CharMoral: A Character Morality Dataset for Morally Dynamic Character Analysis in Long-Form Narratives

Suyoung Bae[1], Gunhee Cho[1], Boyang Li[2], Yun-Gyung Cheong*[1]
[1] Sungkyunkwan University, South Korea
[2] Nanyang Technological University, Singapore
[1] {sybae01, skate4333, aimecca}@skku.edu, [2] boyang.li@ntu.edu.sg

Abstract

This paper introduces *CharMoral*, a dataset designed to analyze the moral evolution of characters in long-form narratives. *CharMoral*, built from 1,337 movie synopses, includes annotations for character actions, context, and morality labels. To automatically construct *CharMoral*, we propose a four-stage framework, utilizing Large Language Models, to automatically classify actions as moral or immoral based on context. Human evaluations and various experiments confirm the framework's effectiveness in moral reasoning tasks in multiple genres. Our code and the *CharMoral* dataset are publicly available at https://github.com/BaeSuyoung/CharMoral.

1 Introduction

Value alignment, the task of ensuring that language models and agents operate in accordance with human values, is a critical challenge in the development of ethical AI (Russell, 2019; Wolf et al., 2023). As psychological research highlights that narratives function as social simulations, allowing individuals to develop and refine social skills (Oatley, 2008), analyzing the ethical behavior of characters in stories offers valuable insights for deploying ethical language models. By examining both the moral actions of characters and the reactions of others within the narrative, agents can more effectively learn to align with human ethical standards, enhancing their capacity for contextually appropriate language understanding and generation.

Previous research on character analysis has mainly focused on three key areas: character identification (Chen and Choi, 2016; Brahman et al., 2021; Yu et al., 2022), social network analysis (Lee and Jung, 2019; Fischer, 2021; TARASEVICH et al., 2023), and the exploration of characters' personas or personalities (Bamman et al., 2013; Kim et al., 2019; Sang et al., 2022; Yu et al., 2023). However, the analysis of characters' moral stances remains an under-explored area of research, primarily due to several significant challenges. First, moral reasoning of actions is highly dependent on context (Pyatkin et al., 2022). For example, the action of *'throwing a bomb at the door'* might generally be considered unethical, but in the context of rescuing trapped individuals from aliens, it could be perceived as morally justified. Second, a character's morality is not static and may evolve throughout the narrative, influenced by unexpected events and interactions with other

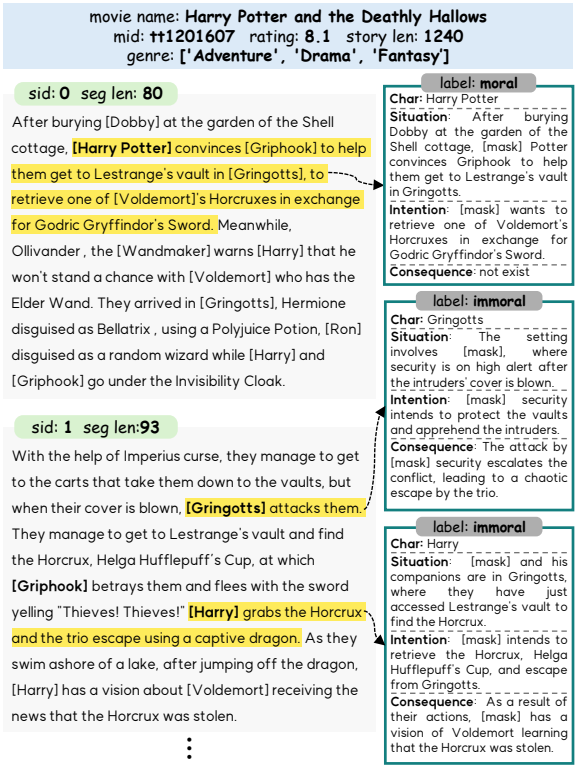


Figure 1: An example of *CharMoral* dataset with annotated actions' morality considering contexts. In each segment, we extract character "actions", and related "contexts". Then, we annotate the morality of the actions, considering each context.

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8809–8818
January 19–24, 2025.

Dataset	#Stories	#Words per story	#Characters	Genres	Morality Labels	#Annotations
Storium (Akoury et al., 2020)	5,743	19,278	25,955	✗	✗	-
TVSTORYGEN (Chen and Gimpel, 2021)	29,000	1,868	34,300	✓	✗	-
Story2Personality (Sang et al., 2022)	507	1,381	3,543	✗	✗	-
Moral stories (Emelin et al., 2021)	12,000	90	12,000	✗	✓	12,000
CharMoral	1,337	1,665	9,389	✓	✓	103,836

Table 1: Comparison of the *CharMoral* dataset with previous story datasets used for character analysis. **#Annotations** refers to the number of morally annotated actions in the entire **#Stories**.

characters who maintain their moral principles. Despite the importance of this dynamic, existing datasets do not facilitate the analysis of moral evolution in long-form narratives. Current datasets, such as Moral Stories (Emelin et al., 2021), focus on short narratives with isolated moral situations. Constructing a dataset that captures moral development over time is particularly challenging, as it requires human annotators, which can be both time-consuming and costly.

To overcome these limitations, we introduce a novel dataset called *CharMoral*, **Char**acter **Moral**ity dataset designed to analyze characters with dynamic moral stances in long-form narratives. This dataset enables the study of how morally dynamic characters influence readers' understanding and affect story engagement. *CharMoral* consists of a curated collection of 1,337 stories extracted from movie synopsis datasets. As outlined in Figure 1, each story is annotated with detailed information, including story segments, character names, actions within segments, contextual information surrounding these actions, and corresponding morality labels.

Our framework operates in four stages. First, we segment the story based on key events involving the main characters. Second, we utilize an LLM to extract characters' actions from each segment. Third, the LLM is employed again to extract contextual information–the situation, intention, and consequence–related to each action. Finally, we classify the morality of the character's actions within their contexts using a fine-tuned expert model, which is optimized to accurately predict moral judgments by factoring in contextual information. This approach ensures a comprehensive and scalable annotation process, facilitating the analysis of moral dynamics across a large corpus of long-form narratives.

We evaluate our dataset and annotation framework through a human evaluation to measure how closely it aligns with human moral judgments. Furthermore, we assess the quality and effectiveness of the dataset in training morality classifiers by performing a range of tasks, including zero-shot, few-shot, fine-tuning, and cross-domain evaluations.

The contributions of our work are as follows:

- We present a novel dataset for analyzing character morality in long-form narratives.
- We develop a framework leveraging LLMs to extract character actions and contextual information (situation, intention, consequence), enabling large-scale moral analysis.
- We developed a classification model, *MAD*, which outperforms GPT-3.5 and GPT-4 in moral reasoning tasks.
- We propose a new metric to track the moral dynamics of characters and explore its correlation with story engagement.

This paper is structured as follows. Section 2 reviews related work, Section 3 presents the dataset and framework, Section 4 details the experimental setup and morality classification, Section 5 analyzes morally dynamic characters, and Section 6 discusses key findings and future directions.

2 Related Work

2.1 Character-Centric Narrative Understanding

Character analysis is essential for narrative understanding, as characters are pivotal in driving the story forward (Bower and Morrow, 1990). The previous study of characters has largely concentrated on three main areas. First, character identification, which aims to identify and associate various mentions of characters throughout a text (Chen and Choi, 2016; Brahman et al., 2021; Yu et al., 2022). For example, Chen and Choi (2016) trained a model to identify characters mentioned in TV show series datasets and determine which character a pronoun refers to.

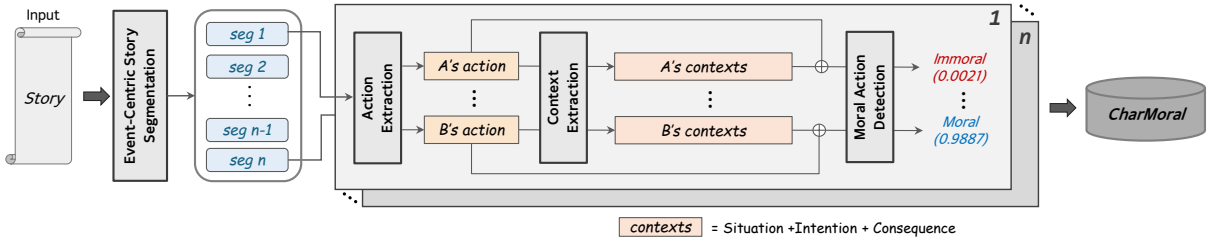


Figure 2: Overview of the *CharMoral* Dataset Construction Framework for annotating the moral stances of characters' actions in long-form narratives.

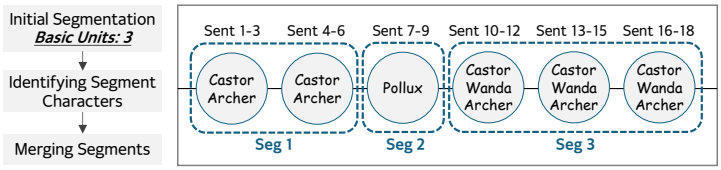


Figure 3: The illustrations of the Event-Centric Story Segmentation Process.

Second, social network analysis uses the occurrence and co-occurrence of characters within the story to construct social networks for analyzing the narrative (Lee and Jung, 2019; Fischer, 2021; TARASEVICH et al., 2023). Through social networks, it's possible to analyze the relationships and intimacy levels among characters, extracting information about key characters, close allies of the protagonist, and antagonists. Recent research has been actively exploring the use of character personas and personality analysis (Bamman et al., 2013; Kim et al., 2019; Sang et al., 2022; Yu et al., 2023). However, analyzing the morality of characters is under-explored.

2.2 Character Morality Datasets

Gert and Gert (2020) defines 'Morality' as "certain codes of conduct put forward by a society or a group (such as religion and nationality), or accepted by an individual for her own behavior." Prior research has focused on creating annotated datasets for machine ethics, such as Social Chemistry (Forbes et al., 2020), Scruples (Lourie et al., 2020), Moral Stories (Emelin et al., 2021), ETHICS (Hendrycks et al., 2021), and COMMONSENSE NORM BANK (Jiang et al., 2022). However, most datasets focus on morality classifications in a short text, typically one sentence long, and fail to capture the complexity of context needed for moral reasoning. In contrast, *CharMoral* offers longer, context-rich stories for deeper moral analysis.

Several datasets are available for character analysis in narrative (Akoury et al., 2020; Chen and Gimpel, 2021; Sang et al., 2022) as shown in Table 1. However, these datasets lack annotations related to the moral personalities of characters. Whereas Emelin et al. (2021) have attempted to annotate moral aspects of character actions in a story, their dataset consists of short stories with only single situations, making it difficult to observe changes in a character's morality throughout the narrative.

3 Building the *CharMoral* Dataset

The *CharMoral* dataset is reconstructed by executing the four steps illustrated in Figure 2: Event-Centric Story Segmentation, Action Extraction, Context Extraction, and Moral Action Detection. *CharMoral* is built based on the IMDB Spoiler Dataset [1], which consists of 1,572 movie synopses and meta-information from IMDB, providing detailed descriptions of characters' actions and their outcomes. For this study, we use 1,337 narratives, eliminating fewer than 10 words.

3.1 Event-Centric Story Segmentation

The first step segments each story based on key events involving the main characters. This method draws upon earlier work, as described in Kim et al. (2022), highlighting the significance of the main characters in these pivotal narrative moments.

As shown in Figure 3, the first step divides the story into segments of `s` sentences. [2] The

[1] https://www.kaggle.com/datasets/rmisra/imdb-spoiler-dataset

[2] In this paper, we manually set `s` to 3 through a pilot experiment.

main characters are then identified as the five most frequently mentioned names in the story. To extract these characters, we utilize Stanford CoreNLP (Manning et al., 2014) for co-reference resolution and Named Entity Recognition (NER) to identify character names. Consecutive segments that feature the same set of main characters are merged into a single segment. For example, if *'Ryan'* and *'Clark'* appear in two consecutive segments but are absent in the next, those segments are combined.

3.2 Action Extraction

Since the story lacks meta-information regarding the characters and their corresponding actions, we extract this by using GPT-4o (OpenAI, 2023). Each segment and the names of the characters mentioned are fed into the LLM using a prompt manually engineered for extracting characters' actions. If there is no action associated with the character, the output is *'no action'* and excluded from our dataset. The prompt is detailed in Table 19.

3.3 Context Extraction

The moral reasoning behind a character's actions is highly context-dependent (Pyatkin et al., 2022), and moral assessments can vary based on situations and intentions (Emelin et al., 2021). To assess morality accurately, we extract the action's situation, intention, and consequence, prompting LLM which was the same model used for character action extraction. If the corresponding context is missing, the output is *'not exist'*. The context extraction prompt is shown in Table 20.

3.4 Action Morality Detection

To accurately annotate whether the actions of the characters are morally justified based on their contextual information, we employ an expert model, the *Moral Action Detector (MAD)*. To provide reliable moral evaluations and ensure accurate predictions that account for relevant context, we fine-tuned the BERT-large model (Devlin et al., 2019) using the Moral Stories dataset (Emelin et al., 2021), following the same fine-tuning settings as described in Emelin et al. (2021)

Using the *MAD*, we predict whether the character involved in the action has performed a moral or immoral act by inputting the four extracted sentences (situation, intention, consequence, and action). In addition to the moral classification, we store the softmax logit scores for further analysis.

After Segmentation	
#Story	1,337
#Segment	30,616
Story Mean length	1,665
Segment Mean length	85
After Context Extraction	
#Character	9,389
#Annotations	103,836
#Action exists	103,836 (100%)
#Situation exists	103,813 (99.98%)
#Intention exists	92,627 (89.2%)
#Consequence exists	82,076 (79.04%)
#All exists	75,724 (79.93%)
Label Distribution	
#Moral	50,717
#Immoral	53,119

Table 2: The statistics of *CharMoral*

3.5 The Statistics of *CharMoral*

The statistics of *CharMoral* are presented in Table 2. **#Character** refers to the total number of characters appearing in our dataset, **#Annotations** refers to the number of morally annotated actions in the entire **#Stories**, and **#Action, #Situation, #Intention, and #Consequence exist** indicate the number of datasets where the respective component is present and it's proportion. Finally, **#All exists** refers to the number of datasets where the character's action and context have all been extracted. The *CharMoral* labels consist of moral and immoral. Details about the dataset are described in the Appendix A.

4 Evaluation

4.1 Human Evaluation

We evaluate the effectiveness of our framework by comparing its predictions with human assessments and recruiting nine proficient English-speaking graduate students for the task. Every annotator evaluates the moral dimensions of 50 randomly sampled sentences (character actions with context). We then use majority voting based on the annotation results of 9 annotators in each sentence. The Inter-Annotator Agreement, measured using Fleiss' kappa (Cohen, 1960), is 0.55, showing moderate agreement among evaluators (Appendix C for details).

Table 3 presents the human evaluation results, showing that our framework achieves approximately 64% agreement with human assessments when using only the action itself (A) to predict

Model	MAD								GPT 3.5	GPT 4
Setting	A	SA	IA	CA	SIA	SCA	ICA	**SICA**	(Segment) + A	
Acc	0.64	0.6	0.7	0.76	0.6	0.76	0.72	**0.92**	0.78	0.86

Table 3: The results of human assessments compare the moral action classification performance of *MAD* and LLMs. For *MAD*, we evaluate using different context settings as outlined in Table 4. For GPT 3.5 and GPT 4, morality is assessed by providing the entire text segment along with the character's action.

morality. This agreement significantly increases to 92% when full context (SICA) is considered.

We also compare the performance of two LLMs (GPT-3.5 and GPT-4) with human assessments to show that using the extracted context factors directly related to actions yields more accurate morality predictions than relying solely on the LLMs' internal knowledge. We provide LLMs with the text segment and action to evaluate the morality of the actions. *MAD* significantly improves moral action predictions, outperforming GPT-3.5 and GPT-4, which achieve 78% and 86% agreement, respectively. This highlights the superior performance of our model in making the predictions more closely aligned with human judgment.

4.2 Action Morality Classification

4.2.1 Experimental Setups

We conduct experiments to evaluate the effectiveness of our automatically building *CharMoral* dataset across various moral reasoning tasks. Additionally, we analyze the impact of varying contextual information on classification performance.

We define eight settings with varying levels of contextual information, as outlined in Table 4, including one setting with no context provided (A). In each case, the model classifies actions as moral or immoral based on the available context.

We use the BERT-large model (Devlin et al., 2019) trained using *CharMoral* to assess the morality classification performance. To verify the model's transferability, we also test cross-domain classification on two out-of-domain datasets, Moral Stories (Emelin et al., 2021) and Social Chemistry (Forbes et al., 2020). Additionally, we evaluate classification performance in zero-shot and few-shot settings to determine whether *CharMoral* provides high-quality demonstrations for accurate moral reasoning. We compare the performance of two LLMs, GPT-3.5 and GPT-4o [3], using the SICA input setting. In the few-shot setting, examples are sampled from *CharMoral* to create demonstrations, showing that our dataset effectively supports few-shot tasks by providing high-quality examples. Additional details can be found in Appendix D.

Setting	Context
A	-
SA	Situation
IA	Intention
CA	Consequence
SIA	Situation + Intention
SCA	Situation + Consequence
ICA	Intention + Consequence
SICA	Situation + Intention + Consequence

Table 4: Eight input settings in varying amounts of context information for moral action classification tasks. For all classification tasks, the model input is formatted as <CLS>context<SEP>action<SEP>.

4.2.2 Results

As shown in Table 5, the classification model trained on an in-domain setting demonstrates exceptional performance, achieving an accuracy of 94.7% when all contextual information is considered (SICA). The model also maintains strong performance in cross-domain settings, with an accuracy of 96.9% when tested on the Moral Stories dataset and 70.7% on Social Chemistry.

Table 6 presents the results of zero-shot and few-shot evaluations. The results demonstrate that the performance in both zero-shot and few-shot settings is reasonable, with GPT-3.5 achieving an accuracy of 61.7% in zero-shot, and GPT-4o achieving 66.7%. In the few-shot setting, the accuracy in the 3-shot setting improved to 62.7% for GPT-3.5 and 68.4% for GPT-4o. We also observe that as the number of demonstration examples increases, performance generally improves. This demonstrates that the *CharMoral* dataset is effective in creating high-quality demonstrations, leading to strong performance in few-shot tasks.

[3] https://openai.com/.

	In-domain		Cross-domain			
Train → Test	Ours → Ours		Ours → MS		Ours → SC	
Metric	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
A	0.692	0.712	0.689	0.716	0.823	0.778
SA	0.708	0.722	0.697	0.719	0.810	0.739
IA	0.709	0.718	0.684	0.701	**0.858**	0.823
CA	0.929	0.928	0.967	0.967	0.850	**0.829**
SIA	0.725	0.733	0.693	0.706	0.795	0.710
SCA	0.933	0.930	0.966	0.966	0.783	0.758
ICA	0.939	0.938	0.968	0.968	0.819	0.801
SICA	**0.947**	**0.947**	**0.969**	**0.969**	0.707	0.655

Table 5: The results of fine-tuning moral action classification tasks: The first row shows the *accuracy* and *F1* score when trained and tested on the *CharMoral*. The second and third rows show the cross-domain classification results, where the model is trained on the *CharMoral* and tested on the *Moral Stories* **(MS)** and *Social Chemistry* **(SC)**, respectively. The **bold** is the best score in each input setting.

Model	**GPT-3.5**		**GPT-4o**	
Metric	*Acc*	*F1*	*Acc*	*F1*
zero-shot	0.617	0.685	0.667	0.680
1-shot	0.619	0.688	0.667	0.690
3-shot	**0.627**	0.703	**0.684**	**0.693**
5-shot	0.622	**0.708**	0.683	0.692

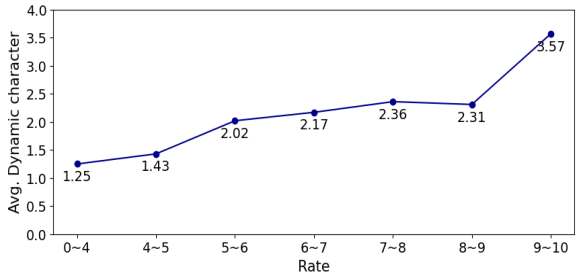
Table 6: The results of zero-shot and few-shot context-aware moral action classification task: The **bold** is the best score in each setting.

5 Morally Dynamic Character Analysis

5.1 Moral Dynamic Score

Through the *CharMoral*, we can observe how characters' morality changes throughout long-form narratives. By analyzing the moral dynamics of characters, we can enhance people's understanding of the story and provide information to automatically identify stories that are both instructive and engaging.

We define a **morally dynamic character** as a character whose morality shifts dynamically as events unfold in a long story and a **morally static character** as one who acts consistently according to a stable set of moral values. To observe the moral dynamics of characters within a story, we introduce a new score, which we refer to as the *moral dynamic score* of a character (c). The moral dynamic score is defined as follows:



Rate	0∼4	4∼5	5∼6	6∼7	7∼8	8∼9	9∼10
# Story	4	14	112	375	531	289	7
# Character	11	90	409	1,547	2,397	1,279	55
# Dynamic Character	5	20	226	812	1,253	668	25

Figure 4: The **above** line plot represents the average number of dynamic characters per story by rate and the **bottom** table represents related statistics by rate.

Moral dynamic score:

$$score_c = \frac{(\#pass_c)}{(\#segment_c) - 1}$$

Morally dynamic character:

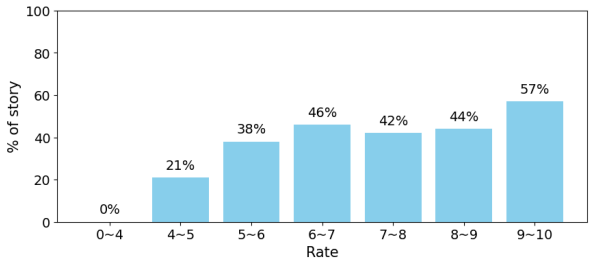
$$dynamics_c = \begin{cases} dynamic & \text{if } score_c \geq 0.5, \\ static & \text{otherwise} \end{cases}$$

CharMoral records morality prediction values for each character's actions as softmax logit scores between 0 to 1. To track changes in a character's morality over time, we calculate how often a character's logit score crosses the 0.5 midpoint (The number of times MAD's prediction switches from immoral to moral or from moral to immoral during segment progression.), referred to as *the number of passes* ($\#pass_c$). We then divide the number of passes by the total number of segments where the character c appears, minus one to calculate the rate of the character's morality change. The $score_c$ can range from 0 to 1. If the $score_c$ is 0.5 or higher, we classify the character c as a **morally dynamic character**, and otherwise, as a **morally static character**.

5.2 Analysis

Using the *moral dynamic score* defined in section 5.1, we draw three conclusions (**A1**, **A2**, and **A3**) from the analysis of story characters' moral dynamics and its correlation with story interest.

A1: A greater number of morally dynamic characters positively contributes to the story's overall interest. We investigate the relationship between the number of morally dynamic characters



Rate	0∼4	4∼5	5∼6	6∼7	7∼8	8∼9	9∼10
# Story	4	14	112	375	531	289	7
# Dynamic main character	0	3	42	171	221	128	4

Figure 5: The **above** box plot represents the percent of the story where there is a dynamic main character by rate and the **bottom** table represents the statistics by rate.

and a story's level of interest by comparing the average number of dynamic characters across different rating groups. The average number of dynamic characters per story for each rating group is depicted as a line graph in Figure 4. The graph shows that when the rating is between 0 and 4, the average number of dynamic characters is 1.25, while as the rating increases, this number gradually rises, reaching an average of 3.57 in stories with ratings between 9 and 10. These results suggest that a higher number of morally dynamic characters is linked to greater story engagement.

A2: The moral evolution of the main character plays a key role in increasing the story's overall interest. We examine whether the moral evolution of the main character (the most frequent character), rather than just the total number of morally dynamic characters, affects story engagement. Specifically, we look at the proportion of stories in each rating group where the main character is morally dynamic. Figure 5 presents this proportion as a bar graph, showing the ratio of stories with a morally dynamic main character to the total number of stories for each rating group. In stories with a rating between 0 and 4, none of the main characters are dynamic, whereas in stories with a rating between 9 and 10, 57% of the main characters are dynamic. These results suggest that stories with higher ratings tend to have morally dynamic main characters, highlighting the importance of the main character's moral evolution in making a story more engaging.

A3: Morally dynamic main characters have a significant impact on genre-specific story engagement and ratings. We examine the impact

Genre	% Dynamic main character	Avg. Rate
Musical	60.00	7.28
Western	60.00	7.96
War	56.52	8.15
Thriller	50.95	7.22
Action	50.90	6.87
Film-Noir	50.00	8.23
Fantasy	49.38	6.96
Adventure	47.67	7.03
Sci-Fi	43.89	6.99
Mystery	42.65	7.47
Crime	42.63	7.38
History	41.82	7.63
Drama	41.42	7.47
Family	40.71	6.79
Horror	40.20	6.74
Biography	40.00	7.71
Romance	39.06	7.10
Comedy	37.68	6.85
Animation	35.62	7.45
Sport	28.00	7.11
Music	21.74	7.04

Table 7: Proportion of morally dynamic main characters and average rating across story genres.

of morally dynamic main characters on genre-specific engagement and ratings. Table 7 shows the proportion of stories in each genre with morally dynamic main characters, alongside their average ratings.

The Western genre has the highest proportion of morally dynamic main characters (60%) and a strong average rating of 7.96. Genres like Musical, War, and Thriller also show high proportions of dynamic main characters, with Musical (60%) and War (56.52%) achieving average ratings of 7.28 and 8.15, respectively. This suggests that the moral evolution of the main character plays a key role in enhancing story interest and ratings. Similarly, Film-Noir, History, and Biography feature high proportions of morally dynamic characters, with average ratings between 7.63 and 8.23. These genres often explore complex moral dilemmas, where the protagonist's internal conflicts are central to the narrative. In contrast, genres like Sport (28%) and Music (21.74%) have lower proportions of dynamic main characters and correspondingly lower ratings, between 7.04 and 7.11.

In summary, genres with a higher proportion of morally dynamic main characters tend to receive higher ratings, suggesting that audiences are more engaged with stories that feature significant character growth or moral conflict. In genres where moral dynamics are less important, like Sport and

Action	Situation	Intention	Consequence	A	SICA	Human
[Scar] manages to knock Simba down and leaps at him	**[Scar]** manages to knock Simba down and leaps at him, creating a tense confrontation at Pride Rock.	**[Scar]** intends to defeat Simba and assert his dominance.	As a result of **[Scar]**'s action, he finds himself surrounded by the hyenas, who are now turning against him.	I	IM	IM
[Lupin] has Harry test himself out on a Boggart.	After Professor Snape discovers Harry out of bed, the map is confiscated by **[Lupin]** who meets them in a setting filled with tension and secrecy.	**[Lupin]** wants Harry to learn how to generate a Patronus by testing himself against a Boggart.	By having Harry confront the Boggart, **[Lupin]** aims to help him improve his magical skills and ultimately succeed in conjuring a Patronus.	IM	M	M
[Maurice] tells Malcolm Ellie and Alex to run.	**[Maurice]** tells Malcolm, Ellie, and Alex to run as the other apes become frightened and wild.	**[Maurice]** wants to ensure the safety of Malcolm, Ellie, and Alex by urging them to escape.	The action may lead to the group avoiding danger and potentially surviving the chaotic situation.	IM	M	M

Table 8: Comparison of context-aware moral action classification results: This table shows three examples of narrative segments, presenting the action and context (situation, intention, consequence) extracted through our framework, along with the predicted action morality and human evaluation results for two settings (`A` and `SICA`). If the prediction in each setting is moral, it is marked as **M**, and if it is immoral, it is marked as **IM**.

Music, the ratings tend to be lower.

5.3 A Case Study

Section 4.1 demonstrates the effectiveness of our dataset and framework in predicting a character's morality, confirming that context enhances judgment accuracy. Table 8 showcases three randomly selected examples of human evaluations, comparing our framework's annotation results for two settings: `A` (action only) and `SICA` (situation, intention, consequence, and action).

These examples show that annotations based solely on actions (`A`) are often inaccurate and misaligned with human judgments. In contrast, including full context (`SICA`) consistently improves prediction accuracy, demonstrating the significant role of context in moral assessment.

6 Conclusion

In this paper, we introduced *CharMoral*, a novel dataset designed to analyze the moral dynamics of characters in long-form narratives. Using a four-stage framework that leverages large language models, we annotated character actions and their morality, considering contextual information such as situation, intention, and consequence. Our experiments show that including context significantly improves the accuracy of moral predictions, aligning more closely with human judgments.

We further demonstrate that morally dynamic characters, particularly main characters, play a key role in increasing story engagement and ratings, particularly in genres where moral dilemmas are central, such as Western, War, and Film-Noir.

7 Limitation

Although our ultimate goal is to develop a method that objectively assesses the morality of character actions based on story context, moral concepts are often shaped by cultural, social, and personal beliefs. As a result, biases inherent in the LLMs used by our framework can affect assessments.

In future work, we aim to refine the framework by accounting for cultural and individual differences in moral perspectives and to expand the dataset to include a broader range of narrative forms. These efforts will further enhance our understanding of character morality and its impact on narrative engagement.

8 Ethic Consideration

We believe our research contributes ethically to the fields of NLP and storytelling without causing harm. To maintain ethical standards during the human evaluation phase, our study was reviewed and approved by our institution's Institutional Review Board (IRB). We collected participant responses without gathering any personally identifiable information. Before participation, individuals were fully informed about the study's objectives and procedures, and we obtained their informed consent.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, Artificial Intelligence Graduate School Pro-

gram(Sungkyunkwan University)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00357849).

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

Gordon H. Bower and Daniel G. Morrow. 1990. Mental models in narrative comprehension. *Science*, 247(4938):44–48.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Conference on Empirical Methods in Natural Language Processing*.

Mingda Chen and Kevin Gimpel. 2021. Tvstorygen: A dataset for generating stories with character descriptions. *arXiv preprint arXiv:2109.08833*.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniil Fischer, Frankand Skorinkin. 2021. *Social Network Analysis in Russian Literary Studies*, pages 517–536. Springer International Publishing, Cham.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Bernard Gert and Joshua Gert. 2020. The definition of morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2020 edition. Metaphysics Research Lab, Stanford University.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment. *Preprint*, arXiv:2110.07574.

Eunchong Kim, Taewoo Yoo, Gunhee Cho, Suyoung Bae, and Yun-Gyung Cheong. 2022. The CreativeSumm 2022 shared task: A two-stage summarization model using scene attributes. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 51–56, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hannah Kim, Denys Katerenchuk, Daniel Billet, Jun Huan, Haesun Park, and Boyang Li. 2019. Understanding actors and evaluating personae with gaussian embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6570–6577.

O-Joun Lee and Jason J. Jung. 2019. Modeling affective character network for story analytics. *Future Generation Computer Systems*, 92:458–478.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv e-prints*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Keith Oatley. 2008. The mind's flight simulator. *The Psychologist*, 21:1030–1032.

OpenAI. 2023. Gpt-4. https://openai.com/research/gpt-4.

Valentina Pyatkin, Jena D Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2022. Clarifydelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. *arXiv preprint arXiv:2212.10409*.

Stuart J. (Stuart Jonathan) Russell. 2019. *Human compatible : artificial intelligence and the problem of control*. Viking, New York.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. MBTI personality prediction for fictional characters using movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6715–6724, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

YURI Y. TARASEVICH, ANNA V. DANILOVA, and OLGA E. ROMANOVSKAYA. 2023. Network analysis of verbal communications in the novel the master and margarita by m. a. bulgakov. *Advances in Complex Systems*, 26(01).

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Mo Yu, JiangNan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. *ArXiv*, abs/2305.10156.

Mo Yu, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Jing Li, Yue Yu, and Jie Zhou. 2022. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. *ArXiv*, abs/2211.04684.