

Iterative Dual-Model Alignment for Story Evaluation

Bruce Qin
Purdue University
qin184@purdue.edu

Dan Goldwasser
Purdue University
dgoldwas@purdue.edu

Abstract

Large language models (LLMs) can both evaluate and explain text quality; however, most existing evaluators operate as static classifiers and lack the ability to refine their reasoning through interaction. We propose an **Iterative Alpha-Beta Learning** framework that jointly trains two complementary 8B models: an Alpha (α) classifier that assesses pairwise story engagement, and a Beta (β) generator that produces structured, rubric-guided comparative explanations. The two models co-evolve within a closed feedback loop: α provides probabilistic preference signals to guide β 's Direct Preference Optimization (DPO), while β 's improved explanations are reintegrated to retrain α via a KL-based contrastive objective. This dual optimization enables mutual learning: α gains interpretability and robustness from β 's textual rationales, while β acquires stronger alignment and discriminative precision from α 's confidence deltas. Experiments on human-annotated story-pair datasets (HANNA) show that the proposed system consistently outperforms strong single-model baselines in both accuracy and explanation quality across multiple iterative rounds¹.

1 Introduction

Story evaluation has become an increasingly important and substantively challenging problem in natural language processing (NLP), particularly in the context of creative writing (Chhun et al., 2024; Fein et al., 2025). Unlike tasks with well-defined notions of correctness, stories admit multiple valid realizations for a given prompt, and their quality is judged along subjective, high-level dimensions such as coherence, engagement, emotional resonance, and narrative flow. This inherent open-endedness makes evaluation both scientifically non-trivial and practically consequential for analyzing

and guiding modern large language models (LLMs) that generate long-form text. Accordingly, a central question is whether existing automated metrics can meaningfully capture narrative-level qualities in a manner that aligns with human judgments.

In practice, most widely used evaluation metrics fall short of this goal. Standard surface-form measures, including n-gram overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), are efficient but correlate poorly with human assessments of story quality. Learned and embedding-based metrics, such as BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020), improve semantic sensitivity but remain limited in capturing higher-level narrative properties like coherence, salience, and engagement. More recent work explores story-specific and reference-free evaluators, including specialized models (Guan et al., 2021) and LLM-based judges (Liu et al., 2023), yet concerns around calibration, robustness, and interpretability persist, particularly in creative writing settings (Fein et al., 2025).

These evaluation challenges are tightly coupled with the growing use of large language models as creative writing assistants. LLMs are increasingly deployed to support ideation, drafting, and feedback, yet the effectiveness of such systems depends critically on how narrative quality is assessed and communicated. Prior work shows that structured assistance can improve long-range storytelling (Mirowski et al., 2022), while the form and granularity of model interventions substantially shape writer outcomes and experience (Dhillon et al., 2024). More recently, research has begun to examine automated feedback for narratives, finding that although LLM-generated critiques are often fluent and plausible, they frequently miss the most consequential story-level issues (Rashkin et al., 2025). Together, these findings highlight a gap between the increasing use of LLMs in creative writing and our limited ability to reliably evaluate

¹Our code, models, datasets, and used prompts will be released on GitHub.

and guide story generation.

Despite rapid progress, evaluation and writing assistance have largely evolved along separate trajectories. Story evaluators, including automated metrics, LLM-based judges, and supervised preference models, can rank or score candidate stories, yet they often fail to produce reliable, human-interpretable rationales that remain stable across prompts and data distributions (Guan et al., 2021; Fein et al., 2025). Conversely, writing assistants and feedback generators can provide detailed critiques, but their outputs are typically not grounded in a consistent comparative preference model and are therefore difficult to use as training signals for robust engagement evaluation (Rashkin et al., 2025). As a result, a unified framework that jointly improves preference prediction and explanation quality in a mutually reinforcing manner remains lacking.

Rather than treating evaluation and explanation as separate components, we propose a joint modeling framework that integrates both within a closed feedback loop. The core idea is to enable an evaluator (α) and an explainer (β) to iteratively refine one another: evaluative signals guide the generation of more informative explanations, while reasoning over explanations yields more robust and interpretable evaluations. We implement this framework by training two task-specialized, moderately sized language models instead of relying on repeated calls to commercial LLM APIs. This approach substantially reduces computational and monetary costs while allowing precise control over model behavior, training dynamics, and evaluation protocols.

Our main contributions are summarized as follows:

- We introduce the **Iterative Alpha–Beta Learning Framework**, a dual-model architecture that jointly optimizes a preference classifier and an explanation generator via round-based mutual feedback.
- We evaluate the α evaluator on the HANNA dataset with human engagement labels and assess the β explainer through its effects on α 's confidence and accuracy, expert human judgments, and downstream story revision performance. Across multiple iterative training rounds on human-annotated story-pair datasets, we observe consistent improvements in preference accuracy and explanation quality relative to strong single-model baselines,

with aligned explanations preferred by expert annotators and leading to higher-quality downstream A/B story revisions.

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 et al., 2021; Guan and Huang, 2020). Embedding-based metrics, including BERTScore (Zhang et al., 2019) 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 et al., 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 (Guan and Huang, 2020). Reference-free and discriminator-based models, such as ADEM (Lowe et al., 2017), RUBER (Tao et al., 2018), and UNION (Guan and Huang, 2020), assess generation quality without reference texts. UNION uses contrastive learning on perturbed examples to train a 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 (ter Hoeve et al., 2020). 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 (Liu et al., 2023). Nonetheless, prompt sensitivity and limited transparency in their decision-making remain key challenges (Liu et al., 2024).

Large language models (LLMs) are increasingly used as writing assistants, supporting tasks ranging from idea generation and drafting to editing and stylistic refinement. Systems such as *Wordcraft* and *GhostWriter* enable users to co-author stories with AI-generated suggestions, thereby enhancing ideation and writing productivity (Yuan et al., 2022; Yeh et al., 2024). However, research on human

AI co-writing reveals important trade-offs: while LLM assistance often improves fluency and efficiency, it can also shape authorial style and reduce writers’ sense of ownership over the text (Gero et al., 2023). Recent work has begun to address these concerns by designing interfaces that provide users with greater control over LLM contributions. For example, some systems allow authors to specify the AI persona or writing mode to better align generated content with authorial intent (Lee et al., 2022). At the same time, the widespread adoption of LLMs in professional and academic writing has intensified debates around plagiarism and responsible use, motivating calls for clearer guidelines and more robust detection methods. Most closely related to our work, Rashkin et al. (2025) introduces a dedicated task and dataset for evaluating LLM-generated feedback on story writing. Their findings show that, although contemporary LLMs often produce feedback that is specific and superficially plausible, they frequently fail to identify the most salient writing issues and struggle to appropriately balance critical and positive comments.

3 Proposed Approach

3.1 Notation

Let the tuple define each training sample:

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

where:

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

We define two task-specialized language models with complementary roles:

- The α model is a pairwise evaluator that predicts which of the two stories is more engaging according to human preferences. Given a story pair, it outputs a preference label.
- The β model is an explainer that generates structured, rubric-guided comparative explanations describing how the two stories differ along engagement-related dimensions, without explicitly stating which story is preferred.

3.2 Overall structure

We introduce an iterative framework that jointly improves story engagement evaluation and generates informative explanations that guide the production of higher-quality stories. Figure 1 illustrates the overall system architecture. The system operates on a pair of prompt story inputs, denoted as (p_A, s_A) and (p_B, s_B) , where p represents the writing prompt and s the corresponding story. Given a story pair, the β model generates a structured, rubric-guided comparative explanation that characterizes differences between the two stories along engagement-related dimensions. The α model then predicts which story is more engaging, optionally conditioning on the generated explanation, and we evaluate the extent to which the explanation improves both prediction confidence and accuracy.² Each explanation is further assessed by the α model by comparing predictions made with and without the explanation and measuring the resulting confidence shift. These confidence shifts are used to rank explanations. High-impact explanations are used to update β via preference optimization, while the strongest explanation for each pair is also incorporated as auxiliary supervision to update α . Through this closed feedback loop, the generator learns to produce more informative explanations and the classifier learns more robust and interpretable preference judgments, with both models improving jointly across iterations.

3.3 α classifier

Overall structure. α is a causal language model employed as a binary classifier over $\{A, B\}$. Classification is performed by extracting the last-token logits at a designated answer position and applying a softmax over the corresponding label tokens.

Input of α . For each example with fields (p_A, s_A, p_B, s_B, y) , we construct two variants: a no-explanation prompt and a with-explanation prompt, where the latter inserts a candidate explanation E from β into the system preamble.

Intermediate steps in α . α tokenizes the input, identifies the string “Answer :” to locate the answer position, truncates the input at that point, performs a single forward pass, and extracts the logits at the final position. Applying a softmax over the label token IDs yields $p_\theta(y | \cdot)$. For any explanation E ,

²All prompts used during training are reported in Appendix A.1.

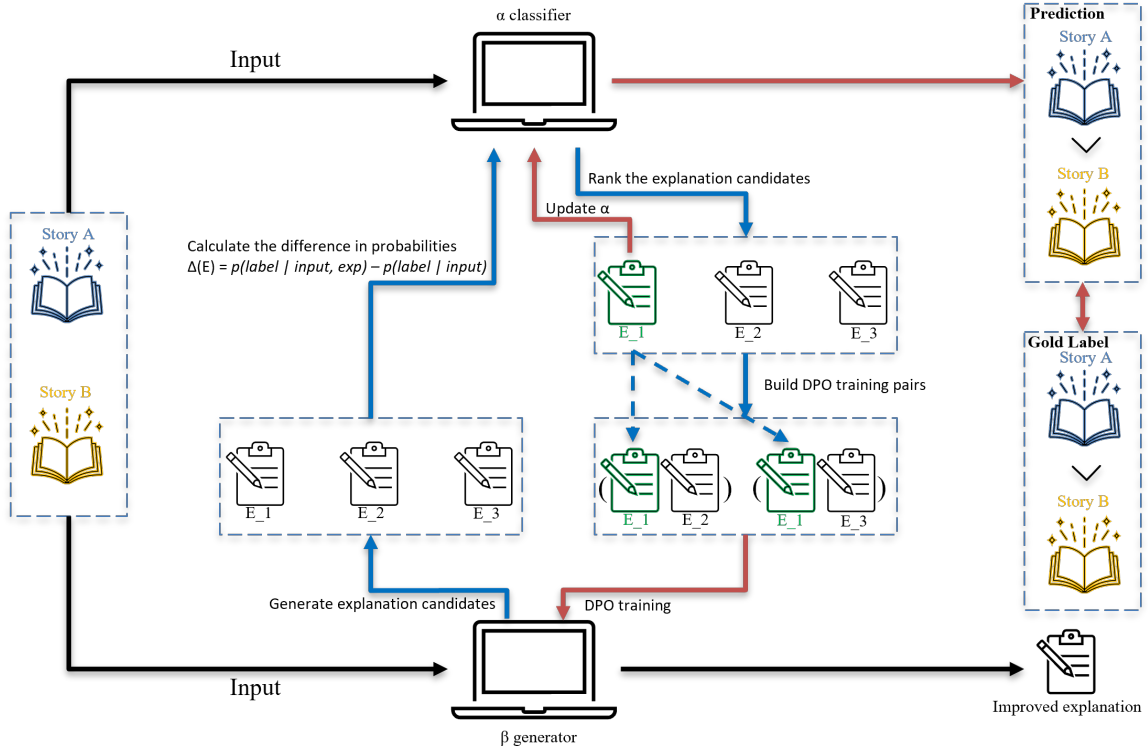


Figure 1: Overall model architecture. Blue arrows indicate data flow, while red arrows denote parameter updates. E_x represents the x th explanation candidate generated by β , and the explanation highlighted in green corresponds to the top-ranked candidate selected by the α classifier.

we define the confidence uplift as

$$\Delta(E) = p_{\theta}(y | E) - p_{\theta}(y | \emptyset), \quad (1)$$

computed using the paired with and without variants described above.

Final output of α . At inference time, α predicts the preferred option by comparing the label-conditioned probabilities at the answer position (“Answer: A” versus “Answer: B”).

Loss of α . α is trained using a KL-style cross-entropy objective at the answer token. Specifically, let $q(y)$ denote the target distribution, defined as a one-hot vector over the gold label, and let $p_{\theta}(y)$ denote the predicted two-class distribution produced at the final token. The per-example loss is

$$\begin{aligned} \mathcal{L}_{\text{KL}}(\theta) &= \text{KL}(q(y) \parallel p_{\theta}(y)) \\ &= - \sum_{y \in \{A, B\}} q(y) \log p_{\theta}(y) + \text{const.} \quad (2) \end{aligned}$$

3.4 β generator

Overall structure. β is a causal language model that generates structured, rubric-guided comparative explanations with seven fixed section headings.

It is designed to be label-blind while remaining evidence-grounded in both stories.

Input and Output of β . We construct a prompt containing (p_A, s_A, p_B, s_B) and instruct β to generate a comparative explanation that contrasts the two stories along engagement-relevant dimensions, without expressing a preference or making a prediction. The explanation is constrained to consist of exactly seven sections: Emotional Resonance, Mystery and Sensory Intrigue, Tension and Conflict, World Building, Pacing, Character, and Clarity and Structure³.

Intermediate steps in β . We sample m candidate explanations for each item. These samples are sent to α for evaluation and ranking. The resulting rankings are then used to construct DPO pairs for subsequent training of β .

Loss of β . We train β using DPO on (chosen, rejected) explanation pairs. Based on ranking scores produced by α , the top-ranked explanation is treated as the chosen explanation, while all re-

³The seven fixed headings are determined based on clustering results of LLM-generated explanations. Further details are provided in Appendix A.2.

maining candidates are treated as rejected explanations. Let ϕ denote the parameters of β and γ the inverse temperature. The DPO objective encourages a higher relative preference for the chosen explanation over the rejected explanation:

$$\mathcal{L}_{\text{DPO}}(\phi) = -\log \sigma(\gamma \Delta_\phi), \quad (3)$$

where $\Delta_\phi = \log p_\phi(E^+) - \log p_\phi(E^-)$ and $\sigma(\cdot)$ denotes the logistic function.

4 Experiments

4.1 Datasets

We evaluate the proposed method and baseline models on the HANNA dataset (Chhun et al., 2022), which contains 1,056 narratives annotated by three independent raters along six quality dimensions using a five-point Likert scale. Our experiments focus on the engagement dimension, which captures the extent to which a reader is engaged with a story. For each narrative, the engagement score is computed as the average of the three human ratings. Based on these averaged scores, we construct a pairwise comparison dataset in which each instance consists of two prompts and story pairs. For each instance, two stories are randomly sampled from the corpus, and a preference label is assigned according to their average engagement scores. To prevent over-representation, each story appears in at most 25 comparison instances.

To simulate realistic supervision for knowledge distillation, we randomly sample 16,000 prompt and story pairs from the comparison dataset for training. We additionally sample 2,000 pairs for validation and 2,000 pairs for testing. Stories included in the training, validation, and test sets are mutually exclusive.

4.2 Benchmarked LLM Performance

For reference, we evaluate several large language models (LLMs) on the test split of the pairwise comparison dataset (approximately 2K samples). As shown in Table 1, most models perform poorly without task-specific supervision, despite strong general capabilities, whereas GPT-5.2-thinking achieves substantially better results. These findings underscore the limitations of out-of-the-box LLMs for story evaluation and motivate dedicated training, which informs our model design. We further evaluate few-shot learning for Llama-3.3-70B-Instruct with five in-context examples, obtaining an accuracy of 0.58 and an F1 score of 0.57. This

| Model | Accuracy | F1 |
|------------------------|----------|------|
| Llama-3.3-70B-Instruct | 0.48 | 0.46 |
| Llama-4-latest | 0.62 | 0.57 |
| GPT-4 | 0.57 | 0.55 |
| GPT-5.2-thinking | 0.87 | 0.84 |

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

result aligns with Chhun et al. (2024), who show that LLMs are more reliable at selecting the better storyteller than at evaluating individual stories in isolation.

4.3 Baselines and Comparisons

All models are fine-tuned on a single A100 GPU using Llama-3.1-8B-Instruct as the base model. Fine-tuning is performed with a LoRA configuration using rank $r = 64$, $\text{lo_ra_alpha} = 16$, a LoRA dropout rate of 0.5, and 4-bit quantization. Both Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct are obtained from Hugging Face.

Pairwise only Baseline This model is trained solely as a discriminator based on Llama-3.1-8B-Instruct. Given two independently generated story prompt pairs, it predicts which narrative better satisfies its prompt in terms of engagement. The input consists only of the two story prompt pairs, and the model outputs a binary decision (“A” or “B”) indicating the preferred example. No explanatory text is generated. Model parameters are optimized using a standard binary classification loss.

Multi-task Generative Model We reformulate the evaluation task as a unified sequence generation problem. Inputs are prefixed with a marker (for example, [cls]) to indicate the classification phase, during which the model generates a single token (“A” or “B”) denoting the preferred story. A second marker (for example, [gen]) then prompts the model to generate a natural language explanation aligned with the explanation labels. These explanation labels are produced by the Llama-3.3-70B model and describe why one story is more engaging than the other.

5 Evaluation

5.1 Evaluating the Explanation Model (β)

We evaluate whether explanations E generated by β improve the labeler α and overall story quality. Each item i comprises a prompt and two stories (A_i, B_i) with gold label $y_i \in \{A, B\}$. The labeler α outputs class probabilities with and without the explanation E_i :

$$\begin{aligned} p_i^{\text{wo}}(y) &\equiv \Pr_{\alpha}(y \mid A_i, B_i), \\ p_i^{\text{w}}(y) &\equiv \Pr_{\alpha}(y \mid A_i, B_i, E_i). \end{aligned} \quad (4)$$

Predictions are defined as $\hat{y}_i^{\text{wo}} = \arg \max_y p_i^{\text{wo}}(y)$ and $\hat{y}_i^{\text{w}} = \arg \max_y p_i^{\text{w}}(y)$. We define the baseline confidence as $c_i = \max\{p_i^{\text{wo}}(A), p_i^{\text{wo}}(B)\}$.

5.2 Log-odds lift

We quantify how E_i updates α 's calibrated belief about the true label. Let the log odds before and after observing E_i be

$$\begin{aligned} \ell_i^{\text{wo}} &= \log \frac{\tilde{p}_i^{\text{wo}}(y_i)}{1 - \tilde{p}_i^{\text{wo}}(y_i)}, \\ \ell_i^{\text{w}} &= \log \frac{\tilde{p}_i^{\text{w}}(y_i)}{1 - \tilde{p}_i^{\text{w}}(y_i)}, \end{aligned} \quad (5)$$

where $\tilde{p} = \min(\max(p, \epsilon), 1 - \epsilon)$ with $\epsilon = 10^{-6}$ to ensure numerical stability. The per item log odds lift is

$$\Delta \ell_i = \ell_i^{\text{w}} - \ell_i^{\text{wo}}. \quad (6)$$

We report $\mathbb{E}[\Delta \ell]$, $\text{median}(\Delta \ell)$, and $\Pr(\Delta \ell > 0)$. A positive lift indicates that E_i increases α 's confidence in the correct label.

5.3 Win-rate@ k by baseline confidence

We partition items by α 's baseline confidence c_i and quantify improvements ("rescues") and harms within each bin. Let the bin boundaries be $\tau_0 < \tau_1 < \dots < \tau_K$, defined by fixed thresholds or quantiles. We define bin $b_k = \{i : \tau_{k-1} < c_i \leq \tau_k\}$.

Rescue Win-rate@ k . Among instances where α is initially incorrect, we measure the fraction that are corrected after incorporating E_i :

$$\text{RescueWR}_k = \frac{\sum_{i \in b_k} \gamma_i}{\max\left(1, \sum_{i \in b_k} \delta_i\right)}, \quad (7)$$

where $\gamma_i = \mathbf{1}[\hat{y}_i^{\text{wo}} \neq y_i \wedge \hat{y}_i^{\text{w}} = y_i]$ and $\delta_i = \mathbf{1}[\hat{y}_i^{\text{wo}} \neq y_i]$. This metric captures whether β improves performance on challenging instances characterized by low c_i .

Harm rate & Net Win-rate. We further report the harm rate,

$$\text{HarmRate}_k = \frac{H_k}{\max(1, C_k)}, \quad (8)$$

where $H_k \equiv \sum_{i \in b_k} \mathbf{1}[\hat{y}_i^{\text{wo}} = y_i \wedge \hat{y}_i^{\text{w}} \neq y_i]$ counts previously correct predictions that become incorrect, and $C_k \equiv \sum_{i \in b_k} \mathbf{1}[\hat{y}_i^{\text{wo}} = y_i]$ counts originally correct predictions in bin b_k .

We also report the net improvement,

$$\text{NetWR}_k = \frac{|G_k| - |L_k|}{|b_k|}. \quad (9)$$

Here, $G_k = \{i \in b_k : \hat{y}_i^{\text{wo}} \neq y_i, \hat{y}_i^{\text{w}} = y_i\}$ denotes corrected errors, and $L_k = \{i \in b_k : \hat{y}_i^{\text{wo}} = y_i, \hat{y}_i^{\text{w}} \neq y_i\}$ denotes newly introduced errors.

5.4 Matched vs. mismatched explanations

To verify that performance gains arise from the item-specific signal, we compare the effect of the matched explanation E_i with mismatched explanations E_j drawn from other items. Let Δ_i^{match} denote a chosen per-item effect, such as probability lift $p_i^{\text{w}}(y_i) - p_i^{\text{wo}}(y_i)$ or log odds lift $\Delta \ell_i$, and let $\Delta_i^{\text{mismatch}}$ denote the same effect when pairing (A_i, B_i) with an explanation E_j for $j \neq i$, averaged over M randomly selected indices j for stability. We report the average treatment effect (ATE):

$$\text{ATE} = \mathbb{E}\left[\Delta_i^{\text{match}} - \Delta_i^{\text{mismatch}}\right]. \quad (10)$$

A positive ATE indicates that matched explanations provide item-specific utility beyond generic text. Statistical significance is assessed using a paired t test or a permutation test over per-item differences.

5.5 A/B revision win-rate (extrinsic quality)

We evaluate whether explanations improve story rewriting quality. For each prompt, we generate a baseline draft $S_{0,i}$ and a revised draft $S_{1,i}$ guided by E_i (e.g., using the same generator with and without conditioning on E_i). A judge compares $(S_{0,i}, S_{1,i})$ and returns a preference. Define $I_i = \mathbb{1}[S_{1,i} \text{ is preferred to } S_{0,i}]$. The A/B revision win rate is

$$\text{WinRate} = \frac{1}{N} \sum_{i=1}^N I_i, \quad (11)$$

with ties assigned a value of 0.5.

| Model | Accuracy | F1 | Accuracy (H) | F1 (H) |
|------------------|-------------|-------------|--------------|-------------|
| Baseline (P) | 0.81 | 0.78 | 0.69 | 0.68 |
| Multi-task | 0.85 | 0.83 | 0.74 | 0.71 |
| Iterative System | 0.91 | 0.90 | 0.84 | 0.82 |

Table 2: Qualitative performance of the α model, where ‘P’ denotes the pairwise-only model and ‘H’ indicates harder instances with less than a one-point difference in human engagement scores.

| Metric | Value |
|------------------------|----------------|
| Mean Δp (test) | +0.0252 |
| Median Δp | +0.0187 |
| $\Pr(\Delta p > 0)$ | 64.1% |

Table 3: Probability lift from conditioning α on β explanations on the test set.

5.6 Human Preference Evaluation of Explanations

In addition to model-based metrics, we evaluate explanation quality using human preference judgments. For each prompt story pair, annotators compare two explanations, one generated by the trained β model and one by an untrained baseline (Llama-3.1-8B-Instruct without task-specific fine-tuning), and select the explanation that better justifies the story preference.

Annotations are provided by six NLP PhD annotators⁴, each independently evaluating explanation pairs. We report the fraction of comparisons in which each explanation source is preferred⁵.

6 Results & Analysis

We report a qualitative comparison of model performance on the dataset in Table 2, using accuracy and F1 score as evaluation metrics. All reported scores are averaged over the test set across five runs with different random seeds during training.

6.1 General Performance

Table 2 summarizes the performance of different models relative to baselines. The pairwise baseline achieves 81% accuracy and 0.78 F1, substantially outperforming GPT-4 (57%, 0.55) and Llama-3.3-70B Instruct (48%, 0.46), as well as its 5-shot variant (58%, 0.57). The multitask model further improves performance to 85% accuracy and 0.83 F1. Our full model achieves the best results, with 91% accuracy and 0.90 F1. On the harder sub-

⁴The related prompts are reported in Appendix A.1.3.

⁵An example of explanation comparison is provided in Appendix A.3.

set, where human-annotated engagement scores differ by less than 1, baseline performance degrades sharply, while our model exhibits only a modest decline, demonstrating greater robustness. All results are evaluated against expert human annotations.

6.2 Explanation-Induced Probability Lift (§5.1)

We evaluate how conditioning on β ’s explanations influences the probability that α assigns to the correct label. Table 3 reports the probability lift $\Delta p = \tilde{p}^w(y_i) - \tilde{p}^{wo}(y_i)$ on the test set. On average, explanations increase α ’s confidence by 2.52 percentage points, suggesting that β provides informative evidence rather than merely stylistic content.

The distribution of Δp indicates that improvements arise from many small, consistent shifts rather than a few large changes. Positive lifts occur in a majority of test cases, while large negative effects are rare, suggesting that β explanations refine α ’s confidence instead of destabilizing predictions. This behavior is consistent with §6.4, where explanation-induced gains concentrate in low-confidence cases and remain largely neutral elsewhere.

6.3 Log-Odds Lift and Stability (§5.2)

To assess whether the observed gains reflect stable calibration shifts rather than brittle decision changes, we compute the log odds lift $\Delta \ell = \ell_i^w - \ell_i^{wo}$ (Eq. (5) in §5.2). Table 4 reports a consistently positive shift, indicating that explanations tend to refine α ’s beliefs rather than destabilize its decisions.

| Statistic | Value |
|------------------------|---------------|
| Mean $\Delta \ell$ | +0.114 |
| Median $\Delta \ell$ | +0.072 |
| $\Pr(\Delta \ell > 0)$ | 66.8% |

Table 4: Log-odds lift from conditioning on β explanations.

6.4 Win-Rate@k vs. Baseline Confidence (§5.3)

Table 5 decomposes rescue and harm⁶ as functions of α ’s baseline confidence $c_i = \max(\tilde{p}_i^{wo}(A), \tilde{p}_i^{wo}(B))$. Explanations are most effective when α is uncertain: in the lowest confi-

⁶Rescue and harm rates are computed conditional on baseline wrong predictions within each confidence bin, ensuring consistency with the overall baseline accuracy.

dence bin, β achieves the highest rescue rate with limited harm, resulting in the largest net win rate. As confidence increases, both rescue and harm decline, leading to near-neutral effects for already easy instances. These patterns support the intended role of β as targeted assistance rather than indiscriminate perturbation.

| Conf. bin | Rescue WR | Harm | NetWR |
|--------------|-------------|------|--------------|
| (0.50, 0.60] | 0.34 | 0.06 | +0.28 |
| (0.60, 0.70] | 0.27 | 0.05 | +0.22 |
| (0.70, 0.80] | 0.18 | 0.04 | +0.14 |
| (0.80, 0.90] | 0.09 | 0.03 | +0.06 |
| (0.90, 1.00] | 0.03 | 0.02 | +0.01 |

Table 5: Win-rate decomposition by α baseline confidence (Eq. (8)–(10) in §5.3).

6.5 Matched vs. Mismatched Explanations (§5.4)

We examine whether performance gains arise from item-specific explanations rather than generic evaluative text. Following §5.4, we compare matched explanations to mismatched ones, defined as random explanations drawn from other items, and report the average treatment effect (ATE). Table 6 shows that matched explanations yield substantially larger lifts than mismatched controls, supporting the claim that β learns content-grounded rationales.

| Measure | Matched | Mismatched | ATE |
|--------------------|----------------|------------|----------------|
| Mean Δp | +0.0252 | +0.0061 | +0.0191 |
| Mean $\Delta \ell$ | +0.114 | +0.031 | +0.083 |

Table 6: Matched vs. mismatched explanation effects (§5.4).

6.6 Extrinsic Utility: Story Revision Win-Rate (§5.5)

We evaluate downstream usefulness by testing whether β -guided revisions yield higher quality stories than their originals. As shown in Table 7, β revised stories are preferred significantly more often, indicating that β provides actionable feedback that improves perceived story quality under expert human judgment⁷.

6.7 Statistical Significance

We evaluate explanation effects using paired, non-parametric tests to avoid distributional assumptions. The matched versus mismatched differences in Table 6 are significant under a paired permutation test

⁷A story revision example is provided in Appendix A.4.

| Comparison | Win-Rate |
|---------------------------------|-------------|
| Revised w/ β vs. Original | 0.61 |

Table 7: A/B preference win-rate for β -guided revisions against the original story (§5.5), annotated by six NLP PhD evaluators. Ties counted as 0.5.

($p < 0.01$). The mean lifts in Δp and $\Delta \ell$ reported in Tables 3 and 4 are also significant under a two-sided Wilcoxon signed rank test ($p < 0.01$). For revision win rate in Table 7, a 95% bootstrap confidence interval indicates a significant improvement over 0.5.

6.8 Human Preference for β Explanations (§5.6)

We report human preference results for explanation quality in Table 8. Across evaluated prompt–story pairs, explanations generated by β are preferred more frequently than those produced by the untrained baseline model. This indicates that iterative alignment improves not only the utility of explanations for α , but also their perceived quality according to expert human judgment.

| Explanation Source | Preference Rate |
|----------------------|-----------------|
| β (trained) | 60.6% |
| Baseline (untrained) | 39.4% |

Table 8: Human preference rates for explanations, comparing the trained β model against an untrained Llama-3.1-8B-Instruct baseline. Preferences are annotated by six NLP PhD evaluators.

Across all metrics (§5.1–§5.6), β explanations provide consistent, item-specific improvements: they lift α ’s confidence and log-odds most in uncertain cases, outperform mismatched controls, and translate into better downstream revisions. These results jointly support iterative dual-model alignment as a practical way to learn explanations that are both discriminative and useful.

6.9 Explanation Comparison Result

In general, the aligned β model produces stronger explanations because it consistently grounds its reasoning in concrete textual evidence, rather than relying on broad but weakly supported thematic descriptions. In particular, the model more effectively identifies salient linguistic cues, defined as specific words or phrases that indicate key narrative elements, and uses them to distinguish between examples and generate interpretations that are directly traceable to the underlying stories. As a result, its

explanations are more specific, discriminative, and interpretable. By contrast, the base model more frequently produces fluent but generic commentary that appears plausible at a high level while remaining loosely grounded in the text. To facilitate understanding of the framework, we present a subset of an explanation comparison example.⁸

Prompt A and Story A.

Story A: the smell of the dry air was strong, but it wasn't really the smell of death. i had never smelled anything so strong. i had never tasted anything so strong before. it was just that i was getting used to it. the smell of the dry air was strong, but it wasn't really the smell of death. i 'd smelled it before, but it wasn't really the smell of death. it was the smell of the dry air. i was getting used to it. i had never smelled anything so strong before. i had never smelled anything so strong before.

Prompt B and Story B.

Story B: do you want to play football again? no, michael replied. i'm playing football. what? you think i'm going to be able to play? yeah, but i don't think i'm going to be able to play football. yeah, that's just a matter of time. do you think i'm going to be able to play football? i don't think so. but i don't think i'm going to be able to play football, michael said.

Explanation from base model.

1) Emotional Resonance & Thematic Connection

Both stories evoke a sense of familiarity and comfort, but in vastly different contexts. In Story A, the narrator's matter-of-fact descriptions of the zombie apocalypse create a sense of resigned acceptance ("i was getting used to it"), while in Story B, Michael's trash talk with President Obama creates a sense of playful camaraderie ("yeah, that's just a matter of time"). In Story A, the theme is survival, while in Story B, it's the absurdity of the situation.

Explanation from aligned β model.

1) Emotional Resonance & Thematic Connection

In Story A, the theme of desensitization to the zombie apocalypse is explored through the protagonist's increasingly casual attitude toward the smell of death ("I was getting used to it."), eliciting emotional detachment. In Story B, youthful arrogance is expressed through Michael's confident trash-talking ("Yeah, that's just a matter of time."), creating a more energetic emotional tone.

7 Conclusion

We propose an iterative dual-model framework that jointly learns story preferences and explanations

⁸The complete example is provided in Appendix A.3.

without explanation-level supervision. The resulting explanations improve preference prediction, especially in low-confidence cases, are preferred by expert evaluators, and provide actionable guidance that improves story revisions. These results show that iterative alignment effectively trains explanations as integral components of evaluation systems rather than post hoc justifications.

8 Limitations

Task and domain scope. Our study focuses on story evaluation with pairwise preferences, and the proposed framework may not directly generalize to other domains or evaluation tasks without task-specific adaptation.

Data limitations. Our study relies on a single human-annotated dataset. The stories in this dataset are relatively short and may exhibit low linguistic quality, including repetition or incoherent content generated by large language models, which can reduce internal consistency. These limitations restrict the generalizability of our findings and motivate future validation on higher-quality datasets.

Human evaluation scale. Human evaluations are conducted by a small pool of expert annotators. While this helps ensure consistency and informed judgments, it may not fully reflect the preferences of broader or more diverse reader populations.

Explanation properties beyond utility. We evaluate explanations primarily through their utility for preference modeling and story revision. Assessing additional properties such as factuality, faithfulness, or longer-term effects on downstream generation remains an important future work direction.

Ethical considerations. As with other explanation-generating systems, there is a risk that explanations may appear persuasive or authoritative even when imperfect or incomplete. Care should be taken when deploying such models in settings where explanations could influence user judgment, decision-making, or creative outcomes without appropriate transparency or safeguards.

Acknowledgments

We thank the anonymous reviewers for their time and insightful comments. This work was partially supported by an NSF CAREER award IIS2048001.

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.
- Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping human-ai collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. [Litbench: A benchmark and dataset for reliable evaluation of creative writing](#). *arXiv preprint arXiv:2507.00769*.
- Katy Itonka Gero, Tao Long, and Lydia B. Chilton. 2023. [Social dynamics of AI support in creative writing](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.
- Jian Guan and Minlie Huang. 2020. Union: An un-referenced metric for evaluating open-ended story generation. *arXiv preprint arXiv:2009.07602*.
- Jian Guan, Zhixin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. [OpenMEVA: A benchmark for evaluating open-ended story generation metrics](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.
- Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81.
- Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. 2024. Mind’s mirror: Distilling self-evaluation capability and comprehensive thinking from large language models. 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 6748–6763.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Ryan Lowe, Michael Noseworthy, Iulian Vlad 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 (Volume 1: Long Papers)*, pages 1116–1126.
- Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. In *Proceedings of the Workshop on Machine Learning for Creativity and Design*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2025. [Help me write a story: Evaluating LLMs’ ability to generate writing feedback](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25827–25847, Vienna, Austria. Association for Computational Linguistics.
- 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)*.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleu: 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*.
- Maartje ter Hoeve, Julia Kiseleva, Maarten de Rijke, and 1 others. 2020. What makes a good summary? reconsidering the focus of automatic summarization. *arXiv preprint arXiv:2012.07619*.
- Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency. *arXiv preprint arXiv:2402.08855*.
- Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large

language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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 Prompts

A.1.1 Alpha (α) Classifier Prompts

The α model is trained as a pairwise engagement classifier that outputs a binary decision over $\{A, B\}$. For each training instance (p_A, s_A, p_B, s_B, y) , we construct two input variants: a no-explanation prompt and a with-explanation prompt, where the latter injects an explanation E generated by β .

With explanation.

You are a helpful chatbot tasked with evaluating story engagement.

Use the following *analysis* as support (it may or may not be helpful): $\{E\}$

Prompt A: $\{p_A\}$

Prompt B: $\{p_B\}$

Story B: $\{s_B\}$

Which story is more engaging? Answer with A or B.

Answer: $\{y\}$

Without explanation.

You are a helpful chatbot tasked with evaluating story engagement.

Prompt A: $\{p_A\}$

Story A: $\{s_A\}$

Prompt B: $\{p_B\}$

Story B: $\{s_B\}$

Which story is more engaging? Answer with A or B.

Answer: $\{y\}$

Here, $\{p_A\}, \{p_B\}$ denote the two prompts, $\{s_A\}, \{s_B\}$ the corresponding stories, $\{E\}$ the candidate explanation (when present), and $\{y\} \in \{A, B\}$ the gold preference label.

A.1.2 Beta (β) Explanation Generator Prompt

The β model is trained to generate label-blind, rubric-guided comparative explanations with seven fixed section headings. It is explicitly instructed *not* to predict or reveal which story is better, and instead to provide evidence-grounded comparisons.

You are evaluating **engagement** strictly by the rubric below. Do **NOT** guess or state which story is better. Do **NOT** reveal any labels. Write a comparative analysis that follows the exact 7 sections and bullets.

RUBRIC SECTIONS (use these exact headings and order):

- 1) Emotional Resonance & Thematic Connection
- 2) Mystery & Sensory Intrigue
- 3) Tension & Conflict Dynamics
- 4) Immersive World-Building
- 5) Pacing & Flow
- 6) Character Development & Relatability
- 7) Narrative Clarity & Structure

For each section: - Cite 1–2 concrete spans from **each** story (short quotes or paraphrases).

- 1–2 crisp comparative sentences: specify how A vs B differ on THIS dimension.

- Avoid vague language and avoid claiming an overall winner.

Return 2–4 sentences per section. Keep the total length concise.

Prompt A: $\{p_A\}$

Story A: $\{s_A\}$

Prompt B: $\{p_B\}$

Story B: $\{s_B\}$

Your structured comparative analysis (with the 7 headings exactly):

A.1.3 Human Annotation Prompt

For annotations related to Section 6.6, the prompt used to instruct annotators is as follows:

Given the prompt story pair inputs, you are asked to judge which explanation provides a more comparative analysis with sufficient detail based on the predefined seven components for evaluating story engagement. Engagement refers to the extent to which a reader is engaged with the story content.

Prompt A: $\{p_A\}$

Story A: $\{s_A\}$

Prompt B: $\{p_B\}$

Story B: $\{s_B\}$

Explanation X: $\{E_x\}$

Explanation Y: $\{E_y\}$

For annotations related to Section 6.8, the prompt used to instruct annotators is as follows:

Given a prompt story pair along with an explanation of why one story is better than the other and a revised version of the weaker story, you are asked to judge which story the original weaker story or the revised version is more engaging. Engagement refers to the extent to which a reader is engaged with the story content.

Prompt (good): $\{p_{\text{good}}\}$

Story (good): $\{s_{\text{good}}\}$

Prompt (bad): $\{p_{\text{bad}}\}$

Story (bad): $\{s_{\text{bad}}\}$

Revised bad story: $\{s_{\text{revised}}\}$

A.2 Beta (β) Explanation Headings

The seven headings are derived by prompting Llama-3.3-70B-Instruct to produce comparative explanations of why one story is more engaging than another, using a comparison dataset constructed from HANNA, followed by clustering the generated explanations. These explanations reveal recurring criteria employed by the model to assess story engagement. Applying k-means clustering identifies seven salient dimensions: Emotional Resonance, Mystery and Sensory Intrigue, Tension and Conflict, World Building, Pacing, Character, and Clarity and Structure. We adopt these headings as the basis for modeling and evaluating story engagement. The definitions of these headings are as follows:

- **Emotional Resonance:** The extent to which a story evokes emotional responses and sustains affective engagement through mood, empathy, or emotional progression.
- **Mystery and Sensory Intrigue:** The degree to which a story maintains curiosity and immersion through suspense, vivid sensory detail, or the gradual revelation of information.
- **Tension and Conflict:** The presence and development of stakes, obstacles, or opposing forces that drive narrative momentum and reader investment.
- **World Building:** The coherence and richness of the story’s setting, including the establishment of rules, context, and background elements that support the narrative.
- **Pacing:** The effectiveness with which the story allocates narrative time and transitions, balancing progression and detail to maintain engagement.
- **Character:** The clarity, consistency, and development of characters, including their motivations, actions, and interpersonal dynamics.
- **Clarity and Structure:** The logical organization and readability of the narrative, including clear progression, coherence, and absence of confusion or unnecessary complexity.

A.3 Explanation Comparison Example

We present a representative example comparing explanations generated by a base model and the aligned β explainer. Although both explanations are structured around the same seven rubric dimensions, the base model frequently relies on generic or weakly grounded descriptors that do not clearly explain *why* one story is preferred. For instance, in discussing emotional resonance, the base model characterizes Story A as conveying “resigned acceptance” and Story B as exhibiting “playful camaraderie”, but does not connect these labels to narrative stakes or character motivation. Similarly, the base explanation notes that Story B is “incredibly unlikely” and therefore mysterious, without explaining how this surprise functions within the narrative.

In contrast, the β model consistently grounds its judgments in explicit narrative mechanisms. When discussing emotional resonance, it links Story A’s detachment to desensitization caused by immortality (“desensitization to the zombie apocalypse”) and contrasts this with Story B’s “youthful arrogance” expressed through competitive trash talk. For tension and conflict, the β explanation explicitly attributes Story A’s lack of suspense to the absence of stakes (“immortality removes meaningful stakes”), while identifying Story B’s conflict as arising from a “competitive gameplay” scenario against a high-profile opponent. These explanations not only describe surface properties but articulate causal relationships between story premises, character motivation, and reader engagement.

Overall, by repeatedly tying abstract rubric dimensions to concrete textual evidence (e.g., repetition-induced monotony in Story A versus “faster pacing and varied dialogue” in Story B), the β model produces explanations that are more faithful, internally consistent, and interpretable. This qualitative behavior aligns with the quantitative results in Sections 5.6 and 6.8, where β -generated explanations are preferred by human annotators and yield stronger downstream effects.

Prompt A and Story A.

Prompt A: You are an immortal during the zombie apocalypse. During the apocalypse the zombies ignore you and you try to live a normal life during the outbreak.

Story A: the smell of the dry air was strong, but it wasn't really the smell of death. i had never smelled anything so strong. i had never tasted anything so strong before. it was just that i was getting used to it. the smell of the dry air was strong, but it wasn't really the smell of death. i 'd smelled it before, but it wasn't really the smell of death. it was the smell of the dry air. i was getting used to it. i had never smelled anything so strong before. i had never smelled anything so strong before.

Prompt B and Story B.

Prompt B: Michael, a 15 year old Call of Duty player, realizes the greatest trash-talking opportunity of all time. Over his headset, he recognizes his opponent's voice as that of President Obama.

Story B: do you want to play football again? no, michael replied. i'm playing football. what? you think i'm going to be able to play? yeah, but i don't think i'm going to be able to play football. yeah, that's just a matter of time. do you think i'm going to be able to play football? i don't think so. but i don't think i'm going to be able to play football, michael said.

Explanation from base model.

1) Emotional Resonance & Thematic Connection

Both stories evoke a sense of familiarity and comfort, but in vastly different contexts. In Story A, the narrator's matter-of-fact descriptions of the zombie apocalypse create a sense of resigned acceptance ("i was getting used to it"), while in Story B, Michael's trash talk with President Obama creates a sense of playful camaraderie ("yeah, that's just a matter of time"). In Story A, the theme is survival, while in Story B, it's the absurdity of the situation.

2) Mystery & Sensory Intrigue

Story A's descriptions of the dry air and smell of death are vivid and immersive, but lack a sense of mystery ("i had never smelled anything so strong before"). In contrast, Story B is shrouded in mystery: Michael and President Obama's conversation is incredibly unlikely, and the reader is left wondering how this is possible.

3) Tension & Conflict Dynamics

Story A's narrative is remarkably tension-free, with the narrator simply going about their day despite the zombie apocalypse. In contrast, Story B's tension comes from the absurdity of the situation: Michael and President Obama's banter creates a sense of awkwardness and unease.

4) Immersive World-Building

Both stories create a vivid sense of their respective worlds, but in different ways. Story A's descriptions of the zombie apocalypse are matter-of-fact, while Story B's world is whimsical and surreal, with President Obama as a trash-talking opponent.

5) Pacing & Flow

Story A's narrative is meandering and repetitive, with the narrator's descriptions becoming increasingly circular. In contrast, Story B's pacing is snappy and engaging, with quick dialogue creating a sense of rapid-fire exchange.

6) Character Development & Relatability

Story A's narrator is enigmatic and underdeveloped, while Story B's Michael is relatable and well-defined. We get a sense of Michael's personality through his trash talk, while the narrator in Story A remains a mystery.

7) Narrative Clarity & Structure

Story A's narrative is non-linear and difficult to follow, while Story B is clear and structured, with a recognizable beginning, middle, and end.

Explanation from aligned β model.

1) Emotional Resonance & Thematic Connection

In Story A, the theme of desensitization to the zombie apocalypse is explored through the protagonist's increasingly casual attitude toward the smell of death ("I was getting used to it."), eliciting emotional detachment. In Story B, youthful arrogance is expressed through Michael's confident trash-talking ("Yeah, that's just a matter of time."), creating a more energetic emotional tone.

2) Mystery & Sensory Intrigue

In Story A, the sensory detail of the "smell of the dry air" creates an immersive atmosphere, but the straightforward narration limits surprise. In Story B, the unexpected reveal of President Obama's voice introduces a strong element of surprise and intrigue.

3) Tension & Conflict Dynamics

Story A lacks tension because immortality removes meaningful stakes, leaving the protagonist largely unchallenged. In contrast, Story B introduces external conflict through competitive gameplay against a high-profile opponent, generating excitement and suspense.

4) Immersive World-Building

Story A focuses on the apocalyptic environment but provides limited specificity beyond repeated sensory cues. Story B constructs a vivid gaming world through details such as the headset interaction and competitive framing.

5) Pacing & Flow

Story A's pacing is slow and repetitive, reinforcing a sense of monotony. Story B's faster pacing and varied dialogue create momentum and narrative energy.

6) Character Development & Relatability

The protagonist in Story A remains underdeveloped, with limited insight into internal motivation. In contrast, Michael's personality, confidence, and competitive drive are clearly expressed, making him more relatable.

7) Narrative Clarity & Structure

Story A's repetitive structure leads to a diffuse and unclear narrative trajectory, whereas Story B follows a clearer progression with a defined setup and interaction, resulting in greater narrative coherence.

A.4 Story Revision Example

We present a representative example to illustrate how rubric-guided explanations translate into concrete narrative improvements. In the original weaker story, the narrative is fragmented and repetitive, with limited emotional grounding, unclear character motivation, and little sense of tension or progression. The explanation explicitly diagnoses these deficiencies across multiple dimensions, including emotional resonance, narrative clarity, and character development, noting the absence of conflict, sensory detail, and purposeful structure. Guided by this feedback, the revised story introduces a focused conversational scene between the restaurant owner and the reporter, establishing clear stakes centered on the restaurant's survival, a coherent narrative arc, and a consistent point of view. Emotional engagement is strengthened through the owner's internal responses, while tension emerges naturally from the reporter's probing questions. The revision further improves pacing and coherence by replacing disjointed dialogue with a structured exchange that advances the narrative toward a credible resolution. Overall, the example demonstrates that explanation-guided feedback supports substantive narrative revision rather than superficial editing.

Prompt.

You are the owner of a small restaurant that is struggling to stay open. Suddenly, your restaurant is on the front page of the newspaper.

Original weaker story.

The newspaper was already in the news. The reporter had been there so long that it was hard to tell what to write about. Hey, I'm the owner of a small restaurant. oh, I'm good, said the reporter. This is the first time I've ever been to a restaurant, and I'm having a hard time believing it. this is the first time I've ever been to a restaurant, said the reporter. i'm sorry, I'm not a big fan of the name, but I'm pretty sure this is the first time I've ever been to a restaurant. That's a bit weird, said the reporter. I think I'm a bit weird, but I'm not very bad at anything. oh, I'm sorry, said the reporter.

Rubric-guided explanation

(1) Emotional Resonance & Thematic Connection. In contrast to Story A, Story B revolves around a mundane and unexpected event, lacking a clear emotional resonance. For example, "this is the first time i've ever been to a restaurant, and i'm having a hard time believing it" conveys bewilderment rather than emotional connection.

(2) Mystery & Sensory Intrigue. Story B is set in a restaurant but lacks sensory details, making it feel less immersive.

(3) Tension & Conflict Dynamics. Story B lacks conflict and tension, as the reporter's visit is unexpected but not necessarily tense. For example, "i think i'm a bit weird, but i'm not very bad at anything" downplays potential tension.

(4) Immersive World-Building. Story B takes place in a familiar setting but provides little world-building beyond generic conversation.

(5) Pacing & Flow. Story B moves quickly through repetitive dialogue without building toward a clear turning point.

(6) Character Development & Relatability. The owner and reporter feel one-dimensional; their dialogue lacks distinct motivations or traits.

(7) Narrative Clarity & Structure. Story B lacks a clear narrative structure, and the reporter's visit feels like a random event rather than a directed plot.

Revised story (guided by the explanation).

The newspaper reporter sat at a small table in the back of my restaurant, scribbling notes on a pad of paper. "i'm writing a story about this place," she said, looking up at me with a mixture of curiosity and skepticism. "it's the first time i've ever been to a restaurant, and i'm having a hard time believing it." i smiled, feeling a mix of emotions. "it's true," i said, "we're a small restaurant, and we're struggling to stay open." the reporter raised an eyebrow. "struggling to stay open?" she repeated. "that's an understatement." she leaned in, her voice taking on a more conspiratorial tone. "tell me more." i hesitated, unsure of how much to reveal. "it's just been tough," i said finally. "the economy's been tough, and we've been trying to stay afloat." the reporter nodded, her eyes narrowing. "i see," she said. "and what makes you think this restaurant is worth saving?" i felt a surge of defensiveness, but the reporter's tone was not unkind. "i think it's worth saving because of the people who work here," i said. "they're good people, and they deserve a chance to succeed." the reporter nodded again, this time with a hint of a smile. "i think i'm starting to understand," she said. "but i have to ask, what's the most surprising thing that's happened here?" i thought for a moment before responding. "i think it's the fact that a newspaper reporter has shown up at my doorstep," i said, smiling wryly. The reporter laughed, a small, mirthless sound. "i guess that's true," she said. But I'm glad I did." as she finished her story, I couldn't help but feel a sense of hope. Maybe, just maybe, this small restaurant would make it after all.