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

Anonymous ACL submission

Abstract

Despite the subjective nature of storytelling, past works on automatic story generation (ASG) have relied on limited ground truths for training and evaluation. In this work, we explore reinforcement learning (d-RLAIF) as a post-training alternative to supervised fine-tuning (SFT). We first apply Todorov's Theory of Narrative Equilibrium to establish principles that define desirable ASG qualities. We prompt 7B and 14B LLM-as-judge models with our principles to test alignment with human annotators and provide reward signals during d-RLAIF. We use Gemini-3-Flash to evaluate the output of our post-trained models and compare them to human-written stories from the Time-Travel dataset. We show that d-RLAIF offers a viable alternative to supervised fine-tuning (SFT)–producing stories that are more diverse and aligned with human narrative conventions. Our paper demonstrates the promise of reinforcement learning for linguistically grounded post-training for subjective tasks such as ASG.

1 Introduction

Automatic Story Generation (ASG) involves the selection and representation of sequence(s) of events (Li et al., 2013), including the editing and reimagining of existing stories. Desirable ASG outputs are often required to display qualities such as coherence, creative diversity and alignment with human narrative conventions while also meeting contextually dependent criteria, especially for applications across various domains such as gaming (Kumaran et al., 2023), education (Han and Han, 2025) and mental health (Vieira Sousa et al., 2024).

While statistical metrics based on reference outputs have been a norm in natural language processing (NLP), they have shown to correlate poorly or even negatively with human judgements at evaluating AI-generated stories (Qin et al., 2019). Intuitively, storytelling's subjectivity precludes a single, universally preferable ground truth. Instead, this paper answers calls to incorporate literary and narrative theory from linguistics and allied areas when designing theory-informed systems of training and evaluation (Alhussain and Azmi, 2021).

Due to their generative ability, large language models (LLMs) offer practical solutions to applying narrative theories with minimal cost, expertise and setup (Carroll, 2024) while outperforming earlier models on ASG tasks (Wang et al., 2023). Current approaches commonly apply structuralist theories from Classical Narratology[1] (Meister, 2011), which find common formal and structural patterns across all textual narratives[2]. While the path of prompt-engineering has found success at applying narrative theories to improve performance (Tian et al., 2024), the path of post-training is comparatively unexplored–where the lack of common ASG datasets, metrics and benchmarks form unavoidable roadblocks in model training and evaluation.

We believe a narrative theory-informed approach to post-training offers potential and practicality in moving forward with these challenges. By examining data from the training process and evaluating the performance of our post-trained models, our paper finds evidence to answer the research question:

"How can a narrative theory be applied in LLM post-training to achieve desirable ASG qualities?"

To replace the pairwise-preference (standard in RLHF) data-driven assessment of narrative quality with a narrative theory-driven assessment, we use direct reinforcement learning from AI feedback (d-RLAIF). We use an LLM-as-judge, prompted with Todorov (1971)'s Theory of Narrative Equilibrium, to produce reward signals for group relative policy optimisation (GRPO) (Shao et al., 2024).

[1] In comparison, Postclassical Narratology refers to a more diverse and interdisciplinary practice involving contextual, cognitive and transmedial/transgeneric perspectives that apply narratology outside of literary and textual narratives.

[2] A well-known example being Campbell (2008)'s *The Hero with a Thousand Faces*.

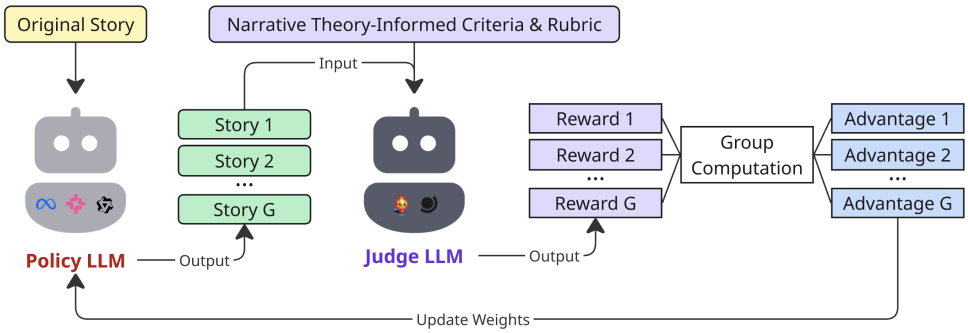

Figure 1: Our architecture that utilises d-RLAIF (Lee et al., 2024) w/ a narrative theory-informed LLM-as-judge generating the reward signal for GRPO (Shao et al., 2024)

The novelty of the paper lies in its utilisation of reinforcement learning (Figure 1) for narrative theory-informed ASG, based on Qin et al. (2019)'s TimeTravel dataset. To address the research question, we (1) Develop an annotation scheme and evaluation criteria by applying Todorov's theory. (2) Compare how humans and LLMs annotate and evaluate a dataset (n=200) of human and AI-generated stories. (3) Train LLMs (7-8B parameters) using our d-RLAIF pipeline. (4) Evaluate and compare the performance of our d-RLAIF models to instruction-tuned models, supervised fine-tuned (SFT) models, and human crowd-workers.

2 Preliminaries

Typically, ASG tasks are evaluated by multidimensional criteria. For example, Chakrabarty et al. (2024) designed a scheme to evaluate stories on the desirable ASG qualities measured using 14 criteria. To focus our efforts primarily on the application of narrative theory through reinforcement learning, we simplify our task by choosing a specific type of storytelling that can be evaluated across significantly fewer criteria–the task of story retelling.

Specifically, we choose Qin et al. (2019)'s TimeTravel dataset and task of rewriting the ending of a short story after a provided 'counterfactual' event[3] replaces the 'initial' event. For example:

Premise: Alex and Blair were classmates.
Initial: They secretly liked each other.
Original Ending: Alex gave in to desire and asked Blair on a date. They got married after graduation.
Counterfactual: They secretly hated each other.
Edited Ending: Alex decided to speak up and confronted Blair. Surprisingly, they resolved their issues. Since then, they've become lifelong friends.

[3]This is actually a counterfactual antecedent while the 'Edited Ending' is a counterfactual consequent.

The original authors show that an AI-generated output's similarity[4] to the human-written ground truths correlates weakly and sometimes negatively with desirable ASG qualities of coherence and relevance. Intuitively, evaluating/training models on their ability to generate a specific string seems unsuitable for storytelling tasks. Instead, we aim to utilise a particular narrative theory, described hereafter, for a linguistically grounded approach.

2.1 Todorov's Theory of Narrative Equilibrium

To develop suitable principles to guide the LLM-as-judge, we first establish our formal definition of a narrative. We choose to adopt Todorov (1971)'s Theory of Narrative Equilibrium, which describes narratives as sequences of causally and contextually connected stages and transformations categorised by 5 types of 'Todorovian stages':

Equilibrium: An initial status quo.
Disruption: Something changes the equilibrium.
Recognition: Awareness of the disruption.
Attempt: Action taken to address the disruption,
New Equilibrium: A new status quo.

We choose Todorov's theory because it is compatible with the structure of the retelling task:

Equilibrium: Alex and Blair were classmates.
Disruption: They secretly liked each other.
Recognition: Alex gave in to desire
Attempt: and asked Blair on a date.
New Equilibrium: They got married after graduation.

We also adapt Todorov's theory to formalise a simplistic criterion of narrativity: the scalar quantity of something being a narrative (Abbott, 2014).

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

Todorov believes that the reader must be able to interpret most types of Todorovian stages from a narrative text. i.e. A text which contains all 5 types of Todorovian stages is higher in narrativity than a text which contains 3, while a text which contains 1 type would be unrecognisable as a narrative.

3 Data Annotation

3.1 Criteria

With myriad ways to make a desirable retelling, it is more concise to define what it should not be, rather than what it should be. Using Todorov's theory, we focus our evaluation criteria on what a desirable retelling should avoid. All desirable retellings need to satisfy the criteria:

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 make the text remains as narratively complete as the original.[5]
min$_{\mathbf{LRC}}$: Since any desirable story needs to satisfy $\frac{3}{3}$ 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.

3.2 Curation

To assess alignment between LLM-as-judge and human annotators, we curate a pool of human and AI-generated retellings. Starting from the first 3000 items in the TimeTravel supervised training split (Qin et al., 2019), where each item provides an 'input' consisting of the premise, initial state, original ending, and counterfactual, and the ground truth human-written edited ending. For each input, we produce three AI-generated 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 models typically produce coherent, contradiction-free, and structurally similar outputs, we introduce a filter in order to diversify the score distribution in the annotation. For each input's 3 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:

$$\text{diversity} = \min(\text{distances}) \times \text{mean}(\text{distances})$$

We then select the 50 inputs with the highest diversity and, for each, retain four retellings (the three LLM endings plus the human retelling), yielding an annotation set of $n = 200$.

Metric	Rating	A (%)	B (%)
Logical	Agree	85.5	95.5
	Neutral	7.0	0.0
	Disagree	7.5	4.5
Rational	Agree	60.5	67.5
	Neutral	19.5	0.0
	Disagree	20.0	32.5
Complete$_{\text{N}}$	Agree	56.0	49.0
	Neutral	7.5	0.0
	Disagree	36.5	51.0
min$_{\text{LRC}}$	Agree	36.5	32.5
	Neutral	18.0	0.0
	Disagree	45.5	67.5
Narrativity	Agree+	49.0	14.0
	Agree	31.5	42.0
	Neutral	8.0	25.0
	Disagree	0.0	2.5
	Disagree-	11.5	16.0

Table 1: Distribution: 2 human annotators (A and B), percentages shown (n=200 per metric). Agree+ is Strongly Agree; Disagree- is Strongly Disagree.

Criteria	AC2	% Agree	κ
Logical	0.8986	0.9075	0.2751
Rational	0.7396	0.8063	0.4906
Complete$_{\text{N}}$	0.5659	0.6813	0.3403
min$_{\text{LRC}}$	0.6197	0.7250	0.3739
Narrativity	0.7072	0.8519	0.3460

Table 2: Inter-annotator agreement measured by Gwet's AC2, % Agreement, and Quadratic Weighted κ.

[5]Consider this retelling: 'Alex and Blair were classmates. They secretly hated each other. They never spoke again.' While the retelling is plausible and coherent, it is hard to see it as a desirable retelling when it 'loses' some of the original's narrativity.

LLM Annotator	Logical	Rational	$\text{Complete}_{\text{N}}$	Overall	min_{LRC}	Narrativity
#1 Reasoning then Score						
Selene-1-mini	0.90/0.07	*0.70/0.23*	**0.66/0.34**	**0.47/0.19**	**0.62/0.27**	**0.77/0.27**
M-Prometheus	**0.91**/0.06	0.68/0.11	0.54/0.24	0.29/0.05	0.38/0.16	0.68/0.17
Gemini-3-Flash	0.88/**0.28**	**0.71/0.23**	0.60/0.34	0.37/0.15	0.48/0.27	0.68/0.04
#2 Score only						
Selene-1-mini	0.91/-0.02	0.68/0.10	0.61/0.20	0.44/0.12	0.55/0.13	0.77/0.26
M-Prometheus	**0.93**/0.08	0.68/0.16	0.53/0.17	0.39/0.12	0.39/0.15	0.68/0.16
Gemini-3-Flash	0.84/**0.27**	**0.71/0.33**	**0.62/0.41**	**0.52/0.29**	**0.55/0.31**	**0.76/0.40**

Table 3: Agreement between LLM and human annotators reported by *Gwet AC2 / Cohen's weighted kappa*.

3.3 Human Annotation

Two human annotators of English-speaking backgrounds evaluated the human and AI-written stories following this tag-and-evaluate process:

- Tag parts of text with Todorovian stages. The variety of tags is automatically counted and computed into a 1-5 Narrativity score. (Section 2.1)
- Use 3-point Likert scales to evaluate on the Logical/Rational/$\text{Complete}_{\text{N}}$ criteria.

Since our LLMs generated logical and rational stories in most cases, our rating distributions are skewed (Table 1), leading to the known prevalence/bias paradox (Zec et al., 2017) for Cohen's κ (e.g. $\kappa = 0.28$ despite 91% agreement). We, therefore, report Gwet (2014)'s AC2 alongside κ for inter-annotator reliability (Table 3), which is more robust to prevalence and marginal bias.

Considering both AC2 and κ, the annotators reach fair to moderate agreement at worst and moderate to substantial agreement at best, making different but equally faithful interpretations of Todorovian story structures. Consider the original 'Alex and Blair' story from section 2: it would be faithful to tag 'they secretly liked each other' as the disruption and it would also be faithful to tag 'Alex asked Blair on a date' as the disruption. The 'Logical' and 'Rational' criteria also draw on psychological and philosophical considerations, placing them beyond the scope of what's definable by this paper.

3.4 LLM-as-judge

We utilise 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 specialising in evaluation tasks, while Gemini-3-Flash is chosen as a 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 try 2 different output formats:

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

We evaluate the models' accuracy based on AC2 & κ measured against the average of the human annotators (Table 3). We found that the smaller 8B Selene-1-Mini is more accurate with reasoning, Gemini-3-Flash is more accurate without, while the 14B M-Prometheus is mostly indifferent.

4 Harnessing d-RLAIF for ASG

To study the effects of our d-RLAIF 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.

We use GRPO (Shao et al., 2024) and LoRA (Hu et al., 2022) to optimise the d-RLAIF training process with TimeTravel's unsupervised learning split as input. The policy model generates 16 outputs for each input, 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 of

[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.

the policy model's LoRA adapter (Figure 1). Both the policy and LLM-as-judge model fit on a single H200 GPU with 141GB VRAM.

Using the results from Table 3, we initially choose Selene-1-mini without reasoning as our fast LLM-as-judge and reward signal generator.[7] While the 'overall' score is less accurate than aggregating $\min_{\text{LRC}}$ from 3 separate inferences, we consider it to be an acceptable degradation at $\frac{1}{3}$ the token cost and use it for our reward signal R_O. We also conduct d-RLAIF with an alternative reward score R_N based on narrativity (Figure 4).

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 of 3 or 5, indicating no further measurable improvement under the current reward signal. 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.

However, we find that when Qwen and Olmo are trained on Selene-1-mini's R_O signal, they converge[8] before and without reaching the goal reward plateau (Figure 2). To ensure the reward plateau is reached, we train all 3 models on M-Prometheus's R_O signal. They all reach the reward plateau within 500 global optimiser steps. (Figure 3) We subsequently use M-Prometheus to train Llama on 2 additional reward signals, R_N and a modified R_{O5}, which changes the rubric to use a 5 point likert scale instead of 3 to test the effects of changing the ordinal scale.

To summarise, the reward signals we use are the:

- R_O signal from the 'Overall' prompt. (1-3)
- R_{O5} signal from the 'Overall' prompt. (1-5)
- R_N signal from the 'Narrativity' prompt.

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 (Figure 5). All models converge faster and with lower number of generations and optimiser steps during SFT than d-RLAIF.

[7]We choose to save the evaluation criteria–Logical, Rational & Complete_N–and our most reliable LLM-as-judge–Gemini-3-Flash without reasoning for final evaluation.

[8]For the gradient norm graph, see figure 8.

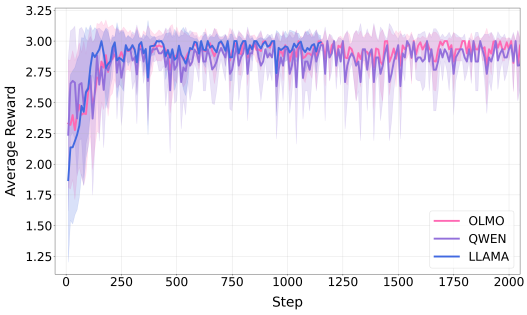

Figure 2: Mean and standard deviation of the reward signal R_O generated by Selene-1-mini during d-RLAIF.

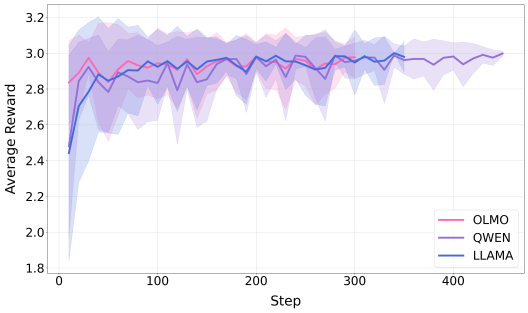

Figure 3: Mean and standard deviation of the reward signal R_O generated by M-Prometheus during d-RLAIF.

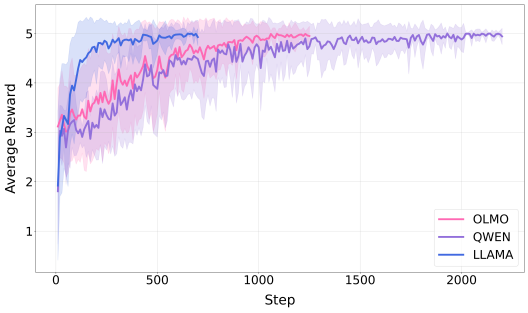

Figure 4: Mean and standard deviation of the reward signal R_N generated by Selene-1-mini during d-RLAIF.

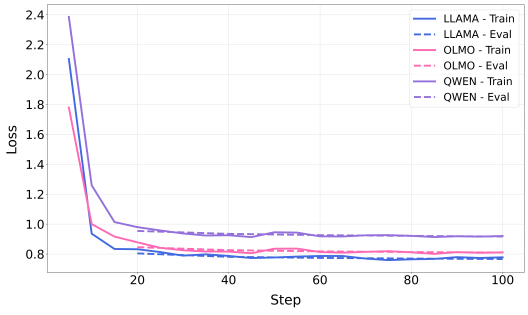

Figure 5: Training and evaluation loss during SFT.

Reteller	Logical	Rational	Complete_N	min_LRC	Narrativity	BLEU-4	ROUGE-L
TIMETRAVEL 'GROUND TRUTHS'							
Human Crowd Workers	2.884	2.927	2.933	2.811	4.718	1.000	1.000
LLAMA-3.1-8B-INSTRUCT							
BASE	2.735	2.385	2.528	2.080	4.347	0.340	0.502
SFT	2.320	2.501	2.934	2.260	4.692	0.810	0.835
d-RLAIF-R_O-Prometheus	2.801	2.825	2.594	2.400	4.534	0.179	0.386
d-RLAIF-R_O5-Prometheus	2.917	2.943	2.059	2.019	4.53	0.007	0.194
d-RLAIF-R_N-Prometheus	2.655	2.717	2.246	2.058	4.219	0.163	0.357
d-RLAIF-R_O-Selene	1.820	2.080	2.779	1.761	4.489	0.174	0.458
d-RLAIF-R_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
SFT	2.402	2.532	2.925	2.317	4.693	0.790	0.823
d-RLAIF-R_O-Prometheus	2.717	2.336	2.566	2.024	4.356	0.337	0.491
d-RLAIF-R_N-Selene	2.781	2.890	2.705	**2.516**	4.894	0.003	0.192
OLMO-3-7B-INSTRUCT							
BASE	2.762	2.812	2.400	2.222	4.328	0.097	0.315
SFT	2.295	2.441	2.910	2.211	4.647	0.787	0.821
d-RLAIF-R_O-Prometheus	2.770	2.857	2.491	2.325	4.460	0.090	0.309
d-RLAIF-R_N-Selene	2.793	2.893	2.595	**2.434**	4.899	0.002	0.178

Table 4: Evaluation of post-trained LLMs using the test split (n=1871) from TimeTravel dataset.

5 Results

We evaluate the performance of our models on TimeTravel's test split (n=1871) by prompting Gemini-3-Flash for evaluation based on our criteria. 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 4. We also use BLEU-4 and ROUGE-L to calculate the average similarity of each model's generation to the nearest human-written ending to measure linguistic similarity.

Table 4 shows that while human crowd workers do not receive the highest scores for any of the first 3 criteria, their stories perform significantly better than LLMs when considering min_LRC, which is more faithfully represents story quality and model performance than the average, as all desirable stories must score perfectly across the 3 criteria.

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 were highest in linguistic similarity to the human-written stories, as well as the most structurally similar to the original stories (measured by Complete_N). Meanwhile, d-RLAIF models trained on R_N produced more linguistically different stories compared to those trained on R_O.[9]

We highlight that d-RLAIF with M-Prometheus's R_O improves the overall performance of all tested models, where Selene-1-mini's narrativity-based reward signal R_N results in the best min_LRC performance, second only to humans. Their outputs also contain the highest narrativity.

From the additional models post-trained from 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.

[9] We do not evaluate Qwen and Olmo trained on Selene-1-mini's R_O since they do not reach the reward plateau during d-RLAIF.

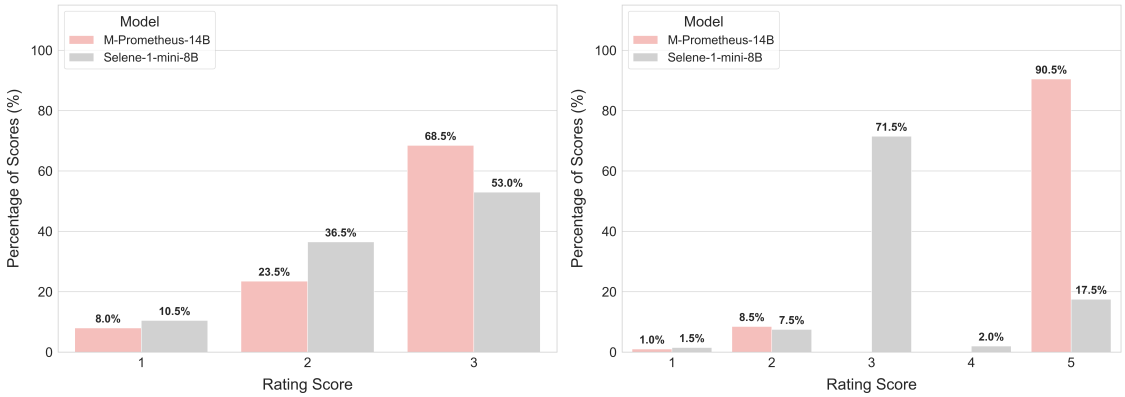

Figure 6: LLM-as-judge on the annotation set. Left: 'Narrativity' distribution. Right: 'Overall' distribution.

6 Discussion

In this section we discuss the performance of our models, the practical implications of post-training for ASG, interpretability derived from this experiment, and future implications of utilising this pipeline for narrative theory-informed ASG.

The results of our R_N-rewarded training suggest that abstract and theoretical narrative structure can be learned through reinforcement learning, and is beneficial in producing logical, rational, structurally complete and linguistically different stories. Comparatively, a qualitative examination of the SFT models' outputs shows that they often generate edited endings that are phrased identically to the original endings. This suggests that it is harder for SFT to achieve syntactically and creatively diverse storytelling. Intuitively, it would 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. d-RLAIF is a more intuitive choice for linguistically driven ASG since humans naturally learn storytelling as a communicative practice through reinforcement learning (Hineline, 2018).

Since both of our d-RLAIF and SFT training processes use less than a single epoch from the dataset to reach convergence, it suggests that massive datasets are not be necessary for LLM post-training. Efforts to curate datasets for ASG tasks may instead focus on comparatively smaller and query-only datasets compared to other tasks.

We also show that a larger model is not always needed to teach a smaller model. Our experiments show that ASG d-RLAIF can be carried out using an 8B student and an 8B teacher–it does not always require pairing a larger teacher model with a smaller student model to achieve noticeable improvements.

The crucial role of LLM-as-judge in d-RLAIF suggests that the path of Automatic Story Generation (ASG) through LLM reinforcement learning will be influenced by findings from Automatic Story Understanding (ASU). Using d-RLAIF, methods of effective ASU can offer great assistance in providing the reward signals needed for the training of future ASG models. We believe it to be beneficial to have unified knowledge and methods between generation and understanding. Additionally, we show that for the evaluation of effective ASU models, it is important to curate balanced datasets of high and low quality stories, which is a task that invites collaboration from ASG.

Our findings support claims that the most accurate LLM-as-judge (Selene-1-mini) is not necessarily the best teacher for reinforcement learning (Razin et al., 2025). Looking at the distribution of scores given to the annotation set in figure 6, we see that Selene-1-mini is a harsher judge than M-Prometheus. While harshness may have prevented models trained by Selene-1-mini's R_O from reaching the goal reward plateaus or improve performance as opposed to those trained by the more lenient M-Prometheus, the same harshness may have contributed to being a better teacher of narrativity. This shows we need additional measures of LLM-as-judge beyond accuracy and alignment.

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 learn-

ing. 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.

The application of narrative theory has been crucial for our experiment. At the same time, while we had initially hoped that Todorov's narrative theory would provide a foundation for usable story metrics that could be shared across tasks for evaluation and training, our annotation process only reach fair/moderate agreement. It indicates that the subjectivity of storytelling remains a quality that is perhaps incompatible with common metrics/benchmarks; it also suggests that Todorov's structural theory alone is not sufficient at modelling all of narrative's qualities. e.g. The 'Logical' and 'Rational' criteria also drawing on psychological and philosophical considerations. This echoes the viewpoints of Postclassical Narratology, which emphasise the contextual and cognitive perspectives of storytelling as a practice beyond structuralist and formalist traditions (Meister, 2011).

Nevertheless, the significant performance improvement and variance in training times for different instruction-tuned models when using narrativity as the reward signal R_N during d-RLAIF (compared to SFT or d-RLAIF with R_O) suggests that narrativity may be a non-arbitrary quality that is exhibited differently in models' behaviours and embeddings. This phenomenon is worth investigating by computer scientists and narratologists alike. In addition to being an insight into model behaviour and interpretability, it is worthwhile to consider an LLM itself as a narrative artefact–how a computational model trained on narrative texts exhibit narrative behaviour.

7 Related Work

Qin et al. (2019)'s TimeTravel dataset and its 'counterfactual story retelling' task have been used by a number of past works, experimenting with: unsupervised learning (Qin et al., 2020; Chen et al., 2022), Program-of-Thought prompting (Liu et al., 2023) and ConceptNet (Ashwani et al., 2024). All these past works position storytelling as a task that tests the ability of a large language model (LLM) to reason and understand causality. In this paper we choose to distance away from viewing the task as a benchmark of causal inference and rather position the task as primarily a narrative task that requires an approach informed by literary narrative theories.

d-RLAIF (Lee et al., 2024) has been shown to perform on par with RLHF at tasks such as summarisation, helpful dialogue generation, and harmless dialogue generation. While peer-reviewed papers of d-RLAIF for narrative generation has been lacking, Wei et al. (2025)'s paper showed that it can perform well at improving creative-writing for generating Chinese greetings, and their findings gave us the confidence to proceed with our experiments.

8 Conclusion and Future Work

Our paper is a novel exploration at the intersection of LLMs and narrative studies. We applied Todorov (1971)'s Theory of Narrative Theory of Equilibrium for Qin et al. (2019)'s counterfactual story retelling task. We used the theory to define evaluation principles and criteria which appear as reward signals. We tested the alignment between humans and LLMs-as-judge using a dataset (n=200) of human and AI-written stories and find fair to moderate agreement. We employed d-RLAIF optimised by GRPO and LoRA using the suboptimally accurate but efficient Selene-1-mini-8B and M-Prometheus-14B LLM-as-judge models to generate various reward signals. After training 3 open-weight LLMs (Llama-3.1, Qwen-3 & Olmo-3) for a comparative study, we find that d-RLAIF generally performed better than SFT and instruction-tuned models, while training using the narrativity reward score R_N produced the best-performing models–as evaluated using Gemini-3-Flash prompted with our Todorovian principles.

Our results show that reinforcement learning–namely, d-RLAIF–is a viable post-training approach for training LLMs in ASG. A key part of our methodology is the principle-guided LLM-as-judge. While our results show that d-RLAIF can achieve improvements even with a sub-optimal LLM-as-judge, it is still the centrepiece of this promising methodology. We believe that the interactive practice of storytelling is inseparable from story understanding. Our paper sets the promise of linguistic theories as the basis of evaluation and reward models for future research in ASG.

Limitations

We do not compare our models with those developed by others (Li et al., 2023; Liu et al., 2023; Ashwani et al., 2024) using the TimeTravel dataset. This is primarily because none of them are reasonable baselines for our experiments, particularly in the context of a theory-informed ASG. We only focused on training 7/8B models, the effectiveness of our method for smaller/larger LLMs is mostly untested. Also, we did not compare the final accuracy of Gemini-3-Flash's evaluations with human judgements. 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.

Ethics Statement

We use benchmark datasets, 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 know 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.

Storytelling is a highly effective communication tool that can be used for influence and manipulation. Development of ASG methods with the goal of matching human storytelling qualities raise potential malicious or unintended harm.

Acknowledgments

Masked for blind review.

References

H. Porter Abbott. 2014. Narrativity. In Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert, editors, *The Living Handbook of Narratology*. University of Hamburg, Hamburg.

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*.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

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.

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.

Joseph Campbell. 2008. *The hero with a thousand faces*, volume 17. New World Library.

Claudia Carroll. 2024. Towards an ai narratology: the possibilities of llm classification for the quantification of abstract narrative concepts in literary studies. In *The Routledge Handbook of AI and Literature*, pages 288–298. Routledge.

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.

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.

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.

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.

Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. Scenecraft: Automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pages 86–96.

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.

Dandan Li, Ziyu Guo, Qing Liu, Li Jin, Zequn Zhang, Kaiwen Wei, and Feng Li. 2023. Click: Integrating causal inference and commonsense knowledge incorporation for counterfactual story generation. *Electronics*, 12(19):4173.

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.

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.

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.

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*.

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*.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

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

José Pedro Vieira Sousa, Pedro Campos, and Paulo Bala. 2024. College tales: Pilot study on large language models generated narratives for mental health literacy. In *Proceedings of the 27th International Academic Mindtrek Conference*, pages 270–275.

Yuxin Wang, Jieru Lin, Zhiwei Yu, Wei Hu, and Börje F. Karlsson. 2023. Open-world story generation with structured knowledge enhancement: A comprehensive survey. *Neurocomput.*, 559(C).

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, and was allocated $50 in credits.

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. Unless otherwise specified, 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 set as follows:

- 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 are:

- Rank $r = 64$
- Alpha $\alpha = 128$
- Target modules: `q_proj, k_proj, v_proj, o_proj`
- Dropout $= 0.05$
- Bias mode: `none`

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 Annotation

While we hypothesised that there would be a correlation between narrativity and desirable qualities of 'Logical' and 'Rational', we only found weak correlation. (Figure 7)

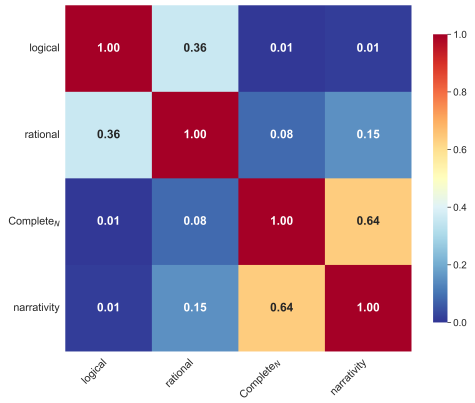

Figure 7: Correlation between each criteria in the human annotations.

F Training

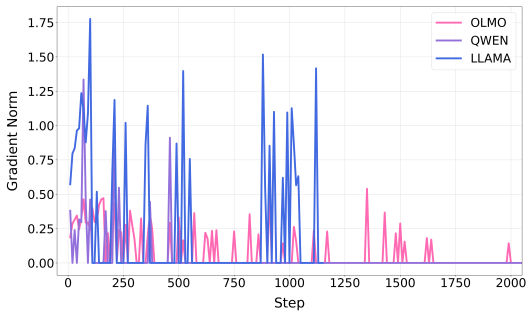

Figure 8: Policy model gradient norm during d-RLAIF using the reward signal R_O generated by Selene-1-mini.