Align to Structure: Aligning Large Language Models with Structural Information

Zae Myung Kim[1]*, **Anand Ramachandran**[2], **Farideh Tavazoee**[2], **Joo-Kyung Kim**[2], **Oleg Rokhlenko**[2], **Dongyeop Kang**[1]

[1]University of Minnesota Twin Cities
[2]Amazon
{kim01756, dongyeop}@umn.edu, {anramac, fayt, jookyk, olegro}@amazon.com

Abstract

Generating long, coherent text remains a challenge for large language models (LLMs), as they lack hierarchical planning and structured organization in discourse generation. We introduce *Structural Alignment*, a novel method that aligns LLMs with human-like discourse structures to enhance long-form text generation. By integrating linguistically grounded discourse frameworks into reinforcement learning, our approach guides models to produce coherent and well-organized outputs. We employ a dense reward scheme within a Proximal Policy Optimization framework, assigning fine-grained, token-level rewards based on the discourse distinctiveness relative to human writing. Two complementary reward models are evaluated: the first improves readability by scoring **surface-level textual features** to provide explicit structuring, while the second reinforces deeper coherence and rhetorical sophistication by analyzing **global discourse patterns** through hierarchical discourse motifs, outperforming both standard and RLHF-enhanced models in tasks such as essay generation and long-document summarization.

Code — https://github.com/minnesotanlp/StructAlign

Datasets — https://huggingface.co/datasets/zaemyung/writing_prompts_collection

Extended version — https://arxiv.org/abs/2504.03622

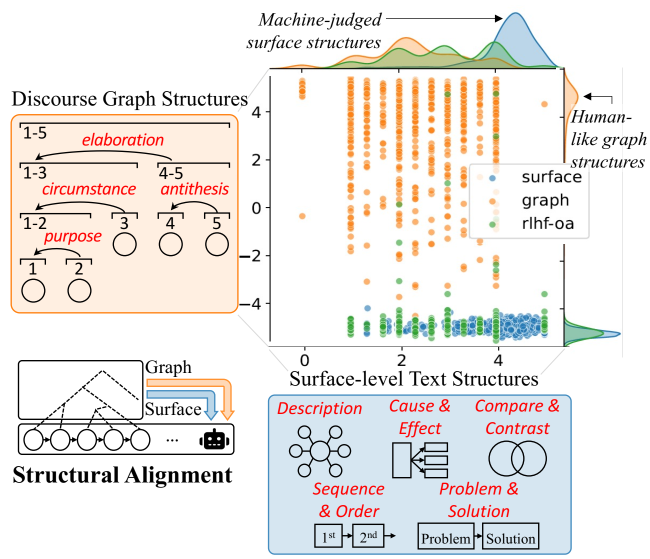

Figure 1: Structural alignment extends conventional LLM alignment by incorporating well-defined textual and discourse-level structures, guiding models toward more coherent and human-like text generation.

1 Introduction

The rapid advances in large language models (LLMs) have sparked significant interest in aligning these models with desired behaviors and outputs (Casper et al. 2023; Kaufmann et al. 2024). While alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017; Ouyang et al. 2022), have effectively improved the "helpfulness" and "harmlessness" of generated text, most such work focuses on *what humans prefer*, rather than *how humans structure* an eloquent and coherent discourse.

Human writers naturally employ established text structures such as problem-solution, cause-effect, or descriptive schemes (Meyer 1975), to maintain local coherence and logical flow. Theoretical frameworks (Mann and Thompson 1987) further illuminate how arguments, explanations, and narratives link together to ensure global consistency. Incorporating these structural principles into LLM training could enhance text generation, enabling models to produce more human-like, coherent discourse and better align with human writing conventions.

To this end, we introduce **structural alignment**, a novel alignment objective that extends beyond aligning models with human values to incorporating linguistically-driven document structures. We explore two avenues of research. First, we align LLMs with established *surface-level text structures* (cf. Figure 1) by leveraging an external LLM to guide or evaluate generation within a Proximal Policy Optimization (PPO) framework (Schulman et al. 2017) under a Reinforcement Learning from AI Feedback (RLAIF) paradigm (Bai et al. 2022).

*Work done during internship at Amazon.

Second, we parse documents using the RST discourse framework to construct *hierarchical discourse trees* and derive discourse motifs–characteristic substructures that highlight patterns distinguishing human-authored from LLM-generated texts. By integrating an authorship classifier that leverages these motifs to discriminate between human and AI texts, we obtain a (soft) binary label serving as a reward in the PPO framework, effectively guiding the generation process toward more structurally coherent and human-like discourse.

A key motivation for structural alignment lies in **long-form text generation**, where ensuring logical progression and thematic consistency poses major challenges. Although existing alignment efforts optimize user preferences, they often overlook the "rhetorical scaffolding" essential for a comprehensive and logically consistent composition (Mann and Taboada 2006). In contrast, structural alignment focuses on organizing content according to recognized discourse relations, thereby facilitating continuity and depth over extended passages. This emphasis proves particularly beneficial for longer texts—such as multi-paragraph essays or creative narratives—where abrupt topic shifts, undergeneration, or superficial cohesion can profoundly affect the overall reading experience (Mirowski et al. 2023). To enhance stability in long-form generation, we introduce a dense reward scheme for PPO that allocates rewards to words based on their relative distinctiveness in human-written and LLM-generated text. Our contributions are as follows:

1. Introduce a novel alignment framework by systematically incorporating linguistically grounded structures into the alignment process, enabling LLMs to generate more well-structured, logically consistent text.
2. Propose a dense reward scheme based on linguistic structures to enhance the stability of RLHF training, particularly for long-form generation (more than 1,000 tokens).
3. Both quantitative and qualitative evaluations show that structurally aligned LLMs generate discourse structures that more closely mirror human writing, leading to essays with improved clarity, coherence, and overall quality.

2 Related Work

2.1 Aligning Large Language Models

Various alignment frameworks have emerged recently, such as reinforcement learning from human feedback (RLHF) (Christiano et al. 2017; Ouyang et al. 2022) or from AI Feedback (RLAIF) (Bai et al. 2022; Ankner et al. 2024; Ye et al. 2025), using different optimization techniques such as Direct Preference Optimization (DPO) (Rafailov et al. 2023), Group Relative Policy Optimization (GRPO) (Shao et al. 2024), among others. Inspired by RLAIF that leverages LLM judges as reward models (RMs) (Ankner et al. 2024; Ye et al. 2025), we adopt a similar approach by utilizing an off-the-shelf LLM as the RM, eliminating the need for additional fine-tuning. Although these alignment techniques differ in their underlying mechanisms, their primary objective remains very similar: shaping model outputs to align with user preferences, social norms, and ethical constraints, ensuring they are perceived as appropriate and acceptable by humans. In contrast, our work proposes an alternative alignment objective—training LLMs to generate text with well-structured, human-like discourse. Our concept of structural alignment defines an alignment goal rather than prescribing a specific training technique, allowing for flexible implementation across different learning paradigms.

2.2 Challenges in Long-Form Text Generation

Recent advances in memory-efficient architectures (Hu et al. 2022; Dao 2024) and extended context handling (Press, Smith, and Lewis 2022; Su et al. 2024) have improved LLMs' ability to process long-form inputs. However, producing high-quality long-form text remains challenging (Deng, Kuleshov, and Rush 2022; Mirowski et al. 2023), due to the scarcity of large-scale human-annotated datasets (Köksal et al. 2024), the inherent subjectivity of extended text evaluation (Tan et al. 2024; Wu et al. 2025), and the complexities of RL-based alignment (Greenberg and Mannor 2021).

To mitigate the dataset gap, we leverage linguistically grounded discourse structures and LLM-based evaluation, building on the premise that verifying the correctness of long-form text is easier than generating it (Zhang et al. 2024). Prior work has addressed RL training instability by employing dense reward shaping (Cao et al. 2024; Chan et al. 2024). In a similar vein, we introduce a discourse-aware dense reward scheme (§3.2) that assigns rewards based on the structural alignment of generated text with human-like discourse patterns (Kim et al. 2024), precomputed from a large corpus.

2.3 Structural Organization in Texts

Text structures—such as problem-solution, cause-effect, comparison-contrast, and description—are foundational in shaping coherent, reader-friendly discourse (Meyer 1975). They guide the flow of information, highlight relationships between ideas, and ensure logical progression. Meanwhile, hierarchical frameworks like Rhetorical Structure Theory (RST) (Mann and Thompson 1987) and Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher 2007) add deeper layers of structure by explicitly defining the roles and connections among discourse segments.

In this work, we adopt RST (Mann and Thompson 1987) to define a hierarchical discourse tree, wherein Elementary Discourse Units (EDUs) represent basic phrases, and higher-level nodes merge these units into increasingly complex structures. Each connection (edge) carries a discourse label, such as elaboration, and contrast, revealing how ideas fit together. Crucially, these labeled nodes indicate the text's main ideas (nuclei) and supporting details (satellites), illustrating the text's overall rhetorical organization. Such higher-level structures function as meta-level signals that convey speaker intent, pragmatic cues, and rhetorical strategies (Haugh and Jaszczolt 2012; Degen 2023). By aligning LLMs with these discourse structures, we enable them to maintain thematic continuity, manage contextual shifts, and generate text that is more consistent, context-sensitive, and human-like.

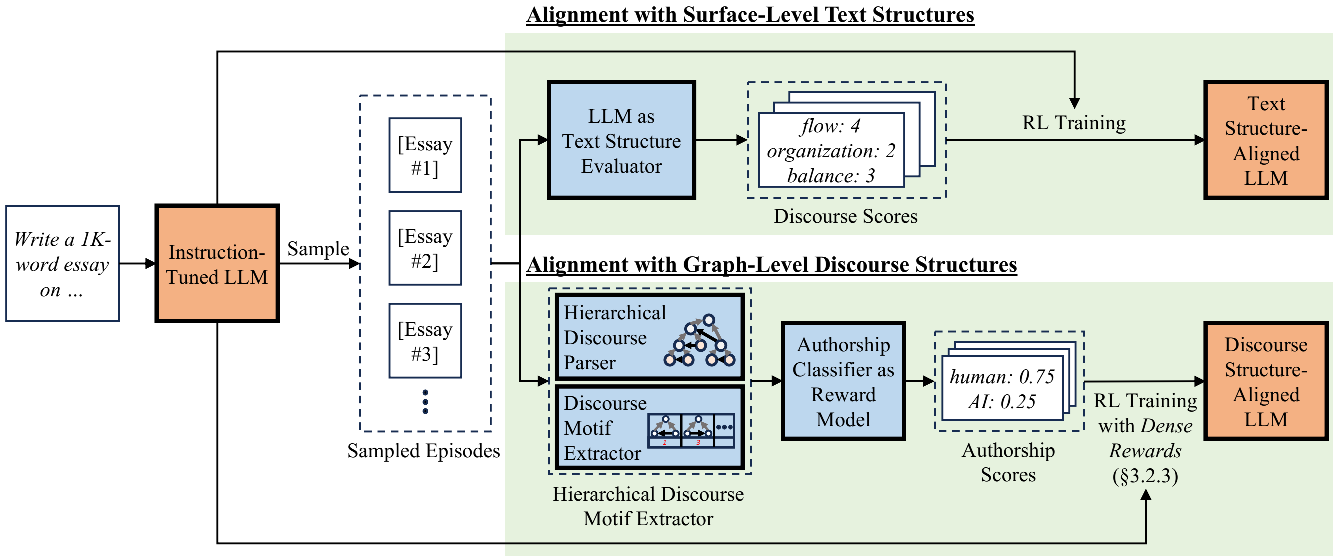

Figure 2: An illustration of our approach to structural alignment. We explore two types of reward modeling: (1) scoring surface-level text structures, and (2) scoring human-like discourse via authorship classifier. Components highlighted in blue remain frozen during RL training, while those in orange are trained.

3 Structural Alignment

The goal of structural alignment is to enable LLMs generate more logically coherent and human-like text. We emphasize that we are introducing a new alignment "task," not a new algorithm, as methods like DPO or GRPO can be readily integrated into our framework. In this study, we achieve this objective through the RLAIF approach within the PPO framework (§3.1, §3.2), exploring two research avenues: (1) employing an evaluator LLM to score surface-level text structures (§3.2), and (2) evaluating human-like discourse via an authorship classifier grounded in hierarchical discourse motifs (§3.2). Additionally, we propose a reward-shaping scheme that leverages dense discourse motifs to enhance training stability (§3.2).

3.1 Preliminaries for RLAIF

Reinforcement Learning from AI Feedback (RLAIF) extends the paradigm of using human feedback by employing AI systems to provide evaluative signals for training a policy model (Bai et al. 2022). Under the Proximal Policy Optimization (PPO) framework (Schulman et al. 2017), RLAIF optimizes a policy $\pi_\theta(a_t|s_t)$, parameterized by θ, to produce actions a_t (such as generated stories) given states s_t (such as writing prompts). The AI-generated feedback is translated into scalar reward signals r_t that quantify the quality of the actions.

The optimization objective in PPO is designed to maximize expected rewards while ensuring that policy updates are stable and do not deviate excessively from the previous policy. This is achieved by maximizing a clipped surrogate objective function:

$$L^{\text{PPO}}(\theta) = \mathbb{E}_t \Big[\min\Big(r_t(\theta)\hat{A}_t, \text{clip}\ (r_t(\theta), 1-\epsilon, 1+\epsilon)\ \hat{A}_t\Big)\Big] \tag{1}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ represents the probability ratio between the new and old policies, and ϵ is a hyperparameter that limits the extent of policy updates. The advantage function $\hat{A}_t$ estimates the relative benefit of an action compared to the average, and can be computed as:

$$\hat{A}_t = r_t + \gamma V_{\theta_{\text{old}}}(s_{t+1}) - V_{\theta_{\text{old}}}(s_t) \tag{2}$$

with $V_{\theta_{\text{old}}}(s_t)$ being the value function of the old policy and γ the discount factor. By iteratively performing gradient descent on $L^{\text{PPO}}(\theta)$, the policy parameters θ are updated to maximize the expected rewards:

$$\theta_{\text{new}} = \theta_{\text{old}} + \alpha \nabla_\theta L^{\text{PPO}}(\theta) \tag{3}$$

where α is the learning rate. This iterative process allows the policy to learn from AI feedback, progressively improving its performance while maintaining stability through the clipping mechanism in the PPO objective. Consequently, the policy becomes adept at generating higher-quality outputs as evaluated by AI feedback, reducing the reliance on human annotations and enabling scalable training.

3.2 RLAIF for Aligning to Structure

Directly applying RLAIF to long-form text generation poses unique challenges, primarily due to the difficulty of *measuring* coherence and organization at scale. Longer outputs require complex thematic progression and consistent rhetorical flow, making it hard to distill quality into a single, easily computed reward. Even powerful off-the-shelf LLMs can struggle to provide reliable "one-shot" scores for coherence when dealing with multi-paragraph or multi-page content.

Consequently, we adopt two distinct yet complementary scoring approaches, as illustrated in Figure 2. The first, introduced in Section 3.2, uses an evaluator LLM as a Text Structure Evaluator that assigns a 1–5 rating based on surface-level text structures—though an imperfect proxy, it still captures key signals such as logical transitions and topical balance. The second, described in Section 3.2, offers a more explicit framework by segmenting extended text and extracting theory-grounded "discourse motifs" via RST, which are then fed into an Authorship Classifier for binary classification.

Surface-Level Text Scoring Our first approach to reward modeling employs a powerful off-the-shelf LLM to provide three surface-level text scores grounded in pragmatics and discourse analysis: Logical Flow and Structure; Hierarchical Organization; and Balance and Emphasis. These scores guide the LLM to generate text that is not only grammatically correct but also pragmatically effective in conveying its intended message within the discourse context.

- *Logical Flow and Structure* assesses whether ideas progress logically and the overall organization is coherent, ensuring that the discourse is easy to follow.
- *Hierarchical Organization* evaluates how effectively content is structured, transitioning from general concepts to specific details, with each section building upon the previous one to support the main argument.
- *Balance and Emphasis* examines whether key ideas are appropriately highlighted and whether different points receive balanced coverage, aligning with pragmatic principles of relevance and informativeness.

Each score is assigned an integer value from 0 (lowest quality) to 5 (highest quality), and the final reward is computed as their average. The full prompt instruction for this scoring mechanism is provided in supplementary materials.

Graph-Level Discourse Scoring Our approach to aligning LLMs with human-like text structure builds on Kim et al. (2024), who developed a methodology for distinguishing human-written from machine-generated text using hierarchical discourse analysis. Their approach, illustrated in Figure 3 employs the RST framework to parse texts into discourse trees, converts these trees into recursive hypergraphs, and extracts discourse motifs—recurring structural patterns—as distinguishing features. These motif distributions serve as inputs for classifiers that effectively detect machine-generated text, even in out-of-distribution and paraphrased cases.

Building on this framework, we extend the RST discourse parser (Liu, Shi, and Chen 2021) to accommodate longer contexts by segmenting texts at the paragraph level, with each chunk containing approximately 400–512 tokens—a technical limitation that future work can address. Discourse parsing is applied to each segment to compute distributions of distinctive discourse motifs, capturing global discourse patterns beyond surface-level text. To prevent over-representation of motifs, segments are non-overlapping. These aggregated motif distributions are fed into a Longformer-based (Beltagy, Peters, and Cohan 2020) authorship classifier (Li et al. 2024), where they are concatenated with the model's [CLS] embedding to integrate textual content with explicit structural signals and produce binary classification scores.

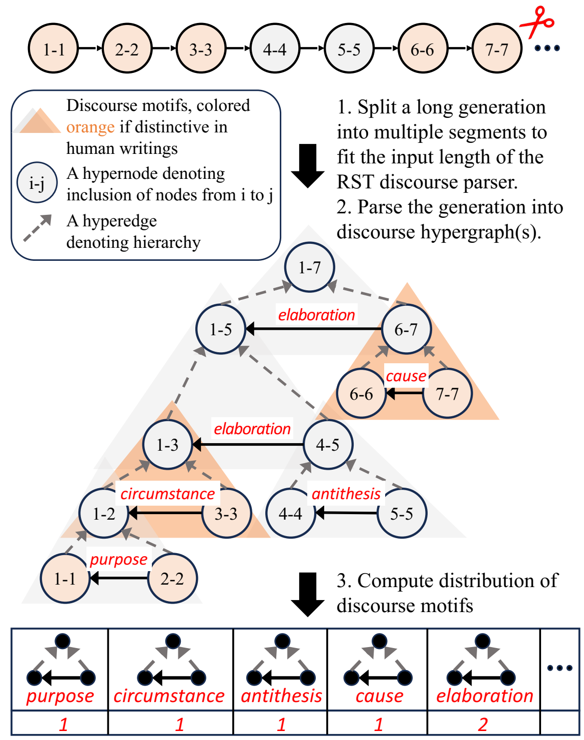

Figure 3: The process of extracting discourse motif vectors from generated text, as illustrated in Kim et al. (2024). We extend the original approach by introducing a segmentation step for texts exceeding the input length of the RST parser. Motif vectors are then aggregated across multiple segments.

Aside from segment-level adaptations, minimal architectural changes are made. The classifier is trained on existing datasets supplemented by curated human-written essays described in Section 4.1. By capturing how human-authored texts typically display richer and more varied structural patterns, while machine-generated texts often rely on uniform, sequential prediction strategies, this method remains effective even against paraphrasing attacks that obscure surface cues. The classifier's binary output—differentiating human from machine-authored text—becomes the reward signal in our structural alignment task. This setup allows the model to harness hierarchical discourse structures, thereby retaining global coherence patterns and enhancing authorship classification for long-form text generation.

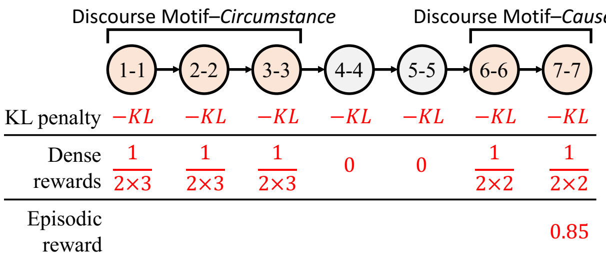

Figure 4: In addition to the episodic scalar reward, tokens contributing to human-distinctive discourse motifs receive a reward of $\frac{1}{2\cdot\text{num\_tokens}}$.

Reward Shaping with Discourse Motifs In the standard RLHF framework with PPO, the episodic reward is assigned only at the end of an episode, specifically, to the final token of the model's generation. This sparsity of rewards limits the training signal's granularity, providing no direct feedback on the quality of individual actions (i.e., token selections) throughout the sequence.

Consequently, as RL episode lengths increase, training stability declines due to the delayed credit assignment, higher exploration complexity, and compounding execution errors (Pignatelli et al. 2024). Reward variance also grows, making policy updates noisier and less effective, particularly in policy gradient methods. While techniques like curiosity-driven exploration (Pathak et al. 2017), imitation learning (Reddy, Dragan, and Levine 2019), or curriculum learning (Chen et al. 2021) can help, we instead employ reward shaping to provide more fine-grained signals—taking advantage of the inherently granular structure of language, in particular, discourse trees.

We identify specific tokens within each sequence that should receive non-zero token-level rewards, enabling more fine-grained guidance for the policy model. Specifically, in addition to this episodic scalar reward, we assign $\frac{1}{2\cdot\text{num\_tokens}}$ reward scores to tokens that contribute to a discourse motif that is more distinctively characteristic of human writing compared to those frequently observed in LLM-generated text (cf. Figure 4). These distinctive discourse motifs are identified across entire corpora by computing Motif Frequency-Inverse Document Frequency (MF-IDF) scores and selecting motifs that exceed a threshold of at least one standard deviation, following the approach of Kim et al. (2024). To preserve finer granularity, we compute discourse motifs over EDUs rather than full sentences; each EDU contains, on average, 20 subword tokens.

While the overall PPO training objective remains unchanged, the implementation involves adding these token-level rewards to the corresponding positions in the reward tensor for each sequence. This design ensures gradient updates capture the localized contribution of each token, improving credit assignment throughout generation. As a result, policy improvements typically occur more quickly and with greater stability than when relying on a single end-of-sequence reward (Chan et al. 2024).

3.3 Length-Penalty Normalization of Rewards

Noting that existing LLMs often fail to adhere to the desired response length specified in prompts (Wu et al. 2025), we adjust the original score by applying a penalty proportional to the shortfall in response length, but only if the response is shorter than the desired length.

The normalized score, S_n, based on the response length can be represented as:

$$S_\text{n} = S_\text{o} \times \left[1 - \alpha \times \max\left(0, \frac{L_\text{d} - L_\text{r}}{L_\text{d}}\right)\right] \quad (4)$$

where S_o is the original score; L_r is the length of the generated response; L_d is the desired length of the response; and α is the length penalty factor.

4 Experiments

As shown in Figure 2, we investigate two reward modeling approaches within the RLAIF framework using PPO: (1) surface-level text structure scores and (2) graph-level discourse structure scores. To support this, we construct a dataset of prompts by collecting and refining writing prompts for essay generation (§4.1) and validate the quality of the LLM-based Text Structure Evaluator (§4.2). Finally, we evaluate structure-aligned LLMs against baseline models trained with standard RLHF using an off-the-shelf general-purpose RM (§4.3).

Throughout our experiments, we use QWEN2-72B-INSTRUCT-AWQ as the Text Structure Evaluator and QWEN2-1.5B-INSTRUCT as the initial policy model for structural alignment (Yang et al. 2024). Due to resource constraints, we selected a 1.5B parameter model as it strikes a balance between being sufficiently powerful and small enough to enable efficient PPO training in data parallelism mode. At the time of our experiments, Qwen 2 models offered state-of-the-art performance in the LLM landscape. For hierarchical discourse analysis, we adopt the DMRST parser (Liu, Shi, and Chen 2021) as our Hierarchical Discourse Parser, while the Discourse Motif Extractor and Authorship Classifier are based on Kim et al. (2024). Details on implementation are provided in Appendix A.

4.1 Constructing an Essay Prompt Dataset

Our work focuses on formal essay writing, as its well-defined rhetorical conventions provide a strong foundation for evaluating structural alignment. To ensure diversity in formal essays, we collected prompts from various domains emphasizing clear logical structures and rhetorical coherence.

Specifically, our PPO training dataset is constructed by refining and integrating multiple existing datasets containing essay prompts, including those from an English proficiency exam, persuasion corpus (Crossley et al. 2023), and an argumentation source, namely, the Change My View (CMV) subreddit (Tan et al. 2016).

For CMV-derived samples, we employ an LLM[1] to generate essay instructions that require students (or LLMs)

[1]Qwen/Qwen2-72B-Instruct

to either support or oppose the original poster's opinion. Each CMV discussion yields two types of instructions: one prompting the student to argue in favor of the given claim and the other requiring them to argue against it. Given the nature of CMV discussions, many of these prompts may be inherently controversial.

To ensure an effective evaluation of model performance, 30% of the dataset is randomly sampled as the test set, except for CMV, where a predefined test split is already provided. The final dataset consists of 26,013 samples for training and 4,096 samples for testing.[2]

4.2 Verification of Text Structure Evaluator

To validate the quality of the text scores generated by our LLM-based RM, Text Structure Evaluator, we examined their correlation with human scores using the "Hewlett Foundation: Automated Essay Scoring" dataset (Hamner et al. 2012).[3] This dataset contains high school student essays along with scores assigned by expert human graders. For each essay, students responded to a writing prompt, and human annotators assessed their quality.

When analyzing the Pearson correlation between the discourse scores from the surface-level RM and human-assigned scores, we observe moderate positive correlations across key dimensions: coherence (0.47), organization (0.44), balance (0.39), purpose (0.37), and variability (0.39). These values indicate the discourse scores align reasonably well with human judgments in these areas. Additionally, using linear regression, we measured the mean squared error of the predicted scores to be 0.79 within the 0-4 range, translating to an average deviation of approximately 0.88 points from human scores. We determined the correlation results were strong enough to justify using the LLM-based discourse scoring as the RM and proceeded with PPO training.

4.3 Structural Alignment

We evaluate the effectiveness of structural alignment by assessing its impact on the quality of generated essays and the similarity of generated summaries to human-written long summaries.

We use "Base" to denote the base model, QWEN2-1.5B-INSTRUCT. Models SA_S and SA_G represent structurally aligned variants of the base model, where alignment is assessed using surface-level textual scores and graph-level discourse structures, respectively. Lastly, $RLHF_{OA}$ denotes the base model aligned with an off-the-shelf RM from OpenAssistant[4], which was trained to predict human-preferred responses given a question.

Exp #1: Assessment of Generated Essay Quality

Setup. We configure the models to generate sequences between 400 and 2K tokens. As training progresses within each epoch, the required generation length is gradually increased using a length penalty reward shaping mechanism (§3.3). For this experiment, we introduce two additional models by further training SA_S and SA_G using the alternate method in a two-step process, resulting in $SA_{S\rightarrow G}$ and $SA_{G\rightarrow S}$. We evaluated generation quality by randomly sampling 1,000 writing prompts from the test set and generating 700-token essays with our models. Because no gold-standard reference essays exist for these prompts, reference-based evaluation is not feasible. Additionally, assessing such lengthy text manually can be both time-consuming and prone to inconsistencies at longer contexts. In response, we adopted an LLM-as-judge approach, which is increasingly common in the field (Chiang and Lee 2023). Specifically, we employed GPT-4o-2024-08-06 to assess overall essay quality, presenting all candidate generations in a randomized order to reduce potential bias in ordering.

Figure 5: Pairwise evaluation results for six models. Base is the unaligned Qwen2-1.5B-Instruct model. SA_S and SA_G are Base aligned with surface-level and graph-level structural rewards, respectively. $RLHF_{OA}$ is Base aligned with an OpenAssistant RLHF reward model. $SA_{S\rightarrow G}$ and $SA_{G\rightarrow S}$ are the two-stage variants, applying the opposite structural reward in a second alignment step.

Results. Figure 5 presents evaluation results comparing generation quality across the various alignment settings. Overall, results show that models trained with structural alignment consistently outperform their baselines, with graph-level alignment providing the most substantial improvements. The two-step alignment process yields slightly higher performance; however, the improvement appears marginal relative to the additional computational cost. Since the two structural alignment methods exhibit low correlation (§4.4), we believe their effects are complementary. Developing more effective multi-reward alignment strategies could further improve performance, making this a promising avenue for future research. We provide two samples of generated essays with target lengths of 350 and 1K tokens in supplementary materials. Overall, the learned policy models effectively utilize explicit discourse connectives and frequently attempt to structure the text by dividing it into sections.

Exp #2: Long-Document Summarization

[2] Included in the supplementary material.

[3] We exclude verification of the graph-level discourse scoring approach, as it follows the validated method proposed by Kim et al. (2024).

[4] OpenAssistant/reward-model-deberta-v3-large-v2

Model	R-1	R-2	R-L
Base	53.21	20.13	50.39
Base+RLHF$_{OA}$	53.25	20.25	50.47
Base+SA$_S$	55.45	21.43	52.30
Base+SA$_G$	55.86	21.72	52.81

Table 1: Evaluations on long document summarization (GovReport). R-1 and R-2 measure unigram and bigram overlap, respectively. R-L leverages the longest common subsequence to assess sentence-level structural similarity.

Setup. We evaluate the aligned models on the downstream task of summarizing a long document. Specifically, we utilize the GovReport dataset (Huang et al. 2021) which is a large-scale dataset comprising around 19.5K US government reports, each paired with expert-written abstractive summaries. The dataset contains significantly longer documents (averaging 9.4K words) and summaries (553 words) compared to datasets like PubMed and arXiv. We randomly selected 5K reports from the U.S. Government Accountability Office (GAO), ensuring that each report selected for summarization does not exceed 14K tokens in length. To guide the LLM via prompting, we identify key aspects to summarize by analyzing the section titles of the gold summaries ("Highlights"), which typically focus on two main elements: **why** the report was conducted and **what** it found.

Results. Table 1 presents the ROUGE scores (Lin 2004) for the generated summaries, demonstrating that structural alignment significantly enhances long-document summarization. Among the approaches, graph-based alignment once again delivers the best overall performance.

4.4 Analyses

By incorporating structured reward signals, we observe significant improvements in discourse organization and human-like text structuring.

Effects of Surface-Level Structural Alignment We qualitatively observe that aligning LLMs with surface-level text structure scores results in (1) increased use of discourse connectives (e.g., *therefore*, *however*, *in contrast*, and *moreover*), (2) improved logical progression and argumentative flow, and (3) more frequent inclusion of section headings (e.g., *Introduction*, *Claim 1*, *Claim 2*, and *Conclusion*). A similar pattern is indirectly observed in Figure 7, where surface-level text structure alignment leads to a significant increase in discourse motifs that exemplify hierarchical structures, such as "Joint" and "Hyperedges."

Effects of Graph-Level Structural Alignment Unlike surface-level alignment, graph-level discourse structure alignment focuses on **global discourse organization** rather than local discourse markers (e.g., connectives). Figure 6 tracks the proportion of human-distinctive discourse motifs identified in each training batch, illustrating how these motifs evolve throughout the generation process. The plot shows that (i) their proportion steadily increases during training, and (ii) the trend plateaus after approximately 50

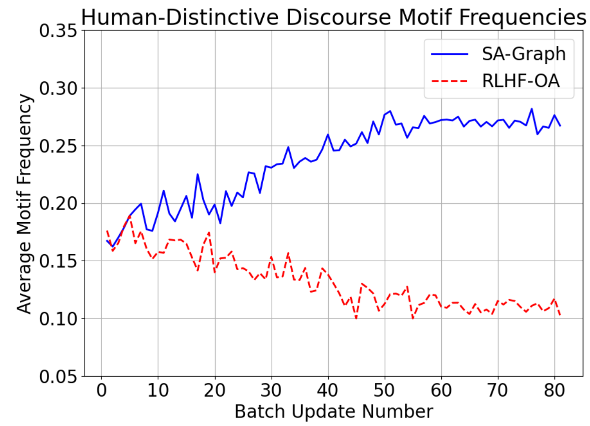

Figure 6: Comparison of human-distinctive discourse motif frequencies across batch updates. The blue line (our method) shows an increasing trend, while the red dashed line (standard PPO) fluctuates downward.

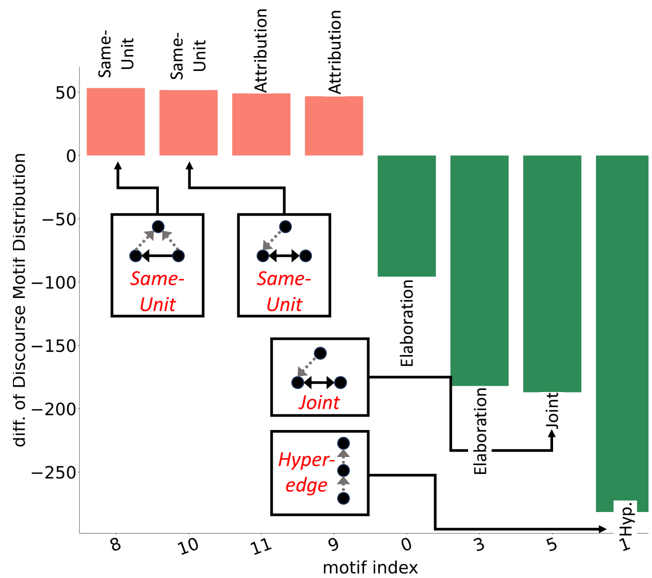

Figure 7: Difference in discourse motif distributions before (red) and after (green) surface-level text structure alignment. Notably, we observe a significant increase in the usage of hierarchical relations such as "Joint" and "Hyperedges (Hyp.)". Full version is in supplementary materials.

batch steps, suggesting discourse structuring reaches an optimal threshold. Notably, standard RLHF training without structural alignment shows a slow decline in such motif proportions, suggesting that conventional RMs do not reinforce human-like discourse organization as effectively.

Correlation between Surface and Graph Scores Figure 1 visualizes the relationship between surface and graph-level logits for essays generated by the two structurally aligned models and the standard RLHF-aligned model. While the structurally aligned models tend to achieve higher scores within their respective evaluation criteria, there appears to be little to no correlation between surface-level and graph-level scores. Additionally, we observe that essays generated by the standard RLHF-aligned model exhibit wide variance in surface scores and are predominantly classified as machine-generated, as indicated by negative graph logits.

In summary, our findings suggest that (1) Surface-level alignment enhances readability and clarity via explicit structuring techniques; (2) Graph-level alignment fosters deeper coherence and rhetorical sophistication by reinforcing global discourse structure; and (3) These two levels of alignment appear to be complementary, with the two-step process improving overall quality.

5 Conclusion

We present an RL approach that aligns LLMs with human-like discourse by utilizing two RMs—one for surface fluency, another for global discourse structure—resulting in clearer, more coherent long-form text that outperforms RLHF baselines. Future work includes scaling to larger models and diverse datasets for greater robustness, and incorporating other linguistic frameworks to refine reward signals and enhance performance.

A Experiment Details

RL Pipeline Setup Our Text Structure Evaluator is a 72B-parameter language model, which made an efficient serving framework essential. We used SGLang (Zheng et al. 2024) to load the QWEN2-72B-INSTRUCT-AWQ model, where AWQ stands for Activation-aware Weight Quantization (Lin et al. 2024). This quantized model can be loaded onto a single NVIDIA A100 80GB GPU. We ran eight separate evaluator instances, each on its own A100 GPU, and communicated with these evaluators through HTTP requests. SGLang handled these batch requests efficiently, allowing us to process multiple queries in parallel.

For training, we used the TRL library (von Werra et al. 2020) to perform Proximal Policy Optimization (PPO). Although TRL provides a strong base for reinforcement learning from human or AI feedback, we made extensive modifications to support our online reward mechanism, which involved querying the Text Structure Evaluator in real-time. The modified code, which enables discourse alignment via remote evaluation, will be made publicly available for further research and development.

The actual PPO-based reinforcement learning (RL) training took place on another server with eight NVIDIA A100 80GB GPUs. We set the per-device train batch size to 2, gradient accumulation steps to 4, local rollout forward batch size to 12, and KL coefficient to 0.03. Despite having large VRAM, we could not increase these batch sizes further due to the large output generation lengths (up to 2K tokens) which places substantial memory demands on the system.

Graph-Level Discourse Structure Scoring We closely followed the approach outlined by Kim et al. (2024) to implement our Graph-Level Discourse Structure Scoring pipeline. Using their curated list of distinctive discourse motifs, we constructed a Discourse Motif Extractor and trained a Longformer-based (Beltagy, Peters, and Cohan 2020) authorship classifier as described in the original paper. However, we extended the training set for the classification by adding 5K essay generations from QWEN2-1.5B-INSTRUCT as negative samples, in addition to the datasets originally used in Kim et al. (2024).

B Training Plots

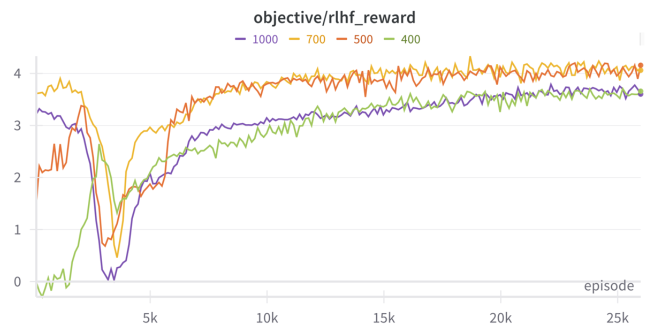

Figure 8: The curves represent the average reward scores provided by the RM over multiple steps or episodes, with each curve corresponding to models targeting different output lengths.

Figure 8 shows the mean RLHF reward for training via the surface-level text structure scoring. The mean reward is the average feedback score over multiple episodes, reflecting how well the model aligns with desired outcomes. We note that different curves represent separate training of the same baseline models with different target lengths of generation. We can see that the policy models learned to optimize its generation after around 7K episodes.

Further supplementary materials, including LLM prompts and generated samples, are available in our arXiv paper at https://arxiv.org/abs/2504.03622.

Acknowledgments

We are grateful to the researchers at Amazon and members of the intern cohort for their valuable feedback during weekly research discussions. We also thank members of Minnesota NLP for their insightful input during group meetings. ZMK is generously supported by the 3M Science and Technology Fellowship and the Doctoral Dissertation Fellowship at the University of Minnesota.

References

Ankner, Z.; Paul, M.; Cui, B.; Chang, J. D.; and Ammanabrolu, P. 2024. Critique-out-Loud Reward Models. arXiv:2408.11791.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150.

Cao, M.; Shu, L.; Yu, L.; Zhu, Y.; Wichers, N.; Liu, Y.; and Meng, L. 2024. Enhancing Reinforcement Learning with Dense Rewards from Language Model Critic. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9119–9138. Miami, Florida, USA: Association for Computational Linguistics.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T. T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P. J.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Biyik, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Chan, A. J.; Sun, H.; Holt, S.; and van der Schaar, M. 2024. Dense Reward for Free in Reinforcement Learning from Human Feedback. In *ICML*.

Chen, J.; Zhang, Y.; Xu, Y.; Ma, H.; Yang, H.; Song, J.; Wang, Y.; and Wu, Y. 2021. Variational Automatic Curriculum Learning for Sparse-Reward Cooperative Multi-Agent Problems. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 9681–9693. Curran Associates, Inc.

Chiang, C.-H.; and Lee, H.-y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631. Toronto, Canada: Association for Computational Linguistics.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Crossley, S. A.; Baffour, P.; Tian, Y.; Franklin, A.; Benner, M.; and Boser, U. 2023. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Available at SSRN 4795747*.

Dao, T. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *ICLR*.

Degen, J. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9(1): 519–540.

Deng, Y.; Kuleshov, V.; and Rush, A. 2022. Model Criticism for Long-Form Text Generation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11887–11912. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Greenberg, I.; and Mannor, S. 2021. Detecting rewards deterioration in episodic reinforcement learning. In *International Conference on Machine Learning*, 3842–3853. PMLR.

Hamner, B.; Morgan, J.; lynnvandev; Shermis, M.; and Ark, T. V. 2012. The Hewlett Foundation: Automated Essay Scoring. https://kaggle.com/competitions/asap-aes. Kaggle.

Haugh, M.; and Jaszczolt, K. M. 2012. Speaker intentions and intentionality. *The Cambridge handbook of pragmatics*, 87: 112.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Huang, L.; Cao, S.; Parulian, N.; Ji, H.; and Wang, L. 2021. Efficient Attentions for Long Document Summarization. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1419–1436. Online: Association for Computational Linguistics.

Kaufmann, T.; Weng, P.; Bengs, V.; and Hüllermeier, E. 2024. A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925.

Kim, Z. M.; Lee, K.; Zhu, P.; Raheja, V.; and Kang, D. 2024. Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5449–5474. Bangkok, Thailand: Association for Computational Linguistics.

Köksal, A.; Schick, T.; Korhonen, A.; and Schütze, H. 2024. LongForm: Effective Instruction Tuning with Reverse Instructions. In *EMNLP Findings*.

Lascarides, A.; and Asher, N. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, 87–124. Springer.

Li, Y.; Li, Q.; Cui, L.; Bi, W.; Wang, Z.; Wang, L.; Yang, L.; Shi, S.; and Zhang, Y. 2024. MAGE: Machine-generated Text Detection in the Wild. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 36–53. Bangkok, Thailand: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. In *MLSys*.

Liu, Z.; Shi, K.; and Chen, N. 2021. DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing. In Braud, C.; Hardmeier, C.; Li, J. J.; Louis, A.; Strube, M.; and Zeldes, A., eds., *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, 154–164. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics.

Mann, W. C.; and Taboada, M. 2006. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies*, 8(3): 423–459.

Mann, W. C.; and Thompson, S. A. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Meyer, B. 1975. *The Organization of Prose and Its Effects on Memory*. North-Holland studies in theoretical poetics. North-Holland Publishing Company. ISBN 9780720493009.

Mirowski, P.; Mathewson, K. W.; Pittman, J.; and Evans, R. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *ICML*.

Pignatelli, E.; Ferret, J.; Geist, M.; Mesnard, T.; van Hasselt, H.; and Toni, L. 2024. A Survey of Temporal Credit Assignment in Deep Reinforcement Learning. *Transactions on Machine Learning Research*. Survey Certification.

Press, O.; Smith, N. A.; and Lewis, M. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *ICLR*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 53728–53741. Curran Associates, Inc.

Reddy, S.; Dragan, A. D.; and Levine, S. 2019. SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards. In *ICLR*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Su, J.; Lu, Y.; Pan, S.; Murtadha, A.; Wen, B.; and Liu, Y. 2024. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 613–624. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.

Tan, H.; Guo, Z.; Shi, Z.; Xu, L.; Liu, Z.; Feng, Y.; Li, X.; Wang, Y.; Shang, L.; Liu, Q.; and Song, L. 2024. PROXYQA: An Alternative Framework for Evaluating Long-Form Text Generation with Large Language Models. In *ACL*.

von Werra, L.; Belkada, Y.; Tunstall, L.; Beeching, E.; Thrush, T.; Lambert, N.; Huang, S.; Rasul, K.; and Gallouédec, Q. 2020. TRL: Transformer Reinforcement Learning. https://github.com/huggingface/trl.

Wu, Y.; Hee, M. S.; Hu, Z.; and Lee, R. K.-W. 2025. LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs. In *ICLR*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Ye, Z.; Greenlee-Scott, F.; Bartolo, M.; Blunsom, P.; Campos, J. A.; and Gallé, M. 2025. Improving Reward Models with Synthetic Critiques. In *NAACL Findings*.

Zhang, L.; Hosseini, A.; Bansal, H.; Kazemi, M.; Kumar, A.; and Agarwal, R. 2024. Generative Verifiers: Reward Modeling as Next-Token Prediction. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Zheng, L.; Yin, L.; Xie, Z.; Sun, C.; Huang, J.; Yu, C. H.; Cao, S.; Kozyrakis, C.; Stoica, I.; Gonzalez, J. E.; Barrett, C.; and Sheng, Y. 2024. SGLang: Efficient Execution of Structured Language Model Programs. In *NeurIPS*.