Supervised Contrastive Distillation for Enhanced Story Engagement Evaluation

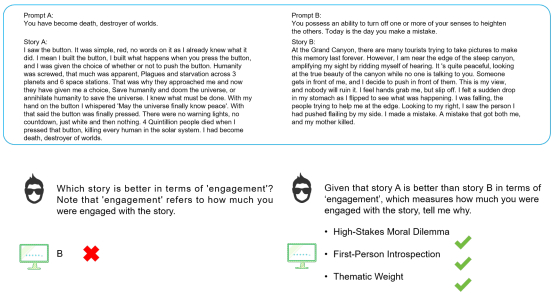
Anonymous ACL submission

Abstract

LLMs have demonstrated strong performance across a range of tasks—from sentiment analysis to factual verification—and are increasingly used to generate high-quality annotations, such as assessing story quality. Most existing quality metrics focus on objective properties rather than subjective aspects such as *engagement*, which captures how much a reader is drawn into a story. We introduce a Supervised Contrastive Distillation (SCD) framework that distills fine-grained pairwise judgments—sourced from human annotations—and explanatory knowledge from powerful teacher models into more efficient student models for evaluating story engagement. Our approach leverages a contrastive loss that aligns predicted preferences with human judgments while penalizing confidence mismatches. We validate our framework on HANNA, a human-annotated benchmark derived from the WritingPrompts corpus, and demonstrate its effectiveness in producing accurate and computationally efficient comparative evaluations. Our distilled student model achieves 40% higher accuracy than GPT-4 while reducing inference costs by approximately 80%, offering a compact yet precise evaluator.

1 Introduction

Story evaluation, a critical component of natural language generation (NLG), underpins the assessment and improvement of generative models. Unlike constrained tasks such as translation or summarization, story generation is inherently open-ended: multiple valid narratives may respond to the same prompt, differing in style, structure, creativity, and content. This diversity challenges conventional reference- and rule-based metrics, which fail to capture the full spectrum of acceptable outputs. Thus, effective evaluation methods are essential not only for benchmarking but also for guiding model development toward more coherent, engaging, and human-like narratives.



Prompt A:
You have become death, destroyer of worlds.

Story A:
I saw the button. It was simple, red, no words on it as I already knew what it did. I mean I built the button, I built what happens when you press the button, and I was given the choice of whether or not to push the button. Humanity was screwed, that much was apparent, Plagues and starvation across 3 planets and 6 space stations. That was why they approached me and now they have given me a choice, Save humanity and doom the universe, or annihilate humanity to save the universe. I knew what must be done. With my hand on the button I whispered 'May the universe finally know peace'. With that said the button was finally pressed. There were no warning lights, no countdown, just white and then nothing. 4 Quintillion people died when I pressed that button, killing every human in the solar system. I had become death, destroyer of worlds.

Prompt B:
You possess an ability to turn off one or more of your senses to heighten the others. Today is the day you make a mistake.

Story B:
At the Grand Canyon, there are many tourists trying to take pictures to make this memory last forever. However, I am near the edge of the steep canyon, amplifying my sight by ridding myself of hearing. It 's quite peaceful, looking at the true beauty of the canyon while no one is talking to you. Someone gets in front of me, and I decide to push in front of them. This is my view, and nobody will ruin it. I feel hands grab me, but slip off. I felt a sudden drop in my stomach as I flipped to see what was happening. I was falling, the people trying to help me at the edge. Looking to my right, I saw the person I had pushed flailing by my side. I made a mistake. A mistake that got both me, and my mother killed.

Which story is better in terms of 'engagement'? Note that 'engagement' refers to how much you were engaged with the story.

B ✗

Given that story A is better than story B in terms of ‘engagement’, which measures how much you were engaged with the story, tell me why.

- High-Stakes Moral Dilemma ✓
- First-Person Introspection ✓
- Thematic Weight ✓

Figure 1: Example from GPT-4

Despite its importance, story evaluation presents several challenges. First, the subjective and multidimensional nature of narrative quality complicates the definition of evaluation criteria. Attributes such as coherence, creativity, emotional resonance, and engagement are difficult to formalize and often rely on human interpretation. Second, traditional absolute scoring approaches suffer from low inter-annotator agreement, and story prompts frequently lack a canonical correct answer—making relative comparisons more informative. In particular, for engagement evaluation, pairwise comparisons offer a more grounded and interpretable framework.

Large language models (LLMs) are capable of recognizing which author produces higher-quality writing overall (Chhun et al., 2024). They are also proficient at explaining why one story is better than another when provided with a ground-truth label. However, their performance degrades when they are required to make such comparative judgments without any reference. For example, as shown in Figure 1, GPT-4 incorrectly identifies Story B as more engaging than Story A, yet it still can produce a reasonable and reliable explanation supporting the correct (ground-truth) preference. We hypothesize that such explanations can serve as effective guidance for improving evaluation models.

To address these limitations and leverage the aforementioned characteristics of LLMs, we propose a novel framework that integrates explanatory signals from LLMs into preference modeling. Specifically, we extract high-quality explanations by querying a powerful LLM (LLaMA-3.1-70B)[1]. These explanations are incorporated into a preference model trained to learn fine-grained pairwise preferences from human annotations. While multi-task learning provides a straightforward means of incorporating explanations, we instead leverage guidance via KL divergence regularization on the total loss, aligning predicted preferences with human judgments while penalizing confidence mismatches. Our approach captures the nuanced preferences of human annotators while leveraging the interpretability of LLM-generated explanations. We evaluate our framework on the HANNA dataset (Chhun et al., 2022), a human-annotated benchmark derived from WritingPrompts (Fan et al., 2018), using accuracy against human annotations as the primary metric. Experimental results show that student models guided by LLM explanations effectively approximate human preferences, achieving 40% higher accuracy than GPT-4, thus providing a scalable and reliable solution for story assessment [2].

2 Related Works

Early work on story evaluation relied on overlap-based metrics such as BLEU, ROUGE, and chrF, which assess surface-level similarity to reference texts. However, these metrics often fail to capture true narrative quality in open-ended tasks like story generation, where diverse but valid outputs exist (Guan and Huang, 2021; Wang et al., 2020). Embedding-based metrics, including BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019), compare contextual embeddings instead of raw tokens. While improving semantic sensitivity, they remain limited in detecting higher-level narrative aspects such as coherence and plot progression (Guan and Huang, 2021). Learned metrics, like BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020), are trained on human annotations, often from summarization or translation domains. These correlate better with human judgments than traditional metrics but lose effectiveness when applied to creative generation (Wang et al., 2020). Reference-free and discriminator-based models, such as ADEM (Lowe et al., 2017), RUBER (Tao et al., 2018), and UNION (Wang et al., 2020), assess generation quality without reference texts. UNION uses contrastive learning on perturbed examples to train classifier identifying low-quality stories, improving robustness but requiring careful error design. Pairwise human comparison has emerged as a preferred evaluation strategy due to higher reliability over Likert-style scoring (See et al., 2019). ActiveEval (Liu et al., 2022) builds on this by efficiently selecting story pairs via a dueling-bandit approach, substantially reducing annotation costs. Recent studies leverage LLMs as evaluators, with models like GPT-3.5 and GPT-4 achieving strong agreement with human judgments in story and summarization tasks (Zhang et al., 2023; Liu et al., 2023). Nonetheless, prompt sensitivity and limited transparency in their decision-making remain key challenges (Liu et al., 2024).

3 Proposed Approach

We propose a Supervised Contrastive Distillation (SCD) framework to train student models that mimic fine-grained pairwise quality assessments derived from human annotations, with explanations provided by large language model (LLM) teachers. Unlike prior work limited to prompt-specific comparisons, our method supports cross-prompt evaluation, broadening narrative-quality coverage. During training, the student conditions on both explanation-augmented and explanation-free inputs, enabling robust, interpretable inference without requiring explanations at test time.

3.1 Notation and Input Construction

Let each training sample be defined by the tuple:

$$(p_A,\ s_A,\ p_B,\ s_B,\ e,\ y)$$

where:

- p_A, p_B are distinct prompt texts,
- s_A, s_B are the corresponding story generations,
- e is the teacher-provided natural-language explanation,
- $y \in \{0, 1\}$ is the binary preference label ($y = 0$ if s_A preferred; $y = 1$ if s_B preferred).

---

[1] Details of the explanation generation process are provided in Appendix A.2

[2] The limitation of our method is described in Section 7.

The model outputs a probability distribution over the two candidate labels ("Story A" and "Story B"), denoted by $\hat{y}_{\mathrm{w}}$ and $\hat{y}_{\mathrm{wo}}$ for inputs with and without LLM-provided explanations, respectively.

3.2 Objective Function

We optimize a composite loss that combines supervised cross-entropy and distributional consistency:

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{w}} = -\sum_{c\in\{A,B\}} \nVdash[c=y] \log \hat{y}_{\mathrm{w}}(c), \qquad (1)$$

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{wo}} = -\sum_{c\in\{A,B\}} \nVdash[c=y] \log \hat{y}_{\mathrm{wo}}(c), \qquad (2)$$

$$\mathcal{L}_{\mathrm{KL}} = \sum_{c\in\{A,B\}} \hat{y}_{\mathrm{w}}(c) \log \frac{\hat{y}_{\mathrm{w}}(c)}{\hat{y}_{\mathrm{wo}}(c)}. \qquad (3)$$

The total loss is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}}^{\mathrm{w}} + \mathcal{L}_{\mathrm{CE}}^{\mathrm{wo}} + \lambda\,\mathcal{L}_{\mathrm{KL}}, \qquad (4)$$

with hyperparameter λ controlling the strength of the consistency penalty.

3.3 KL Divergence Regularization

To enforce consistency between explanation-augmented and explanation-free predictions, we introduce a KL divergence regularizer. Given

$$\hat{y}_{\mathrm{w}} = \big(\hat{y}_{\mathrm{w}}(A),\, \hat{y}_{\mathrm{w}}(B)\big),$$
$$\hat{y}_{\mathrm{wo}} = \big(\hat{y}_{\mathrm{wo}}(A),\, \hat{y}_{\mathrm{wo}}(B)\big),$$

The KL divergence is defined as

$$\mathrm{KL}\big(\hat{y}_{\mathrm{w}} \parallel \hat{y}_{\mathrm{wo}}\big) = \sum_{c\in\{A,B\}} \hat{y}_{\mathrm{w}}(c)\, \log \frac{\hat{y}_{\mathrm{w}}(c)}{\hat{y}_{\mathrm{wo}}(c)}.$$

Adding

$$\lambda\,\mathrm{KL}\big(\hat{y}_{\mathrm{w}} \parallel \hat{y}_{\mathrm{wo}}\big)$$

to the total loss penalizes shifts in confidence caused by omitting explanations; here, $\lambda > 0$ controls regularization strength. By applying this term to the total loss, the model is encouraged to align its output distribution for inputs without explanations with that for inputs with explanations—enabling it to learn from the explanatory signal without reproducing the explanation itself.

4 Experiments

We evaluate our proposed method and baselines on the HANNA dataset (Chhun et al., 2022)[3]. The explanations are produced by LLaMA-3.1-70B-Instruct[4].

4.1 Benchmarked LLM Performance

For reference, we evaluate several state-of-the-art large language models (LLMs), including GPT-4 and LLaMA-3.1-70B-Instruct, on a held-out subset comprising 10% of the pairwise comparison data ($\tilde{2}$K samples). As shown in Table 1, despite their strong general capabilities, these models exhibit limited performance on this task without task-specific supervision. This underscores the need for dedicated training and motivates the development of our student models.

Model	**Accuracy**	**F1**
LLMs (LLaMA-3.1-70B-Instruct)	0.57	0.54
LLMs (GPT-4)	0.51	0.51

Table 1: Performance of LLMs on the held-out subset of the dataset. Reported metrics include accuracy and F1 score.

4.2 Baselines and Comparisons

We evaluate two distinct student models that differ in how they leverage the teacher's judgments and explanatory signals. All models are fine-tuned from LLaMA-3.1-8B-Instruct using a LoRA configuration with $r = 64$, lora_alpha $= 16$, 4-bit quantization, and a LoRA dropout of 0.5.

Pairwise-only Baseline This model is trained exclusively as a discriminator based on LLaMA-3.1-8B-Instruct. Given two independently generated story-prompt pairs, it predicts which narrative better fulfills its prompt for engagement. The input consists solely of the two story-prompt pairs, and the model produces a binary output ("A" or "B") indicating the preferred example. No explanatory text is generated. Model parameters are optimized using a standard classification loss with binary labels provided by the teacher LLM.

Multi-task Generative Model We reformulate the evaluation task as a unified sequence-generation problem. Inputs are prefixed with a marker (e.g., '[cls]') to signal the classification phase, during which the model generates a single token—"A" or "B"—to indicate the preferred story. A second marker (e.g., '[gen]') then prompts the model to continue generating the natural-language explanation provided by the teacher.

---

[3] Dataset details are provided in Appendix A.1

[4] Explanation generation details are reported in Appendix A.2

Model	Accuracy	F1	Accuracy (H)	F1 (H)
Multi-task	0.85	0.83	0.74	0.71
Baseline (P)	0.81	0.78	0.69	0.68
+KL Regularization	**0.91**	**0.91**	**0.87**	**0.86**

Table 2: Qualitative performance, where ‘P’ denotes the Pairwise-only model and ‘H’ indicates harder instances with $<$1-point difference in human engagement scores.

5 Results & Analysis

We report the qualitative comparison of model performance on the dataset in Table 2, using accuracy and F1 score as evaluation metrics. All scores are averaged over five runs with different random seeds for each model.

5.1 GENERAL PERFORMANCE

Table 2 reports the performance of distilled student models compared to baselines. The pairwise student achieves 81% accuracy and an F1 score of 0.78, substantially outperforming GPT-4 (51% / 0.51) and LLaMA-3.1-70B-Instruct (57% / 0.54) on the HANNA engagement comparison task. The multi-task student, which jointly predicts pairwise preferences and generates explanations, further improves to 85% accuracy and 0.83 F1. Our full proposed approach attains the highest performance, with 91% accuracy and 0.91 F1. On the harder subset—samples with two stories having human-annotated *engagement* scores differing by less than 1—baseline models suffer significant drops, whereas our model’s performance declines only slightly, demonstrating the effectiveness of leveraging guidance from large LLMs. All evaluations are conducted against gold-standard judgments from expert human annotators.

Model	Inconsistency Rate (%)
Pairwise-only baseline	17.4
Pairwise + KL regularization	4.1

Table 3: Prediction inconsistency rates across input formats.

5.2 The KL Regularization

To enhance the robustness of the pairwise-only student, we introduce a KL divergence regularization term enforcing consistency between predictions with and without LLM-generated explanations. This regularization penalizes confidence shifts caused by absent explanation input, encouraging the model to internalize stable decision boundaries. The KL-regularized student achieves significant gains, with accuracy increasing from 81% to 91% and F1 rising from 0.78 to 0.91 (see Table 2). Notably, this performance exceeds that of the multi-task model, which benefits from explicit explanation supervision.

To better understand the effect of KL regularization, we measure prediction agreement between explanation-present and explanation-absent inputs, as reported in Table 3. Without the KL term, 17.4% of test instances yield differing predictions under the two formats. This discrepancy decreases to 4.1% with KL regularization, indicating substantially improved prediction stability.

Furthermore, the KL-regularized model achieves consistent accuracy gains across all confidence bins, with the most notable improvement in the highest bin (0.8–1.0), where accuracy rises from 82% to 88%. This suggests not only improved calibration, where confidence scores better reflect true correctness, but also stronger performance on high-confidence predictions, reducing overconfidence and enhancing model trustworthiness.

These results demonstrate KL regularization effectively aligns student predictions across input variants. Although the model does not require explanations during inference, training robustness to their presence fosters stronger inductive bias and improved generalization. In the absence of full explanation supervision (as in multi-task setups), KL regularization provides a lightweight yet powerful alternative to leverage explanation-derived signals.

Confidence Bin	Baseline Accuracy	KL-regularized Accuracy
0.4–0.6	50%	55%
0.6–0.8	68%	72%
0.8–1.0	82%	88%

Table 4: Calibration by confidence bin.

6 Conclusion

We distill human narrative preferences into a compact student model using pairwise supervision and KL regularization, aligning its predictions with LLM-guided reasoning. The model is trained to match its output distribution on inputs without explanations to those with explanations, enabling it to benefit from explanatory signals without reproducing them. This yields 40% higher accuracy than GPT-4 while cutting inference costs by 80%, offering an efficient and reliable evaluator.

7 Limitation

- **Dependence on Teacher Model Quality**: The student model's performance hinges on the accuracy and bias of the teacher LLM's explanations, which may propagate errors or subjective tendencies.
- **Generalization Constraints**: Training on the HANNA dataset risks overfitting to its specific engagement criteria, limiting applicability to diverse genres or cultural contexts.
- **Explainability Trade-off**: While explanations guide training, their absence during inference reduces decision transparency, complicating trust in real-world deployment.
- **Scalability Costs**: Generating high-quality explanations for large datasets requires substantial computational resources, offsetting some efficiency gains from distillation.
- **Ethical Risks**: Biases in human annotations or teacher models could perpetuate unfair evaluations, particularly for subjective metrics like engagement.

References

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M. Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of COLING*.

Cyril Chhun, Fabian M Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

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.

Juncen Guan and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson Liu, Urvashi Khandelwal, Sean Welleck, Samuel R Bowman, and Omer Levy. 2022. Activeeval: An efficient human evaluation framework for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson Liu, Sean Welleck, Ximing Zhou, Caiming Xiong, and Samuel R Bowman. 2024. Mind's mirror: Llm self-evaluation via causal and counterfactual reasoning. *arXiv preprint arXiv:2402.02929*.

Nelson Liu, Sean Welleck, Ximing Zhou, Mo Yu, Caiming Xiong, and Samuel R Bowman. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2309.00624*.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Meta AI. 2025. LLaMA 3.1: Open foundation language models.

Ricardo Rei, Ana Farinha, Alon Lavie, and Andre FT Martins. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Abigail See, Peter J Liu, and Christopher D Manning. 2019. What makes a good summary? reconsidering the focus of automatic summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chongyang Tao, Wei Wu, Can Xu, Yansong Zhang, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qingyun Wang, Xu Hua, Yichi Yang, Xiaojun Wan, and Xiaodong He. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Yujia Zhang, Qingxiu Deng, Bill Yuchen Lin, and Xiang Ren. 2023. Benchmarking large language models as evaluators. *arXiv preprint arXiv:2305.14795*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 563–578.

A Appendix

A.1 Dataset

We evaluate our method on the HANNA dataset (Chhun et al., 2022), which contains 1,056 narratives rated by three independent annotators across six quality dimensions using a five-point Likert scale. Our experiments focus on the *engagement* dimension, measuring how well raters understood characters' emotions, regardless of agreement. Each story's *engagement* score is the average of its three ratings.

To simulate realistic distillation supervision, we randomly sample 20,000 prompt–story pairs from HANNA and use LLaMA-3.1-70B to generate pairwise explanations justifying why one story is superior (Meta AI, 2025). The augmented dataset is split into 80% training, 10% development, and 10% test subsets.

A.2 Explanations from LLMs

In our framework, LLM-generated explanations act as intermediate supervision signals, guiding student models to better capture the rationale behind story preferences. These explanations clarify why one story is preferred over another in terms of the engagement dimension, highlighting aspects such as emotional resonance, character believability, and narrative immersion. To generate these explanations, we prompt LLaMA-3.1-70B-Instruct with two prompt–story pairs and a binary label indicating which story is more engaging. The LLM then produces a concise, natural-language justification focused exclusively on engagement-related attributes. The exact prompting template used to elicit these explanations is provided below. The bold text is fixed, while the regular text is used to fill in different prompts and stories.

System:
You are a helpful chatbot who can evaluate the quality of stories in terms of specific metrics. Given two prompts followed by two generated stories and an indication of which story is better, your task is to explain why one story is better than the other based on their level of engagement, which measures how much the reader is drawn into or emotionally involved with the story. Your evaluation should focus solely on engagement, without considering any other metrics. Below are the prompts and the stories.
User:

Prompt A:
When you die the afterlife is

an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.
Story A:

3,000 years have I been fighting.

Every morning, the raccoons scratch at my eyes. Every evening, the skunks spray me while the opossums chew at my feet. I have never had any tools. I have only my hands. I don't remember the place I came from before this. All I remember is the daily fight between me and these animals. No matter how many times I kill them, they come back the next day.
Prompt B:
When you die the afterlife is

an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.
Story B:

First of all, not everyone was

entitled to be an exterminator. But the ones that were - maybe were, like, *genius*, because, yes, I had once belonged to a less fortunate class of people - had all the opportunity to work for the damn plant killer, and it's hard work. And the horrifying truth is, once you die, and the entire planet turns into a glade that contains a golden fish that would've been crushed by a million million goldfish just moments ago, you're not really good enough for heaven.

Given that story A **is better than story** B **in terms of 'engagement', which measures how much you were engaged with the story, tell me why.**

Assistant:

Story A places the reader immediately into a vivid, immersive, and emotionally charged scenario. The protagonist's daily, brutal struggle in the afterlife arena is described with concrete sensory details—raccoons scratching at eyes, skunks spraying, opossums chewing feet—which evoke a visceral reaction. The repetition of the torment over 3,000 years, combined with the stark image of barehanded combat and the loss of memory, creates a haunting atmosphere that pulls the reader in emotionally and raises compelling questions about perseverance, identity, and punishment.

In contrast, Story B is abstract and disjointed. Its ideas are harder to follow and the imagery (e.g., a glade with a golden fish crushed by goldfish) feels surreal without clear emotional stakes. The tone is also more rambling and less immediate, making it harder for the reader to connect with the narrator's experience or feel invested in the outcome. As a result, Story B lacks the emotional immediacy and narrative clarity that make Story A more engaging.