Probing Narrative Morals: A New Character-Focused MFT Framework for Use with Large Language Models

Luca Mitran[*1] **Sophie Wu**[*2] **Andrew Piper**[2]
[1]Faculty of Arts and Sciences, McGill University
[2]Department of Languages, Literatures, and Cultures, McGill University
luca.mitran@mail.mcgill.ca sophie.wu@mail.mcgill.ca andrew.piper@mcgill.ca

Abstract

Moral Foundations Theory (MFT) provides a framework for categorizing different forms of moral reasoning, but its application to computational narrative analysis remains limited. We propose a novel character-centric method to quantify moral foundations in storytelling, using large language models (LLMs) and a novel Moral Foundations Character Action Questionnaire (MFCAQ) to evaluate the moral foundations supported by the behaviour of characters in stories. We validate our approach against human annotations and then apply it to a study of 2,697 folktales from 55 countries. Our findings reveal: (1) broad distribution of moral foundations across cultures, (2) significant cross-cultural consistency with some key regional differences, and (3) a more balanced distribution of positive and negative moral content than suggested by prior work. This work connects MFT and computational narrative analysis, demonstrating LLMs' potential for scalable moral reasoning in narratives.[1]

1 Introduction

Across all human cultures and time periods, stories have been used to disseminate moral lessons, cultural norms, and core values (Haidt and Joseph, 2004). While the content, medium, and style used in storytelling practices can vary across different cultural contexts, all stories require an agent (i.e. character) who engages in actions and experiences (Piper et al., 2021). Through the role of each character's actions over the course of a narrative, stories implicitly or explicitly communicate ethical frameworks, drawing the distinction between virtuous and transgressive behavior (Vercellone and Tedesco, 2020).

As narrative theory has long posited, characters are the focal point of stories through which forms of social cognition and reader identification develop (Zunshine, 2006; Mar, 2018; Oatley, 2016). By focusing on the described behaviors and choices of characters rather than implicit moral judgments or themes of a story, we can examine how moral values are embedded in storytelling through one of its most essential structural components. We propose that Moral Foundations Theory (MFT) – a framework which categorizes moral values into basic universal foundations (Haidt and Joseph, 2004) – can be used to categorize character actions according to their alignment with cross-culturally salient moral dimensions, allowing for comparisons of moral expression across different narrative contexts at scale.

To do so, we produce a novel MFT assessment tool: a Moral Foundations Character Actions Questionnaire (MFCAQ) which can be used by readers (both human and machine) to evaluate character actions and motivations as associated with the five original moral dimensions posited by MFT: *authority, care, fairness, loyalty*, and *purity*. We then compare LLM annotations on our questionnaire against human annotations for models of different sizes and across each moral foundation. We find reasonable to strong correlations between frontier models and at least one smaller open-weight model and human annotations, indicating that LLMs may be reasonably well-suited to this task. We also note an over-prediction bias where LLMs see stronger moral sentiments than human raters and provide recommendations for downstream adjustments.

After validating LLM performance on this task, we demonstrate the utility of our approach for cultural analytics through a case study evaluating the distribution of predicted moral foundations across a collection of 2,697 folktales drawn from 55 countries covering all major geographical regions of the world. Our analysis offers three key insights: first, moral foundations are widely distributed across stories, suggesting the relevance of MFT for evaluating culturally diverse narratives. Second, our anal-

[*]Equal contribution.
[1]All relevant data for our project can be found at https://doi.org/10.5683/SP3/BCUIXD

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28514–28529
November 4-9, 2025

ysis reveals substantial cross-cultural consistency in moral foundations. While certain values are emphasized differently across the regions represented in our dataset, the overall distribution shows significant overlap. Third, we test the claim that folktales exhibit a positivity bias in moral messaging (Wu et al., 2023). While we observe a modest overall skew toward positive character portrayals, this bias does not hold uniformly across moral foundations. Instead, we find significant variation, indicating that folktales encode both positive and negative moral content in more complex and differentiated ways than previously assumed. Taken together, these insights have important implications for how we understand the didactic function of storytelling.

Outside of cultural studies, our work provides other downstream NLP applications. As LLMs increasingly participate in cultural production, understanding their moral reasoning capabilities and how their moral assessments align with or diverge from human judgements becomes essential for responsible deployment. Through this paper, we also propose a structural way that LLM performance can be evaluated against human judgements.

2 Related Work

2.1 Moral Foundations Theory

We ground our moral analysis of character actions using Moral Foundations Theory (MFT). MFT posits that different forms of human moral reasoning can be systematically categorized into universal foundations (Haidt and Joseph, 2004), and that differences in moral systems across cultures can be interpreted as different emphases on virtues and vices that arise from these foundations (Haidt and Graham, 2007). Although the foundations in the framework are intended to be open for modification, we focus on the five foundations initially proposed by the original authors, since they have been the most widely applied in cultural studies and validated as consistently identifiable categories across cultures (Doğruyol et al., 2019). This cross-cultural validity suggests MFT's suitability for analyzing narratives from diverse cultural backgrounds. The five pillars of MFT are:

Authority: Involves respect for tradition, legitimate authority, and social hierarchy. The vice *(Subversion)* involves behaviors that challenge or undermine authority.

Care: Involves empathy, compassion, and the prevention of harm. The vice *(Harm)* involves indifference to suffering or active harm.

Fairness: Involves justice, reciprocity, and equitable treatment. The vice *(Cheating)* involves unfair bias, exploitation, and dishonesty.

Loyalty: Involves allegiance to one's group, including family, community, or nation. The vice *(Betrayal)* involves disloyalty or favoring outsiders over the in-group.

Purity: Concerns physical and moral cleanliness, often linked to religious or cultural norms. The vice *(Degradation)* includes impurity, defilement, or moral corruption.

2.2 Computational Approaches to Moral Analysis in Narratives

Prior computational applications of MFT in text analysis have focused primarily on social media and political discourse, often relying on lexicon-based resources such as the Moral Foundations Dictionary for social media analysis (Rezapour et al., 2019). Johnson and Goldwasser (2018) created supervised learning frameworks for moral foundation classification in political tweets. Roy and Goldwasser (2021) extended this work by applying lexical methods to analyze how moral foundations shape perceptions of political figures.

While these lexicon-based approaches can be effective for analyzing the language associated with moral foundations, they have notable limitations in dealing with context, ambiguity, and the subtleties of moral expression. For example, in Wu et al. (2023), the only other prior work applying MFT to narrative analysis, a lexicon-based method is used to assess the prominence of moral foundations across stories and the top words associated with authority include "father," "emperor," and "servant," which may reflect common characters in folktales rather than positive alignment with "Authority."

LLMs have emerged as potentially suitable candidates for more complex moral annotation tasks. Stammbach et al. (2022) showed that LLMs can extract character roles from narratives without domain-specific training, achieving significant improvements over dictionary-based methods in identifying archetypal roles like heroes and villains. Hobson et al. (2024) explored the extraction and validation of story morals across various narrative genres using GPT-4. The authors developed a multi-step prompting sequence to derive morals and validate them through automated metrics and human assessments, highlighting the potential of LLMs to approximate human interpretations of story morals.

Studies have also investigated LLMs' abilities to identify and classify moral values across contexts. Roy et al. (2022) developed few-shot learning methods for moral frame identification, showing how in-context learning could improve the efficiency of moral value classification. Liscio et al. (2022) advanced this work by examining cross-domain classification of moral features, providing insights into how moral concepts manifest differently across contexts. Similarly, Chiu et al. (2024) developed the DailyDilemmas dataset to evaluate LLMs' handling of moral scenarios, revealing inherent preferences for certain moral dimensions.

Our work looks to incorporate these abilities to inform more robust moral analysis in narrative studies. Recent work categorizing the morality of characters in stories using LLMs has focused on categorizing character actions as 'moral' or 'immoral' (Bae et al., 2025). However, this binary ignores the diverse moral intuitions that can inform a character's moral significance in narrative. While LLMs have been shown in previous work to be suitable for both moral reasoning in complex categorization problems and character analysis, these capabilities have not yet been combined to identify the moral foundations of characters in narrative texts.

In line with calls for more explicit theoretical grounding in moral NLP research (Vida et al., 2023), we propose that narrative morals can be understood through the values embedded in characters' actions, and we introduce a novel method for measuring these values using the existing Moral Foundations Theory (MFT) framework.

3 Implementation

3.1 Moral Foundations of Character Actions Questionnaire (MFCAQ)

To adapt Moral Foundations Theory (MFT) for narrative analysis, we introduce the Moral Foundations of Character Actions Questionnaire (MFCAQ)—a novel instrument designed to assess the moral dimensions of fictional character behavior (Table 1). The questionnaire consists of 16 standardized items, each prefaced with: "In this story, are the character's actions..." Questions are crafted to capture both positive and negative valences of each moral foundation, reflecting virtues and vices as they manifest through narrative action.

While the MFCAQ includes both positive and negative valence items for each foundation, these are not always simple inverses (e.g., Care vs. Harm). Instead, we adopt a more flexible valence-based approach—Care(+) and Care(-), for example—to better capture the diversity of morally relevant actions depicted in stories. This design choice reflects how moral foundations manifest asymmetrically in narrative contexts and allows us to disentangle distinct moral expressions that might otherwise be conflated.

While some items are structured as direct opposites (e.g., respecting vs. disrespecting authority), others are intentionally non-mirrored to capture different facets of moral action within a single foundation. For instance, Fairness(-) includes both *cheating or lying* and *selfishness*, which reflect distinct violations that frequently arise in storytelling. Similarly, Care(+) includes both *empathy* and *responsibility*, two conceptually related but narratively distinct expressions of prosocial behavior. This asymmetrical structure enables a more nuanced, fine-grained measurement of character morality than a strict binary framing (e.g., Care/Harm or Loyalty/Betrayal) would allow. Responses are rated on a 5-point Likert scale from "Not at all relevant" to "Extremely relevant.", as shown in Table 2.

Foundation	Valence	In this story, are the character's actions...
Authority	+	Exhibiting respect for authority?
	-	Involving disrespect for authority?
Care	+	Showing care for others?
	+	Exhibiting the importance of responsibility to others?
	-	Causing harm to others?
Fairness	+	Driven by a sense of fairness?
	-	Involving cheating or lying?
	-	Driven by a sense of selfishness or self-interest?
Loyalty	+	Exhibiting loyalty to a group that is not family or country?
	+	Exhibiting loyalty to a tribe or country?
	+	Exhibiting loyalty to a family member?
	-	Involving betraying someone?
Purity	+	Adhering to some moral, religious, or cultural code?
	-	Exhibiting cruelty?
	-	Involving the goal of creating chaos?

Table 1: Moral Foundations Character Action Questionnaire (MFCAQ) used to analyze character actions in our dataset.

Additionally, we assess *character valence* using a separate 5-point scale from very negative (villain) to very positive (hero) portrayal. This allows us to understand the narrative perspective of the agent's actions. It is possible that agents that engage in negative moral actions (e.g. harming others) may be celebrated in the story, just as positive actions (e.g. helping others) may be seen negatively (as naive or foolhardy). We use character valence as a way of grounding the overall valence of the character-centred moral foundations exhibited in stories.

The original Moral Foundations Questionnaire (MFQ)[2] is designed to measure how individuals evaluate abstract moral scenarios in relation to their own values. In contrast, our adaptation focuses on how fictional characters express moral foundations through their actions within narrative contexts. Rather than abstract moral judgment, our questionnaire prompts evaluators to connect concrete narrative events to specific moral dimensions. This approach preserves the theoretical grounding of MFT while extending its applicability to the analysis of storytelling and character behavior. A detailed description of the questionnaire development process is provided in Appendix B.

Score	Description
1	Not at all relevant (the character has no consideration for OR acts against this moral foundation)
2	Not very relevant
3	Somewhat relevant
4	Very relevant
5	Extremely relevant (this moral foundation is one of the most important factors when they make a decision)

Table 2: Moral Foundation Scoring Scale

3.2 Data and Model Selection

Our analysis uses the Kaggle Folk Tales dataset, containing 2,697 traditional stories from 55 countries with an average length of 1,916 words. Appendix A provides a breakdown of the dataset by major world regions.

Folktales provide fertile ground for moral analysis as they encode culturally specific values, norms, and didactic structures, often centering on reward and punishment mechanisms (Dundes, 1965). While the reliance on English translations in our data represents an important limitation, research suggests that the principal effect of translation is at the level of style and syntax not meaning (Tirkkonen-Condit, 2002; Wein, 2023). It would therefore be surprising if translations made significant and consistent impacts on the moral foundations of characters, though further work extending our framework using multilingual modeling will be able to more definitively answer this question.

Since narrative studies suggests that the most central character carries the most moral signaling within traditional storytelling (Campbell, 2008), we first prompt GPT-4o to identify the main character(s) for each folktale (temperature = 0). We found zero errors for this step in our manually annotated validation subset. More details on the character identification process can be found in Appendix C.

4 Validation

To validate our approach, we combine human evaluation with construct validation. First, we assess face validity by comparing model predictions to human annotations on a stratified subset of 50 stories and 735 overall questions, measuring alignment with human moral judgments. Second, we evaluate construct validity through two lenses: content relevance, by examining the distribution of moral foundations across the full dataset to confirm broad and balanced coverage; and convergent/discriminant validity, by analyzing intercorrelations among moral scores—expecting positive correlations among related foundations and negative correlations between opposing valences.

4.1 Inter-Annotator Agreement

Five experienced undergraduate annotators from North America with training in literary studies independently rated each story using our adapted MFT questionnaire. Annotators were provided with a detailed codebook, available in our data repository, and participated in multiple rounds of training to ensure familiarity with the Moral Foundations Theory framework and consistent application of labels.

To assess inter-rater reliability, we computed Krippendorff's alpha across all moral foundation dimensions, resulting in $\alpha = 0.44$ based on 735 moral questions. This moderate agreement reflects the inherent subjectivity of moral interpretation in narrative texts, while indicating sufficient reliability for downstream comparison with model predictions. Within-category agreement is shown in Table 3.

[2]https://moralfoundations.org/questionnaires/

We note that lower agreement levels are expected and meaningful for this type of interpretive task. Moral foundation relevance involves inherently subjective judgments that vary across individuals. As such, we also use the inter-annotator agreement as a baseline to compare LLM-human annotation agreement.

4.2 Model Performance

Through iterative testing of various prompting strategies, including contextual prompting and varying formality levels, we found that explicitly referencing MFT in prompts improved model performance while maintaining evaluation consistency. For more details on prompts used see Appendix D. We display results with our best-performing prompt for the remainder of the paper.

On our human annotated subset, we collect LLM responses to the MCAQ using five language models varying in size and accessibility: GPT-4o-08-06-2024 (Number of parameters unknown), and four open-source models - Mistral (7B), Gemma2 (9B), Llama3.1 (8B), and Llama3.2 (3B). This selection enables exploration of performance across models available at different scales while including both proprietary and open-source implementations. In all further experiments, we use default temperature settings for all models.

To evaluate model performance on this task, we use Spearman's rank correlation coefficient (ρ) to compare model outputs to averaged human annotations for each moral foundation using our five-point Likert scale. As shown in Table 3, while GPT demonstrated the strongest overall performance, Gemma2's correlations were competitive despite having far fewer parameters. For both GPT and Gemma2, all correlations with human ratings across the six dimensions were statistically significant at the $p < 0.05$ level.

Due to the subjectivity of this task, identical stories may yield multiple justifiable rankings to the same question. Additionally, slight differences on our Likert scale may still agree on the fundamental moral assessment, as adjacent scores (like 1-2 or 4-5) reflect similar judgments with minor intensity variations rather than contradictory evaluations. The labels for our Likert scale ("1 — Not at all relevant," "2 — Not very relevant," "3 — Somewhat relevant," "4 — Very relevant," "5 — Extremely relevant") create meaningful gradations that capture the nuanced nature of moral relevance judgments. Adjacent categories represent intensity variations within the same general assessment (e.g., both "Very" and "Extremely relevant" indicate strong relevance). Across our validation set, 69.99% of questions have at least one human annotation (out of five) matching GPT-4's annotation (ranging from 61.2%-91.84% across MFT/valence categories), and 96.88% have at least one annotation within ±1 of GPT-4's assessment. Detailed category statistics appear in Appendix E. We also analyze an example story, along with human and MFT annotations that diverge along varying MFT dimensions in Appendix F.

We do note meaningful variance across foundations, with Authority (ρ = 0.32) and Purity (ρ = 0.36) performing notably lower than Care, Fairness, and Loyalty (ρ = 0.56-0.66), mirroring higher levels of human disagreement on these dimensions. We hypothesize that this may be due to greater cultural specificity with respect to Purity and Authority since they are relative to specific cultural codes (tradition in the case of purity, authority figures for Authority). We do not observe differential performance along positive or negative dimensions within these two categories, indicating there is not a bias in terms of moral valence. Nevertheless, we flag this as a possible area for further research on using MFTs to study cross-cultural differences.

Figure 1: Distribution of rating differences between model predictions (GPT and Gemma2) and averaged human annotations, grouped by moral foundation category. Each histogram reflects the extent to which models overestimate (positive values) or underestimate (negative values) the presence of a given moral concern.

One consistent pattern in model behavior is a systematic bias toward over-identifying the presence of moral content (Figure 1). Human annotations are heavily skewed toward the lower end of the Likert

Model	Average	Authority	Care	Fairness	Loyalty	Purity	Valence
Human Avg	0.515	0.409	0.585	0.523	0.508	0.369	0.699
GPT	0.586	0.318	0.648	0.656	0.561	0.355	0.830
Gemma2	0.561	0.418	0.594	0.734	0.545	0.480	0.757
Mistral	0.366	0.168	0.431	0.494	0.340	0.252	0.603
LLaMA 3.1	0.329	0.169	0.393	0.343	0.319	0.287	0.597
LLaMA 3.2	0.178	0.178	0.099	0.164	0.172	0.284	0.170

Table 3: Spearman's Rank correlation between model predictions and average human ratings across all five MFT dimensions and character valence. We also include average pairwise human rater correlation scores for comparison.

scale, indicating that, for most stories, only a handful of questions from our moral questionnaire are relevant. In contrast, models consistently assigned higher relevance scores, suggesting an inflation in moral attribution. On average, our two best models, Gemma2 and GPT, overestimated moral relevance by 0.94 and 0.35 points respectively across categories. To account for these systematic shifts, we use bias-adjusted scores in downstream analyses, calculated by subtracting each model's average overprediction per moral foundation category relative to human ratings from its original predictions. Moving forward we focus our analysis on our best-performing model, GPT.

4.3 Construct Validity

We next evaluate construct validity along two key dimensions with respect to our target dataset: content relevance and convergent/discriminant validity.

Content Relevance As part of our content validity assessment, we examine the breadth and richness of moral foundation coverage across the full dataset (see Figure 2 & Figure 3). A sparse or narrowly distributed presence of foundations would suggest limitations in either our adapted questionnaire or the model's sensitivity to moral content. Instead, we observe a robust distribution: 98.63% of stories include at least one moral foundation, and 76.83% exhibit three or more. This indicates not only that moral foundations are broadly detected across the corpus, but also that stories frequently engage with multiple moral dimensions.

Convergent / Discriminant Validity Figure 4 presents pairwise correlations between moral foundations, disaggregated by valence. We find that conceptually related behaviors co-occur: for example, positive Care is most strongly correlated with positive Fairness, while negative Fairness aligns

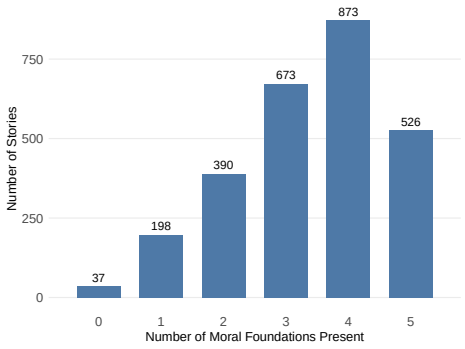

Figure 2: Distribution of moral foundations across stories, where a foundation is considered present if rated as at least "somewhat relevant" (score > 2) to the character's actions.

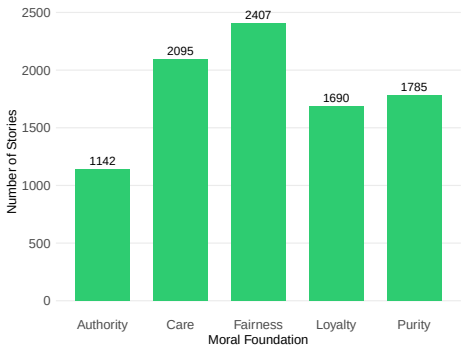

Figure 3: Distribution of each moral foundation with at least one contributing question with a score > 2.

most closely with negative Loyalty. These associations suggest that characters who exhibit compassion also tend to value fairness, while those who cheat are also likely to betray their group. Substantial negative correlations between positive and negative versions of the same foundation further confirm that our method captures moral polarity, not just presence. Finally, character valence is highly correlated with both positive ($\rho = 0.68$) and negative ($\rho = 0.45$) foundation scores, indicating that moral valence also reflects perceived

character morality—good characters behave well, while bad characters do not.

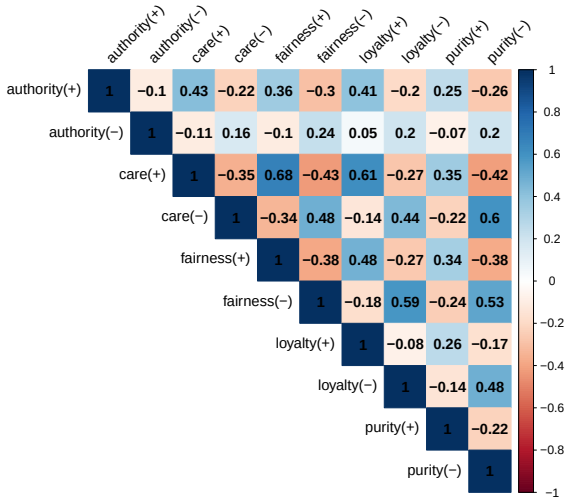

Figure 4: Pairwise correlations of individual foundations distinguished by valence.

5 Results

5.1 Do Folktales Exhibit a Positivity Bias?

Prior work using dictionary-based models has suggested that folktales exhibit a positivity bias with respect to moral foundations, i.e. they focus more on positive examples than negative when communicating moral messages (Wu et al., 2023). If true, this insight has significant implications for how we understand the cultural role of traditional narratives as didactic resources. It would help shore up the common-sense belief that human beings are more predisposed to learn from positive examples.

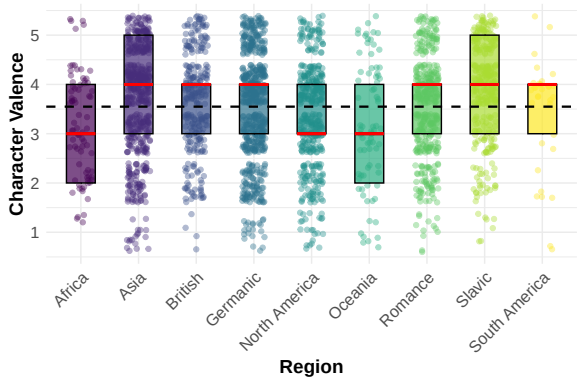

Figure 5: Distribution of character valence by region. Red lines indicate regional medians and the black dashed line the population mean.

To assess whether folktales exhibit a positivity bias in their portrayal of moral character, we begin by examining the distribution of character valence across regions Figure 5. Observationally, we can see a high degree of spread across cultures. To test this more rigorously, we use a cumulative link mixed-effects model (CLMM) with a logit link function. Because the valence ratings are ordinal and the data is hierarchically structured by continent, this approach allows us to model the probability of a character being rated more positively while accounting for variation attributable to region. The model specification is as follows:

$$\text{Character Valence} \sim 1 + (1|\text{Region})$$

Unlike simple frequency counts, this method respects the ordinal nature of the valence scale and permits partial pooling across regions, providing a more nuanced estimate of overall tendencies.

Using this model, we find a modest but statistically significant tendency toward positively evaluated characters: 54.6% receive a valence rating of 4 or 5 ($p < 0.001$). This confirms a general positivity bias in character portrayal across the dataset. At the same time, over a quarter of characters (26.0%) are rated as morally neutral (valence = 3), suggesting that folktales often depict characters whose moral alignment is ambiguous—a nuance frequently overlooked in prior literature. We find no significant regional differences in the overall positivity of character portrayals.

While character portrayals overall tend to skew positive, we next examined whether the relationship between perceived moral value and moral valence differs across cultural contexts and types of moral concern. To do so, we employed a linear mixed-effects model (LMM), with moral foundation valence—a continuous variable ranging from 1 (strongly negative) to 5 (strongly positive)—and continent (a categorical variable with nine levels) as fixed effects. To account for heterogeneity across moral foundations, we included both a random intercept and a random slope for valence by foundation. This allowed us to test whether specific moral foundations respond differently to shifts in moral valence. We specified the model as:

$$\text{Value} \sim \text{Valence} + \text{Region} + (1 + \text{Valence}|\text{MF})$$

Interestingly, the analysis revealed no significant fixed effect of valence ($\beta = 0.251$, p = 0.209), suggesting that, once over-prediction bias is corrected, there is no uniform positivity bias across foundations. However, the random slope variance

remained substantial (SD = 0.374), indicating that certain foundations are more sensitive to valence than others.

5.2 Regional Variations in Moral Foundations

Prior work has emphasized notable regional differences in the expression of moral values in folktales (Wu et al., 2023) and in attitudes among the general population of different regions (Doğruyol et al., 2019). In order to study regional variations in our data we first employed bootstrapping with 1,000 resamples to calculate 95% confidence intervals around regional mean scores for each moral foundation's valence (Care+, Care-, etc.). This approach enabled identification of significant regional deviations from global means while accounting for sample size variability.

Our bootstrapping analysis revealed meaningful regional variation in moral foundation emphasis, with approximately one-third of all region-foundation pairs showing statistically significant deviation from global means based on 95% confidence intervals (see Appendix G for a visualization of these results). This finding indicates that while there is indeed regional differentiation among moral foundations, as expected, a significant amount of consistency also exists across cultures in the behavioral focus of traditional stories.

To better understand regional variations, we applied Principal Component Analysis (PCA) to the scaled moral foundation scores, with regions as observations and moral foundations as variables. The resulting biplot representation in Figure 6 indicates that polarity is itself a key differentiator of regional behevior. Asian and Slavic stories, for example, exhibit notably higher scores on 4 and 5 foundations respectively, while African and Oceanic folktales exhibit notably higher scores on 3 and 4 negative foundations respectively.

These findings are important because they indicate, first, a more general insight about the association between moral foundations and folktales: i.e. that positive and negative foundations are used in highly correlated ways when it comes to cultural narratives. We also gather insights about particular cultures that complement prior work. For example, the oft-noted positive alignment of Asian cultures towards authority (Wu et al., 2023) can also be observed in our data. However, we also observe that other cultures may exhibit this as well, i.e. Slavic folktales according to our dataset, and that this statement overlooks the ways in which pro-

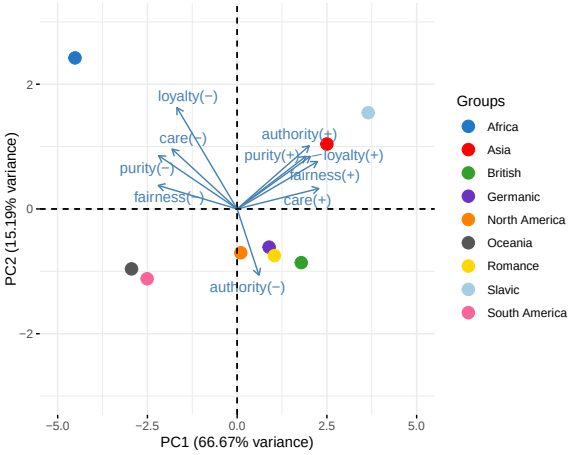

Figure 6: PCA biplot displaying the relationship between regions and moral foundations. Arrows indicate the contribution of each moral foundation.

authority stories are also pro-caring, pro-fairness, and pro-purity as well. As we saw in Figure 2, folktales do not focus on a single moral message, but are instead vehicles of multi-dimensional moral communication, often communicating the value of multiple foundations simultaneously.

6 Conclusion

This study introduces a character-centric framework for analyzing moral values in narratives by applying Moral Foundations Theory (MFT) to the actions and motivations of fictional characters. Building on the premise that characters are the primary vehicles through which stories convey social cognition, we shift focus away from abstract moral themes, and instead anchor moral analysis within explicitly described character behaviour. To facilitate this, we introduce the Moral Foundations Character Actions Questionnaire (MFCAQ), a novel instrument designed to evaluate how character actions align with the five original moral dimensions of MFT.

Our validation results reveal moderate to strong correlations between frontier language models and human judgments, along with encouraging alignment from at least one smaller open-weight model. Notably, we observe a systematic over-prediction bias across models, suggesting the need for further analysis. However, because this bias introduces only modest deviations, a simple correction method proves sufficient to enable reliable downstream inference.

When applied to our folktale dataset, our method

offers a few salient findings. First, we provide empirical support for the prevalence and cross-cultural consistency of moral foundations in folktales, aligning with the universality hypothesized by Moral Foundations Theory (MFT). Our questionnaire indicates that moral foundations are widely spread across cultures and multiply present within stories.

Second, contrary to prior claims of a strong positivity bias in folktales, our results show no significant difference in the portrayal of positive versus negative moral content. This challenges the prevailing view that traditional narratives primarily promote positive moral instruction, revealing instead a more balanced and differentiated moral landscape.

Future work could apply this framework to more complex narrative forms such as novels or films. Experimenting with more localized windows of actions and aggregating over narrative time may also generate insights regarding the moral consistency of characters.

This character-centered approach offers a scalable method for examining how moral values are embedded in storytelling. By focusing on characters as the primary agents of moral action, it provides an alternative to thematic analysis and highlights how ethical frameworks are enacted through narrative. Leveraging LLMs for this task can help deepen our understanding of how societies encode and transmit values through the stories they tell.

Limitations

We highlight here some important limitations to our study. The dataset, which focuses exclusively on folktales, inherently emphasizes shorter, more morally explicit narratives. While folktales are well-suited for analyzing cultural values and didactic functions, they lack the character complexity and narrative depth often found in longer, modern texts. This raises questions about the generalizability of our findings to extended narratives with multiple character arcs and evolving moral themes. There is also a methodological question of how to apply our prompting framework to longer narratives where character actions may be more diverse. Identifying adequate ways to account for character diversity ought to be a central focus of future work.

Our original dataset was drawn from an openly available online collection of folktales from around the world, all presented in English. Because these stories were originally composed in various languages, we acknowledge that the process of translation may have altered key moral emphases and cultural nuances. As a result, the moral signals in our dataset may differ from those in the original versions.

The geographical and temporal scope of the dataset also imposes constraints. Although the dataset covers 55 countries, regional biases in story selection and translation limit its representativeness. While we can be confident about the inter-cultural diversity of the data overall, we cannot assume that our data is fully representative of any single region. Inferences made with respect to specific regional behavior thus need to be made with considerable caution and require further work.

The strong correlations between moral foundations in our results Figure 4 indicate that much of the observed variance in our data is captured by general positive or negative polarity rather than foundation-specific differences. While this could reflect genuine patterns in how moral lessons are conveyed through folktales, it may also stem from a methodological constraint: the condensed nature of folktales means that character actions and plot events are limited, potentially causing our foundation-specific queries to reference the same story events. These events, while potentially multidimensional in their moral implications, may not provide sufficient granularity to distinguish between different moral foundations. Future work should consider prompting frameworks that explicitly tie moral judgments to specific character actions, allowing for more nuanced differentiation between moral foundations even in shorter narratives.

A further limitation lies in the design of our adapted MFQ. While it effectively captures five foundational dimensions of morality, future work will want to explore changes in outcomes with respect to alternative phrasing and prompting of character attributes. This also holds for future LLMs. We expect LLMs to continue to evolve and therefore the outputs and moral reasoning of LLMs may not remain constant. Our data thus offers a useful benchmark to understand this "moral drift" of models.

Further reflection may also be warranted to consider latent moral dimensions not included in the MFT categories considered in this paper. Finally, diversifying the annotator pool and breadth of stories that are manually coded would allow for richer cultural calibration.

Although our work evaluates models of different

parameter sizes and sources, our study does not rigorously evaluate the effect of model scaling or architecture on performance on our proposed task. Rather, our intention was to evaluate a range of publicly accessible LLMs of varying sizes and design decisions to identify which model(s) perform best on a culturally grounded task. Future work could more rigorously investigate dimensions like model scaling in isolation, e.g. using models like Qwen3 of varying sizes—would better isolate the effect of scale from other confounding factors like training data or alignment procedures.

References

Suyoung Bae, Gunhee Cho, Yun-Gyung Cheong, and Boyang Li. 2025. Charmoral: A character morality dataset for morally dynamic character analysis in long-form narratives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8809–8818.

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

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life. *arXiv preprint*. ArXiv:2410.02683.

Burak Doğruyol, Sinan Alper, and Onurcan Yilmaz. 2019. The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, 151:109547.

Alan Dundes. 1965. *The study of folklore*. Englewood Cliffs, N.J.: Prentice-Hall.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story Morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Kristen Johnson and Dan Goldwasser. 2018. Classification of Moral Foundations in Microblog Political Discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Enrico Liscio, Alin E. Dondera, Andrei Geadău, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Raymond A Mar. 2018. Stories and the promotion of social cognition. *Current Directions in Psychological Science*, 27(4):257–262.

Keith Oatley. 2016. Fiction: Simulation of social worlds. *Trends in cognitive sciences*, 20(8):618–628.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311.

Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the Measurement of Social Effects by Capturing Morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, USA. Association for Computational Linguistics.

Shamik Roy and Dan Goldwasser. 2021. Analysis of Nuanced Stances and Sentiment Towards Entities of US Politicians through the Lens of Moral Foundation Theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13, Online. Association for Computational Linguistics.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2022. Towards Few-Shot Identification of Morality Frames using In-Context Learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 183–196, Abu Dhabi, UAE. Association for Computational Linguistics.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Sonja Tirkkonen-Condit. 2002. Translationese—a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.

F Vercellone and S Tedesco. 2020. Folktale, morphology. *Glossary of Morphology*, page 185.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research. *arXiv preprint*. ArXiv:2310.13915 [cs].

Shira Wein. 2023. Human raters cannot distinguish english translations from original english texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12266–12272.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Lisa Zunshine. 2006. *Why We Read Fiction: Theory of Mind and the Novel*. Ohio State UP.

Appendix

A Distribution of stories by region

Table 4 displays the regional distribution of stories in our dataset.

Region	Number of Stories
Germanic	603
Asia	500
North America	407
Romance	395
Slavic	327
British	279
Oceania	78
Africa	76
South America	32
Total	2,697

Table 4: Regional distribution of folktales in the dataset.

B Moral Foundations Questionnaire Development

Our questionnaire was developed through an iterative, interactive process involving annotators, large language models (LLMs), and the research team. We began by designing a core set of items that directly operationalized each of the five original moral foundations from Moral Foundations Theory (MFT), using foundational keywords (e.g., care, fairness) to formulate initial prompts positively and negatively valenced character actions (Full prompts shown to LLMs can be found in Appendix D).

These initial items were then tested through multiple rounds of annotation and model prompting across a diverse sample of folktales. Annotators provided feedback on ambiguous or underspecified questions, while LLM outputs were analyzed for patterns of misclassification or semantic overlap between foundations. Based on these observations, we refined the questionnaire by expanding or splitting items to better capture distinct subtypes of moral expression within a given foundation (e.g., distinguishing responsibility to others from empathic care within the Care foundation, or separating cheating from self-interest under Fairness). This process ensured both conceptual clarity and empirical responsiveness to variation in narrative data.

The final version of the Moral Foundations Character Actions Questionnaire (MFCAQ) consists of 16 items that collectively represent a balanced and interpretable measure of how moral values are embedded in character behavior across narrative contexts. Valid responses to the questionnaire (presented to both human and LLM annotators) are ranked on a 1-5 Likert scale as shown in Table 2.

C Character Identification Process

We use the following prompt to ask GPT to choose the main character(s) of each story. We set temperature to 0 for replication purposes since this is a relatively deterministic task.

"Read the following story and identify the main character. The character will most likely only have one main character, but if there are multiple main characters that you are sure are equally important, list their names separated by commas. If there are no main characters, respond with 'None'. Do not provide any additional text or explanation. Story: story_text"

As part of the validation exercise with the human annotators, each annotator reviewed the choice of main character(s) for appropriateness. We did not find any examples where the model had erred.

D Model Prompting

The exact prompt shown to all models in the experiments for all MFCAQ questions is shown below:

```
Please read the following story.
With respect to the character
{CHARACTER_NAME}, in this story,
are the characterś actions
{MFCAQ_QUESTION}? Please provide
your answer on a 1-5 scale with
the following criteria.
Do not put anything other than
your numerical response for each
question. Note that the questions
reflect scores for a particular
moral foundation under Moral
Foundation Theory, so you should
use that theory to guide your
analysis.
1 = not at all relevant (the
character has no consideration
for OR acts against this moral
foundation)
2 = not very relevant
3 = somewhat relevant
4 = very relevant
5 = extremely relevant (this
moral foundation is one of the
most important factors when they
```

```
make a decision)
Here is the story: {STORY}
```

On the validation set, we found that explicitly referencing Moral Foundation Theory in the prompt improved model alignment with human scores on MFCAQ questions.

The prompt used to determine character valence in all experiments is:

```
Please read the following story.
I want you to tell me, on
a 1-5 scale, how positively
or negatively the character
{CHARACTER_NAME} is portrayed in
the story. Please answer on a 1-5
scale with the following criteria.
Do not output anything other than
your numerical response for this
question.
1 = very negative portrayal
(villain)
2 = somewhat negative portrayal
3 = neutral portrayal
4 = somewhat positive portrayal
5 = very positive portrayal
(hero)
Here is the story: {STORY}
```

E Minimum Match with Annotators

In Table 5, we show the number of questions across all samples (where each sample is a question from the MFCAQ asked on a specific story/character pair) in our validation set which contain at at least one human annotation that matches the GPT-4 score exactly. Table 6 shows the number of samples where at least one human annotation is within ± 1 of the GPT-4 annotation.

Category	Exact Match (%)
AUTHORITY	61.22
CARE	69.39
FAIRNESS	63.95
HARM	77.55
LOYALTY	71.43
PURITY	75.51
CHARACTER VALENCE	91.84
Overall	**69.99**

Table 5: Percentage of samples where at least one human annotation matches the GPT-4 score exactly.

Category	Within ± 1 (%)
AUTHORITY	95.92
CARE	96.94
FAIRNESS	97.28
HARM	100.00
LOYALTY	96.94
PURITY	95.92
CHARACTER VALENCE	100.00
Overall	**96.88**

Table 6: Percentage of samples where at least one human ranking is within ± 1 of the GPT-4 annotation.

F Example For Model Disagreement with Human Score

We show a relevant example here for a story where GPT-4 diverges from human annotations. A summary of the story, along with human and GPT-4 annotations for the MFCAQ, can be viewed in Table 7. We compare and analyze some of the human and model annotations across different categories below:

LOYALTY: we see significant divergence under the question *"Are the following character's actions involving betraying someone?"* (Diff. = -2.8), with human ratings significantly lower than the model rating. Human annotators may have not interpreted the princess's selfishness as a 'betrayal' since the story does not explicitly state prior loyalty, and also may have focused more on the final actions of the princess (who eventually earns the chance to reunite with her parents by learning compassion and hard work) while the model focuses on the Princess initially banishing her parents from the palace. The spread of human ratings (ranging from 1 to 4) underscores this interpretive ambiguity, and the model's high rating may reflect a plausible reading aligned with some human judgments.

CARE: In "Showing care for others?" and *"Exhibiting the importance of responsibility to others?"* show large differences (Diff. = +2.0), with GPT-4 assigning lower scores than the human average. Here too, we observe human ratings ranging from 2 to 4 and from 1 to 5 respectively, suggesting disagreement about whether the Princess's final actions (living with the fairy, making gifts for her parents) constitute genuine care or mere restitution.

FAIRNESS: In *"Involving cheating or lying?"* under FAIRNESS shows a more clear-cut model-human divergence (Diff. = –1.8), despite very low variance among human annotators (four 1s and one 2). The model's rating of 3 suggests that it may

have interpreted symbolic deception in the story (e.g., the fairy in disguise or the Princess's transactional thinking) as morally deceptive, diverging from the human consensus that saw little to no dishonesty.

AUTHORITY: The model gives low scores to both *"Exhibiting respect for authority?"* and *"Exhibiting disrespect for authority?"*, while humans leaned toward moderate to high levels of disrespect (Diff = +1.4, Diff = +2.2). Since there are different characters which could be interpreted as possible 'authority' figures to the princess (i.e. either the parents she banishes or the fairy she later learns compassion from), and the main character exhibits disrespect for some characters (the parents) but deference to others (the fairy), different valid answers could be produced by focusing on different elements of the story. This is also reflected in the diversity of human answers (ranging from 2 to 4).

PURITY: While the model gives similar answers to most human annotators on *"Adhering to some moral, religious, or cultural code?"* (Diff = 0.6), with three human rankings matching the GPT-4 ranking of 1, we see more significant divergence on this sample on *"Exhibiting cruelty?"* (Diff = -1.4). GPT-4's ranking of 4 agrees with one human annotator, while the other four annotators give lower scores, likely implying that they interpret the princess's selfishness as something other than cruelty since it is not explicitly motivated by a desire to hurt other characters.

CHARACTER VALENCE: This question shows moderate disagreement (Diff = +1.0), with GPT-4 assigning a lower score (2 vs. human average of 3). The variance among human ratings (from 2 to 4) hints at divergent interpretations of the protagonist's redemption arc—some annotators may have been influenced by the Princess's eventual change of heart, while others may have focused more on her earlier behavior.

Overall, this example illustrates some common features of narratives that can make interpretation of moral actions of characters highly subjective, such as the development of a character's moral foundations throughout the course, or different characters to which the main character(s) may exhibit different behavior. However, in this example, we can also see that even the model's 'divergent' rankings align with interpretations valid to some human annotators.

G MFT estimates by region

Figure 7 displays an overview of moral foundation estimates by region along with confidence intervals.

Story: What You Shall Give Me?
CHARACTER: The Princess
SUMMARY: A spoiled Princess, raised to expect everything she wanted, demands so much from her parents that they give her everything they own—even their crowns—and she banishes them from the palace. Left alone, she becomes trapped inside the King's heavy crown and, in desperation, agrees to give up everything to a mysterious old woman (a fairy in disguise) in exchange for help. Humbled and changed, the Princess goes to live and work with the fairy, learns compassion and hard work, and eventually earns the chance to reunite with her parents by preparing heartfelt gifts for them.

Question: In the story, are the character's actions...	**Category**	**H1**	**H2**	**H3**	**H4**	**H5**	**Avg (H)**	**GPT-4**	**Diff.**
Showing care for others?	CARE	3	3	2	3	4	3.0	1	2.0
Exhibiting the importance of responsibility to others?	CARE	5	1	3	3	3	3.0	1	2.0
Causing harm to others?	HARM	3	1	4	4	3	3.0	4	-1.0
Driven by a sense of fairness?	FAIRNESS	1	1	1	1	2	1.2	1	0.2
Involving cheating or lying?	FAIRNESS	1	1	1	1	2	1.2	3	-1.8
Driven by a sense of selfishness or self-interest?	FAIRNESS	3	3	5	5	4	4.0	5	-1.0
Exhibiting loyalty to a group that is not family or country?	LOYALTY	1	1	1	1	1	1.0	1	0.0
Exhibiting loyalty to a tribe or country?	LOYALTY	1	1	1	1	1	1.0	1	0.0
Exhibiting loyalty to a family member?	LOYALTY	3	3	3	2	4	3.0	1	2.0
Involving betraying someone?	LOYALTY	2	1	1	4	3	2.2	5	-2.8
Exhibiting respect for authority?	AUTHORITY	3	3	1	2	3	2.4	1	1.4
Exhibiting disrespect for authority?	AUTHORITY	3	2	3	4	4	3.2	1	2.2
Adhering to some moral, religious, or cultural code?	PURITY	1	1	1	2	3	1.6	1	0.6
Exhibiting cruelty?	PURITY	3	2	1	4	3	2.6	4	-1.4
Involving the goal of creating chaos?	PURITY	1	1	1	1	2	1.2	1	0.2
CharacterValence	VALENCE	3	3	2	3	4	3.0	2	1.0

Table 7: Story with associated moral question, human ratings (H), the average human rating and model ratings (GPT-4), the average human The difference between the average human rating and the model rating for each question (human annotated score subtracted from GPT-4 score) is displayed under Diff.

Figure 7: Overview of moral foundation estimates by region. Red indicate statistically significant foundations per region with Bonferroni correction. Dashed blue line indicate sample mean for each foundation.