Retell, Reward, Repeat: Reinforcement Learning for Narrative Theory-Informed LLM Story Retelling

Anonymous ACL submission

Abstract

While current post-training paradigms (SFT, RLHF) degrade creativity in LLM storytelling by overfitting and reinforcing stylistic biases, we propose a paradigm shift that prioritises 'narrativeness'. We synthesise Todorov's Structuralist Narratology with Prince's concept of scalar narrativity and apply the annotation scheme–validated by blinded annotators who reliably rank human texts above LLM generations–to a counterfactual story retelling task. We use the annotations to benchmark three LLM evaluators, isolating the optimal model for final evaluation while using the remaining two to generate training signals during direct Reinforcement Learning from AI Feedback (d-RLAIF) in our theory-informed Retell Reward Repeat (RRR) pipeline. Our evaluations show that 7B/8B models trained via RRR outperform few-shot and SFT baselines in narrativity, logic, rationality, and completeness.

1 Introduction

Storytelling differs from many standard Natural Language Processing (NLP) tasks due to the lack of a universally preferable output. Existing post-training paradigms such as Supervised Fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) limit storytelling diversity in large language models (LLM) (Sui, 2026). While statistical metrics[1] based on reference outputs have been a norm in NLP, they align poorly/negatively with human judgements of Automatic Story Generation (ASG) outputs (Netisopakul and Taoto, 2023); preference-based RLHF reduces output diversity to maximise alignment (Lu et al., 2025), where models converge towards clichéd and unnecessary linguistic styles [2] (Chakrabarty et al., 2025).

[1] e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and BERTSCORE (Zhang et al., 2020).

[2] ASG literature defaults to reductive universal criteria such as complexity, suspense, aesthetics and entertainment: but soothing bedtime stories are not complex or suspenseful; narratives depicting genocide are not aesthetic or entertaining.

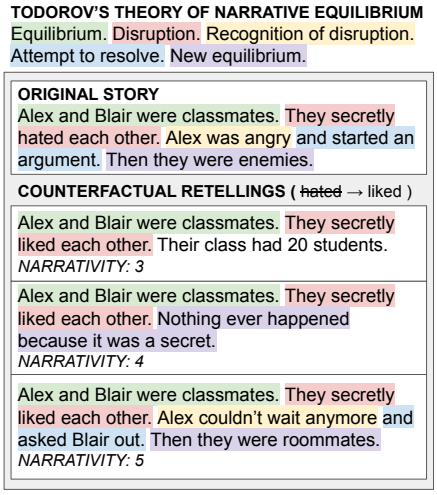

Figure 1: Todorov (1971)'s narrative theory applied to Qin et al. (2019)'s counterfactual storytelling task. Narrativity measures a text's 'narrativeness'.

Moving towards a narrative-aligned LLM post-training starts by redefining how we measure storytelling abilities. When we formulate the problem of ASG as selecting and representing a sequence of events that (A) can be told as a story and (B) meet a set of specific criteria (Li et al., 2013), we can see that current evaluation frameworks (Chhun et al., 2022; Chakrabarty et al., 2024) disproportionately prioritise (B). This drives an impossible search for 'universally good storytelling', as evidenced by the persistent absence of standardised benchmarks (Liu et al., 2026) and conflicting subjective human evaluations of ASG outputs (Marco et al., 2025). Proposing an alternative path, we shift our priority from (B) to (A) and seek to measure the 'narrativeness' of a text–measured using fundamental qualities shared by all narratives that provide an agnostic metric across diverse storytelling forms:

RQ1: How can narrativeness in LLM outputs be measured with a set of definable properties?
RQ2: How can post-training align LLMs with narrativeness and other storytelling criteria?

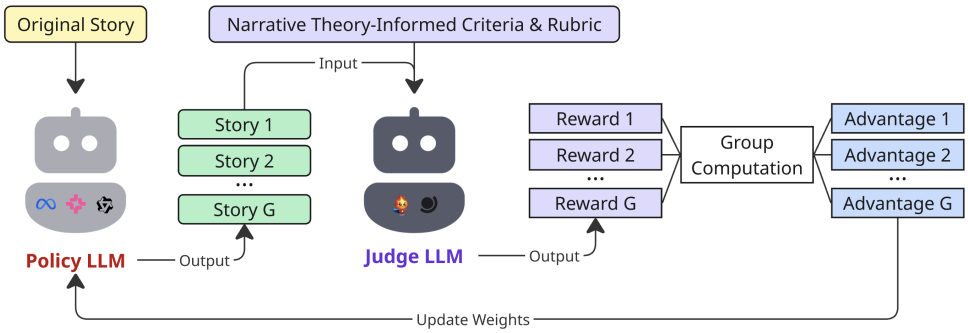

Figure 2: The RRR post-training pipeline uses d-RLAIF (Lee et al., 2024) with a narrative theory-informed LLM-as-judge to generate the reward signal for GRPO (Shao et al., 2024).

Atmakuru et al. (2024) show that LLMs tend to retell narratives from memorised training data during prompt-to-story tasks which makes it difficult to evaluate storytelling abilities, whereas constrained story retelling tasks force the LLM to dynamically select and represent sequences of events in adherence to logical and rational reasoning–which empirically increases difficulty. To discourage memorisation, we select Qin et al. (2019)'s counterfactual story retelling task (Figure 1) for this case study, where our contributions are as follows:

- We construct a scheme that synthesises Todorov (1971)'s Structuralist Narratology with Prince (2003)'s scalar concept of narrativity to measure a text's narrativeness, and additional define 3 additional specific to counterfactual story retelling tasks. We show that, on average, human-authored stories satisfy the criteria more than LLM-generations.
- We release a dataset (`https://anonymous.4open.science/r/CounterfactualTodorov`) of 200 counterfactual retellings tagged and evaluated by three annotators, showing the usefulness and limitations of Todorov's theory. We also identify Gemini-3-Flash (Google, 2025) as the evaluator most aligned with human judgements.
- We introduce a theory-informed 'Retell Reward Repeat' (RRR; Figure 2) pipeline adapted from d-RLAIF(Han and Han, 2025). We show that 7B/8B models trained with RRR achieve the highest narrativity and quality, whereas SFT models mainly copy surface syntax.

Through engaging with debates in literary theory, our study synthesises established theoretical narrative frameworks to confront the fundamental challenges of aligning LLM post-training with storytelling abilities. While existing evaluation criteria for storytelling remain valuable, our work demonstrates the necessity of exploring alternative pathways to overcome limitations in current training paradigms. Overall, our findings show that narrative theory-informed reinforcement learning is a promising new direction for LLM storytelling.

2 Preliminaries

2.1 Structuralist Narratology and Narrativity

Todorov (1969) first coined the term *narratologie* to describe a science of narrative which seeks to discover the abstract structure and essential features that universally and exclusively define a narrative. Structuralist Narratology treats an individual story as just one possible manifestation of narrative's underlying universal structure. For example, (Todorov, 1971)'s plot-focused Theory of Narrative Equilibrium claims that narrative's underlying universal structure can be represented by sequences of causally and contextually connected transformations, categorisable into 5 types of narrative stages. We illustrate these 'Todorovian stages' in Figure 1 where each stage is colour-coded by type.

While Classical Narratology treated a text's narrativeness as a boolean property, Narratology has increasingly embraced the concept of scalar narrativity (Prince, 2003), where narrativeness is a matter of degree rather than an absolute. Todorov (1971) implied this continuum in his original work–a reader must be able to recognise most types of the five equilibrium stages for a text to be considered a narrative–but did not explicitly propose his theory as a metric for narrativity. We synthesise these perspectives to operationalise a measurement of narrativity to evaluate the completeness of a counterfactual retelling–shown in Figure 1–where a text containing 5 types of Todorovian stages exhibits

higher narrativity than one containing 4 or 3.

2.2 Debates in Literary Theory

We engage with debates in literary theory by acknowledging that Todorov's core assumption of universal conditions for narrativity is contentious. Post-structuralists like Barthes (1967) argue that a text exists within fluid structures inhabited and contested by ideas, where interpretation is constructed from innumerable sources of language and culture.

A theory claiming that no condition is universal is difficult to operationalise but we do not ignore this critique. Our experiment engages with this debate by using Todorov's framework as a methodological probe. By providing definable conditions for human and LLM-as-judge annotators, we test the limits of these structuralist assumptions.

While our introduction established that narrative *quality* is subjective, **RQ1** investigates how *narrativity* itself can be measured. Ultimately, our annotation results serve as empirical evidence that the boundaries of narrativity are as contested and fluid in practice as post-structuralist theory suggests.

3 Data Annotation

Qin et al. (2019) show that for the task of rewriting the ending of a short story after a provided 'Counterfactual' [3] event replaces the 'Initial' event (Table 1), an AI-generated output's similarity[4] to the human-written ground truths correlates weakly/negatively with human given ratings of coherence and relevance. To overcome the unreliability of reference-based evaluation, our case study applies Todorov's theory to provide a linguistically grounded approach to training and evaluation.

The functional purpose of the annotation dataset is solely evaluative: i.e., to test agreement between human annotators and LLM-evaluators to inform our decisions in selecting LLM-as-judge models for training and evaluation. The annotated data is not directly used for model training. Additionally, through quantitatively and qualitatively examining the inter-annotator agreement, we can empirically observe whether and how we can measure narrativeness using definable conditions (**RQ1**).

[3]The 'Counterfactual' is a counterfactual antecedent while the 'Edited Ending' is a counterfactual consequent.

[4]Measured using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) & BERTScore (Zhang et al., 2020).

3.1 Curation

We curate a pool of human and AI-generated retellings to reflect the distribution of texts the LLM-evaluator will see during the reinforcement learning process and final evaluation.

We start the curation process from the first 3000 items in the TimeTravel supervised training split (Qin et al., 2019), where each item (Table 5) provides a 'query' consisting of the premise, initial state, original ending, and counterfactual, and a 'response' which is the edited ending (human-written ground truth). For each query, we produce three AI-generated responses (edited endings) using Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Olmo-3-7B-Instruct (Ettinger et al., 2025), and Qwen-3-8B (Yang et al., 2025) with thinking disabled.

Because the three modern instruction-tuned LLMs typically produce similar responses to the same query, we introduce a filter aiming to diversify the data. We do not include the human ground truth in the calculation of our diversity metric to avoid biasing our selection towards 'harder' queries where the LLM output is distant from the ground truth (even if it is a weak correlation).

For each query's three AI-generated responses, we use DeBERTa-v3-small (He et al., 2023) to compute all pairwise distances (defined as $1 - \text{BERTScore}_{F1}$) and define the diversity metric as $\text{diversity} = \min(\text{distances}) \times \text{mean}(\text{distances})$. We then select the 50 inputs with the highest diversity and, for each, retain four retellings (the three LLM responses plus the ground truth human response), yielding an annotation set of $n = 200$.

3.2 Scheme

We operationalise our measurement of narrativity by counting the number of recognisable Todorovian stages in a text (Table 1). Since disruption (i.e., change) is essential for narrative transformation, any text lacking a disruption is assigned a score of 1. For all other cases, the narrativity score is defined as $Narrativity = \min\{s + 1, 5\}$, where s is the number of identifiable Todorovian stages. Our conditions only require 4 recognisable narrative stages to achieve the maximum scores since we do not wish to penalise stories which begin in medias res or end on a cliffhanger (Appendix E.2).

We also define the specific criteria we use to measure quality for the counterfactual story retelling task, which we use to investigate **RQ2**. These criteria are designed to align with those originally

Narrativity
A: (4) Evan was ready to buy himself a bicycle. He went to the store but forgot his money. Then he looked around some more, trying to find something else he could afford or a way to get home.
B: (4) Evan was ready to buy himself a bicycle. He went to the store but forgot his money. Then he looked around some more, trying to find something else he could afford or a way to get home.
C: (3) Evan was ready to buy himself a bicycle. He went to the store but forgot his money. Then he looked around some more, trying to find something else he could afford or a way to get home.
Logical
A: (3) I interpret 'but forgot his money' as Evan forgetting a large sum of money intended for buying a bicycle. Evan may still have the usual money he carries everyday, and is able to logically find something else he could afford.
B: (3) The edited text is logical. Forgetting his money and then looking for an alternative or a way home is internally consistent.
C: (1) There is a logical impossibility in the story because: 1. Evan forgets his money, but 2. proceeds to look around the store for something he can "afford," which would be nothing.

Table 1: A retelling annotated by A, B and C (humans). More examples are in Appendix J.

proposed by Qin et al. (2019), which primarily measure relevance and coherence. Our approach differs by grounding these definitions in Todorov's narrative theory wherever possible:

Logical: The text is free from inconceivable scenario(s) that contradict the text's internal logic.
Rational: The text can be rationally interpreted using Todorovian stage(s) in its entirety without contextual/causal disconnection.
Complete$_{\mathbf{N}}$: The text includes all key Todorvian stage(s) from the original that ensure the text is as narratively complete as the original.
min$_{\mathbf{LRC}}$: Since any desirable story needs to satisfy all 3 of the aforementioned criteria, we use $\min(\text{Logical}, \text{Rational}, \text{Complete}_N)$ to represent a text's overall quality instead of the average. This reflects the principle that failing even one criterion makes the story completely undesirable.

3.3 Human Annotation

Three human annotators, blind to whether the stories were human- or LLM-authored, annotate the texts using a tag-and-evaluate process: (1) Tag parts of text with Todorovian stages. The variety of tags is automatically counted and computed into a 1-5 Narrativity score, then (2) use 3-point Likert scales to evaluate on the Logical/Rational/Complete$_{\text{N}}$ criteria. (Additional instructions and examples given to the annotators are included in Appendix E.)

Since the collection of human-authored and LLM-generated stories is logical and rational in most cases, the rating distributions are skewed (e.g., the average rating for Logical is 2.7/3 while the average for narrativity is 3.9/5. See Appendix Table 6), leading to the well-known prevalence/bias paradox (Zec et al., 2017) for κ (e.g. $\kappa = 0.18$ despite 80% agreement, see Appendix Table 7).

We want the dataset to test whether LLM-evaluators can distinguish between genuine human-level narratives and modern LLM generations. To accurately reflect the expected distribution of scores that the LLM-evaluator will see during d-RLAIF and evaluation, we do not alter the distributions and instead report Gwet (2014)'s AC2 alongside κ for inter-annotator reliability (top row of Table 2). Unlike κ, which mathematically conflates high class prevalence with a high probability of chance agreement, AC2 estimates chance agreement under the assumption that raters do not guess randomly when they are certain of a prevalent class, thereby providing a useful complementary metric.

Inter-annotator agreement across the evaluation criteria ranged from fair to moderate (Table 2), with AC2 scores between 0.55–0.76 and κ values between 0.18–0.41 (or 0.30-0.41 if we treat the Logical criteria as an outlier since distribution is skewed with an average of 2.7/3 leading to a low κ).

After acquiring some additional rationales from the annotators, a qualitative review of the disagreements (Table 1) shows that annotators often formed differing yet valid interpretations of the same text. e.g, readers frequently disagreed on structural starting points—such as whether a narrative begins in a state of equilibrium or in disruption—and applied subjective assumptions to resolve ambiguities.

Story quality and narrativity arise from the subjective interaction between reader and text (Marco et al., 2025). This inherent subjectivity, consistent with reader-centric Post-structuralist literary theory (Barthes, 1967), accounts for the fair to moderate agreement observed from human annotators.

However, when we categorise the ratings by the author of the text (Table 3), which the annotators are blind to during annotation, we can see that humans annotators agree on the ranking of

Annotator	Logical	Rational	Complete_N	Overall	min_LRC	Narrativity
Humans Only	0.76/0.18	0.71/0.41	0.56/0.32		0.55/0.30	0.74/0.38
w/ Selene-1-mini	0.80/0.09	0.71/0.27	0.58/0.25	0.49/0.20	0.54/0.20	0.73/0.31
w/ M-Prometheus	0.81/0.12	0.71/0.30	0.56/0.25	0.47/0.21	0.48/0.23	0.73/0.28
w/ Gemini-3-Flash	0.78/0.22	0.71/0.36	0.59/0.36	0.53/0.29	0.55/0.31	0.77/0.40

Table 2: Inter-annotator agreement, reported by *AC2/*κ. "w/" indicates when an LLM-evaluator is added.

Author	**Logical**	**Rational**	$\textbf{Complete}_N$	$\textbf{min}_{LRC}$	**Narrativity**
human	2.98 / 2.54 / 2.96	2.72 / 2.66 / 2.52	2.82 / 2.72 / 2.28	2.58 / 2.30 / 1.96	4.70 / 4.62 / 3.72
Olmo	2.80 / 2.50 / 2.96	2.52 / 2.66 / 2.48	1.92 / 2.04 / 2.00	1.70 / 1.84 / 1.72	3.84 / 4.00 / 3.36
Qwen	2.64 / 2.40 / 2.92	2.36 / 2.52 / 2.36	2.50 / 2.54 / 2.12	2.12 / 2.12 / 1.68	4.22 / 4.32 / 3.50
Llama	2.70 / 2.36 / 2.80	2.02 / 2.16 / 2.04	1.54 / 1.92 / 1.52	1.24 / 1.60 / 1.24	3.50 / 3.68 / 2.84

Table 3: Human and LLM author comparison across evaluation dimensions with 3 human annotators (A / B / C).

authors across Logical, Rational, Complete_N and Narrativity. e.g., Olmo is rated higher in Logical and Rational but lower in Complete_N and Narrativity. This signals to us that our scheme is able to differentiate between human ground truths and LLM generations–even potentially between different LLMs–where human authored stories are higher in both narrativity and quality (measured by Logical, Rational and Complete_N) than LLM generations.

The annotations suggest that while Todorov's theory can be operationalised to measure narrativity with some degrees of alignment, it does not address other parameters that affect interpretation without a model of reader cognition. It does not completely and universally model narrative structure. However, since it is still a 'usable' measurement, we continue with using human annotations to select the appropriate LLM-as-judge models for training and evaluation.

3.4 LLM-as-judge

We test 3 LLM-as-judge evaluators: Selene-1-mini-8B (Alexandru et al., 2025), M-Prometheus-14B (Pombal et al., 2025) and Gemini-3-Flash (Google, 2025). The first two are recently released open-weight models that specialise in evaluation tasks, while Gemini-3-Flash is chosen as a current state-of-the-art (SOTA) proprietary model. Since the two specialised models were trained to give only a single rating, we run inference independently for each criteria. Hoping to maximise speed and token efficiency during d-RLAIF, we also instruct each model to output an additional 'overall' score with identical criteria to the min_LRC score to see if we can approximate the rating with less tokens.

We evaluate the models' alignment based on AC2 & κ, measured by examining the agreement scores when the LLM-as-judge is included in a 4-way calculation with the 3 human annotators (Table 2). We observe that Gemini is the only model which is able to consistently improve alignment measured by both metrics in Complete_N, Overall, min_LRC and Narrativity. Additionally, when we again categorise the ratings by the author, we see that only Gemini rates human authored stories with the highest scores, while Selene and Prometheus only rank human authored stories first in narrativity (Appendix Table 9). We choose to save Gemini for final evaluation, given the evidence that it is our most reliable and aligned LLM-as-judge.

4 Retell, Reward, Repeat (d-RLAIF)

We adapt Wei et al. (2025)'s implementation of d-RLAIF to use an LLM-as-judge to transform an input story into a theory-informed reward signal. We use GRPO (Shao et al., 2024) and LoRA (Hu et al., 2022) to optimise the RRR training process using TimeTravel's unsupervised training split. The policy model generates 16 responses for each query, which are then given to the LLM-as-judge to evaluate and generate 16 reward scores. The relative advantages are then calculated from the reward scores and used to update the weights (Figure 2) of the policy model's LoRA adapter.

To study the effects of our RRR training on a wide range of models, 3 instruction-tuned models (Llama-3.1-8B, Qwen-3-8B, and Olmo-3-7B) are chosen as policy models for generating retellings.

Using the results from Table 2, we initially choose Selene-1-mini (since the measured alignments are comparable with Prometheus) as the

Model	Method	Logical	Rational	Complete_N	min_LRC	Narrativity	BLEU-4	ROUGE-L
GROUND TRUTHS	Crowd Workers	2.884	2.927	2.933	2.811	4.718	-	-
LLAMA-3.1-8B-INST.	BASE	2.735	2.385	2.528	2.080	4.347	0.340	0.502
	few-shot	2.768	2.615	2.505	2.213	4.303	0.306	0.473
	SFT	2.320	2.501	**2.934**	2.260	4.692	**0.810**	**0.835**
	RRR_O-Prometheus	**2.801**	2.825	2.594	2.400	4.534	0.179	0.386
	RRR_N-Selene	2.785	**2.860**	2.740	**2.554**	**4.733**	0.087	0.299
QWEN-3-8B	BASE (non-thinking)	2.429	2.132	2.759	1.940	4.551	0.572	0.667
	few-shot	2.543	2.219	2.686	2.006	4.559	0.492	0.607
	SFT	2.402	2.532	**2.925**	2.317	4.693	**0.790**	**0.823**
	RRR_O-Prometheus	2.717	2.336	2.566	2.024	4.356	0.337	0.491
	RRR_N-Selene	**2.781**	**2.890**	2.705	**2.516**	**4.894**	0.003	0.192
OLMO-3-7B-INST.	BASE	2.762	2.812	2.400	2.222	4.328	0.097	0.315
	few-shot	2.745	2.837	2.377	2.208	4.415	0.072	0.287
	SFT	2.295	2.441	**2.910**	2.211	4.647	**0.787**	**0.821**
	RRR_O-Prometheus	2.770	2.857	2.491	2.325	4.460	0.090	0.309
	RRR_N-Selene	**2.793**	**2.893**	2.595	**2.434**	**4.899**	0.002	0.178
CODEX	Program Prompt	2.452	2.484	2.758	2.181	4.514	0.617	0.695

Table 4: Evaluation of post-trained LLMs using the test split (n=1871) from the TimeTravel dataset using Gemini-3-Flash. For each base LLM, the highest score in each column is bolded. Significance testing in Appendix I.

LLM-as-judge and reward signal generator. We conduct d-RLAIF with the reward score R_N based on narrativity (Appendix Figure 5). While the 'overall' score is measured as less aligned to human preferences compare to aggregating min_LRC from 3 separate inferences, we consider it to be an acceptable degradation at $\frac{1}{3}$ the token cost and use it for our alternate reward signal R_O.

We apply early stopping once training reaches a reward plateau, defined as 200 consecutive optimiser steps for which the average reward remains within 0.1 of the maximum achievable reward, indicating no further measurable improvement. Additionally, we introduce a length-based penalty for retellings that exceed 3 times the length of the original to prevent reward hacking by lengthy outputs.

When Qwen and Olmo are trained on Selene-1-mini's R_O signal, they converge prematurely without reaching the goal reward plateau (Appendix Figures 4 and 6). To ensure the reward plateau is reached, we train all 3 models on M-Prometheus's R_O signal and they all reach the reward plateau within 500 global optimiser steps (Appendix Figure 7) . For comparison, we conduct standard SFT with LoRA using TimeTravel's supervised training split, also with early stopping at convergence–defined by when loss decreases < 0.01 for 3 consecutive training steps (Appendix Figure 8). All models converge faster and with less generations and optimiser steps during SFT than d-RLAIF.

5 Results

We evaluate our models on TimeTravel's test split (n=1871) using the most aligned LLM-as-judge we found in Section 3.4: Gemini-3-Flash. Each item in the test split contains 3 slightly different human-written edited endings–we evaluate each and use their average for the results in Table 10. We also use BLEU-4 and ROUGE-L to calculate the average similarity of each model's generation to the nearest-of-3 human-written endings to measure similarity. Human crowd workers authored stories receive higher average ratings than LLMs when considering min_LRC, which more better represents overall quality than the average of the 3 scores, as all desirable stories must score perfectly across the 3 criteria.

For baseline comparisons against our method, we compare zero-shot and few-shot for each LLM. We also evaluate the outputs produced by the Program-Prompt method (Liu et al., 2023). Per base model, few-shot does not result in consistent or noticeable performance over zero-shot. SFT achieves the highest Complete_N, BLEU-4 and ROUGE-L for all models. SFT results in an increase of min_LRC performance for Qwen and Llama but degrades the performance of the best-performing base model Olmo. Stories generated by the SFT models are highest in similarity to the human-written ground-truths, as well as the most structurally complete when compared to the corresponding original story (measured by Complete_N).

When using RRR, we find that d-RLAIF with M-Prometheus's R_O improves the overall perfor-

(query) Original Story + Counterfactual	
Premise	Max and May were at a tavern.
Initial	They saw the dartboard and challenged each other to a game.
Original Ending	Max played very well. But May thru even better than him. May won the game.
Counterfactual	They didn't play darts there because they were not competitive.
Edited Ending	???
(response) Edited Ending	
Ground Truth	Max played pool very well. But May thought she was even better than him. May won the game, but neither really cared who won.
Qwen	
Base	Max played very well. But May threw even better than him. May won the game.
Few-Shot	They didn't play darts there because they were not competitive. Instead, they talked about their favorite books and shared stories about their travels.
SFT	Max played very well at darts. But May threw even better than him. May won the game.
RRR_O	They decided to play a different game instead, one that allowed them to enjoy each other's company without competition.
RRR_N	They ordered drinks, then debated chess. Max won, but May vowed to train. They vowed to rematch.

Table 5: Example test outputs from models trained on Qwen. Additional examples in Appendix K.

mance of all tested models, where Selene-1-mini's narrativity-based reward signal R_N reliably results in the best min_{LRC} performance, second only to humans (significance testing in Appendix I). Their outputs also contain the highest narrativity and dissimilarity to human ground truths.[5]

6 Discussion

6.1 Reinforcement Learning vs SFT

Intuitively, it does not make sense to teach a student storytelling by assigning them a task then evaluate what they write based on the teacher's own ground truth. When we consider that humans naturally learn storytelling as a communicative practice through reinforcement learning (Hineline, 2018), our RRR pipeline using d-RLAIF is an intuitive choice for Automatic Story Generation (ASG).

The increased performances in min_{LRC} and Narrativity across all models trained using our RRR_N method (Table 10) indicate that Todorov's abstract and theoretical narrative structure can be learned through reinforcement learning (since Narrativity is measured by the presence of Todorovian structures) and is beneficial in outputting logical, rational, and structurally complete stories even though they differ the most to human ground truths (Table 5).

While SFT models achieve the highest $Complete_N$, the highest similarity to ground truths, and the second highest Narrativity, this does not translate to higher min_{LRC} scores—our established indicator of overall story quality. Outputs show that SFT models often generate high-narrativity story endings by illogically and irrationally duplicating the original story. We do not consider their increased narrativity scores to indicate meaningful narrative alignment. It appears that SFT forces LLMs to over-fit to the training data without learning functional narrative abilities for the counterfactual story retelling task.

The results also highlight that, while optimising for narrativity during training (RRR_N) succeeds in our case study, evaluating based on narrativity alone is insufficient. Recalling our formulation of ASG—selecting and representing a sequence of events that (A) forms a coherent story and (B) meets specific criteria (Li et al., 2013)—both components are essential. While our findings demonstrate the value of prioritising (A) during training, evaluation must still assess the specific criteria (B) required by the specific task.

6.2 How d-RLAIF Teaches Storytelling

The lack of consistent or significant performance increases of few-shot over zero-shot (Table 10) shows that, for counterfactual story retelling, post-training is necessary for improving storytelling structure and quality. Since our RRR pipeline uses less than a single epoch from the dataset to reach convergence, it suggests that massive storytelling datasets may not be necessary for post-training. Efforts to curate datasets for d-RLAIF may instead focus on comparatively smaller and query-only datasets. This approach potentially circumvents the bottleneck of limited shared datasets (Liu et al., 2026).

We demonstrate that d-RLAIF can be efficiently executed on a single GPU using 8B student and

[5] We do not evaluate the incompletely trained models that fail to reach the reward plateau during RRR with d-RLAIF.

teacher models. Furthermore, our findings support the claim that accuracy is not the optimal indicator of a good teacher model for reinforcement learning (Razin et al., 2025). Although Selene-1-mini and M-Prometheus achieve comparable accuracy, Prometheus successfully guides its student to convergence while Selene fails. As illustrated by the score distributions on the annotation set (Appendix Figure 3), Selene-1-mini acts as a harsher judge than M-Prometheus. This harshness may have prevented models trained with Selene-1-mini's R_O from reaching target reward plateaus, unlike those trained by the more lenient M-Prometheus. These results underscore the need for additional LLM-as-judge metrics beyond mere accuracy and alignment.

Using d-RLAIF, methods of effective narrative understanding can provide great assistance in providing the reward signals needed for the training of future ASG models. We believe it to be beneficial to move towards unified knowledge and methods between generation and understanding.

6.3 Future Implications

Narrative theory provides a foundation for usable story metrics. While we hoped to contribute narrativity metrics for shared evaluation and training across different tasks, our annotation process only reached fair-to-moderate agreement. This suggests that the inherent subjectivity of storytelling may be incompatible with universal benchmarks, and that Todorov's structural theory alone cannot model all narrative qualities (e.g., the 'Logical' and 'Rational' criteria also draw on cognitive considerations).

Our case study demonstrates both the benefits and limitations of applying Todorov (1969)'s narrative theory to model storytelling, highlighting the need to move beyond textual features. Postclassical narratology expands beyond the limitations of structuralist narratology by emphasizing the historical, cultural, and cognitive parameters of storytelling, as well as the reader's role in constructing narratives from linguistic artifacts (Meister, 2011).

Although we apply RRR solely to the counterfactual story retelling task, the method is highly adaptable. By prompting an LLM-as-judge within the d-RLAIF pipeline, the training process allows the student policy model to learn behaviours compatible with the teacher model's conceptual representations. For instance, models can be trained to adhere to stylistic genre conventions or generate empathetic narratives. This process requires minimal training data, provided the LLM-as-judge possesses a robust representation of the target behavior. In this way, the pipeline itself could potentially as a Postclassical representation of narrative that models the interaction between text and the reader.

7 Related Work

Qin et al. (2019)'s TimeTravel dataset and its counterfactual story retelling task have been approached by a number of past works, experimenting with: inference-time optimisations (Qin et al., 2020; Chen et al., 2022), multi-step generation (Hao et al., 2021), Program-prompting (Liu et al., 2023) and ConceptNet (Ashwani et al., 2024). These past works position storytelling as a task that tests the ability of a LLM to reason and understand causal logic and do not focus on developing a training method which aligns with storytelling principles.

Wei et al. (2025) showed that d-RLAIF can perform well at improving creative-writing for generating Chinese greetings, a tasks which involves open-ended creative writing.

8 Conclusion

We address the persistent limitations of LLM post-training in Automatic Story Generation through a narrative theory-informed approach. First, we demonstrated that narrativeness can be measured through definable (although not universal/objective) properties–synthesising Todorov's Theory of Narrative Equilibrium with Prince's scalar narrativity into a quantifiable metric based on structural stages (**RQ1**). Building on this, we introduced "Retell Reward Repeat" (RRR) to align LLMs with these criteria (**RQ2**) using direct Reinforcement Learning from AI Feedback (d-RLAIF).

Our results across Llama-3.1, Qwen-3, and Olmo-3 show that models trained with our narrativity reward (R_N) significantly outperform SFT and few-shot baselines in output story quality, as measured in the counterfactual story retelling task using the criteria: Logical, Rational and Complete_N. Ultimately, our findings establish d-RLAIF as a highly viable post-training paradigm for story generation and underscore the potential of using established linguistic theories and principles to ground future reward models, as well as the central role of reader cognition in storytelling evaluation.

Limitations

To ensure correct understanding and application of narrative theory, the authors served as annotators in this case study, which introduces potential bias. While a large pool of external annotators is optimal practice, we believe this limitation does not undermine our core findings.

Our research demonstrates that human evaluation of stories is highly subjective. Supported by narrative theory (Barthes, 1967), existing empirical findings (Marco et al., 2025), and our own results, we argue that annotator disagreement is a feature, rather than a bug, of storytelling tasks. Specifically, our data shows: (1) There are inherent disagreements between human annotators regarding story understanding. (2) Story understanding is a fundamentally subjective process. (3) Blinded human annotators consistently rate human-written ground truths higher than LLM generations.

For an expanded annotator pool to contradict our claims, new data would have to demonstrate that story understanding is objective (showing little to no disagreement) or that humans consistently rate LLM generations higher than human-written ground truths. Given existing evidence across multiple disciplines, we consider this unlikely.

Apart from Liu et al. (2023)'s program-prompt, we do not compare our models with those developed by others using the TimeTravel dataset (Qin et al., 2020; Chen et al., 2022; Hao et al., 2021; Liu et al., 2023; Ashwani et al., 2024). This is primarily because none of them are reasonable baselines for our experiments, particularly in the context of post-training with narrative-based metrics (the others primarily optimise for statistical metrics based on reference outputs).

To avoid the massive costs, we did not compare the final accuracy of Gemini-3-Flash's evaluations with human judgements on the test set. While the fact that it rates human-written stories more highly than AI-generated stories provides some evidence of alignment, our reliance on LLM-as-judge to approximate human judgement is an assumption and limitation.

As a novel case study for reinforcement learning for story generation, we do not test how our approach scales to longer stories. We only focused on training 7/8B models, the effectiveness of our method for smaller/larger LLMs is untested. These can be explored in future studies.

Ethics Statement

We use a benchmark dataset (Qin et al., 2019), and do not have any additional ethical considerations to report. Our human evaluators are authors on the paper. GitHub Copilot and Google AI studio were used to assist with coding tasks and debugging. Microsoft Copilot was used for formatting the latex in the manuscript. Outputs generated by these tools were carefully reviewed and validated by the authors to maintain accuracy and correctness.

Our dataset, tasks and generated stories are in English and refer to Western contexts. We use an European narrative theory. It is well known in the narrative studies community that cultural differences can affect how we perceive narratives and their underlying themes and structures (Aziz, 2023; Phillips et al., 2025). As such, our paper's representation of storytelling is biased and does not reflect all cultures and is not universal.

Storytelling is a highly effective communication tool that can be used for influence and manipulation. The development of Automatic Story Generation methods, with the goal of matching human storytelling, introduces the risk of unintended or even malicious harm.

Acknowledgments

Masked for blind review.

References

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, and 1 others. 2025. Atla selene mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9.

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. *arXiv preprint arXiv:2410.04197*.

Sardar Khawar Aziz. 2023. Cross-cultural narratology: A comparative study of storytelling techniques in

eastern and western literature. *Journal of Asian Development Studies*, 12(4):742–753.

Roland Barthes. 1967. The death of the author. *Aspen*, (5–6).

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–33.

Jiangjie Chen, Chun Gan, Sijie Cheng, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. Unsupervised editing for counterfactual stories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10473–10481.

Cyril Chhun, Pierre Colombo, Fabian 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.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Google. 2025. Introducing Gemini 3 Flash: Benchmarks, Global Availability. https://blog.google/products/gemini/gemini-3-flash/. Accessed 2025-12-27.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Ariel Han and Shenshen Han. 2025. Empowering children's ai literacy through co-creating stories with llm. In *Proceedings of the 24th Interaction Design and Children*, pages 994–999.

Changying Hao, Liang Pang, Yanyan Lan, Yan Wang, Jiafeng Guo, and Xueqi Cheng. 2021. Sketch and customize: A counterfactual story generator. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12955–12962.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*.

Philip N Hineline. 2018. Narrative: Why it's important, and how it works. *Perspectives on Behavior Science*, 41(2):471–501.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yash Kumar Lal, Vanya Cohen, Nathanael Chambers, Niranjan Balasubramanian, and Ray Mooney. 2024. Cat-bench: Benchmarking language model understanding of causal and temporal dependencies in plans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19336–19354.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2024. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 26874–26901.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 598–604.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out (Workshop at ACL)*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

David Y Liu, Aditya Joshi, and Paul Dawson. 2026. Narrative theory-driven llm methods for automatic story generation and understanding: A survey. *arXiv preprint arXiv:2602.15851*.

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. The magic of IF: Investigating causal reasoning abilities in large language models of code. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9009–9022, Toronto, Canada. 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*.

Guillermo Marco, Julio Gonzalo, and Víctor Fresno. 2025. The reader is the metric: How textual features and reader profiles explain conflicting evaluations of ai creative writing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25432–25449.

Jan Christoph Meister. 2011. Narratology. The Living Handbook of Narratology. Rev. 19 January 2014.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Ponrudee Netisopakul and Usanisa Taoto. 2023. Comparison of evaluation metrics for short story generation. *IEEE access*, 11:140253–140269.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272.

Louise H Phillips, Louisa Lawrie, Zuzana Suchomelova, Sara Heinämaa, Amy O'Dwyer, and Min Hooi Yong. 2025. Age and cultural differences in the relationship between reading and theory of mind. *Poetics*, 109:101984.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André FT Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *arXiv preprint arXiv:2504.04953*.

Gerald Prince. 2003. *A dictionary of narratology*. U of Nebraska Press.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805, Online. Association for Computational Linguistics.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. 2025. What makes a reward model a good teacher? an optimization perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Peiqi Sui. 2026. Llms exhibit significantly lower uncertainty in creative writing than professional writers. *arXiv preprint arXiv:2602.16162*.

T. Todorov. 1969. *Grammaire du Décaméron*. Approaches to semiotics. Mouton.

Tzvetan Todorov. 1971. The 2 principles of narrative. *diacritics*, pages 37–44.

Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025. Igniting creative writing in small language models: LLM-as-a-judge versus multi-agent refined rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17171–17197, Suzhou, China. Association for Computational Linguistics.

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

Slavica Zec, Nicola Soriani, Rosanna Comoretto, and Ileana Baldi. 2017. High agreement and high prevalence: the paradox of cohen's kappa. *The open nursing journal*, 11:211.

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) 2020*. OpenReview.net.

A License

The TimeTravel dataset is used under the MIT License. Hugging face's Transformers and TRL libraries are used under the Apache License 2.0 license. Llama 3.1 is used under the LLaMA 3.1 Community License. Olmo 3, Qwen 3 and Selene-1-mini are used under the Apache License 2.0.

We use these artifacts in a way that's consistent with intended use for the purpose of academic research.

B TimeTravel Dataset

The TimeTravel dataset is constructed from the ROCStories corpus (Mostafazadeh et al., 2016) which consist of over 100k human written five-sentence short stories. In TimeTravel, there are 1.87k items in the test set, 1.87k items in the validation set, and 16.8k items in the supervised training set, all with human written counterfactual endings. Furthermore, the entire 100k ROC stories can be used for unsupervised training.

The dataset includes narratives that may allude to violent themes, and some generated outputs could be distressing. To the best of our knowledge, these stories are entirely fictional and do not represent real individuals.

C Computational Budget

Inference and d-RLAIF and SFT training was conducted on a single H200 GPU. Each training process took between 30 minutes - 4 hours.

Evaluation using Gemini-3-Flash was done using OpenRouter API, which cost roughly $3 per testset.

D Hyperparameters

We detail the hyperparameters applied for supervised fine-tuning (SFT) and GRPO-based deep reinforcement learning from AI feedback (d-RLAIF). All training and inference were performed using Hugging Face's Transformers and TRL libraries. The same hyperparameter configuration was maintained across all experimental settings. Early stopping was implemented for both SFT and d-RLAIF.

D.1 Supervised Fine-Tuning (SFT)

For SFT, we train the model using the AdamW optimiser with a fixed learning rate and Low-Rank Adaptation (LoRA) parameterisation (Hu et al., 2022). Training is conducted for a single fine-tuning stage, with evaluation performed periodically on a held-out validation set. The best model checkpoint is selected based on validation loss using early stopping.

LoRA-specific hyperparameters are:

- Rank $r = 64$
- Alpha $\alpha = 128$
- Dropout $= 0.05$

Other training hyperparameters are:

- Per-device batch size $= 8$
- Gradient accumulation steps $= 8$
- Learning rate $= 1 \times 10^{-4}$
- Number of epochs $= 1$
- Maximum sequence length $= 640$ tokens

Mixed-precision training is enabled using bfloat16 (bf16). Gradient check pointing is used to reduce memory usage. No extensive hyperparameter tuning was performed; values were chosen heuristically.

D.2 d-RLAIF with GRPO

For d-RLAIF, we employ GRPO with LoRA parameterisation for the policy model. At each training step, 16 candidate completions are generated per prompt, and early stopping is used based on evaluation metrics.

LoRA-specific hyperparameters for d-RLAIF:

- Rank $r = 64$
- Alpha $\alpha = 128$
- Dropout $= 0.05$

Other training hyperparameters are:

- Per-device batch size $= 24$
- Gradient accumulation steps $= 2$
- Learning rate $= 5 \times 10^{-6}$
- Number of epochs $= 1$
- Maximum prompt and completion length $=$ 512 tokens

Mixed-precision training using bfloat16 (bf16) is enabled. No extensive hyperparameter search was performed; hyperparameters were selected heuristically for stability. All experiments use fixed hyperparameters across conditions to ensure comparability. Sensitivity to hyperparameter choices is left to future work.

E Annotator Guidelines

To ensure consistent understanding of definitions, annotators are provided with instructions and examples before continuing into the main workflow, which consist of two primary phases: (1) Reading and Tagging, and (2) Rating.

E.1 Narrative State Tagging

Annotators are instructed to read both the original and rewritten stories, highlighting text according to Todorov's Narrative Theory. They are provided with the following criteria for the five narrative stages:

Equilibrium: An initial state before any transformation that is then changed by a disruption.
Disruption of Equilibrium: The event/condition that disrupts equilibrium and initiates transformations.
Recognition of the Disruption: A character or narrator recognises that the disruption has occurred.
Attempt to Resolve Disruption: Actions taken in an attempt to address or resolve the disruption.
Return to New Equilibrium: A new equilibrium state reached after the transformation.

Key Tagging Guidelines

1: Tagging every single word or sentence is not required; only highlight parts specifically relevant to that stage.
2: Some stages may be implicit and cannot be tagged.
3: Filler or background detail can be left untagged.

E.2 Common Variations and Visual Examples

We provide annotators with annotated examples of different narrative structural variations.

Variation 1: In Medias Res The narrative begins with the Disruption of Equilibrium, skipping the initial equilibrium.
Example: "The kitchen pipe burst, flooding the tiles in seconds. I saw the water reaching the carpet and knew the basement was next. I grabbed a wrench and tightened the main valve. Within an hour, the plumber arrived and the house was dry again."

Variation 2: Non-Linear Narrative Events are revealed out of order, often starting with the resolution or final effort.
Example: "I crossed the finish line and collapsed onto the grass. Years of regular jogs had kept me healthy until a sudden ankle sprain last winter. Sitting in the doctor's office back then, I decided I would run a marathon to prove I could heal. Today, my medal finally felt real."

Variation 3: Cliffhanger Endings The text ends abruptly during an Attempt to Resolve Disruption, with no resolution shown.
Example: "The exam room was quiet as the students wrote. Suddenly, the fire alarm screeched throughout the hall. The proctor stood up and realized there was smoke beneath the door. She grabbed the evacuation log and shouted for everyone to follow her into the stairwell..."

E.3 Rating/Evaluation Guide

After tagging, annotators evaluate the edited text based on its adherence to specific criteria. They are instructed to focus strictly on whether the criteria are met, not on the stylistic quality of the writing.

Logical: The text must contain no internal contradictions. For example:

> **Logical:** "Alex lost her job, but since she's an optimist, she walked home smiling."
>
> **Not Logical:** "Alex lost the job that meant everything to her. She walked home with a smile." (Illogical without context.)

Rational: The text must follow a coherent Todorovian progression, with no contextual or causal disconnection. All events should be interpretable as belonging to compatible narrative stages. For example:

> **Rational:** "The oven broke. I ordered pizza. We had a great meal." (Clear causal and contextual continuity.)
>
> **Not Rational:** "I was eating dinner. The waiter spilt wine. I finally got my car fixed." (Narrative jumps without coherent stage progression.)

Complete$_N$: The text must preserve all key Todorovian stage(s) from the original narrative. An edited text is complete only if it reproduces all essential stages, even if the specific events differ. For example:

Complete: Orig: Health → Flu → Recovery. Edit: Health → Accident → Surgery → Death. (All original stages preserved.)

Incomplete: Orig: Health → Flu → Recovery. Edit: Health → Accident → Surgery. (Missing the New Equilibrium stage.)

E.4 Handling Ambiguity

Annotators were given an **Unsure** (rating=2) option, to be used when the text is ambiguous.

Metric	Rating	A (%)	B (%)	C (%)
Logical	Agree	85.5	72.0	95.5
	Neutral	7.0	1.0	0.0
	Disagree	7.5	27.0	4.5
Rational	Agree	60.5	72.0	67.5
	Neutral	19.5	6.0	0.0
	Disagree	20.0	22.0	32.5
Complete	Agree	56.0	64.0	49.0
	Neutral	7.5	2.5	0.0
	Disagree	36.5	33.5	51.0
$\min_{\text{LRC}}$	Agree	36.5	47.0	32.5
	Neutral	18.0	2.5	0.0
	Disagree	45.5	50.5	67.5
Narrativity	Agree+	49.0	53.5	14.0
	Agree	31.5	27.5	42.0
	Neutral	8.0	9.0	25.5
	Disagree	0.0	1.0	2.5
	Disagree-	11.5	9.0	16.0

Table 6: Distribution: 3 human annotators (A, B, C), percentages shown (200 per metric per annotator). Agree+ is Strongly Agree; Disagree- is Strongly Disagree.

F Few-Shot Method

To select examples for our few-shot method, we use TimeTravel's supervised training set. For each example in the training set, we use DeBERTa-v3-small (He et al., 2023) to calculate the similarity between the retelling and the original (using BERTScore (Zhang et al., 2020)). We then rank the examples from most different to least different. We select the 1st, 500th and 1000th examples to diversify our 3 shot prompt to include examples that require high, medium, and low changes to the original ending.

G LLM-as-judge

While we ultimately only use a 'no-reasoning' prompt for the LLM-as-judge for training and evaluation, we initially tested two different output formats:

1. Generate the reasoning and then the score
2. Generate the score only[6]

We found that while reasoning increases alignment (Table 9 and 8), it increases token costs far too greatly for it to be realistically used in a d-RLAIF setting. So reasoning was excluded from being used in LLM-as-judge models.

H Training

A key challenge we encounter is the early stagnation during d-RLAIF training with GRPO, where policy models converge prematurely without reaching the optimal reward plateaus. To investigate this, we compare the standard 3-point Likert scale (R_O) against a more granular 5-point scale (R_{O5}), hypothesising that increased reward variance would mitigate signal loss in GRPO and improve learning. Interestingly, the R_{O5}-rewarded policy model becomes better at producing Logical and Rational outputs but significantly worse at retelling the narrative structure, degrading overall performance. This suggests that simply increasing scale granularity is insufficient to prevent gradient starvation (Pezeshki et al., 2021) in GRPO setups if the underlying reward distribution remains sparse or saturated, where the LLM-as-judge fails to meaningful differentiate between outputs.

We conduct additional variations of post-training on Llama-3.1-8B-Instruct. Those trained by M-Prometheus achieve better performance when trained using R_O and worse when trained using R_N compared to those trained by Selene-1-mini. When the LLM-as-judge model is instructed to output a 5-point Likert score instead of the original 3-point and the group relative advantage is calculated from R_{O5} instead of R_O, the policy model's outputs improve at the Logical and Rational criteria but significantly degrade at the Complete_N criteria.

[6]We instruct the LLM to output the score then the reasoning because post-hoc reasoning can improve accuracy (Lampinen et al., 2022; Lal et al., 2024). We stop inference after the score is generated, since it is impossible for the subsequent tokens to change the already-generated score.

Criteria	AC2	% Agree	κ
Logical	0.7648 [0.71, 0.82]	0.7967	0.1797 [0.0567, 0.2968]
Rational	0.7082 [0.65, 0.76]	0.7792	0.4068 [0.3024, 0.5035]
Complete_N	0.5611 [0.50, 0.62]	0.6750	0.3180 [0.2322, 0.4129]
min_LRC	0.5544 [0.49, 0.61]	0.6767	0.2981 [0.2116, 0.3851]
Narrativity	0.7415 [0.69, 0.79]	0.8660	0.3764 [0.2690, 0.4796]

Table 7: Inter-annotator agreement measured by Gwet's AC2, % Agreement, and Quadratic Weighted κ. Values in brackets represent the 95% Confidence Interval (CI).

Model	Logical	Rational	Complete_N	Overall	min_{LRC}	Narrativity
human	2.96/2.85/3.00	3.00/2.80/3.00	2.96/2.48/3.00	2.96/2.70/2.96	2.92/2.34/3.00	4.98/4.30/4.92
Qwen3	2.76/2.98/2.94	3.00/2.94/3.00	2.92/2.41/2.86	2.72/2.58/2.98	2.68/2.38/2.80	4.96/4.41/4.74
Olmo3	2.76/2.94/2.94	2.88/2.92/2.96	2.52/1.96/2.60	2.84/2.24/2.94	2.32/1.94/2.54	4.98/4.04/4.42
Llama3.1	2.72/2.74/2.84	2.56/2.34/2.72	2.20/1.71/2.18	2.36/2.00/2.74	1.96/1.62/2.06	4.82/3.22/3.86

Table 8: Model comparison across evaluation dimensions with LLM judges (Gemini / Selene / Prometheus) with reasoning.

Model	Logical	Rational	Complete_N	Overall	min_{LRC}	Narrativity
human	2.80/2.94/3.00	2.92/2.94/2.96	2.96/2.38/2.82	2.52/2.46/2.62	2.72/2.38/2.80	4.90/3.62/5.00
Qwen3	2.60/3.00/2.96	2.80/3.00/3.00	2.36/2.54/2.86	2.40/2.74/2.76	2.04/2.54/2.82	4.78/3.42/4.88
Olmo3	2.72/3.00/3.00	2.80/2.98/3.00	1.88/2.14/2.72	2.20/2.40/2.76	1.76/2.14/2.72	4.38/3.26/4.76
Llama3.1	2.68/2.92/2.94	2.24/2.72/2.52	1.52/1.90/2.22	1.60/2.10/2.28	1.36/1.86/2.12	3.82/2.76/4.18

Table 9: Model comparison across evaluation dimensions with LLM judges (Gemini / Selene / Prometheus) without reasoning.

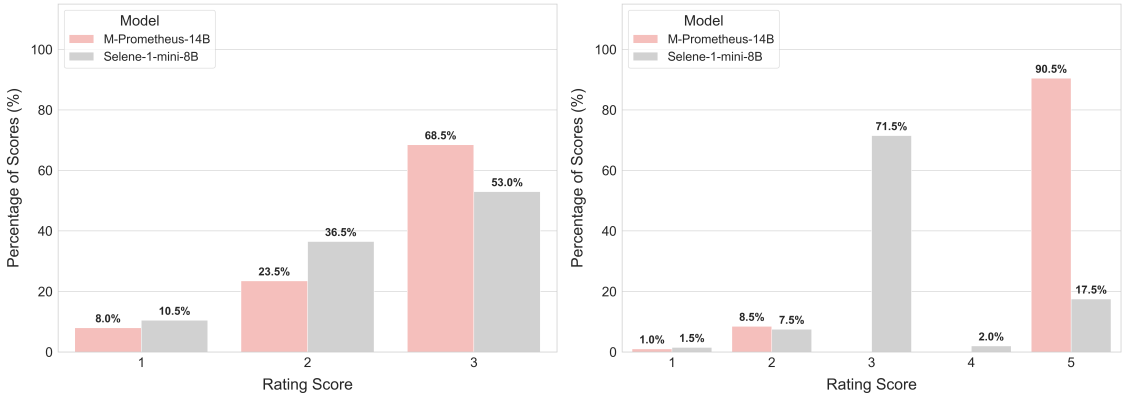

Figure 3: LLM-as-judge on the annotation set. Left: 'Narrativity' distribution. Right: 'Overall' distribution.

Author	Logical	Rational	Complete_N	min_LRC	Narrativity	BLEU-4	ROUGE-L
LLAMA-3.1-8B-INSTRUCT							
BASE	2.735	2.385	2.528	2.080	4.347	0.340	0.502
few-shot	2.768	2.615	2.505	2.213	4.303	0.306	0.473
SFT	2.320	2.501	**2.934**	2.260	4.692	**0.810**	**0.835**
RRR_O-Prometheus	2.801	2.825	2.594	2.400	4.534	0.179	0.386
RRR_O5-Prometheus	**2.917**	**2.943**	2.059	2.019	4.53	0.007	0.194
RRR_N-Prometheus	2.655	2.717	2.246	2.058	4.219	0.163	0.357
RRR_O-Selene	1.820	2.080	2.779	1.761	4.489	0.174	0.458
RRR_N-Selene	2.785	2.860	2.740	**2.554**	**4.733**	0.087	0.299

Table 10: Evaluation of post-trained LLMs using the test split (n=1871) from TimeTravel dataset using Gemini-3-Flash. Additional models trained on Llama-3.1.

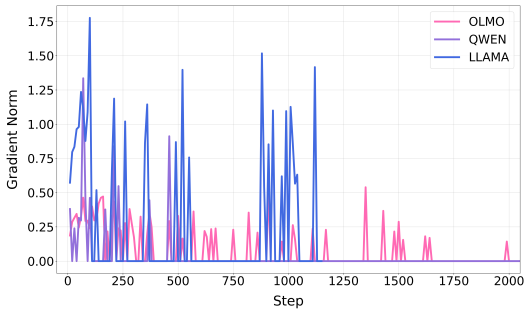

Figure 4: Policy model gradient norm during d-RLAIF using the reward signal R_O generated by Selene-1-mini.

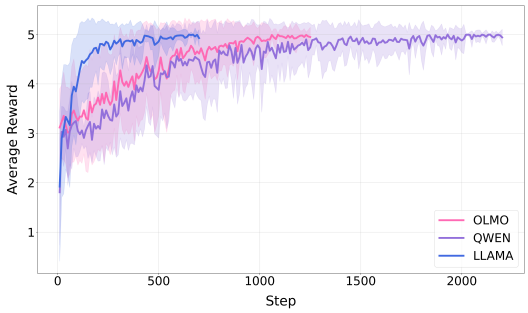

Figure 5: Mean and standard deviation of the reward signal R_N generated by Selene-1-mini during d-RLAIF.

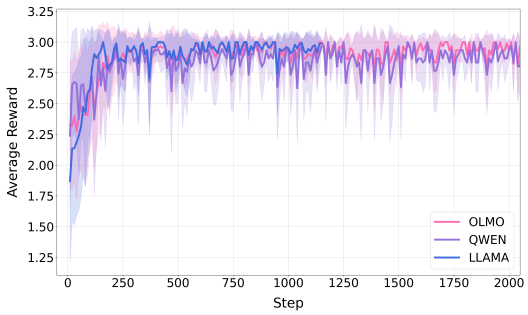

Figure 6: Mean and standard deviation of the reward signal R_O generated by Selene-1-mini during d-RLAIF.

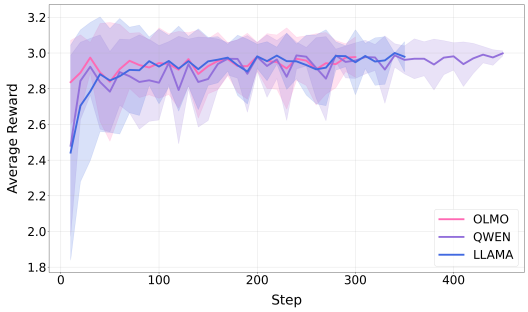

Figure 7: Mean and standard deviation of the reward signal R_O generated by M-Prometheus during d-RLAIF.

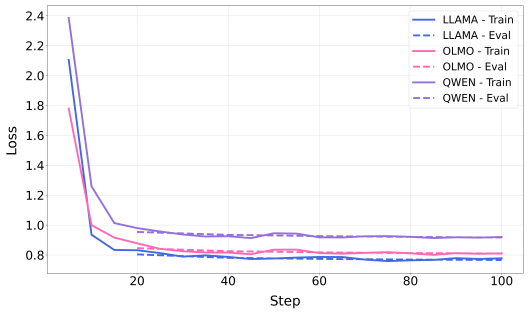

Figure 8: Training and evaluation loss during SFT.

I Significance Testing

Model	Method	Logical	Rational	Complete_N	min_LRC	Narrativity
	BASE	2.735*	2.385*	2.528*	2.080*	4.347*
	few-shot	2.768	2.615*	2.505*	2.213*	4.303*
LLAMA-3.1-8B-INST.	SFT	2.320*	2.501*	**2.934***	2.260*	4.692
	RRR_O-Prometheus	**2.801**	2.825*	2.594*	2.400*	4.534*
	RRR_N-Selene	2.785 ^{Base}	**2.860** ^{Base}	2.740 ^{Base}	**2.554** ^{Base}	**4.733** ^{Base}

Table 11: Evaluation of Llama-3.1-8B variants using the test split ($n = 1871$) from the TimeTravel dataset. Values marked with * indicate a statistically significant difference ($p < 0.05$ via Wilcoxon signed-rank test) compared to the **RRR_N-Selene** baseline. Bolded values represent the best performance for each metric.

Model	Method	Logical	Rational	Complete_N	min_LRC	Narrativity
	BASE	2.429*	2.132*	2.759*	1.940*	4.551*
	few-shot	2.543*	2.219*	2.686	2.006*	4.559*
QWEN-3-8B	SFT	2.402*	2.532*	**2.925***	2.317*	4.693*
	RRR_O-Prometheus	2.717*	2.336*	2.566*	2.024*	4.356*
	RRR_N-Selene	**2.781** ^{Base}	**2.890** ^{Base}	2.705 ^{Base}	**2.516** ^{Base}	**4.894** ^{Base}

Table 12: Evaluation of Qwen-3-8B variants using the test split ($n = 1871$) from the TimeTravel dataset. Values marked with * indicate a statistically significant difference ($p < 0.05$ via Wilcoxon signed-rank test) compared to the **RRR_N-Selene** baseline. Bolded values represent the best performance for each metric within the model family.

Model	Method	Logical	Rational	Complete_N	min_LRC	Narrativity
	BASE	2.762	2.812*	2.400*	2.222*	4.328*
	few-shot	2.745*	2.837*	2.377*	2.208*	4.415*
OLMO-3-7B-INST.	SFT	2.295*	2.441*	**2.910***	2.211*	4.647*
	RRR_O-Prometheus	2.770	2.857*	2.491*	2.325*	4.460*
	RRR_N-Selene	**2.793** ^{Base}	**2.893** ^{Base}	2.595 ^{Base}	**2.434** ^{Base}	**4.899** ^{Base}

Table 13: Evaluation of Olmo-3-7B variants using the test split ($n = 1871$) from the TimeTravel dataset. Values marked with * indicate a statistically significant difference ($p < 0.05$ via Wilcoxon signed-rank test) compared to the **RRR_N-Selene** baseline. Bolded values represent the best performance for each metric within the model family.

J Example Annotations (with Additional Reasoning/Rationale)

A: My nephew had a bad back one day. He took some ibuprofen and was fine. The pills were never even opened, and his back felt better by the next day.	
B: My nephew had a bad back one day. He took some ibuprofen and was fine. The pills were never even opened, and his back felt better by the next day.	
C: My nephew had a bad back one day. He took some ibuprofen was fine. The pills were never even opened, and his back felt better by the next day.	
Logical	
A	**1:** It's illogical and inconceivable that the nephew took ibuprofen yet the pills were never opened.
B	**3:** The edited text is logically coherent. Taking ibuprofen and leaving the prescription pills unopened is fully consistent.
C	**3:** The story does not present an inconceivable scenarios. Whilst at first there appears to be a logical inconsistency... "the pills" do not necessarily refer to ibuprofen.
Rational	
A	**1:** There's causal and contextual disconnection between taking ibuprofen and the pills being never opened.
B	**3:** The edited text is rationally interpretable. The progression from pain to treatment to recovery is clear.
C	**1:** The story cannot rationally be interpreted using Todorov's stages in its entirety because there is a causal disconnection... the most likely case is that "the pills" refer to the ibuprofen in sentence two, meaning that there is a causal disconnection.
Complete$_N$	
A	**1:** There are no recognisable disruptions. The edited text is not as much of a narrative as the original text.
B	**3:** The edited text is complete. It preserves the same broad Todorovian stage structure as the original.
C	**3:** All stages of Todorov's theory presented in the original text is present in the rewritten story.

Table 14: Annotator comments and narrative stage.

Narrativity
A: (4) My nephew had a bad back one day. He took some ibuprofen and was fine. The pills were never even opened, and his back felt better by the next day.
B: (4) My nephew had a bad back one day. He took some ibuprofen and was fine. The pills were never even opened, and his back felt better by the next day.
C: (4) My nephew had a bad back one day. He took some ibuprofen was fine. The pills were never even opened, and his back felt better by the next day.
Logical
A: (1) It's illogical and inconceivable that the nephew took ibuprofen yet the pills were never opened.
B: (3) The edited text is logically coherent. Taking ibuprofen and leaving the prescription pills unopened is fully consistent.
C: (3) The story does not present inconceivable scenarios. Whilst at first there appears to be a logical inconsistency, "the pills" do not necessarily refer to ibuprofen.
Rational
A: (1) There's causal and contextual disconnection between taking ibuprofen and the pills being never opened.
B: (3) The edited text is rationally interpretable. The progression from pain to treatment to recovery is clear.
C: (1) The story cannot rationally be interpreted using Todorov's stages in its entirety because there is a causal disconnection; the most likely interpretation is that "the pills" refer to the ibuprofen in sentence two.
Complete$_N$
A: (1) There are no recognisable disruptions. The edited text is not as much of a narrative as the original.
B: (3) The edited text is complete. It preserves the same broad Todorovian stage structure as the original.
C: (3) All stages of Todorov's theory presented in the original text are present in the rewritten story.

Table 15: An example of valid annotator disagreement when tagging and evaluating stories.

Narrativity

A: (5) I started using coupons. I realized it was too much work, with no real savings. I still collected coupons from the Sunday paper, emails and snail mail, but I didn't put much effort into researching stores that double the value of coupons. In the end, I didn't clip, tear, or cut any coupons, and I didn't end up saving any money last week.

B: (5) I wanted to see if i could save some money so I started using coupons. I realized it was too much work, with no real savings. I still collected coupons from the Sunday paper, emails and snail mail, but I didn't put much effort into researching stores that double the value of coupons. In the end, I didn't clip, tear, or cut any coupons, and I didn't end up saving any money last week.

C: (4) I wanted to see if i could save some money so I started using coupons. I realized it was too much work, with no real savings. I still collected coupons from the Sunday paper, emails and snail mail, but I didn't put much effort into researching stores that double the value of coupons. In the end, I didn't clip, tear, or cut any coupons, and I didn't end up saving any money last week.

Logical

A: (1) It is illogical that the narrator would keep collecting coupons after realising that it was too much work with no real savings.

B: (3) The edited text is logically coherent. The speaker tries coupons, recognises the effort outweighs the savings, reduces that effort, and ends without saving money.

C: (3) The events of the story progress without inconceivable scenarios. A logical inconsistency could be interpreted between sentences three and four... However, this inconsistency could be explained if we viewed the time of narration as a flashback.

Rational

A: (1) There is no rational causal connection between realising that coupons are too much work and continuing to collect coupons.

B: (3) The edited text is rationally interpretable. The causal progression from initial goal to failed outcome is clear.

C: (3) The events progress rationally: the character wants to see if they can save money with coupons, they realise it is too much work, then stop attempting to do so.

Complete$_N$

A: (3) It has a narrative structure of disruption, recognition, attempt and equilibrium. Even though it doesn't make sense, it's as much of a narrative as the original.

B: (3) The edited text is complete. It includes disruption, recognition, attempt, and a new equilibrium, and is at least as stage-complete as the original.

C: (1) The rewritten story is less complete than the original because it omits the crucial "attempt" stage. Even though the rewritten story includes as many Todorovian stages, it is not as narratively complete because character agency isn't represented as fully.

Table 16: An example of valid annotator disagreement when tagging and evaluating stories.

Narrativity

A: (4) Evan was ready to buy himself a bicycle. He went to the store but forgot his money. Then he looked around some more, trying to find something else he could afford or a way to get home.

B: (4) Evan was ready to buy himself a bicycle. He went to the store but forgot his money. Then he looked around some more, trying to find something else he could afford or a way to get home.

C: (3) Evan was ready to buy himself a bicycle. He went to the store but forgot his money. Then he looked around some more, trying to find something else he could afford or a way to get home.

Logical

A: (3) I interpret 'but forgot his money' as Evan forgetting a large sum of money intended for buying a bicycle. Evan may still have the usual money he carries everyday, and is able to logically find something else he could afford.

B: (3) The edited text is logical. Forgetting his money and then looking for an alternative or a way home is internally consistent.

C: (1) There is a logical impossibility in the story because: 1. Evan forgets his money, but 2. proceeds to look around the store for something he can "afford," which would be nothing.

Rational

A: (3) Looking for an alternative after forgetting the money to buy the bicycle is causally and contextually connected.

B: (1) The edited text is irrational in Todorovian terms. The attempt is present, but the narrative does not clearly resolve the disruption or establish a justified new equilibrium.

C: (1) The events of the story are causally disconnected.

Complete$_N$

A: (1) The edited text does not reach a new equilibrium whereas the original text does: "Then he rode home!" so it is not as complete of a narrative as the original.

B: (3) The edited text is complete. It includes disruption, attempt, and a form of new equilibrium, which matches the broad stage structure of the original.

C: (1) The original text presented a disruption in Evan's desire to buy a bicycle; an 'attempt at doing so in looking around the store; and a 'new equilibrium' after he purchases the bicycle he initially desired.

Table 17: An example of valid annotator disagreement when tagging and evaluating stories.

Narrativity
A: (1) I had to go to Boston for a work trip. It was a boring trip filled with conferences and no fun. We attended several conferences and had no memorable experiences, but I learned a lot for my career that I will carry with me.
B: (4) I had to go to Boston for a work trip. It was a boring trip filled with conferences and no fun. We attended several conferences and had no memorable experiences, but I learned a lot for my career that I will carry with me.
C: (4) I had to go to Boston for a work trip It was a boring trip filled with conferences and no fun. We attended several conferences and had no memorable experiences, but I learned a lot for my career that I will carry with me
Logical
A: (3) The scenario is conceivable and breaks no logical rules. "boring/unmemorable" and "learn a lot" are not mutually exclusive.
B: (1) The edited text contains a weak internal contradiction: the trip is described as having "no memorable experiences," yet it also results in a significant and lasting professional benefit. Because this shift is not sufficiently explained, the narrative's internal logic is weakened.
C: (3) At first, it seems there is a logical impossibility because the narrator, after having an uneventful trip "filled with [...] no fun" and "no memorable experiences," expresses in the last line that they learnt a lot. However, the conjunction "but" suggests that even though they had a bad time, the experience nevertheless taught them something. Bad or uneventful experiences can also be informative.
Rational
A: (3) Looking for an alternative after forgetting the money to buy the bicycle is causally and contextually connected.
B: (1) Although the text signals disruption and ends in a new equilibrium, it does not rationally develop the causal steps between them. The transition from a boring, uneventful trip to a meaningful career lesson lacks narrative grounding.
C: (3) Again, there appears to be a causal disconnection within the last sentence (from recognition to equilibrium). However, the word "but" makes the story rational. If the conjunction were "and", the story would be causally disconnected.
Complete$_N$
A: (1) There is no disruption in the edited text. It is not as complete as the original narrative, which has disruption and new equilibrium.
B: (1) The edited text is not fully complete because it excludes the Attempt stage. Without that stage, the sequence of Todorovian development is reduced, making the edited version less narratively complete than the original.
C: (3) Even though the fourth stage of Todorov's theory seems more important to create a complete narrative, the rewritten story feels just as complete because of the final clause "but I learned a lot for my career that I will carry with me."

Table 18: An example of valid annotator disagreement when tagging and evaluating stories.

Narrativity
A: (5) In the last mile of a marathon only two runners were in contention. One runner got cramps on the ground and tripped the other. They both rested on the ground. before they drop out.
B: (5) In the last mile of a marathon only two runners were in contention. They both got cramps and had to drop out of the race. One runner got cramps on the ground and tripped the other. They both rested on the ground. They rested for a few second before they drop out.
C: (3) In the last mile of a marathon only two runners were in contention. They both got cramps and had to drop out of the race. One runner got cramps on the ground and tripped the other. They both rested on the ground. They rested for a few second before they drop out.
Logical
A: (3) Conceivable. The events are not narrated in chronological order. "They both got cramps and had to drop out" is a summary of the explanations that follow.
B: (1) The edited text is not logical. The claim that one runner "got cramps on the ground and tripped the other" is internally unclear and weakens the event sequence.
C: (3) The events of the story progress logically, free from internal contradictions; the two runners both get cramps, drop to the floor, and will eventually drop out.
Rational
A: (3) Everything is contextually and causally connected even if it's not in chronological order.
B: (1) The edited text is not rational. The causal progression from cramping to tripping to resting to dropping out is not presented in a fully coherent way.
C: (3) All events can be rationally interpreted using the stages. The last two sentences are an elaboration of the disruption that I tagged in sentence two.
Complete$_N$
A: (3) It is just as complete of a narrative as the original. Cramps are the disruption, resting is the attempt to resolve, and dropping out is the new equilibrium.
B: (3) The edited text is complete in stage coverage. It still presents equilibrium, disruption, recognition, attempt, and a final outcome.
C: (1) The rewritten story omits an attempt at new equilibrium because there is no activity to overcome the obstacle of falling to the ground with cramps.

Table 19: An example of valid annotator disagreement when tagging and evaluating stories.

K Example Outputs

Source	Text	L	R	C	M	N
Original Story						
Premise	Max and May were at a tavern.					
Initial	They saw the dartboard and challenged each other to a game.					
Original Ending	Max played very well. But May thru even better than him. May won the game.					
Counterfactual Retelling						
Premise	Max and May were at a tavern.					
Counterfactual	They didn't play darts there because they were not competitive.					
Edited Ending	???					
Human Ground Truth						
Crowd Worker	Max played pool very well. But May thought she was even better than him. May won the game, but neither really cared who won.	3	3	3	3	5
Olmo						
Base	Max and May decided not to play since they weren't competitive.	3	3	1	1	1
Few-Shot	Max and May talked and laughed instead of playing. They enjoyed the atmosphere together.	3	3	3	3	1
SFT	Max played very well. But May threw even better than him. May won the game.	1	1	3	1	1
RRR_O	They left the tavern satisfied with their decision not to compete.	3	3	3	3	1
RRR_N	They decided to sing a song instead. The tavern erupted in laughter and applause—both agreed it was far more fun than competition.	3	3	3	3	5
Llama						
Base	Max and May were at a tavern. They saw the dartboard but decided not to play because they were not competitive.	3	1	1	1	1
Few-Shot	They decided to skip the dartboard and instead ordered a couple of beers to relax and enjoy the evening.	3	3	1	1	1
SFT	Max played very well at cards. But May threw even better than him. May won the game.	1	3	3	1	5
RRR_O	They ended up chatting over a pint instead, enjoying each other's company without the competitive edge that darts had brought out in them.	3	3	1	1	5
RRR_N	They ordered a round and struck up a conversation. Enjoying each other's company, they decided to share a story. Max won over May with his tale.	3	3	3	3	5
Qwen						
Base	Max played very well. But May threw even better than him. May won the game.	1	1	3	1	1
Few-Shot	They didn't play darts there because they were not competitive. Instead, they talked about their favorite books and shared stories about their travels.	3	1	1	1	1
SFT	Max played very well at darts. But May threw even better than him. May won the game.	1	1	3	1	5
RRR_O	They decided to play a different game instead, one that allowed them to enjoy each other's company without competition.	3	3	1	1	4
RRR_N	They ordered drinks, then debated chess. Max won, but May vowed to train. They vowed to rematch.	3	3	3	3	5
Codex						
Code-Prompt	Max played very well. But May thru even better than him. They had a good time.	1	1	1	1	1

Table 20: Example outputs from the test set. Ratings are given by LLM-as-judge (Gemini-3-Flash): L=Logical, R=Rational, C=$Complete_N$, M=min_{LRC}, N=Narrativity.

Source	Text	L	R	C	M	N
Original Story						
Premise	Kimmy was on a trip to Paris.					
Initial	Her plane had just landed and she and her classmates her excited.					
Original Ending	They couldn't wait to start seeing museums and churches! They spent lots of time shopping and seeing historical sites. Kimmy was very excited.					
Counterfactual Retelling						
Premise	Kimmy was on a trip to Paris.					
Counterfactual	Kimmy's plane crashed over the ocean and killed everyone on board.					
Edited Ending	???					
Human Ground Truth						
Crowd Worker	Kimmy had just wanted to see the museums and churches! Her parents spent a lot of time planning her funeral. Kimmy was dead.	3	3	3	3	5
Olmo						
Base	Given the tragic counterfactual, there is no living person to experience excitement or see museums and churches, so the premise cannot be fulfilled in the same way as the original ending.	3	3	1	1	3
Few-Shot	There is no edited ending, as the counterfactual scenario is a tragic accident with no possible positive or alternative ending.	3	3	1	1	3
SFT	They couldn't wait to start seeing museums and churches! They spent lots of time shopping and seeing historical sites. Kimmy was very excited.	1	1	3	1	1
RRR_O	Due to the tragic accident, the trip never happened, and Kimmy and her classmates never experienced Paris or its attractions. There were no museums, churches, shopping, or historical sites to see. Kimmy could never have been excited in the same way.	3	3	3	3	4
RRR_N	Grief-stricken, the world mourned. No legacy remained—only memories in the waves. The tragedy ended hope; Paris dreams faded. The story closed with silent sorrow.	3	3	3	3	5
Llama						
Base	They never made it to Paris. Kimmy's plane crashed over the ocean and there were no survivors.	3	1	3	1	3
Few-Shot	Kimmy's plane crashed over the ocean and killed her.	3	3	1	1	3
SFT	They couldn't wait to start seeing museums and churches! They spent lots of time shopping and seeing historical sites. Kimmy was very excited.	1	1	3	1	1
RRR_O	The plane wreckage was found at the bottom of the ocean, a tragic reminder of the devastating crash that took Kimmy and her classmates' lives.	3	3	3	3	4
RRR_N	Panicked, Kimmy clung to debris, the only survivor. She was rescued and returned home, re-evaluating life. Kimmy was forever changed.	1	3	3	1	5
Qwen						
Base	There were no survivors, and Kimmy's adventure in Paris was cut short tragically. The classmates and everyone on board lost their lives in the crash.	3	3	3	3	3
Few-Shot	Kimmy's plane crashed over the ocean and killed everyone on board. There was no one left to see museums and churches, and no one to share the excitement with. Kimmy's trip ended tragically before it could begin.	3	1	3	1	5
SFT	They couldn't wait to start seeing museums and churches. They spent lots of time shopping and seeing historical sites. Kimmy was very excited about the trip.	1	1	3	1	3
RRR_O	Unfortunately, Kimmy and her classmates were never able to explore Paris. The plane crash claimed their lives before they could experience the city's wonders.	3	3	3	3	4
RRR_N	Miraculously, Kimmy survived and was rescued. She mourned her classmates but vowed to honor them. She traveled solo, explored Paris, and wrote a heartfelt tribute. Her story inspired many.	1	3	3	1	5
Codex						
Code-Prompt	Kimmy was very sad. She was very sad. She was very sad.	1	1	3	1	3

Table 21: Example outputs from the test set. Ratings are given by LLM-as-judge (Gemini-3-Flash): L=Logical, R=Rational, C=$Complete_N$, M=min_{LRC}, N=Narrativity.

Source	Text	L	R	C	M	N
Original Story						
Premise	Amy's class went on a field trip to see A Mid Summer Night's Dream.					
Initial	They arrived at the theater and found their seats.					
Original Ending	Amy was in awe of the actors' discipline and talent. After the play she was able to speak with the cast. Amy was thrilled.					
Counterfactual Retelling						
Premise	Amy's class went on a field trip to see A Mid Summer Night's Dream.					
Counterfactual	They forgot their tickets in the classroom.					
Edited Ending	???					
Human Ground Truth						
Crowd Worker	Amy was in awe at their stupidity. She saw the play the next week and then she was able to speak with the cast. Amy was thrilled.	3	3	3	3	5
Llama						
Base	They had to go back to the classroom to get the tickets, but were able to find their seats just in time for the start of the play. Amy was in awe of the actors' discipline and talent. After the play, they were unable to speak with the cast because they had to rush back to the classroom to pick up their forgotten items, but Amy was still thrilled to have seen the performance.	3	3	3	3	5
Few-Shot	Amy was frustrated that they couldn't get in to see the play. They had to go back to the classroom to get the tickets. After finally getting in, Amy was in awe of the actors' discipline and talent. However, she was disappointed that they missed the beginning of the play.	3	3	3	3	5
SFT	Amy was in awe of the actors discipline and talent. After the play she was able to speak with the cast. Amy was thrilled.	2	1	3	3	3
RRR_O	Panicked, they rushed back to the classroom to retrieve their tickets, but unfortunately, the class had already left for the theater. Disappointed, they decided to watch from outside the theater, still managing to catch a glimpse of the performance and hearing the sound of the actors' voices carrying through the windows.	1	3	1	1	5
RRR_N	Panicked, they rushed back to school. Luckily, they found the tickets. They arrived just in time, and Amy was amazed by the performance. She met the cast backstage, beaming with joy.	3	3	3	3	5
Qwen						
Base	Amy was in awe of the actors' discipline and talent. After the play, she realised she had forgotten her tickets in the classroom and couldn't speak with the cast. Amy was disappointed.	1	1	3	1	5
Few-Shot	They forgot their tickets in the classroom. Amy was disappointed and worried they would miss the play. After a frantic search, they found the tickets just in time to enter the theater.	3	1	3	1	5
SFT	Amy was in awe of the actors discipline and talent. After the play she was able to speak with the cast. Amy was thrilled.	2	1	3	1	5
RRR_O	Amy was frustrated when they realized they had forgotten their tickets in the classroom. She tried to find a solution but ultimately had to leave the theater without seeing the play.	3	3	3	3	5
RRR_N	They raced back, found tickets, and rushed in. The play was unforgettable—Amy cheered, then met the cast proudly.	3	3	3	3	5
Olmo						
Base	Due to forgetting their tickets, Amy and her classmates couldn't enter the theater. Everyone was disappointed but decided to talk to the teachers about the situation to see if they could get in later. The experience left them eager to try again.	3	3	1	1	5
Few-Shot	Amy was frustrated that they couldn't go into the show. She wished they had remembered the tickets so they could have enjoyed the performance and maybe even spoken with the cast afterward.	3	3	1	1	4
SFT	Amy was in awe of the actors discipline and talent. After the play she was able to speak with the cast. Amy was thrilled.	2	1	3	1	3
RRR_O	Amy was frustrated by not being able to enter the theater and had to explain the situation to staff before being redirected to a nearby event. Despite disappointment, she appreciated seeing live theatre from a different angle.	3	3	3	3	5
RRR_N	Realising the mix-up, the group rushed back to class, scrambled for tickets, and finally boarded the bus—reassured and racing to the theater. The play began, applause echoed, and relief washed over them.	3	3	3	3	5
Codex						
Code-Prompt	Amy was in awe of the actors discipline and talent. After the play she was able to speak with the cast. Amy was thrilled.	1	1	3	1	3

Table 22: Example outputs from the test set. Ratings are given by LLM-as-judge (Gemini-3-Flash): L=Logical, R=Rational, C=$Complete_N$, M=min_{LRC}, N=Narrativity.