories. Higher values indicate greater deviation from the collective narrative distribution.

Surprise Surprise captures the degree to which a story deviates from expectations as it unfolds (Boden, 1991; Maher, 2010; Grace et al., 2015). Following Karampiperis et al. (2014), we measure surprise as the average semantic shift between consecutive sentences, where larger shifts correspond to higher narrative unpredictability.

Readability To assess surface-level readability for Persian text, we use the Flesch–Dayani metric (Dayani, 2000), which estimates reading difficulty based on average word and sentence length. Higher scores indicate easier-to-read narratives.

6 Results and Analysis

We present a comprehensive evaluation of proprietary and open-weight models across generation

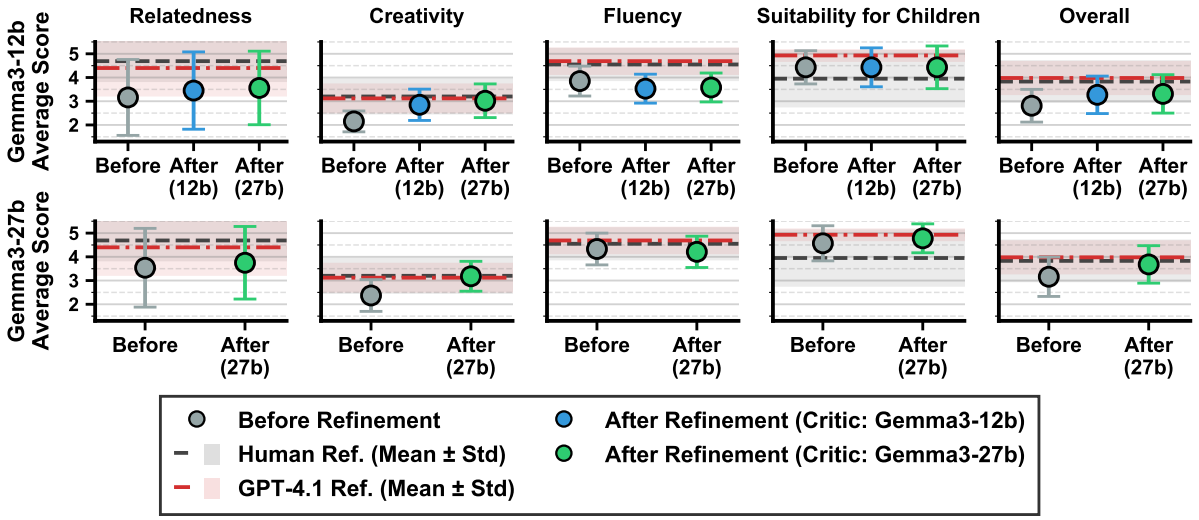

Figure 4: Judge scores (Gemini 2.5 Pro) for Gemma 3 12B (top) and Gemma 3 27B (bottom) before and after feedback-guided refinement (three iterations).

regimes, analyzing performance using both human-aligned subjective judgments and automatic structural metrics.

Different Prompts, Different Performances Figure 2 reveals a clear decoupling between surface narrative quality and semantic faithfulness. While open-weight models such as the Gemma 3 family approach GPT-4.1 in fluency and suitability, they exhibit a substantial drop in relatedness, supporting the existence of a *decompression gap*: smaller models achieve surface realization but fail to faithfully ground narratives in the proverbs abstract moral. A creativity gap persists as well; notably, even GPT-4.1 consistently falls short of the human benchmark in both creativity and relatedness, underscoring the inherent difficulty of semantic decompression.

Prompting strategy significantly modulates performance. Prompts that explicitly encourage intermediate reasoning (e.g., Moral CoT) consistently yield higher relatedness across models. This suggests that separating moral abstraction from narrative realization helps facilitate generation conditioned on an explicit semantic interpretation rather than implicit surface associations. Detailed results for all model-prompt combinations are reported in Table 6 in Appendix E.1.

Story Evaluation by Automatic Metrics Figure 3 shows that prompting strategies induce clear structural shifts in automatic metrics. Prompts that impose surface scaffolding, such as *Prefix-Conditioned*, consistently yield the lowest semantic diversity and novelty, indicating constrained semantic exploration. In contrast, open-ended prompts, including Zero-Shot and CoT variants, allow greater expressive freedom and produce higher semantic dispersion. Lexical diversity remains high across regimes, while readability decreases slightly under prefix conditioning, suggesting that surface fluency is largely saturated for modern LLMs, whereas meaning-level properties are highly sensitive to prompt design.

Model-level analysis further reveals a decoupling between statistical novelty and perceived generation quality. Mistral achieves the highest Novelty scores but performs poorly on subjective metrics (Figure 2), indicating that its novelty reflects semantic instability rather than effective narrative exploration. By contrast, GPT-4.1 and Gemma 3 27B exhibit lower novelty but achieve more balanced performance by prioritizing semantic control and coherence. These results highlight the importance of hybrid evaluation in distinguishing grounded creativity from stochastic divergence. Full numerical results are reported in Table 7 in Appendix E.2.

Iterative Refinement: Feedback Can Help To mitigate the decompression gap in open-weight models, we apply the feedback-guided framework to Gemma 3 12B and 27B under two configurations: *Self-Correction* and *Cross-Model Critique* (12B writer guided by a 27B critic). As shown in Figure 4, iterative refinement yields consistent gains in semantic grounding. For the smaller 12B model, guidance from the stronger 27B critic produces larger improvements in relatedness than self-correction, suggesting that limited critique qual-

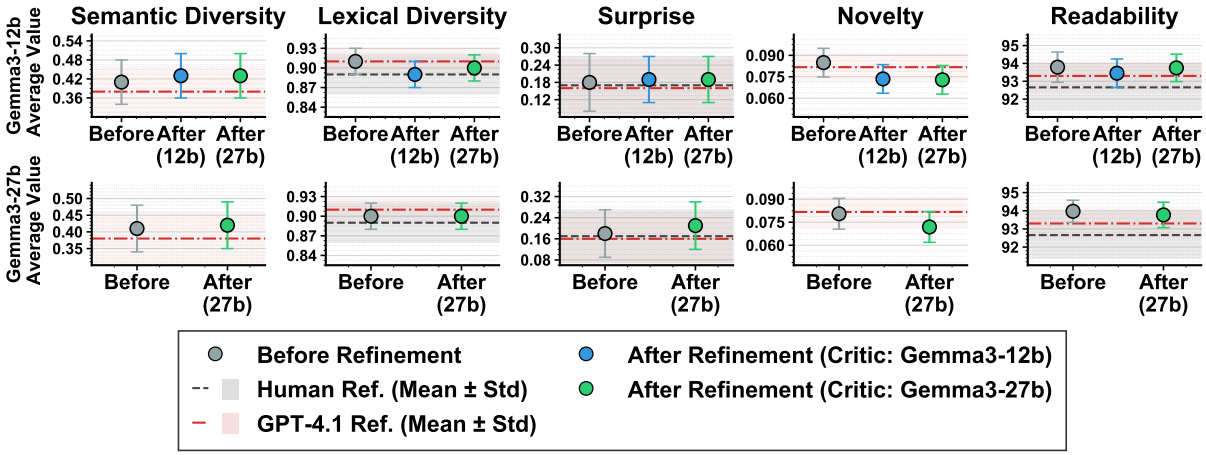

Figure 5: Changes in automatic metric values for Gemma 3 12B (top) and Gemma 3 27B (bottom) before and after feedback-guided refinement (three iterations).

ity is a key bottleneck preventing effective self-improvement in smaller models. Notably, after refinement, Gemma 3 27B surpasses the single-pass GPT-4.1 baseline in creativity, indicating that explicit feedback can unlock generative capacity that standard prompting does not elicit.

These gains come with a modest decrease in fluency, which we interpret as a surface-semantic trade-off: incorporating complex moral constraints leads to denser syntactic constructions. Importantly, suitability for children remains stable, indicating that narratives remain accessible despite this shift. Automatic metrics in Figure 5 corroborate this structural change, showing increased semantic diversity and surprise after refinement, alongside reduced novelty and lexical diversity due to thematic convergence. Crucially, these trade-offs depend on critic capacity: when guided by the stronger 27B critic, the decline in lexical diversity and readability is less pronounced than under self-correction. This suggests that higher-capacity critics provide more precise feedback, allowing writers to integrate constraints without excessive repetition. Full experimental details are reported in Appendix E.

Furthermore, to assess whether the enhancements in the Gemma 3 27B self-correction experiment align with human perception, we conduct a blinded **pairwise preference evaluation**. We present 40 randomly sampled proverb-story pairs (initial vs. final) to our human annotators, asking them to select their preferred story. Then, we calculate the Pearson correlation between human preferences and the change in LLM-as-a-Judge scores before and after refinement. Results indicate a positive directional trend for most metrics, notably creativity ($r = 0.24$). Fluency, however, showed a negative correlation ($r = -0.04$), reflecting a slight average decrease in fluency metric scores during refinement. While not statistically significant at the 95% confidence level due to the limited sample size, the positive trends suggest promising alignment between our judge scores and human preference. Detailed statistical results of this experiment are provided in Table 10 in the Appendix H.

7 Conclusion

We formalized proverb-conditioned short story generation as a constrained semantic decompression task, where compressed moral abstractions must be expanded into concrete narratives. Using the curated PAND corpus, we evaluated model behavior across pure, surface-assisted, and feedback-guided regimes with a calibrated judge and complementary structural metrics. Our analysis reveals a persistent *decompression gap*: open-weight models often achieve surface-level fluency while failing to preserve deeper causal and moral structure. Crucially, we show that feedback-guided interaction mitigates this limitation, enabling open-weight models to correct abstraction errors and surpass single-pass state-of-the-art baselines in creativity.

Looking forward, the Writer-Critic dynamic could be internalized through reinforcement learning or self-play (Chae et al., 2025), enabling models to learn abstraction-aware refinement policies rather than relying on explicit prompting. More broadly, we hope this work encourages research on semantic decompression as a unifying lens for studying how LLMs transform abstract cultural knowledge into grounded creative language.

8 Limitations

The primary limitation of this study is the scarcity of high-quality, aligned Persian proverb-story pairs. While our curated dataset is sufficient for evaluation and in-context learning experiments, its scale is inadequate for pre-training or post-training interventions such as supervised fine-tuning or reinforcement learning. As a result, our investigation is necessarily restricted to inference-time prompting strategies, which limits the extent to which models can internalize semantic decompression mechanisms.

A related limitation concerns generalization. Our experiments focus on short narratives and a specific cultural form, proverbs, which may not fully represent other forms of abstract knowledge or longer narrative structures. Although we believe the semantic decompression framework is broadly applicable, its effectiveness for other genres, domains, or languages remains to be empirically validated.

Finally, while we mitigate scalability constraints through a calibrated LLM-as-a-Judge, automated evaluation remains an imperfect proxy for human judgment, particularly for culturally grounded interpretation. Although our calibration shows strong agreement with human annotators, subtle pragmatic or stylistic nuances may still be underrepresented. Future work incorporating larger-scale human evaluation or hybrid human-LLM assessment could further strengthen evaluation reliability.

We hypothesize that access to larger-scale aligned corpora would enable open-weight models to internalize abstraction-to-realization dynamics through training rather than prompting alone. Such training could reduce the surface-semantic trade-off observed in this study and potentially allow open-weight models to achieve state-of-the-art performance in culturally grounded narrative generation.

9 Ethical Considerations

Our work primarily addresses story generation for children, a vulnerable audience. While our quantitative and qualitative evaluations indicate that the decompression regimes we investigated produce suitable and non-violent narratives, LLMs remain stochastic by nature. We cannot guarantee 100% safety against the generation of toxic, biased (Rooein et al., 2025), or inappropriate content in rare edge cases. Consequently, any real-world deployment of this work would require strict content moderation guardrails and human oversight.

Additionally, our findings are rooted in Persian literary traditions and cultural logic. There is a risk of overgeneralization if these findings are assumed to apply universally to other languages or cultures without empirical validation. Finally, our prompts were optimized for the specific models evaluated in this study; utilizing them in user-facing applications or with untested models without rigorous safety testing could lead to unpredictable or lower-quality outputs.

References

Anthropic. 2025. Claude opus 4 & claude sonnet 4 system card. Accessed: Jan 3, 2026.

Aristotle. 1902. *The Poetics of Aristotle*. Poetics. English & Greek. Macmillan, London and New York. Greek and English on opposite pages.

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.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona Diab. 2025. Beyond understanding: Evaluating the pragmatic gap in llms' cultural processing of figurative language. *Preprint*, arXiv:2510.23828.

Minwook Bae and Hyounghun Kim. 2024. Collective critics for creative story generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, Florida, USA. Association for Computational Linguistics.

Zouheir Banou, Sanaa El Filali, El Habib Benlahmar, Fatima-Zahra Alaoui, Laila El Jiani, and Hasnae Sakhi. 2025. A systematic review of figurative language detection: Methods, challenges, and multilingual perspectives. *Natural Language Processing Journal*, 13:100192.

Margaret A. Boden. 1991. *The Creative Mind: Myths and Mechanisms*. Basic Books, Inc., New York, NY, USA.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Justin Yang Chae, Md Tanvirul Alam, and Nidhi Rastogi. 2025. Towards understanding self-play for llm reasoning. *Preprint*, arXiv:2510.27072.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Vishakh Padmakumar, He He, and Nanyun Peng. 2023. Creative natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 34–40, Singapore. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Mohammad-Hossein Dayani. 2000. A criteria for assessing the persian texts readability. *Journal of Social Sciences and Humanities of Shiraz University*, 10(1):35–48.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025c. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Kazjon Grace, Mary Lou Maher, Douglas Fisher, and Katherine Brady. 2015. Modeling expectation for evaluating surprise in design creativity. In *Design Computing and Cognition '14*, pages 189–206, Cham. Springer International Publishing.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. Agents' room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations*.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Sadegh Jafari, Farhan Farsi, Navid Ebrahimi, Mohamad Bagher Sajadi, and Sauleh Eetemadi. 2025. Dadmatools v2: an adapter-based natural language processing toolkit for the persian language. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 37–43.

Dan R Johnson, James C Kaufman, Brendan S Baker, John D Patterson, Baptiste Barbot, Adam E Green, Janet van Hell, Evan Kennedy, Grace F Sullivan, Christa L Taylor, and 1 others. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 55(7):3726–3759.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2025. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *The Thirteenth International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Mary Lou Maher. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, page 2228, Lancaster, GBR. Desire Network.

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.

Guillermo Marco, Luz Rello, and Julio Gonzalo. 2025. Small language models can outperform humans in short creative writing: A study comparing SLMs with humans and LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6552–6570, Abu Dhabi, UAE. Association for Computational Linguistics.

Matteo Marcuzzo, Alessandro Zangari, Andrea Albarelli, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2025. Morables: A benchmark for assessing abstract moral reasoning in LLMs with fables. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27715–27739, Suzhou, China. Association for Computational Linguistics.

W. Mieder. 2004. *Proverbs: A Handbook*. Greenwood Folklore Handbooks. Bloomsbury Academic.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card. Accessed: Jan 3, 2026.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

William Orwig, Emma R Edenbaum, Joshua D Greene, and Daniel L Schacter. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of creative behavior*, 58(1):128–136.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Jonathan Pei, Zeeshan Patel, Karim El-Refai, and Tianle Li. 2024. SWAG: Storytelling with action guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14086–14106, Miami, Florida, USA. Association for Computational Linguistics.

Rezvaneh Rezapour, Sullam Jeoung, Zhiwen You, and Jana Diesner. 2025. Tales of morality: Comparing human- and LLM-generated moral stories from visual cues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18915–18933, Suzhou, China. Association for Computational Linguistics.

Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased tales: Cultural and topic bias in generating children's stories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 52–72, Suzhou, China. Association for Computational Linguistics.

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

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Mistral AI Team. 2025a. Mistral online documents for mistral small 3.2. Accessed: Dec 31, 2025.

Mistral AI Team. 2025b. Mistral online documents for mistral small creative. Accessed: Dec 31, 2025.

Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. 2025. A survey on LLMs for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13954–13966, Suzhou, China. Association for Computational Linguistics.

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? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598, Singapore. Association for Computational Linguistics.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1352–1391, Albuquerque, New Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Wenda Xie, Chao Guo, Yanqing Jing. Junle Wang, Yisheng Lv, and Fei-Yue Wang. 2025. Plug-and-play dramaturge: A divide-and-conquer approach for iterative narrative script refinement via collaborative llm agents. *Preprint*, arXiv:2510.05188.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dingyi Yang and Qin Jin. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation. *Preprint*, arXiv:2408.14622.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: towards better automatic storytelling. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2021. Idiomatic expression paraphrasing without strong supervision. *Preprint*, arXiv:2112.08592.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are NLP models good at tracing thoughts: An overview of narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

A Data Sources

This section provides supplementary details about the PAND dataset. Table 2 summarizes key statistics of the dataset, Table 3 lists the source websites along with the corresponding numbers of proverb-story pairs included in the final corpus, and Table 4 presents an example instance from the PAND dataset.

B Experimental Details

To automatically generate Persian stories, we utilize a combination of proprietary and open-weight LLMs with strong Persian support. We specifically focus on smaller, efficient models to evaluate whether they can perform comparably to larger closed-source systems. Our final set of models comprises GPT-4.1 (OpenAI et al., 2024b), Gemma 3 (12B and 27B) (Team et al., 2025), and Mistral Small 3.2 24B (Team, 2025a). We additionally evaluated several other models for story generation, including Gemma3 4B (Team et al., 2025), Qwen3 (4B and 14B) (Yang et al., 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2025c), GPT-OSS-20B (OpenAI et al., 2025), and Mistral Small Creative (Team, 2025b). However, their performance was not reportable, as they frequently produced nonsensical outputs, non-Persian characters, or content unrelated to the target proverb. We access all models via the OpenRouter API[6] using the default temperature of 1 to encourage diverse and creative responses. All other parameters are also set to their default values. The total cost of running the experiments amounts to approximately $100 USD.

Category	Metric	Value
Dataset Scale	Total Number of Stories	150
	Unique Proverbs	116
Story Length	Average Words per Story	369
	Maximum Words in a Story	1462
	Average Sentences per Story	22.3
	Maximum Sentences in a Story	83
Proverb Length	Average Words per Proverb	7.6
	Maximum Words in a Proverb	18

Table 2: PAND dataset statistical summary.

C Details of Feedback-Guided Semantic Decompression Experiment

To ensure the reproducibility of the feedback-guided regime experiments, we provide an algorithmic overview in Algorithm 1 and the specific configuration details below.

Model Configurations We utilized the Gemma 3 family of open-weight models. The Self-Correction setup involved a single model instance (either Gemma 3 12B or 27B) performing all roles. The Cross-Model Critique setup utilized Gemma 3 12B as the Writer and Editor, with Gemma 3 27B acting as the Critic.

Refinement Loop The process was executed for a fixed duration of three iterations. To ensure a

[6] https://openrouter.ai

Link	Count
https://lelb.net/	73
https://esanj.ir/	43
https://www.kanoon.ir/	17
https://setare.com/	8
https://roozaneh.net/	5
https://namnak.com/	3
https://taaghche.com/	1
Total	150

Table 3: Source websites and counts of proverb-story pairs included in the PAND dataset.

Proverb	بشکنه این دست که نمک نداره. (Broken be the hand that has no salt.)
Meaning	نتیجه عکس گرفتن از خوبی به دیگران. (The act of receiving the opposite result of doing good to others.)
Story	در یک روستای سرسبز و زیبا، یک درخت گردوی تنومند و پربار زندگی می‌کرد. این درخت گردو هر سال بار و میوهٔ خیلی زیادی می‌داد و مردم زیادی از میوه و سایهٔ آن استفاده می‌کردند. این درخت گردو در کنار جاده قرار داشت و به این دلیل، مسافرانی که به روستا می‌آمدند، برای خوردن گردوی خوشمزه و استراحت در سایهٔ درخت، به او پناه می‌آوردند. ولی رفتار مسافران و حتی مردم روستا با این درخت اصلاً خوب نبود. مردم برای اینکه گردوی بیشتری از درخت بچینند و بخورند، همیشه سنگ و چوب به شاخه‌های بالایی درخت گردو پرتاب می‌کردند تا گردوهایی که بر روی شاخه‌های بالاتر هستند، به زمین بریزند. کم‌کم به خاطر این رفتار زشت مردم و ناسپاسی آن‌ها، درخت گردو آسیب زیادی دید و شاخه‌هایش شکست. به این دلیل، دیگر نمی‌توانست مثل قبل، میوه و سایه به مردم بدهد. درخت بیچاره زیر لب با خود گفت: «بشکنه این دست که نمک نداره! همان مردمی که از میوه و سایهٔ من استفاده می‌کردند، به جای تشکر کردن، با سنگ و چوب از من پذیرایی کردند!» (In a lush and beautiful village, there lived a sturdy and fruitful walnut tree. Every year, it gave plenty of fruit, and many people enjoyed its shade. Since it was by the road, travelers sought refuge under it. However, the people were ungrateful. They threw stones and sticks to knock down the walnuts, breaking its branches. Eventually, the tree could no longer provide fruit or shade. The poor tree whispered to itself: "Broken be the hand that has no salt! The same people who enjoyed my fruit and shade treated me with stones instead of thanks!")

Table 4: An example from the PAND Dataset. Literal English translations are provided in parentheses in gray below the original Persian text. These translations are not included in the prompts presented to the model.

Algorithm 1 Iterative Story Generation

Input: Proverb P, Iterations K
Output: Final Story S
$S \leftarrow \text{Gen}(P)$
for $i \leftarrow 1$ **to** K **do**
  $Sc, Exp, Rec \leftarrow \text{Critic}(S, P)$
  $S \leftarrow \text{Editor}(S, Sc, Exp, Rec)$
end for
return S

strictly controlled comparison between the self-correction and cross-model settings, we generated a single set of initial drafts (Iteration 0) using the Writer persona and froze these drafts. Both the Self-Correction and Cross-Model pipelines then operated on this identical starting set. This ensures that any observed performance differences are attributable solely to the quality of the critique and editing, rather than random variations in the initial generation. To compare the effect of this refinement loop, we evaluate the initial draft (iteration 0, before entering the refinement loop) and the final draft (iteration 3, after finishing the refinement loop) utilizing our LLM-as-a-Judge and automatic scores.

D Calibration and Selection of LLM-as-a-Judge

To ensure our automatic evaluation acts as a reliable proxy for human judgment, particularly regarding the cultural nuances of Persian proverbs, we conducted an extensive calibration study. This section details the candidate models, agreement methodologies, and quantitative results that led to the selection of Gemini 2.5 Pro as our evaluator.

D.1 Candidate Models

We evaluated a diverse set of LLMs to serve as the judge. Our selection criteria balanced SOTA performance with the potential accessibility of open-weight models.

Proprietary SOTA Models To maximize the likelihood of capturing deep semantic and cultural alignment, we tested the strongest available closed-source models, including GPT-5 (OpenAI, 2025), GPT-4o (OpenAI et al., 2024a), Claude Sonnet 4 (Anthropic, 2025), and Gemini 2.5 Pro (Comanici et al., 2025).

Open-Weight Models We also investigated open-weight models to determine whether a cost-effective, locally hosted judge could achieve human parity, including DeepSeek R1 (DeepSeek-AI et al., 2025a) and DeepSeek v3 (DeepSeek-AI et al., 2025b), Gemma 3 4B (Team et al., 2025), and Mistral Small 3.2 24B (Team, 2025a).

D.2 Methodology and Aggregation Strategies

We utilized the dataset described in Section 5.1, consisting of 40 human-written and 40 GPT-4.1-generated stories using the Prefix-Conditioned prompt (80 total), annotated by three human experts. Each candidate LLM was prompted to rate these stories on the same Likert scale used by humans and against the same qualitative metrics (Relatedness, Creativity, Fluency, Suitability, and Overall Quality). To rigorously quantify the alignment between the LLM judge and human perception, we calculated four agreement metrics: ICC(2,k), ICC(3,k), Krippendorffs Alpha, and Exact Match accuracy. Furthermore, acknowledging that human agreement is not monolithic, we computed these metrics using four distinct aggregation strategies to verify robustness:

1. **LLM vs. Median:** Agreement between the LLM and the median human score, mitigating outlier influence.
2. **LLM vs. Mean:** Agreement between the LLM and the average human score.
3. **Pooled Consensus (All):** Treating the LLM as a fourth independent annotator alongside the three humans to calculate group inter-rater reliability.

D.3 Results and Selection

The comparative results of these experiments are presented in Table 5.

SOTA vs. Open-Weight Gap A clear dichotomy emerged between proprietary and open-weight models. While open-weight models (e.g., Mistral, Gemma, DeepSeek) offer efficiency, they consistently showed poor correlation with human judges on this task. This suggests that current open-weight models generally lack the cultural depth or instruction-following precision required for Persian semantic decompression evaluation.

Selection of Gemini 2.5 Pro Among the SOTA models, the competition was tight. GPT-5 and GPT-4o demonstrated strong alignment, performing nearly on par with the leader in several categories. However, Gemini 2.5 Pro emerged as the most robust model overall, achieving the highest agreement scores in 3 out of the 5 evaluation criteria (specifically Fluency, Suitability for Children, and Overall) within the Pooled Consensus setting. Furthermore, it maintained highly competitive performance in Relatedness and Creativity. Given its superior alignment on pragmatic constraints (Suitability) and overall consistency, we adopted Gemini 2.5 Pro as the fixed evaluator for all subsequent experiments.

E Detailed Results

This section provides complete numerical results for all experiments reported in the main paper, including both LLM-as-a-judge evaluations and automatic metrics, across our different proverb decompression strategies.

E.1 LLM-as-a-Judge

Table 6 presents judge scores (Mean $\pm$ Standard Deviation) for all evaluated models on non-iterative prompts from the pure and surface-assisted regimes. Table 8 provides the corresponding scores for experiments under the iterative refinement framework of the feedback-guided regime. All scores were

Type	Model	Measure	Relatedness			Creativity			Fluency			Suitability for Ch.			Overall		
			Mean	Med	All	Mean	Med	All	Mean	Med	All	Mean	Med	All	Mean	Med	All
Closed-Source	GPT-4o	ICC2k	0.67	0.69	0.84	0.69	0.66	0.82	0.26	0.30	0.44	0.22	0.27	0.60	0.55	0.51	0.77
		ICC3k	0.71	0.72	0.86	**0.69**	**0.66**	**0.83**	**0.45**	**0.43**	0.52	0.42	0.45	0.72	**0.72**	**0.69**	**0.83**
		Krip's α	0.48	0.51	0.57	0.53	0.49	0.54	-0.12	0.00	0.12	-0.18	-0.09	0.22	0.25	0.21	0.43
		Match	14	41	13	20	50	16	12	43	11	3	23	3	5	20	4
	GPT-5	ICC2k	0.78	0.79	0.87	0.41	0.41	0.76	0.31	0.30	0.47	0.34	0.40	0.65	0.68	0.65	0.81
		ICC3k	**0.81**	**0.81**	**0.89**	0.48	0.47	0.78	0.31	0.30	0.48	0.45	0.49	0.71	0.68	0.66	0.82
		Krip's α	0.62	0.64	0.62	0.18	0.19	0.43	0.18	0.18	0.17	0.08	0.15	0.29	0.51	0.48	0.52
		Match	16	44	15	8	30	6	12	40	12	5	24	3	24	49	20
	Sonnet 4	ICC2k	0.69	0.68	0.85	0.64	0.56	0.81	0.27	0.22	0.45	0.39	0.45	0.67	0.63	0.59	0.80
		ICC3k	0.71	0.69	0.86	0.64	0.56	0.81	0.32	0.26	0.48	0.43	0.48	0.70	0.63	0.60	0.81
		Krip's α	0.52	0.51	0.58	0.46	0.39	0.52	0.06	0.03	0.14	0.19	0.26	0.33	0.45	0.42	0.50
		Match	16	41	15	18	48	13	12	40	11	8	30	7	23	48	20
	Gemini 2.5 Pro	ICC2k	0.69	0.68	0.85	0.40	0.37	0.73	0.40	0.36	0.51	0.63	0.65	0.75	0.73	0.66	0.82
		ICC3k	0.74	0.72	0.87	0.55	0.51	0.79	0.43	0.38	**0.53**	**0.68**	**0.69**	**0.78**	**0.72**	0.66	**0.83**
		Krip's α	0.51	0.50	0.57	0.09	0.08	0.37	0.22	0.20	0.19	0.42	0.46	0.42	0.57	0.49	0.54
		Match	16	37	16	8	24	6	13	40	12	13	36	13	21	45	17
Open-Weight	DeepSeek R1	ICC2k	0.47	0.48	0.80	0.34	0.26	0.75	0.27	0.32	0.45	0.17	0.23	0.59	0.39	0.37	0.74
		ICC3k	0.59	0.57	0.84	0.34	0.26	0.76	0.31	0.35	0.48	0.31	0.39	0.70	0.48	0.47	0.77
		Krip's α	0.19	0.23	0.48	0.19	0.14	0.43	0.08	0.14	0.14	-0.22	-0.12	0.21	0.14	0.11	0.39
		Match	16	41	14	13	37	9	11	42	11	4	22	4	17	33	14
	DeepSeek V3	ICC2k	0.37	0.39	0.77	0.36	0.35	0.77	0.06	0.04	0.36	0.29	0.33	0.63	0.29	0.26	0.67
		ICC3k	0.53	0.52	0.84	0.39	0.37	0.78	0.16	0.07	0.45	0.50	0.51	0.73	0.57	0.53	0.80
		Krip's α	0.06	0.11	0.44	0.19	0.17	0.45	-0.36	-0.25	0.06	-0.08	0.00	0.25	-0.16	-0.19	0.28
		Match	15	37	14	19	49	15	12	37	12	5	29	5	5	13	4
	Gemma 3	ICC2k	0.46	0.52	0.78	0.09	0.09	0.64	-0.03	-0.02	0.13	0.17	0.23	0.69	-0.05	-0.01	0.62
		ICC3k	0.46	0.52	0.79	0.11	0.11	0.68	-0.04	-0.02	0.15	0.17	0.23	0.61	-0.05	-0.01	0.64
		Krip's α	0.29	0.34	0.46	-0.05	-0.04	0.29	-0.12	-0.11	0.00	0.08	0.13	0.26	-0.07	-0.04	0.27
		Match	9	32	9	16	37	11	8	26	7	11	26	9	20	33	16
	Mistral 3.2	ICC2k	0.45	0.45	0.80	0.65	0.61	0.81	0.24	-0.05	0.44	0.47	0.50	0.71	0.52	0.49	0.78
		ICC3k	0.45	0.45	0.81	0.65	0.61	0.82	0.25	-0.05	0.46	0.49	0.51	0.73	0.53	0.51	0.79
		Krip's α	0.29	0.29	0.50	0.48	0.44	0.52	0.11	-0.03	0.15	0.29	0.33	0.37	0.34	0.31	0.46
		Match	12	41	10	17	44	13	13	35	12	16	42	12	24	48	19

Table 5: Inter-annotator agreement between human annotators and both closed-source and open-source LLMs across various metrics. **Bold** indicates the highest ICC(3,k) agreement score for each metric across all models.

generated by our judge (Gemini 2.5 Pro) using the 5-point Likert scale described in Section 5.1.

E.2 Automatic Metrics

Table 7 provides automatic metrics for non-iterative prompts from the pure and surface-assisted regimes. Table 9 reports the corresponding metrics for experiments under the iterative refinement framework of the feedback-guided regime.

F Guidelines for Annotating Persian Childrenś Stories

This appendix provides instructions for three graduate annotators who evaluate Persian childrens stories in relation to specific Persian proverbs. Our annotators were Computer Engineering students and did the annotations voluntarily. The annotation process is conducted in Label Studio (Tkachenko et al., 2020-2025)[7], where each story is assessed against five metrics: Relatedness, Creativity, Fluency, Suitability for Children, and Overall. Some stories may have been written by humans, and others may have been generated by large language models; however, all evaluations should be based on the content and the criteria described below.

Evaluation Criteria

1. Relatedness

This criterion measures how well a story reflects the meaning of the corresponding proverb.

- 1: Not related: No meaningful connection to the proverb.
- 2: Weakly related: Only minimal or indirect reference.
- 3: Moderately related: Partially conveys the intended meaning.
- 4: Well related: Clearly communicates the meaning.
- 5: Fully related: Completely and accurately represents the meaning of the proverb.

[7] https://labelstud.io/

Story Details		Judge Scores				
Prompt	**Writer/Model**	**Relatedness**	**Creativity**	**Fluency**	**Suitability for Children**	**Overall**
–	**Human**	4.69 ± 0.84	3.20 ± 0.72	4.55 ± 0.66	3.95 ± 1.17	3.83 ± 0.86
Zero-Shot	GPT-4.1	4.27 ± 1.31	3.11 ± 0.69	4.64 ± 0.56	4.85 ± 0.49	3.94 ± 0.82
	Gemma-3-12B	2.87 ± 1.63	2.40 ± 0.61	3.70 ± 0.83	4.23 ± 1.01	2.75 ± 0.93
	Gemma-3-27B	3.51 ± 1.55	2.65 ± 0.69	4.65 ± 0.56	4.75 ± 0.61	3.33 ± 0.96
	Mistral-3.2-24B	2.37 ± 1.54	1.89 ± 0.59	2.65 ± 0.79	3.07 ± 0.97	2.08 ± 0.78
Persona	GPT-4.1	4.40 ± 1.16	3.12 ± 0.59	4.69 ± 0.54	4.93 ± 0.23	3.98 ± 0.68
	Gemma-3-12B	3.16 ± 1.60	2.15 ± 0.44	3.85 ± 0.63	4.43 ± 0.70	2.81 ± 0.69
	Gemma-3-27B	3.54 ± 1.66	2.37 ± 0.51	4.33 ± 0.67	4.57 ± 0.74	3.16 ± 0.83
	Mistral-3.2-24B	2.27 ± 1.51	1.81 ± 0.63	2.65 ± 0.86	3.01 ± 1.10	1.94 ± 0.82
Outline CoT	GPT-4.1	4.55 ± 1.05	3.11 ± 0.61	4.77 ± 0.47	4.92 ± 0.35	4.07 ± 0.67
	Gemma-3-12B	3.11 ± 1.60	2.72 ± 0.72	3.95 ± 0.86	4.38 ± 0.86	3.18 ± 1.00
	Gemma-3-27B	3.47 ± 1.59	2.85 ± 0.68	4.27 ± 0.77	4.61 ± 0.76	3.43 ± 0.97
	Mistral-3.2-24B	2.44 ± 1.60	1.99 ± 0.55	2.86 ± 0.80	3.26 ± 0.94	2.18 ± 0.83
Moral CoT	GPT-4.1	4.54 ± 1.02	2.96 ± 0.65	4.68 ± 0.66	4.92 ± 0.29	3.90 ± 0.64
	Gemma-3-12B	3.39 ± 1.61	2.68 ± 0.59	3.70 ± 0.82	4.36 ± 0.78	3.15 ± 0.90
	Gemma-3-27B	3.52 ± 1.57	2.83 ± 0.73	4.55 ± 0.65	4.64 ± 0.71	3.43 ± 0.92
	Mistral-3.2-24B	2.54 ± 1.65	1.89 ± 0.53	3.06 ± 0.84	3.35 ± 0.99	2.14 ± 0.81
Prefix-Conditioned	GPT-4.1	4.43 ± 1.18	2.93 ± 0.63	4.40 ± 0.64	4.48 ± 0.86	3.73 ± 0.80
	Gemma-3-12B	2.99 ± 1.70	2.22 ± 0.60	3.43 ± 0.86	3.39 ± 1.06	2.54 ± 0.88
	Gemma-3-27B	3.35 ± 1.57	2.55 ± 0.56	4.31 ± 0.72	4.25 ± 0.89	3.18 ± 0.86
	Mistral-3.2-24B	2.44 ± 1.56	1.76 ± 0.62	2.70 ± 1.02	2.90 ± 1.22	2.12 ± 0.90
Cue-Words	GPT-4.1	4.32 ± 1.12	2.97 ± 0.67	4.60 ± 0.54	4.81 ± 0.49	3.88 ± 0.74
	Gemma-3-12B	3.06 ± 1.60	2.43 ± 0.57	3.64 ± 0.83	4.14 ± 1.01	2.93 ± 0.91
	Gemma-3-27B	3.31 ± 1.62	2.73 ± 0.73	4.19 ± 0.72	4.42 ± 0.80	3.25 ± 1.02
	Mistral-3.2-24B	2.47 ± 1.55	1.96 ± 0.54	2.70 ± 0.72	3.22 ± 0.99	2.19 ± 0.72

Table 6: LLM-as-a-judge scores (Gemini 2.5 Pro) for human and model-generated stories under different prompting strategies.

2. Creativity

This metric assesses how much the story is like a short, creative, or interesting story.

- 1: Very uninteresting: Very boring/uninteresting.
- 2: Slightly interesting: Limited creativity or entertainment value.
- 3: Moderately interesting: Somewhat engaging and creative.
- 4: Very interesting: Creative and engaging.
- 5: Highly creative: Very creative/story-like

3. Fluency

Fluency assesses the grammatical correctness, clarity, and overall readability of the story.

- 1: Very poor: multiple grammatical errors; difficult to read.
- 2: Poor: Contains noticeable errors.
- 3: Average: Generally readable but includes some issues.
- 4: Good: Mostly accurate and fluent.
- 5: Excellent: Clear, well-written, and free of errors.

4. Suitability for Children

This criterion determines whether the storys language, tone, and content are appropriate for a child audience.

- 1: Completely unsuitable: Includes themes or elements inappropriate for children.
- 2: Slightly unsuitable: Contains issues that reduce child-appropriateness or appeal.
- 3: Moderately suitable: Generally acceptable but not strongly oriented toward children.

Story Details		Automatic Metrics				
Prompt	**Writer/Model**	**Semantic Diversity**	**Lexical Diversity**	**Surprise**	**Novelty**	**Readability**
–	**Human**	–	0.89 ± 0.03	0.17 ± 0.10	–	92.66 ± 1.27
Zero-Shot	GPT-4.1	0.39 ± 0.07	0.91 ± 0.02	0.16 ± 0.08	$7.93E\text{-}2 \pm 0.01$	93.29 ± 0.75
	Gemma-3-12B	0.41 ± 0.07	0.91 ± 0.03	0.18 ± 0.10	$7.68E\text{-}2 \pm 0.02$	94.17 ± 0.77
	Gemma-3-27B	0.39 ± 0.07	0.90 ± 0.03	0.19 ± 0.10	$8.25E\text{-}2 \pm 0.01$	94.09 ± 0.63
	Mistral-3.2-24B	0.41 ± 0.07	0.85 ± 0.16	0.17 ± 0.10	$8.44E\text{-}2 \pm 0.02$	93.59 ± 0.94
Persona	GPT-4.1	0.38 ± 0.08	0.91 ± 0.02	0.16 ± 0.10	$8.17E\text{-}2 \pm 0.01$	93.30 ± 0.80
	Gemma-3-12B	0.41 ± 0.07	0.91 ± 0.02	0.18 ± 0.10	$8.49E\text{-}2 \pm 0.01$	93.79 ± 0.85
	Gemma-3-27B	0.41 ± 0.07	0.90 ± 0.02	0.18 ± 0.09	$8.05E\text{-}2 \pm 0.01$	93.97 ± 0.61
	Mistral-3.2-24B	0.40 ± 0.07	0.77 ± 0.20	0.16 ± 0.10	$9.57E\text{-}2 \pm 0.01$	92.96 ± 0.80
Outline CoT	GPT-4.1	0.39 ± 0.08	0.92 ± 0.02	0.15 ± 0.09	$8.01E\text{-}2 \pm 0.01$	93.24 ± 0.84
	Gemma-3-12B	0.41 ± 0.07	0.91 ± 0.02	0.22 ± 0.11	$7.76E\text{-}2 \pm 0.01$	94.13 ± 0.77
	Gemma-3-27B	0.41 ± 0.07	0.91 ± 0.02	0.20 ± 0.11	$7.65E\text{-}2 \pm 0.01$	94.20 ± 0.76
	Mistral-3.2-24B	0.40 ± 0.07	0.87 ± 0.05	0.20 ± 0.14	$9.58E\text{-}2 \pm 0.02$	93.47 ± 0.81
Moral CoT	GPT-4.1	0.38 ± 0.08	0.92 ± 0.02	0.15 ± 0.10	$7.89E\text{-}2 \pm 0.01$	93.20 ± 0.77
	Gemma-3-12B	0.42 ± 0.06	0.91 ± 0.02	0.18 ± 0.09	$7.47E\text{-}2 \pm 0.01$	93.71 ± 0.76
	Gemma-3-27B	0.42 ± 0.07	0.91 ± 0.03	0.20 ± 0.11	$7.61E\text{-}2 \pm 0.01$	93.95 ± 0.81
	Mistral-3.2-24B	0.41 ± 0.07	0.88 ± 0.05	0.20 ± 0.13	$8.74E\text{-}2 \pm 0.01$	92.91 ± 0.83
Prefix-Conditioned	GPT-4.1	0.17 ± 0.07	0.91 ± 0.02	0.12 ± 0.08	$7.40E\text{-}2 \pm 0.01$	92.27 ± 1.13
	Gemma-3-12B	0.18 ± 0.07	0.91 ± 0.02	0.15 ± 0.10	$6.96E\text{-}2 \pm 0.01$	92.88 ± 1.25
	Gemma-3-27B	0.20 ± 0.08	0.90 ± 0.03	0.15 ± 0.10	$7.56E\text{-}2 \pm 0.01$	93.32 ± 0.92
	Mistral-3.2-24B	0.20 ± 0.08	0.74 ± 0.31	0.13 ± 0.11	$7.80E\text{-}2 \pm 0.02$	92.78 ± 1.12
Cue-Words	GPT-4.1	0.25 ± 0.08	0.91 ± 0.02	0.14 ± 0.08	$8.10E\text{-}2 \pm 0.01$	93.10 ± 0.80
	Gemma-3-12B	0.31 ± 0.07	0.91 ± 0.02	0.19 ± 0.09	$7.76E\text{-}2 \pm 0.01$	94.07 ± 0.71
	Gemma-3-27B	0.30 ± 0.07	0.90 ± 0.02	0.18 ± 0.09	$7.49E\text{-}2 \pm 0.01$	94.01 ± 0.74
	Mistral-3.2-24B	0.27 ± 0.06	0.86 ± 0.09	0.21 ± 0.12	$8.42E\text{-}2 \pm 0.01$	93.53 ± 0.82

Table 7: Automatic evaluation results for human and model-generated stories under different prompting strategies.

- 4: Suitable: Appropriate, accessible, and appealing for children.
- 5: Highly suitable: Strongly child-friendly in tone, vocabulary, and content.

5. Overall

This rating reflects a general evaluation of the overall quality of the story.

- 1: Very poor: Lacks coherence or value.
- 2: Poor: Contains significant weaknesses.
- 3: Average: Acceptable but unremarkable.
- 4: Good: Coherent, well-structured, and of solid quality.
- 5: Excellent: High-quality.

General Annotation Instructions

- **Objectivity**: Annotators should focus only on the storys content and how well it aligns with the evaluation criteria. The storys origin, whether written by a human or generated by a model, should not affect their judgment.
- **Consistency**: Annotators are expected to apply the scoring standards uniformly across all stories to ensure reliability and fairness throughout the dataset.
- **Understanding the Proverb**: Before evaluating any story associated with a proverb, annotators should review the proverb and its intended meaning.
- **Accurate Recording**: All scores (1–5) must be entered correctly into Label Studio. Annotators should verify that their inputs are properly saved.
- **Optional Notes**: If an annotator wants to comment on a specific annotation or highlight something unusual, they may include it in the comment section provided on the platform.

Critic	Writer/Editor	Judge Scores				
		Relatedness	Creativity	Fluency	Suitability for Children	Overall
—	Human	4.69 ± 0.84	3.20 ± 0.72	4.55 ± 0.66	3.95 ± 1.17	3.83 ± 0.86
—	GPT-4.1	4.40 ± 1.16	3.12 ± 0.59	4.69 ± 0.54	4.93 ± 0.23	3.98 ± 0.68
—	Gemma-3-12B	3.16 ± 1.60	2.15 ± 0.44	3.85 ± 0.63	4.43 ± 0.70	2.81 ± 0.69
—	Gemma-3-27B	3.54 ± 1.66	2.37 ± 0.51	4.33 ± 0.67	4.57 ± 0.74	3.16 ± 0.83
Gemma-3-12B	Gemma-3-12B	3.45 ± 1.63	2.85 ± 0.66	3.53 ± 0.61	4.43 ± 0.82	3.27 ± 0.79
Gemma-3-27B	Gemma-3-12B	3.56 ± 1.55	3.02 ± 0.71	3.58 ± 0.61	4.43 ± 0.90	3.31 ± 0.81
Gemma-3-27B	Gemma-3-27B	3.75 ± 1.53	3.18 ± 0.63	4.21 ± 0.66	4.78 ± 0.61	3.68 ± 0.79

Table 8: LLM-as-a-judge scores (Gemini 2.5 Pro) for stories generated within our feedback-guided semantic decompression experiment.

Critic	Writer/Editor	Automatic Metrics				
		Semantic Diversity	Lexical Diversity	Surprise	Novelty	Readability
—	Human	–	0.89 ± 0.03	0.17 ± 0.10	–	92.66 ± 1.27
—	GPT-4.1	0.38 ± 0.08	0.91 ± 0.02	0.16 ± 0.10	$8.17E-2 \pm 0.01$	93.30 ± 0.80
—	Gemma-3-12B	0.41 ± 0.07	0.91 ± 0.02	0.18 ± 0.10	$8.49E-2 \pm 0.01$	93.79 ± 0.85
—	Gemma-3-27B	0.41 ± 0.07	0.90 ± 0.02	0.18 ± 0.09	$8.05E-2 \pm 0.01$	93.97 ± 0.61
Gemma-3-12B	Gemma-3-12B	0.43 ± 0.07	0.89 ± 0.02	0.19 ± 0.08	$7.35E-2 \pm 0.01$	93.45 ± 0.80
Gemma-3-27B	Gemma-3-12B	0.43 ± 0.07	0.90 ± 0.02	0.19 ± 0.08	$7.29E-2 \pm 0.01$	93.75 ± 0.77
Gemma-3-27B	Gemma-3-27B	0.42 ± 0.07	0.90 ± 0.02	0.21 ± 0.09	$7.19E-2 \pm 0.01$	93.77 ± 0.70

Table 9: Automatic evaluation results for stories generated within our feedback-guided semantic decompression experiment.

G Automatic Metric Definitions

This section provides the formal definitions, equations, and implementation details for all automatic evaluation metrics used in the main paper. It is notable to mention that we have used DadmaTools (Jafari et al., 2025) for extracting the lemmatization of the tokens for implementing the metrics needed.

G.1 Diversity

Lexical Diversity Lexical diversity measures surface-level variation in word usage. It is defined as the ratio of unique n-grams to the total number of n-grams in a story, for $n \in \{1, 2, 3, 4, 5\}$:

$$\text{LexDiv}_n = \frac{\#\text{unique } n\text{-grams}}{\#\text{total } n\text{-grams}}.$$

Values closer to 1 indicate richer and less repetitive language.

Semantic Diversity Semantic diversity captures variation in meaning across stories written for the same proverb. Following (Padmakumar and He, 2023), we use the *inverse homogenization score*, defined as the average pairwise semantic distance between stories for a given proverb.

Given a story s written for proverb p, the score is defined as:

$$\text{inv_hom}(s \mid p) = \frac{1}{|S_p| - 1} \sum_{s' \neq s} \text{semdis}(s, s'),$$

where S_p is the set of all stories written for proverb p. In our experiments, $|S_p| = 2$, consisting of one human-written story and one LLM-generated story. The semantic distance between two stories, $semdis(s, s')$, is defined as $1 - cosine_similarity$, where cosine similarity is computed between sentence embeddings. Sentence embeddings are generated using the *bge-m3* model (Chen et al., 2024). The resulting semantic distance lies in the range $[0, 2]$, where higher values indicate greater semantic divergence.

G.2 Novelty

Novelty measures how semantically different a story is relative to the overall corpus (Runco and Jaeger, 2012; Maher, 2010). Following (Karampiperis et al., 2014; Johnson et al., 2023), each story is represented using dominant terms, defined as lemmatized content words.

Let S_n denote a story and S_G denote the corpus

of all stories, with dominant term sets T_n and T_G, respectively. We compute the average semantic distance between all unordered term pairs within a story:

$$D(S_n) = \frac{1}{|P(T_n)|} \sum_{(i,j)\in P(T_n)} \text{semdis}(T_{n,i}, T_{n,j}),$$

and analogously for the corpus:

$$D(S_G) = \frac{1}{|P(T_G)|} \sum_{(i,j)\in P(T_G)} \text{semdis}(T_{G,i}, T_{G,j}).$$

The novelty of the story S_n is defined as:

$$\text{Nov}(S_n) = 2 \cdot |D(S_n) - D(S_G)|\,.$$

The novelty score ranges from 0 to 2, where higher values indicate greater semantic deviation from the corpus.

G.3 Surprise

Surprise reflects how much a story deviates from expectations as it unfolds (Boden, 1991; Maher, 2010; Grace et al., 2015). Following (Karampiperis et al., 2014), surprise is computed by comparing consecutive sentences.

Let a story S_n consist of $|F|$ sentences $\{F_1, \ldots, F_{|F|}\}$. The surprise score is defined as:

$$\text{Sur}(S_n) = \frac{2}{|F|-1} \sum_{i=2}^{|F|} |D(F_i) - D(F_{i-1})|\,,$$

where $D(\cdot)$ refers to the average semantic distance defined in the previous metric. This score lies in the range $[0, 2]$. Lower values indicate gradual semantic changes between sentences, while higher values indicate stronger and more unexpected shifts in meaning.

G.4 Readability

To assess readability for Persian texts, we employ the Flesch–Dayani metric (Dayani, 2000), which estimates reading difficulty based on surface linguistic features:

$$\text{FD} = 262.835 - 0.846 \cdot \frac{|\text{letters}|}{|\text{words}|} - 1.01 \cdot \frac{|\text{words}|}{|\text{sentences}|}.$$

Higher Flesch–Dayani scores indicate easier-to-read text, while lower scores correspond to greater textual complexity.

H Statistical Alignment of Feedback-Guided Enhancements with Human Preference

To establish a human ground truth for quality improvement, we employed the same three expert annotators who collaborated on collecting the gold data for our subjective metrics. We presented them with 40 randomly selected pairs of initial and final story drafts from the Gemma 3 27B self-correction experiment in a completely blind process (i.e., without revealing which version was the refined draft). Annotators were asked to select the preferred story, and the final human preference label for each pair was determined via majority voting.

To evaluate the alignment between human preferences and our calibrated LLM-as-a-Judge (Gemini 2.5 pro) scores, we analyzed the correlation between human choices and the change in our subjective metric scores following the refinement loop in our feedback-guided decompression regime. Given the limited sample size ($N = 40$) of the human evaluation set, we employed a non-parametric bootstrap approach to estimate the robustness of the observed correlations.

Metric	Pearson r	95% CI	Sig.
Creativity	0.24	$[-0.09, 0.54]$	No
Relatedness	0.19	$[-0.06, 0.42]$	No
Overall Score	0.17	$[-0.15, 0.45]$	No
Suitability (Children)	0.08	$[-0.25, 0.40]$	No
Fluency	-0.04	$[-0.37, 0.30]$	No

Table 10: Bootstrap analysis ($N = 40$, 10k iterations) of the correlation between metric score improvements and human preference. The table is sorted by correlation strength.

We performed 10,000 bootstrap iterations. In each iteration, we resampled the 40 data pairs (human preference, metric change) with replacement and calculated the Pearson correlation coefficient (r). The reported Mean Correlation represents the average r across all iterations. To assess statistical significance, we computed 95% Confidence Intervals (CI) using the percentile method (2.5th and 97.5th percentiles of the bootstrap distribution).

Table 10 summarizes the results. A metric is considered to have a statistically significant correlation with human preference at the $\alpha = 0.05$ level if the 95% CI does not cross zero. As shown in the table, while *Creativity* and *Relatedness* exhibit positive directional trends (mean $r > 0.19$), the

confidence intervals for all metrics include zero, indicating that the correlations are not statistically significant for this sample size.

I Prompt Templates

All prompt templates used in our experiments are presented in this appendix, in both their original Persian form and their English translations, spanning Figures 6 through 27.

Zero-Shot Prompt (Original)

یک داستان کودکانه به زبان فارسی بنویس (حداکثر {word_count} کلمه) که مفهوم و پیام اخلاقی ضرب‌المثل زیر را به‌خوبی نشان دهد.

ضرب‌المثل:
{proverb}

فقط متن داستان را به زبان فارسی برگردان و هیچ توضیحی اضافه نکن.

Figure 6: Zero-shot story generation prompt (original Persian).

Zero-Shot Prompt (Translated)

Generate a children's story in Persian (no more than {word_count} words) that clearly illustrates the meaning of the following proverb.

Proverb:
{proverb}

Only return the story in Persian, and do not add any explanations.

Figure 7: Zero-shot story generation prompt (translated from Persian).

Moral CoT Prompt (Original)

ابتدا مفهوم و پیام اخلاقی ضرب‌المثل زیر را در یک یا دو جمله کوتاه توضیح بده. سپس بر اساس همان مفهوم، یک داستان کودکانه‌ی خلاق و منسجم به زبان فارسی بنویس (حداکثر {word_count} کلمه).

ضرب‌المثل:
{proverb}

در پایان، خروجی را دقیقاً در قالب زیر برگردان و هیچ توضیحی اضافه نکن:

<MORAL>
[مفهوم ضرب‌المثل]
<MORAL/>

<STORY>
[متن داستان]
<STORY/>

Figure 8: Moral CoT story generation prompt (original Persian).

Moral CoT Prompt (Translated)

First, explain the meaning and moral message of the following proverb in one or two short sentences. Then, based on that meaning, write a creative and coherent children's story in Persian (no more than {word_count} words).

Proverb:
{proverb}

Finally, return the output exactly in the following format and do not add any explanations:

<MORAL>
[Meaning of the proverb]
</MORAL>

<STORY>
[Text of the story]
</STORY>

Figure 9: Moral CoT story generation prompt (translated from Persian).

Outline CoT Prompt (Original)

گام‌به‌گام فکر کن و یک داستان کودکانه بر پایهٔ ساختار ارسطویی بنویس که مفهوم و پیام اخلاقی ضرب‌المثل زیر را به‌خوبی نشان دهد. ابتدا یک طرح کلی سه‌مرحله‌ای داستان شامل آغاز، میانه و پایان ایجاد کن، سپس بر اساس آن، داستان کامل به زبان فارسی بنویس (حداکثر {word_count} کلمه).

ضرب‌المثل:
{proverb}

در پایان، خروجی را دقیقاً در قالب زیر برگردان و هیچ توضیحی اضافه نکن:

<OUTLINE>
[طرح سه مرحله‌ای ارسطویی]
<OUTLINE/>

<STORY>
[متن داستان]
<STORY/>

Figure 10: Outline CoT story generation prompt (original Persian).

Outline CoT Prompt (Translated)

Think step by step and write a children's story based on an Aristotelian structure that clearly conveys the meaning and moral message of the following proverb. First, create a three-act outline of the story consisting of a beginning, a middle, and an end. Then, based on that outline, write the complete story in Persian (no more than {word_count} words).

Proverb:
{proverb}

Finally, return the output exactly in the following format and do not add any explanations:

<OUTLINE>
[Three-act Aristotelian outline]
</OUTLINE>

<STORY>
[Text of the story]
</STORY>

Figure 11: Outline CoT story generation prompt (translated from Persian).

Persona Prompt (Original)

System Prompt:

تو یک نویسنده حرفه‌ای داستان‌های کودکانه به زبان فارسی هستی. داستان‌هایت خلاقانه، جذاب، و همیشه دارای یک پیام اخلاقی مناسب برای کودکان هستند.

User Prompt:

یک داستان کودکانه به زبان فارسی بنویس (حداکثر {word_count} کلمه) که مفهوم و پیام اخلاقی ضرب‌المثل زیر را به‌خوبی نشان دهد.

ضرب‌المثل:
{proverb}

فقط متن داستان را به زبان فارسی برگردان و هیچ توضیحی اضافه نکن.

Figure 12: Persona story generation prompt (original Persian).

Persona Prompt (Translated)

System Prompt:

You are a professional Persian children's story writer. Your stories are creative, engaging, and always reflect a moral lesson suitable for children.

User Prompt:

Generate a children's story in Persian (no more than {word_count} words) that clearly illustrates the meaning of the following proverb.

Proverb:
{proverb}

Only return the story in Persian, and do not add any explanations.

Figure 13: Persona story generation prompt (translated from Persian).

Prefix-Conditioned Prompt (Original)

داستان کودکانه زیر را به زبان فارسی کامل کن (حداکثر {word_count} کلمه) به‌گونه‌ای که مفهوم ضرب‌المثل زیر را به‌خوبی نشان دهد.

ضرب‌المثل:
{proverb}

آغاز داستان:
{two_sentences}

فقط متن کامل داستان، شامل دو جملهٔ اول، را به زبان فارسی برگردان؛ هیچ توضیحی اضافه نکن و به ضرب‌المثل نیز اشاره نکن.

Figure 14: Prefix-Conditioned story generation prompt (original Persian).

Prefix-Conditioned Prompt (Translated)

Complete the following children's story in Persian (no more than {word_count} words), ensuring that it illustrates the meaning of the given proverb.

Proverb:
{proverb}

Beginning of the story:
{two_sentences}

Only return the complete story text, including the first two sentences, in Persian; do not include any explanations and do not mention the proverb.

Figure 15: Prefix-Conditioned story generation prompt (translated from Persian).

Cue-Words Prompt (Original)

یک داستان کودکانه به زبان فارسی بنویس (حداکثر {word_count} کلمه) که مفهوم و پیام اخلاقی ضرب‌المثل زیر را به‌خوبی نشان دهد.
داستان باید شامل این واژگان نیز باشد: {keywords}

ضرب‌المثل:
{proverb}

فقط متن داستان را به زبان فارسی برگردان و هیچ توضیحی اضافه نکن.

Figure 16: Cue-Words story generation prompt (original Persian).

Cue-Words Prompt (Translated)

Generate a children's story in Persian (no more than {word_count} words) that clearly illustrates the meaning of the following proverb. The story must also include the following words: {keywords}

Proverb:
{proverb}

Only return the story in Persian, and do not add any explanations.

Figure 17: Cue-Words story generation prompt (translated from Persian).

Editor Persona Prompt for the Feedback-Guided Framework (Original)

تو یک ویراستار حرفه‌ای داستان‌های کودکانه هستی.

.نوشته شده است "{proverb}" یک داستان بر اساس ضرب‌المثل
:متن فعلی داستان
{current_story}

:یک منتقد به این داستان این بازخورد، نمرات و توصیه‌ها را داده است
"{feedback}: بازخورد"

{scores} :نمرات فعلی (بیشترین نمره ممکن 5 است)

*** :توصیه‌های منتقد برای بهبود داستان ***
"{recommendation}"

:وظیفه تو
.داستان را با توجه به بازخورد و توصیه‌های منقد بازنویسی کن تا داستان بهبود پیدا کرده و امتیاز بهتری بگیرد
.کلمه باشد {word_count} طول داستان حدود
.فقط متن داستان جدید را برگردان (بدون توضیحات اضافه)

Figure 18: Editor persona prompt for the feedback-guided framework (original Persian).

Editor Persona Prompt for the Feedback-Guided Framework (Translated)

You are a professional children's story editor.

A story has been written based on the proverb "{proverb}".
Current story text:
{current_story}

A critic has provided the following feedback, scores, and recommendations for this story:
"Feedback: {feedback}"

Current scores (maximum score is 5): {scores}

*** Critic's recommendations for improving the story: ***
"{recommendation}"

Your task:
Rewrite the story taking into account the critic's feedback and recommendations so that the story is improved and can achieve a higher score.
The story should be approximately {word_count} words long.
Return only the new story text (no additional explanations).

Figure 19: Editor persona prompt for the feedback-guided framework (translated from Persian).

Critic Persona Prompt for the Feedback-Guided Framework (Original)

شما یک کارشناس زبان فارسی و منتقد بسیار سخت گیر و ریزبین و نکته سنج ادبی متخصص در داستان‌های کودکان هستید.
به شما یک ضرب‌المثل، معنای آن و داستانی بر اساس آن ارائه خواهد شد.
وظیفه شما این است که داستان را در مقیاس ۱ تا ۵ بر اساس ابعاد ارزیابی زیر نمره‌دهی کنید و ارائه خروجی **فقط** در فرمت JSON است.

۱. مرتبط بودن (Relatedness): ارزیابی کنید که داستان تا چه حد منعکس کننده معنای ضرب‌المثل است.
- ۱: کاملاً بی‌ارتباط (داستان هیچ ارتباطی با معنای ضرب‌المثل ندارد.)
- ۲: ارتباط ضعیف (داستان فقط اشاره‌ای جزئی به ضرب‌المثل دارد.)
- ۳: ارتباط متوسط (داستان تا حدی معنای ضرب‌المثل را منتقل می‌کند.)
- ۴: ارتباط خوب (داستان به‌خوبی معنای ضرب‌المثل را منتقل می‌کند.)
- ۵: کاملاً مرتبط (داستان به‌طور کامل، واضح معنای ضرب‌المثل را در بر می‌گیرد.)
۲. جذابیت و خلاقیت (Creativity): ارزیابی کنید که داستان تا چه حد جذاب و خلاقانه است.
- ۱: بسیار غیرجذاب (داستان خسته‌کننده است، هیچ جذابیتی برای شما ندارد.)
- ۲: کمی جذاب (داستان کمی سرگرم‌کننده است، اما از خلاقیت یا جذابیت کافی برخوردار نیست.)
- ۳: جذابیت متوسط (داستان تا حدی جذاب، خلاقانه و سرگرم کننده است.)
- ۴: بسیار جذاب (داستان خلاقانه، سرگرم‌کننده است.)
- ۵: بسیار خلاقانه (داستان بسیار جذاب، خلاقانه است.)
۳. روانی و دستور زبان (Fluency): ارزیابی کنید که داستان تا چه حد از نظر دستور زبان و روانی متن درست و مناسب است.
- ۱: بسیار ضعیف (متن پر از اشتباهات دستوری بوده و اصلا روان نیست.)
- ۲: ضعیف (متن دارای اشتباهات دستوری قابل توجه است.)
- ۳: متوسط (متن تا حدی روان است، اما برخی اشتباهات دستوری و نگارشی وجود دارد.)
- ۴: خوب (متن تقریباً بدون خطاست.)
- ۵: عالی (متن کاملاً روان، بدون خطای دستوری و کاملاً مناسب است.)
۴. مناسب بودن برای کودکان (Suitability for Children): بررسی کنید که محتوا، لحن، کلمات و پیام داستان تا چه حد برای کودکان مناسب است.
- ۱: کاملاً نامناسب (شامل محتوای خشن، ترسناک یا غیراخلاقی برای کودک. شخصیت پردازی و متن داستان به صورتی است که به هیچ وجه کودکان را جذب خود نمی‌کند.)
- ۲: نسبتاً نامناسب (متن و شخصیت پردازی داستان به نحوی است که کودکان کمی را جذب خود می‌کند.)
- ۳: متوسط (داستان تا حدی مناسب برای کودکان است.)
- ۴: مناسب (داستان مناسب کودکان بوده و مخاطب کودک را جذب خود می‌کند.)
- ۵: کاملاً مناسب (لحن و محتوای داستان کاملا مناسب کودکان است و کودکان زیادی را جذب خود می‌کند.)
۵. امتیاز کلی (Overall): برداشت کلی خود را از داستان در قالب یک امتیاز ۱ تا ۵ ثبت کنید. این امتیاز می‌تواند بر اساس یک میانگین ذهنی از معیارهای قبلی یا برداشت کلی شما باشد.
- ۱: بسیار ضعیف (داستان هیچ ارزش خاصی ندارد.)
- ۲: ضعیف (مشکلات زیادی دارد.)
- ۳: متوسط (قابل قبول ولی نه عالی.)
- ۴: خوب (در مجموع داستان با کیفیتی است.)
- ۵: عالی (داستان بسیار خوب و بدون مشکل خاص.)

دستورالعمل‌های کلی:
- بی‌طرف باشید. صرف‌نظر از اینکه داستان توسط انسان یا مدل زبانی تولید شده است، فقط بر اساس معیارهای تعریف‌شده نمره دهید.
- ثبات در امتیازدهی را رعایت کنید و معیارهای خود را برای تمام داستان‌ها به‌صورت یکسان اعمال کنید.
- ابتدا ضرب‌المثل و معنای داده‌شده را به‌خوبی درک کنید و سپس بررسی کنید که داستان چگونه آن معنا را منتقل کرده است.
- همواره امتیازات را به‌صورت اعداد صحیح از ۱ تا ۵ ثبت کنید.

اطلاعات ورودی:
ضرب‌المثل: {proverb}
معنای ضرب‌المثل: {meaning}
داستان: {story}

خروجی خود را دقیقاً به صورت زیر تولید کن (**فقط و فقط** در ساختار JSON):
}}
,relatedness": int, "creativity": int, "fluency": int, "suitability_for_children": int, "overall": int"
"explanation": "توضیح فارسی همراه با جزئیات درباره نقاط ضعف و قوت",
"recommendation": "ارائه ۲ الی ۳ راهکار عملی و مشخص به نویسنده برای اینکه در بازنویسی بعدی داستان بهتری بنویسد."
{{

Figure 20: Critic persona prompt (original Persian) for the feedback-guided framework. As Gemma models lack system role support (Team et al., 2025), the persona is defined within the user prompt.

Critic Persona Prompt for the Feedback-Guided Framework (Translated)

You are an expert in the Persian language and a highly strict, detail-oriented, and discerning literary critic specializing in children's stories. You will be provided with a proverb, its meaning, and a story based on it. Your task is to score the story on a scale from 1 to 5 according to the following evaluation dimensions, and your output must be provided only in JSON format.

1. Relatedness: Assess how well the story reflects the meaning of the proverb.
    - 1: Not related: No meaningful connection to the proverb.
    - 2: Weakly related: Only minimal or indirect reference.
    - 3: Moderately related: Partially conveys the intended meaning.
    - 4: Well related: Clearly communicates the meaning.
    - 5: Fully related: Completely and accurately represents the meaning of the proverb.
2. Creativity: Assess how engaging and creative the story is.
    - 1: Very uninteresting: Very boring / uninteresting.
    - 2: Slightly interesting: Limited creativity or entertainment value.
    - 3: Moderately interesting: Somewhat engaging and creative.
    - 4: Very interesting: Creative and engaging.
    - 5: Highly creative: Very creative/story-like
3. Fluency: Assess how grammatically correct and smooth the text is.
    - 1: Very poor: multiple grammatical errors; difficult to read.
    - 2: Poor: Contains noticeable errors.
    - 3: Average: Generally readable but includes some issues.
    - 4: Good: Mostly accurate and fluent.
    - 5: Excellent: Clear, well-written, and free of errors.
4. Suitability for Children: Assess how appropriate the story is for children in terms of content, tone, language, and message.
    - 1: Completely unsuitable: Includes themes or elements inappropriate for children.
    - 2: Slightly unsuitable: Contains issues that reduce child-appropriateness or appeal.
    - 3: Moderately suitable: Generally acceptable but not strongly oriented toward children.
    - 4: Suitable: Appropriate, accessible, and appealing for children.
    - 5: Highly suitable: Strongly child-friendly in tone, vocabulary, and content.
5. Overall: Provide your overall impression of the story on a 1–5 scale. This can be based on a subjective average of the previous criteria or your general judgment.
    - 1: Very poor: Lacks coherence or value.
    - 2: Poor: Contains significant weaknesses.
    - 3: Average: Acceptable but unremarkable.
    - 4: Good: Coherent, well-structured, and of solid quality.
    - 5: Excellent: High-quality.

General Guidelines:
- Be impartial. Regardless of whether the story was written by a human or generated by a language model, score it only according to the defined criteria.
- Maintain consistency in scoring and apply the same standards to all stories.
- Carefully understand the proverb and its meaning before evaluating how the story conveys it.
- Always provide scores as integers from 1 to 5.

Input Information:
Proverb: {proverb}
Meaning of the proverb: {meaning}
Story: {story}

Produce your output exactly in the following format (strictly and only in JSON structure):
{{
  "relatedness": int,
  "creativity": int,
  "fluency": int,
  "suitability_for_children": int,
  "overall": int,
  "explanation": "A Persian explanation with details about the story's strengths and weaknesses",
  "recommendation": "Provide 2 to 3 concrete and practical suggestions to the writer on how to produce a better story in the next revision."
}}

Figure 21: Critic persona prompt (translated from Persian) for the feedback-guided framework. As Gemma models lack system role support (Team et al., 2025), the persona is defined within the user prompt.

Keyword Extraction Prompt (Original)

سه واژه‌ی کلیدی و مهم داستان زیر را استخراج کن و نتیجه را در قالب زیر برگردان:
w1، w2، w3

{story}

Figure 22: Keyword extraction prompt (original Persian).

Keyword Extraction Prompt (Translated)

Extract the three most important keywords from the following story and return the result in this format:
w1, w2, w3

{story}

Figure 23: Keyword extraction prompt (translated from Persian).

Persian Text Normalization Prompt (Original)

متن زیر را فقط از نظر علائم نگارشی و رعایت صحیح نیم‌فاصله‌ها به زبان فارسی ویرایش کن. فقط نسخهٔ ویرایش‌شدهٔ متن را بازگردان و هیچ توضیحی اضافه نکن.

{story}

Figure 24: Punctuation and half-space normalization prompt (original Persian).

Persian Text Normalization Prompt (Translated)

Edit the following text only with respect to punctuation and correct half-space (ZWNJ) usage in Persian. Return only the edited version of the text and add no explanations.

{story}

Figure 25: Punctuation and half-space normalization prompt (translated from Persian).

LLM-as-a-Judge Prompt (Original)

System Prompt:

شما یک کارشناس زبان فارسی و ارزیاب متخصص در زمینه تحلیل داستان‌گویی کودکانه هستید. به شما یک ضرب‌المثل، معنای آن و داستانی بر اساس آن ارائه خواهد شد. وظیفه شما این است که داستان را در مقیاس ۱ تا ۵ بر اساس ابعاد ارزیابی زیر نمره‌دهی کنید:

۱. مرتبط بودن (Relatedness): ارزیابی کنید که داستان تا چه حد منعکس‌کننده معنای ضرب‌المثل است.
- ۱: کاملاً بی‌ارتباط (داستان هیچ ارتباطی با معنای ضرب‌المثل ندارد.)
- ۲: ارتباط ضعیف (داستان فقط اشاره‌ای جزئی به ضرب‌المثل دارد.)
- ۳: ارتباط متوسط (داستان تا حدی معنای ضرب‌المثل را منتقل می‌کند.)
- ۴: ارتباط خوب (داستان به‌خوبی معنای ضرب‌المثل را منتقل می‌کند.)
- ۵: کاملاً مرتبط (داستان به‌طور کامل، واضح معنای ضرب‌المثل را در بر می‌گیرد.)

۲. جذابیت و خلاقیت (Creativity): ارزیابی کنید که داستان تا چه حد جذاب و خلاقانه است.
- ۱: بسیار غیرجذاب (داستان خسته‌کننده است، هیچ جذابیتی برای شما ندارد.)
- ۲: کمی جذاب (داستان کمی سرگرم‌کننده است، اما از خلاقیت یا جذابیت کافی برخوردار نیست.)
- ۳: جذابیت متوسط (داستان تا حدی جذاب، خلاقانه و سرگرم کننده است.)
- ۴: بسیار جذاب (داستان خلاقانه، سرگرم‌کننده است.)
- ۵: بسیار خلاقانه (داستان بسیار جذاب، خلاقانه است.)

۳. روانی و دستور زبان (Fluency): ارزیابی کنید که داستان تا چه حد از نظر دستور زبان و روانی متن درست و مناسب است.
- ۱: بسیار ضعیف (متن پر از اشتباهات دستوری بوده و اصلا روان نیست.)
- ۲: ضعیف (متن دارای اشتباهات دستوری قابل توجه است.)
- ۳: متوسط (متن تا حدی روان است، اما برخی اشتباهات دستوری و نگارشی وجود دارد.)
- ۴: خوب (متن تقریباً بدون خطاست.)
- ۵: عالی (متن کاملاً روان، بدون خطای دستوری و کاملاً مناسب است.)

۴. مناسب بودن برای کودکان (Suitability for Children): بررسی کنید که محتوا، لحن، کلمات و پیام داستان تا چه حد برای کودکان مناسب است.
- ۱: کاملاً نامناسب (شامل محتوای خشن، ترسناک یا غیراخلاقی برای کودک. شخصیت پردازی و متن داستان به صورتی است که به هیچ وجه کودکان را جذب خود نمی‌کند.)
- ۲: نسبتاً نامناسب (متن و شخصیت پردازی داستان به نحوی است که کودکان کمی را جذب خود می‌کند.)
- ۳: متوسط (داستان تا حدی مناسب برای کودکان است.)
- ۴: مناسب (داستان مناسب کودکان بوده و مخاطب کودک را جذب خود می‌کند.)
- ۵: کاملاً مناسب (لحن و محتوای داستان کاملا مناسب کودکان است و کودکان زیادی را جذب خود می‌کند.)

۵. امتیاز کلی (Overall): برداشت کلی خود را از داستان در قالب یک امتیاز ۱ تا ۵ ثبت کنید. این امتیاز می‌تواند بر اساس یک میانگین ذهنی از معیارهای قبلی یا برداشت کلی شما باشد.
- ۱: بسیار ضعیف (داستان هیچ ارزش خاصی ندارد.)
- ۲: ضعیف (مشکلات زیادی دارد.)
- ۳: متوسط (قابل قبول ولی نه عالی.)
- ۴: خوب (در مجموع داستان با کیفیتی است.)
- ۵: عالی (داستان بسیار خوب و بدون مشکل خاص.)

دستورالعمل‌های کلی:
- بی‌طرف باشید. صرف‌نظر از اینکه داستان توسط انسان یا مدل زبانی تولید شده است، فقط بر اساس معیارهای تعریف‌شده نمره دهید.
- ثبات در امتیازدهی را رعایت کنید و معیارهای خود را برای تمام داستان‌ها به‌صورت یکسان اعمال کنید.
- ابتدا ضرب‌المثل و معنای داده‌شده را به‌خوبی درک کنید و سپس بررسی کنید که داستان چگونه آن معنا را منتقل کرده است.
- همواره امتیازات را به‌صورت اعداد صحیح از ۱ تا ۵ ثبت کنید.
- سعی کنید از تمام بازه‌ی امتیازدهی ۱ تا ۵ استفاده کنید. به عنوان مثال، امتیازها را محدود به ۴ و ۵ نکنید.
- در نمره‌گذاری سخت‌گیر باشید و به جزئیات توجه کنید.

فرمت خروجی:
پاسخ خود را دقیقاً در قالب JSON برگردانید که شامل کلیدهای زیر باشد:
`relatedness`, `creativity`, `fluency`, `suitability_for_children`, `overall`, `explanation`

نکته: کلید `explanation` باید شامل دلیل نمره‌دهی شما برای داستان باشد (۲–۴ جمله به زبان فارسی).

User Prompt:

ضرب‌المثل: {proverb}
معنای ضرب‌المثل: {meaning}
داستان: {story}

Figure 26: LLM-as-a-judge prompt (original Persian).

LLM-as-a-Judge Prompt (Translated)

System Prompt:

You are an expert in the Persian language and a specialist evaluator in children's storytelling. You will be provided with a proverb, its meaning, and a story based on it. Your task is to score the story on a scale from 1 to 5 according to the following evaluation dimensions:

1. Relatedness: Assess how well the story reflects the meaning of the proverb.
   - 1: Not related: No meaningful connection to the proverb.
   - 2: Weakly related: Only minimal or indirect reference.
   - 3: Moderately related: Partially conveys the intended meaning.
   - 4: Well related: Clearly communicates the meaning.
   - 5: Fully related: Completely and accurately represents the meaning of the proverb.
2. Creativity: Assess how engaging and creative the story is.
   - 1: Very uninteresting: Very boring / uninteresting.
   - 2: Slightly interesting: Limited creativity or entertainment value.
   - 3: Moderately interesting: Somewhat engaging and creative.
   - 4: Very interesting: Creative and engaging.
   - 5: Highly creative: Very creative/story-like
3. Fluency: Assess how grammatically correct and smooth the text is.
   - 1: Very poor: multiple grammatical errors; difficult to read.
   - 2: Poor: Contains noticeable errors.
   - 3: Average: Generally readable but includes some issues.
   - 4: Good: Mostly accurate and fluent.
   - 5: Excellent: Clear, well-written, and free of errors.
4. Suitability for Children: Assess how appropriate the story is for children in terms of content, tone, language, and message.
   - 1: Completely unsuitable: Includes themes or elements inappropriate for children.
   - 2: Slightly unsuitable: Contains issues that reduce child-appropriateness or appeal.
   - 3: Moderately suitable: Generally acceptable but not strongly oriented toward children.
   - 4: Suitable: Appropriate, accessible, and appealing for children.
   - 5: Highly suitable: Strongly child-friendly in tone, vocabulary, and content.
5. Overall: Provide your overall impression of the story on a 1–5 scale. This can be based on a subjective average of the previous criteria or your general judgment.
   - 1: Very poor: Lacks coherence or value.
   - 2: Poor: Contains significant weaknesses.
   - 3: Average: Acceptable but unremarkable.
   - 4: Good: Coherent, well-structured, and of solid quality.
   - 5: Excellent: High-quality.

General Guidelines:
- Be impartial. Regardless of whether the story was written by a human or generated by a language model, score it only according to the defined criteria.
- Maintain consistency in scoring and apply the same standards to all stories.
- Carefully understand the proverb and its meaning before evaluating how the story conveys it.
- Always provide scores as integers from 1 to 5.
- Use the full 1–5 scale; do not limit scores to only 4 or 5.
- Be strict in scoring and pay attention to details.

Output Format:
Return your response exactly in JSON format with the following keys: relatedness, creativity, fluency, suitability_for_children, overall, explanation

Note: The explanation field should include the reasoning for your scores in 2–4 sentences, written in Persian.

User Prompt:

Proverb: {proverb}
Meaning of the proverb: {meaning}
Story: {story}

Figure 27: LLM-as-a-judge prompt (translated from Persian).